DIFFERENTIAL EQUATIONS – LINEAR THEORY AND APPLICATIONS

B SHAYAK

This page intentionally left blank

This page intentionally left blank

Differential Equations – Linear Theory and Applications

"A" edition

B SHAYAK

Original version date : July 2018

Current version date : October 2019

This book is open access – it is FREE to download, read and distribute. SHAYAK BHATTACHARJEE asserts his copyright as the author of this work. Academic usage of results contained here must be accompanied by a citation. The preferred citation format is :

B Shayak, "*Differential Equations – Linear Theory and Applications*," available electronically at www.shayak.in/Shayakpapers/DELTA/DELTA.pdf

The number of the relevant Unit or Tutorial may be included in the citation if appropriate.

Hard copies may be sold only at the cost of printing and binding and NO additional charges. Printing on A5 sheet (148 x 210 mm) or paper of equivalent size is recommended.

Preface

Differential Equations – Linear Theory and Applications (DELTA) originated as a companion textbook to the course MATH2930 (formerly TAM293) at Cornell University for which I was a Tutor/ Teaching Assistant during 2016-18. In its current and more or less invariant form, Part 1 of this book covers material appropriate for a one-semester course on ordinary differential equations (ODE) while Part 2 is appropriate for a one-semester course on partial differential equations (PDE). Both are at approximately the second year undergraduate level.

The scope of this book is as follows. Part 1 : first order ODEs, second and higher order linear ODEs mostly with constant coefficients, boundary value problems and Fourier series. Part 2 : heat equation, Laplace equation and wave equation. The basic approach I have followed is **how, why, where** – for each topic, we typically first see a quick algorithm of *how* to solve problems, then apply the algorithm to some examples and understand *why* it works, and finally go on to applications *where* the topic is relevant. It is these applications which, in my opinion, are the most interesting – they range from pharmacokinetics to gymnastics, from Japanese paper pots to magnetic levitation.

Part 2 here presents an approach to PDEs different from what is found in the current literature. Taking existence of solutions etc. more or less for granted and confining the syllabus to simple functions such as polynomials, exponentials and trigs, I have tried instead to clarify the myriad conceptual intricacies behind solving PDEs and modeling physical phenomena using PDEs. I believe that if the student once absorbs the underlying concepts and techniques without getting distracted by the solutions, s/he will have no difficulty in later learning the properties of special functions and solving equations where they appear. I hope that Part 2 shall thus act as a useful preparation for a variety of courses in physics, applied mathematics and engineering – among them classical electromagnetism, quantum mechanics, heat transfer, fluid flow and elasticity.

Although the book is based on a second year undergraduate course, there are some examples which are at the level of pedagogical publications and a few

at the level of research papers. These examples are more appropriate for higher level courses, and for projects. Two topics in particular have received an especially thorough treatment here . Section 6.3 and the relevant examples from Section 7.4 form the core of a rigorous one-semester course on classical electromagnetic theory. Relevant parts of Sections 2.1 and 2.3, all of Section 3.3, and relevant parts of Chapter 7make up a one-semester senior-level elective course on linear vibrations.

A quick word on what you will NOT find here. There are no special functions, hence there is no method of Frobenius. The emphasis on first order ODEs is less than that of most texts of similar scope and coverage. This is because the applications of first order ODEs are limited, and the solution techniques are not generalizable to higher order ODEs and PDEs. First order PDEs get a miss altogether, for similar reasons. Evaluation of complicated integrals is something I have not shown in detail – for one, they don't come up too often and for another, every doable integral can be found online in a table or evaluated in one minute using computer algebra. Nonlinear dynamics is a fascinating subject which is well served by a standalone book of this size. Hence I haven't gone into any details here beyond linear stability analysis which is just an application of linear ODEs and PDEs. Laplace transforms don't appear in this book because of the difficulty of inverting them. PDEs where the method is essential are beyond our scope, and every ODE which can be solved with this transform can also be solved otherwise. There is no better demonstrator of this than the concluding problem of Part 1.

In this book I have preserved the informal style of the tutorial section. Some examples of this informality are (i) the writing style (the nominal addressee is the undergraduate student), (ii) occasional anecdotes, jokes and exam strategies, and (iii) scarcity of references and absence of discussion regarding novelty of any result presented here. I have found that a relaxed atmosphere provides a perfect setting for students to absorb rigorous argumentation and complicated calculations.

It would have been impossible to accomplish a project of this size on my own. I am very grateful to some students of MATH2930 in the semester 2017-18/II, who read the preliminary draft carefully and highlighted numerous shortcomings. These criticisms led to a major revision. Since lists of contributing students tend to grow with time, I regret that I cannot include

names here. Nevertheless, my sincere thanks to you all. I would also like to thank the students of 2930 in the semesters 2016-17/I and II and the students of MATH2940 Linear Algebra in the semester 2017-18/I for their enthusiastic participation. Many of the issues which I have covered here first arose as queries from these classes.

I am grateful to my doctoral advisor, Professor RICHARD RAND for showing me many interesting features of boundary value problems, including FREDHOLM's theorem. Special thanks go to Professor LAURENT SALOFF-COSTE, instructor of 2940 in 2017-18/I – from his course I learnt a lot of linear algebra which I have used here. I express heartfelt gratitude to Professor T MUTHUKUMAR, instructor of the course MTH203N at IIT Kanpur, where I myself learnt how to solve differential equations. My sincere thanks to the anonymous journal Reviewers for their comments on those problems which I have submitted for publication – the feedbacks have caused improvements to the presentation here. This book also draws inspiration from the erstwhile educational system at Cambridge University (England not Massachusetts) where the coaches who trained the candidates for the Mathematical Tripos were distinct from the examiners who actually set the question papers – this double-blind system ensured that the coaches trained the students no holds barred, and each side tried to outwit the other by devising ever more creative problems. On quite a few occasions, these problems were the professors' original research which later made its way into the journal pages. The comprehensiveness of exposition and originality of problems are things that I have tried to imitate here.

Every few weeks I am producing a revised version of this book – each update responds to suggestions and criticisms I have received from readers, corrects errors and lack of clarity which I have noticed myself, and occasionally introduces replacement problems which I have thought up in the interim and found to be substantially better than the original ones. Nevertheless, the book remains, and will remain, by and large unchanged from its original edition.

Despite all my revision efforts, I know that DELTA contains errors. I have tried to ensure that there are no howlers i.e. incorrect final answers by testing them for plausibility, limiting cases etc. Many of these checks I have also described in the book. As regards typographic and formatting bloomers, that is another story. The only way a book can be purged of those is if the readers inform the author whenever they see one. So, please let me know of anything you find – even if it's something as trivial as an unitalicized variable or a matrix typed in Calisto instead of Times New Roman. Other kinds of feedback and suggestions are also most welcome. My email addresses are sb2344@cornell.edu and shayak.2015@iitkalumni.org.

SHAYAK

Theoretical and Applied Mechanics,
Sibley School of Mechanical and Aerospace Engineering,
Cornell University, Ithaca,
New York State 14853, USA

October 2019

Table of Contents

This book has two Parts (ODE and PDE), which are divided into four and three Chapters respectively. The level rating of each example is indicated in brackets after the title – Read Only examples are denoted by the code 0. Pages are numbered by Chapter – thus the 3rd page of Chapter 1 is designated as 1003 and so on.

INTRODUCTORY MATERIAL

Preface 3
Table of Contents 7
Preliminary. Introduction; note on the Tutorial level ratings 14
General discussion 14
Demo example [0]. 19

PART ONE : ORDINARY DIFFERENTIAL EQUATIONS

Chapter 1 : First order ordinary differential equations

1.1 First order non-generic ODEs 1002
1.1 General discussion 1002
Ex 1.1.1 Train acceleration [1] 1006
Ex 1.1.2 Exact for the exam [1] 1009
Ex 1.1.3 Integrating factor [1] 1012
Ex. 1.1.4 Train again [2]. 1015

1.2 **Linear ODEs : Properties of solutions 1019**
1.2 General discussion 1019
Ex. 1.2.1 LCC lollipop [0] 1023
Ex. 1.2.2 Radiocarbon dating [2] 1027
Ex. 1.2.3 Chemical tank – basic model [1] 1032
Ex. 1.2.4 Chemical tank with imperfect mixing [2] 1035
Ex. 1.2.5 Rocket sled [1] 1038
Ex. 1.2.6 NEWTON's rule of cooling [1] 1041
Ex. 1.2.7 LR circuit [1] 1043
Ex. 1.2.8 Pharmacokinetics and drug dosage [3] 1045
1.3 DIRAC's delta function 1051
L3 General discussion 1051
Ex. 1.3.1 Pharmacokinetics again [2]. 1053
1.4 Introduction to stability analysis 1056
1.4 General discussion 1056
Ex. 1.4.1 Stability analysis verification [2]. 1060
Ex. 1.4.2 Nuclear reactor [2] 1064
Ex. 1.4.3 Population dynamics [1]. 1067
Ex. 1.4.4 Bifurcation [1] 1071
Ex. 1.4.5 Malaria epidemic [2] 1073
Ex. 1.4.6 Fireball [2] 1076
End of Chapter 1 1077

Chapter 2 : Higher order homogeneous initial value problems

2.1 LCCH equations of any order 2002
2.1 General discussion 2002
Ex. 2.1.1 Harmonic oscillator [0] 2005
Ex. 2.1.2 Repeated root and reduction of order [2] 2012
Ex. 2.1.3 Damped harmonic oscillator [0]. 2015
Ex. 2.1.4 Bounded solutions [2] 2020
Ex. 2.1.5 Special IC [1] 2023
Ex. 2.1.6 Approximation analysis [2]. 2025
Ex. 2.1.7 Fourth order ODEs [2] 2030
Ex. 2.1.8 Loco coupling to train [1] 2033

2.2 Linear independence and Wronskian 2036
2.2 General discussion 2036
Ex. 2.2.1 Basic Wronskians [1] 2038
Ex. 2.2.2 High order Wronskian [1] 2040
Ex. 2.2.3 ABEL's theorem [1] 2042
2.3 Linear algebra and LCCH systems 2044
2.3 General discussion 2044
Ex. 2.3.1 Easy eig [1] 2058
Ex. 2.3.2 Self-closing door [2]. 2060
Ex. 2.3.3 Baby heat equation [2]. 2064
Ex. 2.3.4 Coupled tanks [2] 2068
Ex. 2.3.5 Baby wave equation [3] 2071
2.4 Multivariate stability analysis 2076
2.4 General discussion 2076
Ex. 2.4.1 Simple pendulum [1] 2080
Ex. 2.4.2 Lorenz model [2] 2083
Ex. 2.4.3 Brouwer saddle [3] 2087
End of Chapter 2 2092

Chapter 3 : Inhomogeneous initial value problems

3.1 Linear inhomogeneous equations 3002
3.1 General discussion 3002
Ex. 3.1.1 Undet coeff – uncomplicated case [1] 3007
Ex. 3.1.2 Another easy example [2] 3010
Ex. 3.1.3' Homogeneous resonance' [2] 3014
Ex. 3.1.4 Why undet coeff works [3] 3017
Ex. 3.1.5 Undet coeff – tricky cases [2] 3021
3.2 Euler Cauchy equations 3026
3.2 General discussion 3026
Ex. 3.2.1 Euler Cauchy equation [2] 3028
3.3 Modeling with LCC equations 3034
3.3 General discussion 3034
Ex. 3.3.1 Simplified induction motor [2] 3037
Ex. 3.3.2 The Gain function [0] 3041
Ex. 3.3.3 Vehicle on a bumpy road – unicycle model [2] . . 3047

Ex. 3.3.4 Biaxial bus [3] 3051
Ex. 3.3.5 Unsprung mass [3] 3056
Ex. 3.3.6 Lollipop LCR circuit [1]. 3061
Ex. 3.3.7 Vibration isolation [2]. 3063
Ex. 3.3.8 Beats and frequency measurement [2] 3066
Ex. 3.3.9 Rabi oscillations [2] 3070
Ex. 3.3.10 Resonance or no resonance [2] 3073
Ex. 3.3.11 Washing machine [2] 3080
Ex. 3.3.12 Sommerfeld effect [3] 3086
End of Chapter 3 3093

Review of trigonometric identities

Chapter 4 : Boundary value problems, Fourier series and periodic excitation

4.1 BVPs, discrete derivatives, dot products, matrix method . 4002
4.1 General discussion 4002
Ex. 4.1.1 Basic BVP [0] 4007
Ex. 4.1.2 Miscellaneous BCs [2]. 4010
Ex. 4.1.3 Inhomogeneous BVP [2] 4013
Ex. 4.1.4 Matrix method preliminaries [0] 4016
Ex. 4.1.5 First order BVP, matrix method [0] 4022
Ex. 4.1.6 Second order BVP, matrix method [2] 4026
Ex. 4.1.7 FREDHOLM's alternatives [3] 4030
4.2 Simple beam theory 4038
4.2 General discussion 4038
Ex. 4.2.1 Cute little cantilever [1] 4041
Ex. 4.2.2 Not so cute cantilever [3] 4043
Ex. 4.2.3 Point and line loads [3] 4047
Ex. 4.2.4 Cantilever by matrix method [3] 4053
Ex. 4.2.5 Helicopter fan blade [4] 4059
4.3 Engineering Fourier series 4069
4.3 General discussion 4069
Ex. 4.3.1 Square wave [1] 4073
Ex. 4.3.2 Waves with straight lines [1] 4077

Ex. 4.3.3 Chopper wave [2] 4080
4.4 Periodic excitation and matched BCs 4084
4.4 General discussion 4084
Ex. 4.4.1 Trampoline gymnast [3]. 4085
Ex. 4.4.2 Mathieu equation [3] 4089
Ex. 4.4.3 Chopper control of dc motor [2] 4097
Ex. 4.4.4. Direct torque control of induction motor [6] . 4101
End of Part 1 4114

PART TWO : PARTIAL DIFFERENTIAL EQUATIONS

Chapter 5 : The 1+1 Heat equation

5.1 Introduction; separation of variables 5002
5.1 General discussion 5002
Ex. 5.1.1 Good and bad PDE problems [2] 5011
Ex. 5.1.2 Simple sep var [0] 5015
Ex. 5.1.3 Complicated IC [0]. 5020
Ex. 5.1.4 Insulation BC [1] 5027
Ex. 5.1.5 Mixed BC [2] 5031
Ex. 5.1.6 Periodic IC [1] 5035
5.2 Discretization and matrix method 5038
5.2 General discussion 5038
Ex. 5.2.1 Baby and big heat equations [2] 5040
Ex. 5.2.2 Leakage BC [2] 5045
5.3 Inhomogeneous PDE problems 5050
5.3 General discussion 5050
Ex. 5.3.1 Inhomogeneous BC [2] 5052
Ex. 5.3.2 Free space GREEN's function [0] 5055
Ex. 5.3.3 GREEN's function approach [1] 5060
5.4 Modeling with the 1+1 heat equation 5064
5.4 General discussion 5064
Ex. 5.4.1 Cooking a steak [4]. 5069

Ex. 5.4.2 Oven coil [2] 5076
Ex. 5.4.3 Film boiling of glycerine [4] 5080
Ex. 5.4.4 Japanese paper pot [2]. 5089
Ex. 5.4.5 Hot and cold [4] 5093
Ex. 5.4.6 NEWTON's rule again [2] 5102
Ex. 5.4.7 Force-ventilated traction motor [3] 5105
End of Chapter 5 5113

Review of multivariate calculus

Chapter 6 : The Laplacian operator and related equations

6.1 Introduction to Laplace equation 6002
6.1 General discussion 6002
Ex. 6.1.1 Lollipop Laplace equation [0] 6004
Ex. 6.1.2 Variations on the 2D box [3] 6008
Ex. 6.1.3 Laplace in 3D [0] 6013
Ex. 6.1.4 Cylindrical polar [3] 6022
Ex. 6.1.5 Spherical polar [2] 6027
6.2 Many plus one heat equation 6031
6.2 General discussion 6031
Ex. 6.2.1 2+1 Heat equation [3] 6032
6.3 Static and dynamic electricity 6036
6.3 General discussion 6036
Ex. 6.3.1 EARNSHAW's theorem [2] 6051
Ex. 6.3.2 Free space GREEN's function [3] 6054
Ex. 6.3.3 Method of images [0] 6059
Ex. 6.3.4 Currents on a cylinder [3] 6065
Ex. 6.3.5 Current on a sphere [2] 6072
Ex. 6.3.6 Induction motor dynamic model [4] 6076
Ex. 6.3.7 Magnetic fridge ornament [4] 6087
Ex. 6.3.8 Magnetic helicopter [6] 6097
6.4 Steady state fluid mechanics. 6111
6.4 General discussion 6111
Ex. 6.4.1 Facilissimo flows [1] 6117
Ex. 6.4.2 3I Flow past a cylinder [2] 6120
Ex. 6.4.3 The Lift on a rotating cylinder [6] 6124

Ex. 6.4.4 Aeroelastic flutter [2] 6138
End of Chapter 6 6141

Chapter 7 : The Wave equation

7.1 Introduction to wave equation 7002
Ex. 7.1.1 Simple wave equation problem [2] 7004
Ex. 7.1.2 D'ALAMBERT's formula [0] 7009
Ex. 7.1.3 Vibrations of a drumhead [3] 7018
Ex 7.1.4 Heat or wave [2] 7024
7.2 Scenarios involving wave equation. 7031
7.2 General discussion 7031
Ex. 7.2.1 Sound wave [0] 7035
Ex. 7.2.2 Baby and big wave equations [2] 7038
Ex. 7.2.3 Vibrating cantilever [3]. 7044
7.3 Elementary quantum mechanics 7048
7.3 General discussion 7048
Ex. 7.3.1 Easy energy eigs [1] 7052
Ex. 7.3.2 Quantum harmonic oscillator [2] 7055
Ex. 7.3.3 Quantum free particle [3] 7058
7.4 Modeling with the wave equation 7064
7.4 General discussion 7064
Ex. 7.4.1 Stability transition [4] 7065
Ex. 7.4.2 Vibrations of an aircraft wing [3] 7072
Ex. 7.4.3 Catenary cable [3] 7081
Ex. 7.4.4 Stick-slip friction and the violin [6] 7089
Ex. 7.4.5 Resonance box [4] 7108
Ex. 7.4.6 Light interference [1] 7117
Ex. 7.4.7 Localizer – aircraft landing system [2] 7121
Ex. 7.4.8. Dipole radiation and magnetic levitation [6]. . . 7128
Ex. 7.4.9 Train jerks [3] 7142
End of the book 7154

Preliminaries

INTRODUCTION; NOTE ON THE EXAMPLE LEVEL RATINGS

Question : What is common to the following five situations ?

(a) A centrifuge started off with an improper distribution of samples inside its rotor starts oscillating violently at 10 percent of its rated speed and refuses to accelerate further, despite full power being applied by the motor.

(b) A gymnast jumping on a trampoline attains a height five times of what she would have if jumping on hard ground.

(c) A chef cooking a steak on a griddle flips the foodstuff every couple of minutes.

(d) A bar magnet rotating rapidly inside a copper cavity levitates at a stable equilibrium.

(e) A bridge exposed to wind performs very large amplitude oscillations even though other, apparently more fragile structures are completely unaffected.

Answer : All of them are governed by differential equations.

*

Differential equations are here, there and everywhere . It is a fact that the laws of nature (NEWTON, MAXWELL, SCHROEDINGER ……) are all differential equations, and many phenomena governed by empirical rules rather than fundamental laws (say dynamics of stock market) also depend heavily on differential equations . Hence the study of this topic is of invaluable importance.

So what is a differential equation ? Most generally, suppose there is a function f of one or more independent variables. **A differential equation for f is an equation which features the derivatives of f with respect to (w.r.t) these variables.** If f depends on one variable, we call it **ordinary differential equation (ODE)**. If f depends on more than one variable, we call it **partial differential equation (PDE)**.

For ODEs, we usually take the independent variable to be either a spatial coordinate x or a time t. In the former case, it is customary to denote the dependent variable by y (think of a curve in the plane) and the derivative by a prime thus $y' = \mathrm{d}y/\mathrm{d}x$. In the latter case, we usually call the dependent variable as x (think of the position of a moving body) and the derivative by a dot thus $\dot{x} = \mathrm{d}x / \mathrm{d}t$. A typical ODE thus looks like say $\dot{x} + x = 0$. The **order** of the ODE is the maximal number of derivatives present on any term, thus $\dot{x} + x = 0$ is first order while $\dddot{x} + \dot{x} + (\sin t)x = 1 + \cos t$ is third order. A typical PDE has derivatives with respect to either multiple spatial variables or space as well as time. Thus, it looks something like $\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} = \frac{\partial^2 z}{\partial t^2}$. (It's an interesting fact that the pronunciation of ∂ shows a lot of variation – for example in India most of us call it "del" while in USA one prefers plain "dee" or "partial".) Once again, the order of the PDE is the maximal number of partial derivatives present on any term.

To completely define an ODE problem, we need something more than the ODE itself. This is the **initial condition** or **boundary condition**. Likewise, to completely define a PDE problem we need **initial and/or boundary conditions** apart from the equation itself. The details of these, we'll see later. So much for what is a differential equation.

POLICY ON UNITS

Since there are a lot of physical problems in this book, units and dimensions play a big role. The only system of Units that we shall recognize and use is SI. The tremendous advantage of this system is that if all inputs to a calculation are in SI, then whatever the steps involved, the outputs will also be in SI with no effort expended on our part. Most problems will have all quantities specified in SI Units – in case there's a deviation (say speed given in km/hr), then we'll convert it to SI before actually starting the calculation.

THE LEVEL RATINGS OF EXAMPLE PROBLEMS

The 130 example problems are the backbone of this book. All carry one of six possible level ratings – Read Only (abbreviated to 0), 1, 2, 3, 4 and 6. The rating is an indicator of the difficulty you will encounter if you attempt the problem yourself (although all problems carry full solutions, attempts make for excellent practice). Read Only problems cover fundamental aspects of theory and are for learning rather than doing. The non-trivial ratings are best explained by analogy with the game of cricket. 1, 2 and 3 are the runs you get by literally scampering across the pitch – they are the problems which are directly relevant for Cornell 's MATH 2930, IITK 's MTH 203 N or equivalent course. A Level 3 is about the hardest problem which can reasonably be set in the exams of such a course, or slightly harder still. But, as the pianist-cum-music scholar HANS VON BULOW says, "in the study room the pianist must be able to achieve the double of what is required of him in the concert hall". Four is the boundary, a significant step up from the ground runs. These problems are typically at the level of pedagogical publications in a journal like American or European Journal of Physics. Six is the over-boundary, the ultimate big hit, typically at the level of a research publication (but always within our scope).

The number of problems of each level in the book is as follows :

Level	**0**	**1**	**2**	**3**	**4**	**6**	**Total**
Part 1 count	7	21	32	13	1	1	**75**
Part 2 count	8	6	18	12	7	4	**55**

Evidently the second Part is much harder than the first. Since fours and sixes are the highlights, I have marked the pages containing these problems with a grey box in the top corner. Nevertheless, watching the highlights alone is no substitute for watching the whole match, and there are quite a few interesting problems in the lower levels also. No material from a 4 or a 6 acts as prerequisite for a subsequent 1, 2 or 3.

Occasional examples carry "extra" questions after the regular ones. These either require knowledge (for example, of mechanics) beyond the scope of the course or cover an aspect which is considerably more subtle than the bulk of the problem and does not reappear later (except probably in high level problems). Extra questions do not contribute to the level rating. For some of the harder problems on the other hand, I have suggested easier alternatives which are more relevant for an elementary course. I have marked these as "ossia facile" and placed them before the main problem statement. I have also indicated the part of the main problem where you can find the solution to the ossia.

I have chosen the problems so as to avoid repeats – no two examples have exactly the same topic or scope. This enables us to cover the largest amount of material with the smallest number of problems. The emphasis on problems also makes this book a useful study material for graduate qualifying examinations in physics and engineering. Right now there are no drill type homework assignments here – I really don't have the time to make them up. If you would like to create a set of exercises for this book, you are most welcome to do so – please contact me at sb2344@cornell.edu or shayak.2015@iitkalumni.org.

SMALL PRINT

Citation policy : Although this book is free to download, read, distribute and share, academic usage of results contained here must be accompanied by a citation. The preferred citation format is :

B Shayak, "*Differential Equations – Linear Theory and Applications*," available electronically at www.shayak.in/Shayakpapers/DELTA/DELTA.pdf

Honesty of computer programming : Due to the number and length of computer programs involved in generating the solutions, I have found it impossible to include codes as part of the book. I have made every possible effort to ensure that all codes perform exactly the tasks described. Any departure is unintentional and is the result of an error. If you would to like to see the codes, either for verification or for educational purpose, please contact me at sb2344@cornell.edu or shayak.2015@iitkalumni.org.

No advertisements : There are occasional references within this book to products and services which are offered commercially by government or private agencies. Such references are not, and are not to be construed as, advertisements. I have received no incentive, financial or otherwise, from anyone to mention these products and services in this book.

Formatting errors : This 'A' edition is derived from the original edition by editing the PDF file and not the source files. This has resulted in certain formatting errors in this edition, which are completely absent in the original. I apologize for these blemishes.

That's enough of legal prattle, now let's see what a model example looks like.

Example

Demo example

For each of the following five mathematical statements, determine whether it is a differential equation. If yes, write whether it is an ODE or a PDE and state its order.

(a) $\dot{x} + x = 0$

(b) $ax^2 + bx + c = 0$

(c) $\dfrac{\partial^2 V}{\partial x^2} + \dfrac{\partial^2 V}{\partial y^2}$

(d) $\dot{u} = 9u'' + 16$

(e) $\left(\dfrac{\mathrm{d}^5 y}{\mathrm{d}x^5}\right)^2 + y\dfrac{\mathrm{d}y}{\mathrm{d}x} = \cos x$

[Read Only]

(a) YES. This is an ODE as it has $\mathrm{d}x/\mathrm{d}t$ equalling some function of x. The highest derivative is the first, so this is first order.

(b) NO. This is an algebraic equation, not a differential one.

(c) NO. This has partial derivatives but is an expression, not an equation.

(d) YES. This is a PDE as it is an equation featuring derivatives of u with respect to both space (prime) and time (dot). The highest derivative is the double prime (second spatial) so it's second order.

(e) YES. This is an ODE as it has something involving $\mathrm{d}^5y/\mathrm{d}x^5$ equaling stuff involving lower derivatives and the functions themselves. The highest derivative is the fifth, so the ODE is fifth order (NOT tenth – the square doesn't count!).

*

COMMENTS : Often though not always, the solution to an example problem is followed by a commentary, like this one. The most important display equations in each section are labelled by a two digit number.

This page intentionally left blank

PART ONE

****

Ordinary

Differential

Equations

1

FIRST ORDER ORDINARY DIFFERENTIAL EQUATIONS

This Chapter covers the theory and applications of first order equations. Some of the solution methods are specific to first order ODEs, although techniques generalizable to higher order systems have been used wherever possible.

1.1

FIRST ORDER NON-GENERIC ODES

Here we will see four types of first order ordinary differential equations (ODE) which are quite simple and are non-generic in that the concepts are not extendable to higher order ODEs. The four types are separable, exact, $f(y/x)$ form, and integrating factors form. To make a well-defined problem, every first order ODE must come with **one initial condition (IC)** which specifies the value of dependent variable at one particular value of the independent variable.

SEPARABLE

This is an ODE of the form

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{g(t)}{f(x)} \quad ,$$

where f and g are some functions. To solve it, we can treat the differential like a fraction to write

$$f(x)\mathrm{d}x = g(t)\mathrm{d}t \quad .$$

Then integrate the both sides to get

$$F(x) = G(t) + C \quad ,$$

where F and G are antiderivatives of f and g. The arbitrary constant C gets determined from the IC. This process is called **transfer of variables** (the phrase "separation of variables" is also used sometimes but it also refers to another, far more complicated process in PDE theory so I prefer this term).

A slight variation is when you are given some initial data point (t_0,x_0) and asked for x_1 at some point t_1 (or the other way round). Then you must use the fact that the integral on f runs from x_0 to x_1 and that on g from t_0 to t_1, thus

$$F(x_1) - F(x_0) = G(t_1) - G(t_0) \quad .$$

This completes the theory of this type of equation.

EXACT

A differential equation of the form

$$M(x,y)\,\mathrm{d}x + N(x,y)\,\mathrm{d}y = 0 \quad ,$$

where M and N are functions of x and y is called **exact** if there exists a function $\Phi(x,y)$ such that the left hand side (LHS) the above equation is the total differential $\mathrm{d}\Phi$. To test whether an equation is exact you should verify whether

$$\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x} \quad .$$

If the equation passes this test then it is exact (the test comes from equality of crossed partial derivatives on Φ). If the ODE is exact then first integrate M w.r.t x, adding an arbitrary function of y, next integrate N w.r.t y adding an arbitrary function of x, and finally match the functions to get $\Phi(x,y)$. The solution is

$$\Phi(x,y) = C \quad ,$$

where C is an arbitrary constant which you must determine from the IC.

f(y/x) FORM

Here the ODE takes the special form

$$\frac{\mathrm{d}y}{\mathrm{d}x} = f\left(\frac{y}{x}\right) \quad ,$$

for some function f. It is sometimes called "homogeneous," a term I don't like because it means a completely different thing in a far more general context (coming up later). Here, the trick is to replace the dependent variable y with a **new dependent variable** $u(x)$ where $u = y/x$. This substitution converts the original equation to

$$x\frac{\mathrm{d}u}{\mathrm{d}x} = -u + f(u) \quad ,$$

which is separable and can be solved by transferring variables. An example of such an equation is $\mathrm{d}y/\mathrm{d}x = y^2/x^2 + y/x$.

INTEGRATING FACTOR

The ODE for this type of problem has the form

$$A(x)\frac{\mathrm{d}y}{\mathrm{d}x} + B(x)y = C(x) \quad ,$$

for some functions A, B and C. This is called a **linear** ODE. To solve it, we first divide through by A on both sides, getting

$$\frac{\mathrm{d}y}{\mathrm{d}x} + P(x)y = Q(x) \quad . \tag{01}$$

This is the canonical form of this equation, and you **must** bring it to this form before solving.

The trick is to multiply the both sides of (01) by some function $\mu(x)$ such that the LHS, $\mu y' + \mu P y$, becomes a total derivative of the function μy and the equation becomes $\mathrm{d}(\mu y)/\mathrm{d}x = \mu Q$. Here μ is called **integrating factor**. Since $(\mu y)' = \mu y' + \mu' y$, equating this to the LHS of the ODE shows that $\mu y' + \mu P y = \mu y' + \mu' y$, whereby μ must satisfy

$$\mu' = P\mu \quad .$$

Solving this by transferring variables and then integrating both sides gives the integrating factors formula

$$\mu(x) = \exp\int^{x} P(s)\mathrm{d}s \quad . \tag{02}$$

This immediately gives the solution of the ODE as

$$\mu y = \int^{x} \mu(s)Q(s)\mathrm{d}s + C \quad , \tag{03}$$

where C is an arbitrary constant.

It is pedantically correct to write (02) and (03) as I did above, with the integration being carried out over a dummy variable s and the actual variable x being the upper limit; in most cases however the more informal expressions

$$\mu = \exp\int P(x)\mathrm{d}x \quad , \text{ and} \tag{04a}$$

$$\mu y = \int \mu Q \ \mathrm{d}x + C \quad , \tag{04b}$$

are sufficient. It is a good idea to be familiar with the more fancy notation however because in a later stage of our course we'll encounter one problem where it will be essential to write the integrals that way.

You might wonder at linear ODE also being classified under non-generic – surely "linear" is one of the words in the book title. But it's the solution method which makes the thing non-generic. In the integrating factors formula we didn't use any of the fundamental implications of linearity but merely exploited the happy accident that the LHS could be written as the derivative of a product.

NEWTON'S LAWS OF MOTION

Sir ISAAC NEWTON has presented us the following laws of motion.

First Law : An inertial frame is defined to be one where if a body is in the state of rest, it will remain in the state of rest, and if it is in the state of motion, it will continue moving in the same direction with the same speed unless an external force is applied upon it.

Second Law : The rate of change of momentum of a body is directly proportional to the force applied on it, and the change in momentum is in the direction in which the force is applied. That is, $\mathbf{F} = m\mathbf{a}$.

Third Law : To every action there is always an equal and opposite reaction.

The above 'modernized' presentation of the First Law should nip in the bud the foolhardy conception that it is just a trivial special case of the Second Law. This latter law readily gives rise to differential equations if we write acceleration as the second derivative of position or the first derivative of velocity, whichever is appropriate. Since this is not a mechanics book, we'll have only elementary force balancing problems here – let's open the examples with one such problem.

Example 1.1.1

Train acceleration

A certain train can be modelled as a mass of 1130 tons which is subject to the traction of the locomotive and no other forces. The loco has a maximal power rating of 6000 HP, which is used during acceleration . Note that power equals tractive force times train velocity.

(a) Write an ODE for the velocity v of the train as a function of time.

(b) Determine the time taken for the train to accelerate from 76 to 130 km/hr.

[Level : 1]

(a) We have force equals mass times acceleration so the ODE is $m\dot{v} = F$. We are also given the form of F – it equals the power P divided by v. So

$$\frac{\mathrm{d}v}{\mathrm{d}t} = \frac{P}{mv} \quad .$$

As per our Units policy, we must use SI Units for all physical quantities. For power, the SI Unit is Watt so we must multiply the HP rating by 746 – doing so we get

$$\frac{\mathrm{d}v}{\mathrm{d}t} = \frac{3.96}{v} \quad ,$$

which is the ODE asked for.

(b) This ODE is clearly separable, so we solve by transferring variables. The initial point is (t_0, v_0) with v_0 given – we can take the initial time to be zero but any value does. The final point has v_1 given and t_1 is what we wish to find. We have

$$\int_{v_0}^{v_1} v\mathrm{d}v = 3.96 \int_{t_0}^{t_1} \mathrm{d}t \quad ,$$

where v_1 means v_1 etc. The underscore notation, alla Latex, for the subscript when already appearing in a sub- or superscript position might be a bit unconventional but it is easy on the eye especially in printed editions. Doing the integration,

$$t_1 - t_0 = \frac{v_2^2 - v_1^2}{7.92} \quad .$$

Note that the arbitrary initial time t_0 has gone out, leaving behind only the difference which has physical significance.

Converting the values for v_0 and v_1 into SI Units, we find that they equal 21 and 36 m/s respectively; plugging these in, the time taken for the acceleration works out to 1 min 48 s. (It is always a good idea to underline, box or otherwise highlight your answer especially if it is a number or a value. When the examiner checks your copy, the first thing s/he looks for is the answer – if THAT is correct (and there is no obvious sign of your having cheated), then s/he very likely gives you full marks without bothering about the nitty-gritties of your steps.)

*

COMMENTS : The actual time taken by the train for this acceleration is approximately 2 min 55 s. The error here is huge , so the model must be incorrect in a fundamental way. Indeed, it is so. The flawed assumption is that there is no other force on the train other than the locomotive's traction. As per this model, the train will maintain a constant speed if the loco idles. In reality, friction between train and rail is considerable – the power required to maintain 130 km/hr on the train is about 3000 HP. Consistent with the nature of the error, the calculated answer is less than the actual value . Because this was our first problem , I troubleshot the error myself – in subsequent problems however you will have to analyse models and their plausibility on your own.

Here we substituted the numerical values into the equation before solving it (the equation of Part (a) had the coefficient 3.96). We could also have kept the ODE in terms of algebraic parameters, solved it and then plugged in the numbers at the end. The advantage of our process is that the equations look simpler with numbers than with combinations of letters (this is not always true but does hold in this case). Moreover, if we have to solve things numerically, then letter values are useless. Since numerical values are dependent on Units however, we must be careful to stick to SI Units only, all the time.

Example 1.1.2

Exact for the exam

Solve the ODE

$$\frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{6xy + y^2}{3x^2 + 2xy + \cos 2y} \quad .$$

The answer may be left in implicit form.

[Level : 1]

Of course, the question tells us in this case that the equation is exact. In general how do we guess ? We do it by ruling out the other possibilities. The given ODE is obviously not separable and it is not in the integrating factor form. So let's try exact and write the ODE in the form $M\,\mathrm{d}x + N\,\mathrm{d}y = 0$, getting

$$\underbrace{\left(6xy + y^2\right)}_{=M(x,y)}\mathrm{d}x + \underbrace{\left(3x^2 + 2xy + \cos 2y\right)}_{=N(x,y)}\mathrm{d}y = 0 \quad .$$

Next comes the exactness test : check whether $\partial M/\partial y = \partial N/\partial x$. In this case, both equal $6x + 2y$ so the equation is indeed exact (hardly a surprise).

After the test passes, define the function $\Phi(x,y)$ to be $\int M\ \mathrm{d}x$ plus some arbitrary function of y :

$$\Phi = 3x^2 y + xy^2 + f(y) \quad . \tag{01}$$

But Φ must also be $\int N\ \mathrm{d}y$ plus some arbitrary function of x :

$$\Phi = 3x^2 y + xy^2 + \frac{1}{2}\sin 2y + g(x) \quad . \tag{02}$$

The two definitions (01) and (02) are for the SAME Φ, so we must be able to match them. Indeed, if we choose $f(y) = (1/2)\sin 2y$ and $g(x) = 0$, then the two definitions are consistent. Hence we must have

$$\Phi(x,y) = 3x^2 y + xy^2 + \frac{1}{2}\sin 2y \quad . \tag{03}$$

Note again that in the testing step it's the "crossed" partial derivatives – M with respect to y and N with respect to x, but in the integration step it's the "direct" integrals – M with x and N with y.

In an exam, many students put box around (03) and label it as answer. Then they are dumbstruck when the copy comes back with 50 percent marks awarded for the problem. This happens because the question asks for a solution of the ODE, which is y as a function of x (or at least a relation between x and y which leads to a curve in the x-y plane). The question does NOT ask for a function $\Phi(x,y)$ which satisfies such and such properties ! Completing the solution from (03) is a trivial but extremely important step – the solution is $\Phi(x,y) = C$ where C is arbitrary constant. Thus we have

* $$3x^2y + xy^2 + \frac{1}{2}\sin 2y = C \quad ,$$ *

which is in implicit form and is the desired answer.

*

COMMENTS : Many times, as in here, the solution of exact equation is extremely tedious or impossible to express as y equals such and such function of x. So the final answers often stay in implicit form, irrespective of whether the question mentions it explicitly or not. If an IC is given, then C comes out – for example if we were told that $y(0) = \pi$, then we can easily substitute $(x,y) = (0,\pi)$ into our solution and find that $C = 0$.

Suppose you jump to the Φ-finding stage without checking for exactness, and try to hack it on an equation which is not exact – then what happens ? What happens is that the two expressions (01) and (02) become fundamentally inconsistent and it is impossible to match them for any choices of f and g. As an example, try this problem with the '6' on xy replaced by 5. Then there will be a 5/2 instead of 3 in (01) while (02) will remain unchanged – it is not possible to mesh them together into a single Φ under any circumstances whatsoever.

Finally, why is this problem "exact for the exam" ? Because unfortunately, that's about the only place where exact equations occur. Separable equations are reasonably common in science and engineering, integrating factor also makes an occasional appearance but till date I have not seen a physical process etc being governed by an exact equation or one of f(y/x) form. By the way, some problems of this latter form also happen to be exact or solvable by integrating factor, and that solution is invariably quicker. The integrations in f(y/x) often tend to be very messy, so keep that approach as a last resort only when all else fails.

Example 1.1.3

Integrating factor

Solve the following two ODEs :

(a) $xy' + 2y = 3x$

(b) $\dot{x} + \left(1 + a\cos\omega t\right)x = 0$

[Level : 1]

(a) We first bring it into canonical integrating factor form by dividing through by x. Hence our canonical ODE is

$$y' + \frac{2}{x} y = 3 \quad .$$

The integrating factor is $\mu = \exp\left[\int (2/x)\mathrm{d}x\right]$, which is $\mathrm{e}^{2\log x}$, which is $\exp\left(\log x^2\right)$ which is x^2. Please be careful here : writing it as $2x$ is a common, tempting and costly mistake ! The exponential and logarithm can cancel only when they are right next to each other ! Then the solution formula says

$$x^2 y = \int 3x^2 \mathrm{d}x + C \quad \text{, or}$$

$$* \qquad y = x + \frac{C}{x^2} \quad , \qquad *$$

which is the desired answer.

(b) This is already in standard form with $P(t) = 1 + a\cos\omega t$ and $Q(t) = 0$. In the exams, please be careful whether the problem is in terms of $x(t)$ or $y(x)$ or some other variable combination – a mistake here might not be costly from a marks viewpoint but creates a negative impression on the examiner. The integrating factor is $\mu = \exp\int (1 + a\cos\omega t)\,\mathrm{d}t$ and the solution is $\mu x = C$ or $x = C/\mu$. This gives

$$* \qquad x = C\exp\left(-t - \frac{a}{\omega}\sin\omega t\right) \quad , \qquad *$$

which is the desired answer.

*

COMMENTS : Here we used the more informal notation of 1.1.04) instead of the pedantic (1.1.02-3). As I already said, this notation is perfectly acceptable for problems like this, unless your instructor is a mathematician and martinet.

Part (b) shows you how non-generic the integrating factor method really is. As it was, the problem could be solved almost trivially – the solution is a boring function which goes to zero at large time, when the $-t$ inside the exponential dominates the sinusoidal term. But change that dot on the x to a doubledot and all hell breaks loose. The resulting equation is a very famous one, called Mathieu equation (after EMILE MATHIEU who first studied it). Although it is still linear

(just take that from me now, definition coming in the next Unit), it CANNOT be solved exactly at all. The solutions of that equation can either remain bounded or blow up to infinity, and it requires all sorts of approximations and fancy manipulations to find out the regions in the a-ω plane where the two different behaviours occur.

Example 1.1.4

Train again

Again consider the train from Example 1.1.1. Train mass (1130 tons) and loco traction (P/v) remain as they were; this time one is given the extra information that the train experiences a frictional force given by the formula

$$F_D = \frac{m}{1000}\left(10 + 1.0v + 0.012v^2\right) \quad ,$$

where m is the train mass and everything is in SI Units.

(a) Taking $P = 6000$ HP, write an ODE for the velocity v of the train as a function of time.

(b) Solve the above ODE using a computer and hence determine the time taken for the train to accelerate from 76 to 130 km/hr (21 to 36 m/s).

(c) Determine the power P which must be exerted by the loco for the train speed to remain constant at 130 km/hr.

Suppose that the loco is exerting the power obtained above and the train is cruising at 130 km/hr, when a temporary down-gradient pushes up the speed to 132 km/hr.

(d) Qualitatively describe the subsequent temporal behaviour of the train speed.

[Level : 2]

(a) There are two forces acting on the train – traction F_T and friction F_D (D for drag). Traction is positive since it is in the direction of motion while friction is negative since it is opposite to the direction of motion. Thus, NEWTON's Second law looks like $m\dot{v} = F_T - F_D$. Substituting the dependences given in the question, we have

$$m\frac{\mathrm{d}v}{\mathrm{d}t} = \frac{P}{v} - \alpha - \beta v - \gamma v^2 \quad , \qquad *$$

where $m = 1130000$, $P = 4476000$, (in Watts!) $\alpha = 11300$, $\beta = 1130$ and $\gamma = 1130\text{x}0.012 = 13.56$ SI Units.

(b) The equation is separable so we transfer variables. We have

$$m\frac{\mathrm{d}v}{\frac{P}{v} - \alpha - \beta v - \gamma v^2} = \mathrm{d}t \quad .$$

Integrating the LHS from $v_1 = 21$ to $v_2 = 36$ and the RHS from t_1 to t_2 gives the acceleration time as

$$t_2 - t_1 = m\int_{21}^{36} \frac{\mathrm{d}v}{(4476000/v) - 11300 - 1130v - 13.56v^2} \quad .$$

This integral is not analytically doable so the computer comes to the rescue.

We evaluate the integral numerically using the rectangle method (the crudest). For this, we split the v interval into a large number N (say 1000) of lattice points equally spaced between 21 and 36. Then the step size h becomes $(36–21)/(N–1)$. The required integral is the function of v evaluated at each lattice point, multiplied by h and summed over all lattice points. We try different values of N, increasing it until the result no longer changes appreciably. The largest N I've gone for is 1000; the time then evaluates to 2 min 58 s.

(c) For the train speed to remain constant, $\mathrm{d}v/\mathrm{d}t$ must be zero (obviously). Hence the forces on the train should balance, with traction being equal to drag at a speed of $v = 36$. We thus have

$$\frac{P}{36} = \alpha + 36\beta + 1296\gamma \quad ,$$

which we can solve, divide by 746, and get the loco power to be 3357 HP.

(d) We can see that for constant P the traction decreases with increasing speed while the friction always increases with speed. So, at 132 km/hr, friction exceeds traction and the resultant force is negative. This will cause the speed to reduce until it returns to the balancing point of 130 km/hr. Likewise, if a temporary up-slope (or a gust of wind) brings down the speed to say 129, traction will exceed friction and again cause the train to accelerate back up to MPS.

*

COMMENTS : The parameters α, β and γ in this problem are derived phenomenologically and not from a theory – people have got those values after performing measurements on actual trains and then fitting the formula to the data. Indeed, getting their values to even the correct order of magnitude from an ab initio calculation would be a hair-raising exercise. In fact, there is some variation in these parameters even from run to run (for example, a headwind can increase the value of γ), which results in minor fluctuations of the observed acceleration time. The parameter and time values however are always close to what we used and found here.

This becomes the first problem in this book where the computer comes in handy. Indeed, the machine will be a considerable help throughout the book. So if you are unfamiliar with any computer math package (or programming language), now is a good time to parallelly register for a course which teaches elementary computer applications in science and engineering. Matlab is the software which I use, and is quite a popular choice. Mathematica and Maple are equivalent programs. In all these packages, the user's workload is the least. In programming languages (as against packages) like C++ the user apparently gains greater flexibility at the cost of programming difficulty. I am technically not competent to express any more definite or in-depth opinion on this topic.

While first writing the book, I had posed this problem in a different form, where one had to evaluate the integral analytically, using an approximation. However, in today's age it is self-limiting to dogmatically ignore all the calculational conveniences which computers bring. Rather, what we'll try to do here is use the machine liberally but intelligently – rely upon it for mechanical calculations but never use it as a substitute for learning concepts.

Let me also assure you that this book will NOT become a course on numerical methods or worse, assume knowledge of those as a prerequisite. In all cases, we'll either be using only the simplest and most conceptually transparent numerical procedures, such as the rectangle method above, or using the computer as a blackbox, as in eigenvalue evaluations in later Units. Fortunately we won't have to worry about the time taken for the calculations, since the routines will all be pretty simple (the subtleties will lie entirely in the analytical parts). And as regards the accuracy, we'll forever be alert so that we aren't duped into accepting bad answers from the machine.

In Part (d) here I have introduced to you a concept which has long-term implications. I'll not elaborate too much here, but what we saw is that the train comes back to MPS if it is disturbed slightly from there. This need not always have happened. Hypothetically, suppose that traction and friction reverse roles – let the traction be the quadratic polynomial and the friction be $1/v$ (both these functions are physically nonsensical but still). Then also, exactly 130 km/hr would be a balancing speed for the train. But if a down-slope pushed it up to 132, then traction would exceed friction and the train would continue accelerating further in a runaway manner. Likewise, if a gust of wind brought it down to 128, then friction would outpace traction and bring the train headlong to a standstill. For now, just ponder on these two possible situations – we'll formalize the concepts in due course of time.

1.2

LINEAR ODES – PROPERTIES OF SOLUTIONS

After a sneaky guest appearance in the last Unit, linearity now makes its grand entrance onto the stage accompanied by the inevitable horn fanfares. One of the most important features of an ODE of **any** order (the definition holds for PDE also but let's not go into that here) is whether it is linear. The definition of linearity is :

In a linear ODE, the DEPENDENT VARIABLE and ALL its derivatives appear to FIRST OR ZERO POWER alone, with the coefficients being ARBITRARY functions of the independent variable.

I cannot over-estimate the importance of this concept. As we will see, linearity implies a number of statements holding true for the equation, and these often enable us to solve it. In a way, linearity is synonymous with solvability – it's not that every linear equation is solvable (far from it) but if your equation is linear then at least there's a chance; if your equation is nonlinear, the odds are less than one in a thousand (or a lakh) that you can come up with an explicit solution.

Before going further, a few quick examples, all of which are in terms of the variable $x(t)$ with no loss of generality.

$$\dot{x} + 5x = 0 \quad :$$

This is obviously linear – $\dot{x}$ and x both appear to first power.

$$\ddot{x} + 5x = t^2 \quad :$$

This is linear again – the order of the derivative is 2 but that makes the equation **second-order**, NOT nonlinear. The t^2 on the RHS looks like a nonlinearity but that is a function of the independent variable alone and it can be arbitrary.

$$\ddot{x} + 2x\dot{x} = \cos t \quad :$$

This is nonlinear – there's a $x\dot{x}$ term which is a product of dependent variable and its derivative. That's not allowed. Note however that the $\cos t$ term is not an offender – function of the independent variable alone.

$$\ddot{x} + (1 + \cos \omega t)\, x = 0 \quad :$$

This is Mathieu equation which I mentioned in the last Tute. Check against the definition – it's linear despite the $\cos \omega t$ multiplying x.

The big bold definition of linear ODE above is actually one of two ways we can define linearity. It is, in a way, a working definition, in the sense that it tells you what a linear ODE looks like. There's another way of defining linearity which focusses more on the **implications** of linearity than on the appearance. This definition is more mathematically subtle but it is helpful for a better understanding of the concepts and is essential for the more linear algebraically oriented components of this book, which will occur quite often in Chapter 4 onwards.

For this definition, we write the ODE in the following form : we collect everything featuring x in the LHS and everything featuring t alone on the RHS, which we call $f(t)$. Then, this ODE has the form that some derivatives etc act on the function $x(t)$ and produce the new function $f(t)$. We think of these combinations of derivatives etc as an operator $\mathbf{O}$ which converts the function x into the function f. Thus, the ODE looks like $\mathbf{O}x = f(t)$. Now, the operator $\mathbf{O}$ is actually a linear operator $\mathbf{L}$ if and only if it satisfies the three following basic properties :

- $\mathbf{L}0 = 0$
- $\mathbf{L}(cx) = c\mathbf{L}x$ for all constants c
- $\mathbf{L}(x_1 + x_2) = \mathbf{L}x_1 + \mathbf{L}x_2$ for all x_1 and x_2.

These three properties are the A-B-C of linearity – any kind of transformation must satisfy all these properties to be classified as linear.

In operator notation, the letter D is often used to denote differentiation w.r.t the independent variable. Thus, D denotes $\mathrm{d}/\mathrm{d}t$ here, D^2 denotes $\mathrm{d}^2/\mathrm{d}t^2$ and so on. The linear equations in the set I presented above can be written as

$$(D + 5)\, x = 0 \quad ,$$

$$\left(D^2 + 5\right) x = t^2 \quad ,$$

$$\left(D^2 + 1 + \cos\omega t\right)x = 0 \quad .$$

Now another definition : a linear ODE $\mathbf{L}x = f(t)$ is **homogeneous** if $f(t)$ is identically zero, and is inhomogeneous if it is not. Hence in the above examples, the first and third equations are homogeneous while the second is not.

These definitions allow us to introduce a few very important properties of linear ODEs. We again recall that the entire discussion of this Lecture holds true for linear ODEs of **all** order, not just first order.

PROPERTY 1

An n^{th} order linear homogeneous ODE $\mathbf{L}x = 0$ has ***n* fundamental solutions** $x_1(t)$, $x_2(t)$, ….., $x_n(t)$ which are **linearly independent** of each other i.e. no linear combination of any two or more of the fundamental solutions equals any other fundamental solution. Any arbitrary linear combination of these fundamental solutions, i.e. the expression $C_1 x_1(t)+C_2 x_2(t)+\ldots..+C_n x_n(t)$ where C_1, C_2, ….., C_n are arbitrary constants, is also a solution of the ODE. The C's here can be determined from the IC.

Evidently, one function x which always satisfies $\mathbf{L}x = 0$ is the solution $x = 0$. This solution is an immediate consequence of the first characteristic of linearity, and is useless to human or beast. It is called the **trivial solution**.

PROPERTY 2

The solution of an inhomogeneous ODE $\mathbf{L}x = f(t)$ where f is non-zero, and some ICs are given can be decomposed into two parts : a **particular solution** $x_p(t)$ which satisfies $\mathbf{L}x_p = f(t)$ but does not satisfy the IC and a **homogeneous solution** $x_h(t)$ which satisfies $\mathbf{L}x_h = 0$ and can be used to fit the ICs. The combined solution is the sum of the homogeneous and particular solutions.

These properties are completely general. Both of them follow from the three-point definition of linearity of $\mathbf{L}$. As a first step, we combine the second and third features of the definition to see that $\mathbf{L}\left(ax_1 + bx_2\right) = a\mathbf{L}x_1 + b\mathbf{L}x_2$, for all functions x_1 and x_2 and constants a and b. Then, to derive Property 1, if $\mathbf{L}x_1=0$ and $\mathbf{L}x_2=0$ then $\mathbf{L}(c_1x_1+c_2x_2)$ must be zero also. For Property 2, we observe that $\mathbf{L}(x_h+x_p) = \mathbf{L}(x_h)+\mathbf{L}(x_p) = 0+f(t)$, hence the general solution is indeed the sum of the homogeneous and particular components.

Now we come to another solution property which holds true when the equation is homogeneous and the operator $\mathbf{L}$ features only constants and $\mathrm{d}/\mathrm{d}t$'s – no functions of t. These equations are called **linear constant coefficient homogeneous (LCCH)** ODEs. These, together with their inhomogeneous equivalents (which we'll call LCC ODEs, no 'H'), are by far the most important ODEs which arise in practical applications.

PROPERTY 3

The solution of an LCCH ODE acquires the **exponential form** $x = \mathrm{e}^{\lambda t}$ where λ is unknown. Plugging this form of the solution into the ODE yields the **characteristic equation** or **characteristic polynomial** and therefrom the values of λ. This process works because $\mathrm{e}^{\lambda t}$ produces constant multiples of itself upon differentiation.

Only now, we specialize to first order ODEs. Here, the characteristic polynomial always has a single real root – the upcoming Tute shows what to do with it. For higher order ODEs there will be more λ's, and how to handle them, we'll see when the time is appropriate.

Example 1.2.1

LCC lollipop

Solve the LCCH ODE

$$\dot{x} + \gamma x = 0 \quad ,$$

where γ is some positive real number. Basis that result, solve the inhomogeneous ODE with IC

$$\begin{aligned} \dot{x} + \gamma x &= b \\ x(0) &= x_0 \end{aligned} \quad .$$

Discuss some salient features of the solution. Compare with the solution of

$$\dot{x} - \gamma x = b \quad ,$$

starting from the same initial condition.

[Read Only]

The equation $\dot{x}+\gamma x=0$ is LCCH so invoke Property 3 of the Lecture and try $x=\mathrm{e}^{\lambda t}$; plugging this into the ODE gives $\left(\lambda+\gamma\right)\mathrm{e}^{\lambda t}=0$. Now, $\mathrm{e}^{\lambda t}$ is not the zero function so $\lambda+\gamma = 0$. This is a polynomial equation for λ; we call it the CHARACTERISTIC EQUATION. The solution of this simple equation is $\lambda = -\gamma$. Hence the fundamental solution of the ODE is

$$x=\mathrm{e}^{-\gamma t} \quad .$$

Now use Property 1 : an arbitrary linear combination of the fundamental solutions of a homogeneous ODE is itself a solution. For a first-order ODE, arbitrary linear combination simply means multiplication by an arbitrary constant. Hence the solution of the ODE is $\underline{x = C\mathrm{e}^{-\gamma t}}$, where C is an arbitrary constant. We can easily check by direct substitution that this is indeed a solution of the ODE.

Now we consider the case where the RHS is the constant b instead of zero. This makes the ODE inhomogeneous. So, using Property 2 we can write its solution as $x=x_h+x_p$ where x_h satisfies $\dot{x}_h+\gamma x_h=0$ and x_p satisfies $\dot{x}_p+\gamma x_p=b$ without the initial condition. We have already seen what x_h is – it is $C\mathrm{e}^{-\gamma t}$. As for x_p, we just guess it in this case – it is the constant function $x_p=b/\gamma$. Clearly, its derivative is zero so it algebraically satisfies the ODE. So we have

$$x_p=\frac{b}{\gamma} \quad .$$

Thus the GENERAL SOLUTION of the ODE can be written as $x_G=x_h+x_p$, which is

$$x_G=\frac{b}{\gamma}+C\,\mathrm{e}^{-\gamma t} \quad .$$

Note the term : general solution means the particular solution plus the homogeneous solution with all arbitrary constants in place.

Finally use the arbitrary constant to fit the IC : setting $t=0$ into the general solution gives $b/\gamma+C=x_0$ whereby $C=x_0-b/\gamma$. Hence our solution is

* $$x=\frac{b}{\gamma}+\left(x_0-\frac{b}{\gamma}\right)\mathrm{e}^{-\gamma t} \quad . \qquad (01)^*$$

This solution, which satisfies the ODE as well as the IC, is called the SPECIFIC SOLUTION.

We could have derived this solution by a number of other ways. For one, the equation is separable and we could have transferred variables. For another, we could have used integrating factor. In fact, the conventional approach here is to use either of those two methods, and not this. Then why did I do it this way ? That's because, we keep at the back of our minds the possibility that the dot on x might easily become a doubledot. Forthwith, variable transfer, integrating factor and all the other methods will vanish, leaving behind only the approach we used here. You challenge : "Hold your generality talk – you just GUESSED that particular solution ! What kind of generic procedure is that ?" Fortunately however, this challenge is baseless. For LCC equations there is a very standard and fail-safe algorithm for generating particular solutions, called the method of undetermined coefficients. For such an easy problem I won't go into the details, but let me just say that what I did here is an elementary application of that method. If you aren't convinced, look ahead to 3.1.

Several features of the solution (01) are noteworthy. First is the EXPONENTIAL DECAY of the homogeneous solution. This happens on account on the sign on γ. Physically, $\dot{x} = -\gamma x$ means that if x is positive to begin with, it decreases in time, so it decays. The second feature is that at large times ($t \to \infty$ in theory and γt of the order of 3 or 5 in practice) $x(t)$ becomes constant with value b/γ. Since this solution does not change in time, it is called the STEADY STATE solution. In terms of the steady state solution, we can write the complete solution as

$$x = SS + (IC - SS)\mathrm{e}^{-\gamma t} \quad ,$$

where SS and IC of course denote steady state and initial condition. This is a good (and often useful) way to remember the solution of this equation. The exponential behaviour means that the decay has a time constant, which is $1/\gamma$.

If the sign on γ had been negative, then the solution of $\dot{x} - \gamma x = b$ would have been

$$x = -\frac{b}{\gamma} + \left(x_0 + \frac{b}{\gamma}\right)\mathrm{e}^{\gamma t} \quad .$$

Here the homogeneous solution GROWS exponentially. This solution blows up to infinity as time goes on (reaching $x = \infty$ at $t = \infty$ in theory and a very large value at $\gamma t = 10$ or so in practice). Physically, $\dot{x} = +\gamma x$ means that if x is positive to begin with, it keeps increasing, so it grows unboundedly.

The particular solution $x_p = -b/\gamma$ would technically still be a steady state solution because if $x = x_p$ then $\dot{x} = 0$. However you would not be able to see this solution because instead of converging to it, the system diverges from it in time. Suppose you set up a laboratory experiment where the position of a mass follows $\dot{x} - \gamma x = b$ and you try to start it off with the IC exactly equal to the particular solution i.e. $x(0) = x_0 = x_p = -b/\gamma$. Then too, you will never see the mass settling at $x = x_p$. In the lab there will always be a deviation from x_p, either due to the limited accuracy of setting up the IC or from a passing disturbance etc. and this will excite the homogeneous solution. Because of the exponential growth there, even a minuscule error will take just seconds to amplify and become huge. So in this case it is better to call x_p just the particular solution and not a steady state solution.

Note that $\dot{x} + \gamma x = f(t)$ occurs very commonly when modeling reality using ODEs, as we will see in the Tutes coming up. Conversely, the equation with the reversed sign occurs very rarely because of the unphysical divergence of its solution. Hence, if a reality-based problem throws up the latter equation then there is a high probability that you have botched the signs somewhere. From the viewpoint of course exams, an error in the sign here must be avoided at all costs. If you start out on a problem with the wrong sign on this equation and continue solving, then you very likely will lose all the marks for the problem with the steps after the faulty equation not even being checked.

Example 1.2.2

Radiocarbon dating

The mass, or equivalently the number of atoms, of a radioactive specimen decays at a rate proportional to itself through some constant γ. At $t=0$ the initial mass is given as m_0.

(a) Write an ODE problem for the mass of the specimen as a function of time.

(b) Solve the above ODE.

(c) Determine the time taken for the mass to reduce to $m_0/2$. This is called the half-life of the specimen. Then calculate the time for it to become $m_0/4$ and $m_0/8$.

(d) Generalizing the above one can say that the mass of the specimen decreases by equal _____ in equal times (fill in the blank with a suitable word).

A particular radioactive substance which follows this decay law is the isotope Carbon-14. The half-life of C-14 is 5730 years. Assume that when a sheet of paper is manufactured in a factory, it contains 1.336×10^{11} atoms of C-14 per gram of paper. After manufacture, the isotope starts to decay. At a later time (in a later era), if the C-14 concentration of the sheet is measured, an approximate prediction can be made regarding its date of manufacture.

Suppose a collector discovers an old musical score, dating from the late 18^{th} or early 19^{th} century, and signed "WOLFGANG AMADEUS MOZART". Handwriting and musicological analysis has whittled down the work to either a genuine MOZART (relevant dates 1770-90) or a clever forgery (1820-40). The laboratory contains an apparatus which can measure the concentration of C-14 in the manuscript paper accurate to 1 percent.

(e) Assuming that paper quality did not change from 1750-1850, composers wrote on recently manufactured paper etc. can the

apparatus be used to choose with certainty between MOZART and the forger ? Justify your answer.

[Level : 2]

(a) We are given that the rate of change of the mass i.e. $\mathrm{d}m/\mathrm{d}t$ is proportional to m through a constant γ. Further the specimen is decaying, so the mass is decreasing and the rate is negative. Hence

$$\frac{\mathrm{d}m}{\mathrm{d}t} = -\gamma m \quad \text{OR} \quad \dot{m} + \gamma m = 0 \quad , \tag{*}$$

which is the ODE asked for. Be careful of the sign on γ in the ODE. The associated IC is $m(0) = m_0$.

(b) Comparing with 1.2.1, steady state is $SS = 0$ and initial condition $IC = m_0$ hence the solution here is

$$m = m_0 \, \mathrm{e}^{-\gamma t} \quad . \tag{*}$$

(c) When t equals the half life T, m equals $m_0/2$. We have

$$\frac{m_0}{2} = m_0 \, \mathrm{e}^{-\gamma T} \quad .$$

This simplifies to

$$T = \frac{\log 2}{\gamma} \quad . \tag{*}$$

Throughout this book, $\log$ means to base e – nowhere will we need log to base 10 or any other base. By an analogous calculation, the time to become one-quarter is $\log 4 / \gamma$ and to become 1/8 is $\log 8 / \gamma$. But, from properties of logs, $\log 4 = 2 \log 2$ and $\log 8 = 3 \log 2$. Hence the quarter time is $2T$ and the eighth time is $3T$.

(d) The most appropriate choice is "fraction" or equivalently "percentage" or "ratio". The exponential decay curve $\mathrm{e}^{-\gamma t}$ has the unique property that at the end of every interval of a given size Δt (NOT necessarily small), x decreases to a fixed fraction of what its value was at the start of the interval. This follows immediately from the fact that

$$\mathrm{e}^{-\gamma(t+\Delta t)} = \mathrm{e}^{-\gamma \Delta t} \, \mathrm{e}^{-\gamma t} \quad ,$$

for all t; the reduction fraction is thus $\mathrm{e}^{-\gamma \Delta t}$.

The equal fractions equal time property of exponential decay is a very special feature which you should know about. Clever problem-setters (that's me haha) often assign exercises on this topic with the numerical values chosen in such a way that if you can exploit this trick, you can bypass a lot of tedious arithmetic and/or algebra.

(e) The answer is NO. If MOZART was the composer, the manuscript would date from 1770-90 and the C-14 would have had 227-247 years to decay until today. If the imitator was the composer, the manuscript would date from 1820-40 and the C-14 would have had 177-197 years to decay until today. The outcome of concentration measurement is uncertain because the two decay times might not be sufficiently far apart for the apparatus to detect the resulting change in concentration. To evaluate feasibility of the experiment, we must go with the worst possible case, which is that the separation between the two times is the least. This happens if MOZART wrote it in 1790 and the forger in 1820, amounting to a 30 year interval between the two possible dates.

Since the half life is 5730 years, we have

$$\gamma = \frac{\log 2}{5730}\ \mathrm{yr}^{-1} \quad ,$$

for this situation. Now use equal fractions equal time to bypass computation of actual concentrations. From 1820 to today, the C-14 concentration will decrease by the SAME FRACTION irrespective of whether the sample dates from 1790 or 1820. In the earlier case, there will be some extra decay from 1790 to 1820 while in the later case it will start from full concentration at 1820; this will be the ONLY difference between the two cases today. If this difference is greater than 1 percent of the original, which is the limit of sensitivity of the apparatus, then the test will succeed; otherwise it will be inconclusive. So it remains to compute whether the sample decays by less or more than 1 percent of itself over 30 years.

If the initial concentration is 1, then at the end of 30 years the concentration is

$$\exp\left(-\frac{30\log 2}{5730}\right) \quad ,$$

which, since the argument of exponential is small, we can Taylor expand and write as

$$1-\frac{30\log 2}{5730} \quad .$$

Since $\log 2$ is less than 1 (its numerical value is 0.693), the difference from 1 is evidently lower than 1 percent and below the limit of sensitivity of the apparatus. Hence the experiment fails to differentiate between MOZART and the imitator.

*

COMMENTS : Radiocarbon dating was invented by WILLARD LIBBY in 1949, and it earned him the Nobel Prize eleven years later. The Prize of course was not for solving the trivial ODE but for thinking of this remarkable application of the simple maths. What this example shows, more than anything else, is that even the most basic techniques, combined with the right physical insight, can carry you a long way.

Example 1.2.3

Chemical tank – basic model

A tank of volume V is initially filled with pure water. At $t = 0$, water containing concentration C_0 (kg/m^3) of dissolved salt starts entering the tank through an inlet pipe at a rate R (m^3/s). The incoming water mixes instantly and perfectly with the water in the tank. This tank water flows out at rate R through an outlet pipe.

(a) Write an ODE problem for the concentration of salt inside the tank as a function of time.

(b) Solve the ODE and determine the concentration at $t \rightarrow \infty$.

[Level :1]

(a) At first let us make a careful distinction between quantity and concentration. The quantity of salt $q(t)$ inside the tank is the total amount of salt in it, measured in kg. The concentration of salt $C(t)$ inside the tank is the amount per unit volume, measured in kg/m^3. The assumption of instant and perfect mixing inside the tank (an assumption which is sometimes left implicit in this type of problem) means that at any instant, the concentration is constant throughout the tank and the total quantity of salt in the tank is thus concentration times volume i.e. $q = CV$.

Now, we state the basic physical principle which leads to the ODE. This is that, the change in quantity of salt in the tank equals the amount flowing in minus the amount flowing out. In a time interval Δt, let's say that the quantity changes by amount Δq. We have to equate this change to the resultant of the inflow and outflow terms. Since the incoming water has concentration C_0, and is flowing in at rate R, the inflow quantity during this time interval is $RC_0\Delta t$. Note carefully how the dimensions work here : R is m^3/s and C_0 is kg/m^3 so RC_0 is kg/s, and $RC_0\Delta t$ is kg, a quantity. Likewise, the outgoing water has concentration C (that of the tank) and flow rate R so the outflow quantity during the interval is $RC\Delta t$. Equating the change to the flow we have

$$\Delta q = RC_0\Delta t - RC\Delta t \quad .$$

Now, we divide the both sides of this equation by Δt and let Δt and Δq become infinitesimally small so that Δ can be replaced by d. Then we have

$$\dot{q} = RC_0 - RC \quad .$$

Finally, we have to introduce C in place of q. This we do using $q = CV$, getting

$$* \qquad \frac{\mathrm{d}C}{\mathrm{d}t} + \frac{R}{V}C = \frac{RC_0}{V} \quad , \qquad *$$

which is the ODE asked for. Since at $t = 0$ the water in the tank is pure i.e. no salt, we have the initial condition $C(0) = 0$.

(b) We instantly recognize the form 1.2.1 (you don't? go back and see!); here $SS = C_0$ and $IC = 0$ hence

$$C = C_0\left(1 - \mathrm{e}^{-(R/V)t}\right) \quad .$$

Since the tank is being pumped with water at concentration C_0, we also expect on physical grounds that the concentration inside will tend to $\underline{C_0}$ at large times.

*

COMMENTS : The quantity-balancing argument which we used to get the ODE for C is in fact a simple example of a conservation principle. Similar arguments in fact lead to the relation between charge and current, the equation for conduction of heat, and the continuity equation of fluid mechanics. We'll see these examples in later Chapters.

Example 1.2.4

Chemical tank with imperfect mixing

A tank of volume V is initially filled with pure water. At $t=0$, water containing concentration C_0 of dissolved salt starts entering the tank through an inlet pipe at a rate R. Water also flows out of the tank at rate R via the outlet pipe. Due to a faulty design however, the incoming salt water does not mix instantly and perfectly with the water in the bulk tank. Rather, the flow between the inlet and outlet pipes forms a localized channel of volume V_1 (less than V) in the interior of the tank. Within this channel, mixing is instant and perfect due to turbulent flow. The channel exchanges salt with the bulk of the tank via diffusion. In this process, the salt flows into the bulk from the channel at a rate k times the concentration in the channel. Conversely, salt flows into the channel from the bulk at a rate k times the concentration in the bulk. Let C_1, C_2 and V_1, V_2 $(=V-V_1)$ be the concentrations and volumes of the channel and bulk tank respectively.

(a) Write a pair of coupled ODEs for C_1 and C_2 as functions of time. Write the initial conditions also.

(b) Under what relation between the various parameters in the problem do the two ODEs approximately decouple ? Assume that this relation holds true for the subsequent parts of this question.

(c1) Without solving the ODEs, determine the asymptotic concentrations of salt in the channel and the bulk tank.

OR

(c2) Solve the ODEs in Part (b) to determine the asymptotic concentrations of salt in the channel and the bulk tank.

[Level : 2]

(a) The quantity of salt in the channel changes on account of four processes : inflow of salt water from the outside, outflow of channel water to the outside, incoming diffusion of salt from the bulk and outgoing diffusion of salt to the bulk. Combining these four using the logic of the last Tute lead to the ODE

$$\frac{\mathrm{d}q_1}{\mathrm{d}t} = RC_0 - RC_1 + kC_2 - kC_1 \quad .$$

Note carefully the signs on all the terms on the RHS here.

The quantity of salt in the bulk changes on account of two processes : incoming diffusion of salt from the channel and outgoing diffusion of salt to the channel, thus

$$\frac{\mathrm{d}q_2}{\mathrm{d}t} = kC_1 - kC_2 \quad .$$

Finally using $q_1 = C_1 V_1$ and $q_2 = C_2 V_2$ we have the ODEs

* $$\frac{\mathrm{d}C_1}{\mathrm{d}t} + \left(\frac{R+k}{V_1}\right) C_1 = \frac{R}{V_1} C_0 + \frac{k}{V_1} C_2 \quad , \qquad \text{(01a)*}$$

* $$\frac{\mathrm{d}C_2}{\mathrm{d}t} + \frac{k}{V_2} C_2 = \frac{k}{V_2} C_1 \quad , \qquad \text{(01b)*}$$

which is the equation pair asked for.

Since at $t = 0$ the channel as well as the bulk contain pure water, the IC is $\underline{C_1(0) = C_2(0) = 0}$.

(b) The non-trivial assumption under which the ODEs decouple is $\underline{k << R}$. For in this case, the k terms in (01a) become vanishingly small compared to the other terms, leading to a standalone equation for $C_1(t)$ which features only the large terms. In (01b) on the other hand, every term is of order k so nothing drops there, and the solution for C_2 depends on C_1. Note once again that smallness and largeness is a completely relative concept and has NO meaning in absolute terms. If a human grows by 1 mm overnight, that is negligible while if an ant grows by that amount overnight then it is quite significant. Here, the smallness condition is k is much less than R, NOT "k is very small". Setting $k = 0$ in (01a) will create small errors while the same substitution in (01b) will be completely fallacious !

(c1) If we understand the physics behind the chemical tank, then this is a oneliner. The channel is being pumped continuously with water at concentration C_0 so its asymptotic concentration must be C_0. As for the bulk, at long times, the bulk will be in equilibrium with the channel – the salt it gives the channel will equal the salt it takes from the channel. This will be possible only if the bulk concentration is C_0 also. Hence, the asymptotic concentration of both channel and bulk is C_0. The weak coupling assumption of Part (b) is not required here – it only helps to facilitate the maths in the alternative version of this question.

(c2) The standalone C_1 equation is the standard form

$$\dot{C}_1 + \gamma_1 C_1 = \gamma_1 C_0 \quad ,$$

where $\gamma_1 = R_1 / V_1$. Here $SS = C_0$, $IC = 0$ and hence

$$C_1 = C_0 \left(1 - \mathrm{e}^{-(\gamma_1)t}\right) \quad .$$

Now plug this into the C_2 equation to get

$$\dot{C}_2 + \gamma_2 C_2 = \gamma_2 C_0 \left(1 - \mathrm{e}^{-(\gamma_1)t}\right) \quad ,$$

where $\gamma_2 = R_2 / V_2$. This is first order linear and can be solved by integrating factors. You can also do it using undetermined coefficients (take a sneak peek ahead at N°. 09). Using the IC $C_2(0) = 0$, the solution is

$$C_2 = C_0 + \frac{\gamma_2 C_0}{\gamma_1 - \gamma_2} \mathrm{e}^{-(\gamma_1)t} - \frac{\gamma_1 C_0}{\gamma_1 - \gamma_2} \mathrm{e}^{-(\gamma_2)t} \quad .$$

Now to interpret these results.

To get the steady state, we set $t \to \infty$ in the expressions for both C_1 and C_2. Clearly both equal C_0 in the steady state.

Example 1.2.5

Rocket sled

This is a variation on a problem which appeared in the first mid-semester examination in 2016-17/II.

In a rocket sled testing facility, the sled is decelerated at the end of the run by passing through a channel of water. The maximum permissible exit speed from the channel is 16 km/hr. Within the channel, the only force acting on the sled is drag, which is proportional to the square of the velocity.

(a) Write an ODE for the velocity of the sled inside the channel as a function of time.

(b) Use the Chain Rule to write an ODE for the velocity of the sled inside the channel as a function of distance travelled.

If the entry speed of sled is 250 km/hr, the minimum length of the channel required is found to be 270 m.

(c) Determine the minimum length required if the entry speed is 40 km/hr.

[Level : 1]

(a) This is the standard traction-friction problem except that the traction here is zero so

$$* \qquad \frac{\mathrm{d}v}{\mathrm{d}t} = -Cv^2 \quad , \qquad *$$

where C is an unknown proportionality constant. This is the ODE asked for.

(b) For velocity vs. distance, we need the derivative $\mathrm{d}v/\mathrm{d}x$. From the Chain Rule

$$\frac{\mathrm{d}v}{\mathrm{d}x} = \frac{\mathrm{d}v}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}x} = \frac{\mathrm{d}v}{\mathrm{d}t}\Big/\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{1}{v}\frac{\mathrm{d}v}{\mathrm{d}t} \quad \text{, or}$$

$$\frac{\mathrm{d}v}{\mathrm{d}t} = v\frac{\mathrm{d}v}{\mathrm{d}x} \quad .$$

Now, we are given that $\mathrm{d}v/\mathrm{d}t = -Cv^2$, so $v(\mathrm{d}v/\mathrm{d}x) = -Cv^2$ or

$$* \qquad \frac{\mathrm{d}v}{\mathrm{d}x} + Cv = 0 \quad , \qquad *$$

which is the ODE asked for.

(c) For minimum length of the channel, the exit speed must be its maximum permissible value i.e. 16 km/hr. Hence the given information means that that the speed of the sled reduces from 250 to 16 km/hr over a distance of 270 m. The question asks for the distance it takes to reduce the speed from 40 to 16 km/hr. The numbers here suggest equal fractions equal distance : the given data features a speed reduction by $(2/5)^3$ while we are asked for a reduction by (2/5). Clearly the former distance must be thrice of the latter, so the desired answer is 270/3 which evaluates to <u>90 m</u>.

*

COMMENTS : Although individual expectations vary, the answer to Part (c) is likely to be considerably greater than what we would expect basis our gut feeling alone. Based on experience from cars and trains, we generally have a conception that the stopping distance increases significantly with increase in initial speed. Consequently, for a six-fold reduction in entry speed we would tend to expect at least a six-fold reduction in distance, maybe more. This intuition however is based on braking where the deceleration is more or less a constant function of speed. There, $v^2 = 2aS$ so the distance increases quadratically with initial

velocity. With this braking rule for the sled, the answer would have come out to 7.5 m instead of 90 !

With the quadratic drag what happens is that the drag itself depends very strongly on the speed. So at high speed, the drag is huge and it slows the sled down so drastically that it doesn't get to cover that much of distance. At low speed however, the drag is weak and the sled keeps crawling along, covering distance inch by inch while retarding only marginally. This leads to the apparently anomalous answer.

This is the first problem in this book where the Chain Rule has made an appearance. It is in fact one of the most important calculus rules in the study of differential equations, since it arises every time we have to change variable in an ODE (and in a PDE also for that matter). In the course of this book, we'll be needing it quite a few times; so if your memory has become rusty, now is a good time to go back to a calculus book, read up on the Rule and work out a few exercises.

Example 1.2.6

NEWTON's rule of cooling

If a body at temperature T is kept in atmosphere of ambient temperature T^* ($<T$), it loses heat at a rate proportional to the temperature difference between it and the atmosphere. This is called NEWTON's rule of cooling (sometimes also NEWTON's "law" of cooling).

(a) Explain why the word "rule" might be more appropriate than the word "law" in this instance (compare with the laws of motion and other laws of Nature).

(b) Write an ODE for the temperature of the body as a function of time, and solve it if the initial temperature is given as T_0.

(c) Does the temperature become exactly T^* at a finite time ? Justify your answer.

[Level : 1]

(a) The word “rule” is preferable to “law” because this simple statement is only an approximate and empirical observation, which might or might not be satisfied by a body in reality. A body might have substantial variation in internal temperature; to apply the rule we must consider the average temperature. The cooling process has several mechanisms – convection, conduction and radiation are the primary drivers. Each of these is governed by its own set of principles and only when everything is favourable (one mechanism dominant, nice shape of the body etc.) can we see NEWTON’s rule actually working well. It is eminently possible to design bodies and cooling mechanisms where there are substantial deviations from the rule. We will investigate this topic more thoroughly in the later Chapter on heat equation.

On the other hand, “laws” typically refer to fundamental principles of Nature. There are no bodies which violate NEWTON’s laws of motion (of course excluding quantum and relativistic particles where the laws themselves are different). But you can never have a situation where F equals ma^2 rather than ma. Or one where a planet attracts a satellite with a $1/r$ force instead of $1/r^2$. Or a magnetic field whose divergence is non-zero (give me that and I’ll build you a worldwide maglev network tomorrow). You get the point.

(b) The equation of the body is

$$\frac{\mathrm{d}T}{\mathrm{d}t} = -k\left(T - T^*\right) \quad , \tag{*}$$

which is the ODE asked for. To get the sign correct on the RHS either note that if $T > T^*$ the body has to cool, or that the wrong sign would produce $\dot{x} - \gamma x = b$ form which is physically implausible. Once again we have the same old basic equation from 1.2.1 – the last several T’s have been featuring the same thing, but each one in a different context ! The solution is

$$T = T^* + \left(T_0 - T^*\right)\mathrm{e}^{-kt} \quad , \tag{*}$$

which is the desired answer.

(c) Since e^{-kt} does not become exactly zero at finite time, T also does not become exactly T^* at a finite time.

Example 1.2.7

LR circuit

The current $i(t)$ in an LR circuit connected to an ac voltage source satisfies the ODE

$$L\frac{\mathrm{d}i}{\mathrm{d}t} + Ri = V\cos\omega t \quad .$$

Show that the current in the circuit at large times is independent of the initial condition and determine its value.

[Level : 1]

As usual, the current will be composed of a homogeneous solution plus a particular solution. The homogeneous solution must satisfy $Li_h' + Ri_h = 0$ or $i_h' + \gamma i_h = 0$ where $\gamma = R/L$ and we're using prime rather than dot to mean $\mathrm{d}/\mathrm{d}t$ when it applies to i, for obvious reason. This has the solution $i_h = C\,\mathrm{e}^{-\gamma t}$ which decays to zero at large time. So at these times, we can see only the particular solution, which by definition is independent of the initial condition (that goes into fixing the C). Hence the current at large time is independent of IC.

To find i_p while sticking to the methods of this Chapter, we use integrating factors. The ODE is

$$i' + \gamma i = \frac{V}{L}\cos\omega t \quad ,$$

and the integrating factor is $\mu = \mathrm{e}^{\gamma t}$. In the solution formula $\mu i = \int \mu Q\,\mathrm{d}t + C$, the C term contributes the homogeneous solution, which we already know. The particular solution is the other term, which is

$$i_p = \frac{1}{\mathrm{e}^{\gamma t}} \int \frac{V}{L}\mathrm{e}^{\gamma t}\cos\omega t\ \mathrm{d}t \quad .$$

I won't go into the details of evaluating the integral – you can always look it up. If you want to do it yourself, the trick is to integrate by parts twice. The result is

$$i_p = \frac{V}{L\left(\omega^2 + \gamma^2\right)}\left(\gamma\cos\omega t + \omega\sin\omega t\right) \quad ,$$

which is the desired answer.

*

COMMENTS : When you learn the method of undetermined coefficients (3.1) you will see that this can also be done using that method, and with much less work (no integral, only derivative). I didn't present the method in this Chapter itself since it's pointless to discuss that before we are prepared to handle higher order homogeneous equations. But once that trick is up our sleeve, we can always do first order stuff trivially, as a special case.

Example 1.2.8

Pharmacokinetics and drug dosage

A patient consumes 350 mg of a drug every 12 hours. Assume that drug is constantly eliminated from his system at a rate proportional to the quantity inside. In the absence of inflow, the quantity decreases by 25 percent over a six-hour period. After each dose, the quantity jumps up instantaneously by the amount of the dose.

The instantaneous jump is difficult to model within the existing framework so it will be approached via a limit. Assumption is made that during the dose, drug is pumped in at a high but uniform rate over a fixed and very small (compared to 12 hours) time interval Δt. Since the total intake is 350 mg, the inflow rate during the dose must be $350/\Delta t$. Each dose starts at $t = 12n$ hours (n is an integer 0,1,2, …..) and runs until $12n+\Delta t$; in the period thereafter until $t = 12(n+1)$ there is no inflow of drug.

(a) Model the system as a first order linear inhomogeneous ODE for quantity x as a function of time, with the RHS function defined in a piecewise manner.

(b) What happens to the homogeneous solution at large time ?

A particular solution x_p will now be sought which is periodic. This is natural because the excitation is periodic. Such a solution is independent of the number n of the dose so long as n is sufficiently large. Let the value of x_p just before the n^{th} dose i.e. at $t = 12n$ be the unknown quantity x_0.

(c) Solve for x_p during the interval of the dose, and hence determine its value at $t = 12n+\Delta t$. Use the smallness of Δt wherever appropriate.

(d) Solve for x_p in the interval $[12n+\Delta t, 12n+12]$ using the fact that $x_p(t)$ must be continuous at all times, and hence determine x_p at $t = 12n+12$. Again use the smallness of Δt wherever appropriate.

(e) Impose the periodicity condition $x_p(12n+12) = x_0$ to determine x_0.

(f) Qualitatively describe the trajectory of $x_p(t)$ in the limit $\Delta t \rightarrow 0$.

[Level : 3]

Our first Level 3 appears really long and complicated. When faced with such a thing in the exams, the best strategy is to read carefully and proceed bit by bit. What not to do on the other hand is to try taking it all in at once and become confused. Proceeding systematically allows you to gain conceptual clarity from solving the earlier parts and also earn the marks attached to those parts. In fact we'll see as we go along that the bark is worse than the bite, since the individual parts of the question are quite short and easy and the problem literally walks us through to the solution. A Level 3 with just two or three parts is far more demanding than this.

(a) First thing we see is that this is like a tank concentration problem where change equals inflow minus outflow. The outflow is proportional to quantity while input is the dose function, so

$$\dot{x} = -\gamma x + f(t) \quad ,$$

where $f(t)$ is the dosing function. By the conditions of the problem, $f(t)$ is nonzero only in the intervals $[12n,\ 12n+\Delta t]$, where n is every whole number; in these intervals it is constant with value $350/\Delta t$. Hence

$$* \qquad f(t) = \begin{cases} \dfrac{350}{\Delta t} \text{ if } 12n \le t \le 12n + \Delta t \\ 0 \text{ otherwise} \end{cases} \quad . \qquad *$$

Note that this is just a mathematical restatement of what is given in the problem, there's nothing new here !

(b) The homogeneous solution goes as $e^{-\gamma t}$ so it decays to zero in time. This implies that at large times, the only solution we will see is the particular solution. This will be independent of the initial condition since all the IC information is carried by the homogeneous solution.

For the rest of the problem, we have to work in the interval $[12n,\ 12n+12]$ where n is arbitrary so long as it is large. Since the ODE does not feature time as an explicit parameter, we can simply shift the time axis so that $t = 12n$ becomes $t = 0$ on our new scale. Then, we can solve the problem in the time interval $[0,\ 12]$ which is much easier to deal with than $[12n,\ 12n+12]$.

(c) During the interval of the dose we have

$$\dot{x}_p + \gamma x_p = \frac{350}{\Delta t} \quad ,$$

with the IC $x_p(0) = x_0$. As the question tells us, x_0 is unknown at this stage – we are not to worry about this. The steady state solution is $SS = 350 / \gamma\Delta t$ and the complete solution is

$$x_p = \frac{350}{\gamma\Delta t} + \left(x_0 - \frac{350}{\gamma\Delta t} \right) \mathrm{e}^{-\gamma t} \quad .$$

We want to evaluate this at $t = \Delta t$; since this is a small quantity we Taylor expand the exponential to write $\mathrm{e}^{-\gamma\Delta t} = 1 - \gamma\Delta t$. This yields, to first order in Δt,

* $$x_p(\Delta t) = x_0 + 350 \quad ,$$ *

which is the desired answer. This result has a clear physical interpretation – since the dose occurs over a very short time, the system gets negligible opportunity to eliminate the drug and the quantity jumps up by the amount administered during the dose.

(d) We must now solve for x_p in the interval $[\Delta t, 12]$. The equation is $\dot{x}_p + \gamma x_p = 0$ since there is no inflow of drug during this time. Although this is the homogeneous equation and x_p is the particular solution, there is no contradiction here; x_p follows this equation for only PART of the time while x_h satisfies it continuously, for ALL time. The IC for this part comes from the continuity of x_p. Part (c) yielded that $x_p(\Delta t)$ was $x_0 + 350$; this must act as the IC for this phase of the dynamics. We thus have

$$x_p = \left(x_0 + 350 \right) \exp\left[-\gamma \left(t - \Delta t \right) \right] \quad .$$

Into this expression we must substitute $t = 12$; that yields

$$x_p(12) = \left(x_0 + 350 \right) \exp\left[-\gamma \left(12 - \Delta t \right) \right] \quad . \qquad (01)$$

Now we use the decay rate given in the problem – the quantity decreases by 25 percent i.e. becomes 3/4 of itself in 6 hours so it must become 9/16 of itself in 12 hours. Hence $\mathrm{e}^{-12\gamma} = 9/16$. Using this and Taylor expanding the exponential in (01) we get

$$\exp\left[-\gamma\left(12-\Delta t\right)\right]=\mathrm{e}^{\gamma\Delta t}\,\mathrm{e}^{-12\gamma}$$
$$=\frac{9}{16}\mathrm{e}^{\gamma\Delta t}=\frac{9}{16}\left(1+\gamma\Delta t\right) \quad ,$$

so that

$$x_p(12)=\frac{9}{16}\left(x_0+350\right)\left(1+\gamma\Delta t\right) \quad ,$$

which is the desired answer.

(e) This is the step which actually fixes the unknown x_0. Imposing the periodicity condition $x_p(12)=x_0$ we get

$$\frac{9}{16}\left(x_0+350\right)\left(1+\gamma\Delta t\right)=x_0 \quad ,$$

which has the solution

$$x_0=\frac{3150\left(1+\gamma\Delta t\right)}{7+9\Delta t} \quad .$$

This is very close to the value 450, differing from it in terms of order Δt and smaller.

(f) With Δt small but not zero, the trajectory of x_p is a straight line of sharp slope from $(0,\ x_0)$ to $(\Delta t,\ x_0+350)$, followed by a gradual exponential decay over the interval between the doses. As Δt decreases, the slope of the straight part keeps increasing, and in the limit, x_p changes instantaneously from x_0 (which now equals exactly 450) to x_0+350 (which is 800). Then, over the next 12 hours x_p decreases by $(3/4)^2$ back to a value of 450 just before the next dose. This is consistent with the particular solution being a periodic state.

Thus we see that this limiting process indeed captures the phenomenon that x_p increases instantaneously during the dose event. This instantaneous increase is definitely not a trifling matter : x_p is the solution of a differential equation which means that a derivative has to exist – with an instant jump, not only does it not exist but the function itself becomes discontinuous. This is why it is worthwhile to approach this solution carefully, as a limit of a sequence of perfectly well-defined solutions of an ODE.

*

COMMENTS : In most antibiotics and other drugs, the brochure attached to the drug indeed contains a section on pharmacokinetics which describes the rate of elimination (and also absorption – a phenomenon we neglected here) of the drug to and from the patient's system. The dosage is governed primarily by the elimination time constant, and the need to keep the quantity above a minimum level at all times. Thus amoxicillin (Mox 500) has a half life of just above 1 hour and is taken thrice a day while azithromycin (Azithral 500) has a half life of 12 hours or more and is taken once a day.

We assumed here that the particular solution is periodic. Fourier analysis, coming up in Chapter 4, shows that indeed it will be so. This problem is also a simple example of what will later become the method of matched boundary conditions to find a periodic particular solution of a system driven by a non-standard periodic excitation.

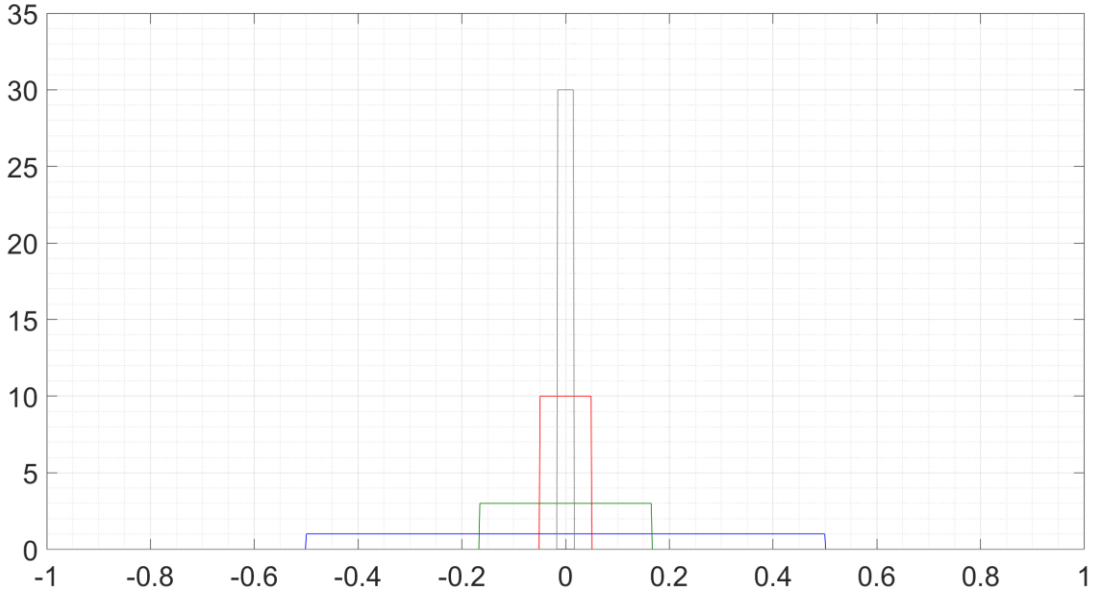
Now suppose we had to solve the complete problem with a realistic IC – the patient consumes the first dose when there is no drug in his bloodstream. Then, the quantities of drug in the bloodstream right after the first few doses are as follows : 350 mg after the first dose (obviously), 547 after the second, 658 after the third, 720 after the fourth, 755 after the fifth and 775 after the sixth. So after six doses, we are at better than 95 percent of the steady state solution.

1.3

DIRAC'S DELTA FUNCTION

Here's a function named after this book. This function was invented by PAUL DIRAC to systematically handle cases of impulsive phenomena like the dosing function in the last example. Although it is not a function in the true sense of the term, the caveats and lacunae in the definition are relevant for mathematicians and not for us. So far as we are concerned, the delta function is as legitimate as any other function like cos or sin and it can be the RHS of an ODE.

Consider the function which is 1 in the region $x = -1/2$ to $x = +1/2$ and zero everywhere else. This is a rectangle of width 1 and height 1 and enclosed area 1. Now let's stretch it and make it 3 in the region $-1/6$ to $+1/6$ and zero everywhere else. This produces a skinnier but taller rectangle with the same area. Likewise we can stretch still further to make it say 10 in the region $-1/20$ to $+1/20$ and zero everywhere else. Four such stages of stretching are shown in the Figure below.



The successive plots are in blue, green, red and grey respectively. (As an aside, this also allows us to introduce the colour code which we'll use more or less consistently throughout this book for presenting multiple plots in the same graph – the colour order will be blue, green, red, grey, cyan, magenta, yellow and black for the first eight plots and we'll not plot more in one Figure.)

At the limit of such stretching, you will get an infinitely tall and infinitesimally narrow spike which encloses a total area of 1. This spike is called **Dirac delta function** $\delta(x)$. If the rectangle is shifted to be centred at say $x = x_0$ rather than $x = 0$ then the limiting function is $\delta(x - x_0)$. Another name for the delta function is **unit impulse function** – this name is more common in engineering.

Mathematically, delta function has the following properties :

- $\delta(x - x_0)$ is zero if x does not equal x_0 and is infinity if x equals x_0.
- The dimension of $\delta(x)$ is "$1/x$".
- $\int_a^b \delta(x - x_0)\,dx$ equals unity if the interval $[a, b]$ includes the point x_0 and zero if the interval does not include x_0. The size of the interval does not matter since δ is nonzero only at one point.
- $\int_a^b f(x)\delta(x - x_0)\,dx$ where f is a continuous function equals $f(x_0)$ if the interval $[a, b]$ includes the point x_0 and zero if it does not. Once again, the size of the interval is immaterial.

To verify these properties, you can simply use the definition of the delta function as a limit of the sequence of rectangles. More importantly, to solve an ODE which has delta function on the RHS, the basic trick is to manually integrate the both sides over an infinitesimal interval spanning the delta function. How to do this is coming up next.

Example 1.3.1

Pharmacokinetics again

Consider the same situation as 1.2.8, which is repeated below. "A patient consumes 350 mg of a drug every 12 hours. Assume that drug is constantly eliminated from his system at a rate proportional to the quantity inside. In the absence of inflow, the quantity decreases by 25 percent over a six-hour period. After each dose, the quantity jumps up instantaneously by the amount of the dose."

The assignments this time are :

(a) Write a first order linear inhomogeneous ODE for quantity x as a function of time. Define the inhomogeneity in terms of the delta function.

The periodic particular solution x_p will now be sought. Focusing on an arbitrary n^{th} dose, the time is shifted to equal zero just before this dose.

(b) Letting x_0 be the quantity just before the dose, integrate the ODE across the time $t = 0$ to determine the quantity just after the dose.

(c) Obtain $x_p(t)$ in the interval from just after this dose to just before the next dose. Impose the condition that this latter value be x_0 to determine x_0 and hence x_p.

[Level : 2]

(a) From 1.2.8 we already know that each dose is the limit of a series of rectangles of width Δt and height $350/\Delta t$ to make a total area of 350. The limit is clearly 350 times the delta function. Hence the n^{th} dose can be written as $350\,\delta(t-12n)$. Since the dosing function is periodic every 12 hours, the complete RHS function $f(t)$ is the sum of the delta at $t=0$ plus that at $t=12$ plus that at $t=24$ etc. Taking the LHS from 1.2.8 we have

$$\dot{x}+\gamma x=\sum_{n=0,1,2,\ldots..} 350\ \delta\left(t-12n\right)\quad , \tag{*}$$

which is the ODE asked for.

(b) This is the step where we actually perform the manipulation using the delta function so watch it carefully. We are given that, just before the dose at $t=0$, $x=x_0$. Now integrate the both sides of the ODE over an infinitesimal time interval spanning $t=0$. We have

$$\int_{-\varepsilon}^{\varepsilon}\dot{x}\mathrm{d}t+\gamma\int_{-\varepsilon}^{\varepsilon}x\mathrm{d}t=350\int_{-\varepsilon}^{\varepsilon}\delta(t)\mathrm{d}t\quad ,$$

where ε is some infinitesimally small quantity. On the RHS, only the delta at $t=0$ is relevant since the integration is centred at $t=0$, so I have dispensed with the summation over the other deltas.

The first term on the LHS is $x(\varepsilon)-x(-\varepsilon)$, which is $x(\varepsilon)-x_0$ as we take the limit $\varepsilon\rightarrow 0$. The second term on the LHS is x integrated over an infinitesimally small interval. Now x itself cannot be infinite – its size is bounded by the amount of the dose. Physically also, an infinite quantity of drug inside the patient's system is utter nonsense. Hence this term is the integral of a FINITE FUNCTION over an INFINITESIMAL INTERVAL, which is infinitesimal, and we drop it from consideration. The integral of the RHS is 350 times unity, by the properties of the delta function. This gives

$$x(\varepsilon)-x_0=350\quad , \tag{*}$$

or, the quantity just after the dose is 350 mg higher than what it was just before the dose.

This of course is the conclusion we arrived at in 1.2.8 Part (c). There we approached it by treating the delta function as the limit of the rectangle functions , while here we did it formally by exploiting the properties of the delta

function. For any equation involving delta function, the trick is always the same – integrate LHS and RHS across an infinitesimal interval and throw out the integrals of finite functions over the said interval to obtain the jump.

(c) From this point onwards it's just Parts (d) and (e) of 1.2.8 with the limit (f) already in place. I will not waste time typing it all out again.

*

COMMENTS : Although this Unit is already over, it would be a pity if DELTA were to contain just this one problem featuring its namesake. Indeed, DIRAC's cute invention is going to keep making guest appearances in the subsequent Chapters. Delta function is a practical way of mathematically modeling impulsive forces, point charges and similar stuff which is huge in a very short interval of time/space and zero everywhere else. Although it looks "pathological" with its definition as ∞ somewhere and 0 elsewhere, it is actually far easier to use than the limit of a sequence of rectangles or any other continuous functions. It is also the foundation of the extremely powerful method of GREEN's functions for differential equations (both ODE and even more PDE), which is coming up in Chapter 4 onwards.

When higher order ODEs are involved, the basic fact is that if the n^{th} derivative of x has delta, then the $n-1^{\text{th}}$ derivative has a finite jump, the $n-2^{\text{th}}$ derivative is continuous but has a change in slope and so on, with the successively lower orders of derivative being more and more tractable. To see this, consider the successive integrals of $\delta(t)$ – the first integral is a step function 0 for negative t and 1 for positive t, the next one is 0 for negative t and t for positive t, the next is 0 for negative t and parabola for positive t and so on. We'll see more of this when actually dealing with higher order ODEs.

1.4

INTRODUCTION TO STABILITY ANALYSIS

Consider a system of the form

$$\frac{dx}{dt} = f(x) \quad , \tag{01}$$

where f is some complicated function. Of course this is a separable equation and you can solve it in theory by transferring variables. But in many cases, the x integral cannot be done explicitly. Even then, it is possible to extract a large amount of information on the asymptotic behaviour of the system by utilizing the concepts of fixed points and stability. In higher order systems, transfer of variables no longer exists and nonlinear equations are almost always impossible to solve. But in very many such situations, stability analysis tells you a lot about how the system is going to behave, literally giving you something out of nothing.

What we do here is as follows : we look for points where $f(x) = 0$, which are called **fixed points (fp)**. If the system somehow reaches a fp then it will stay there for ever after, so these points are understandably of interest. Given a fp, the question is, if you start the system from a nearby point then will it approach the fp or go away from it. To answer this, we must obtain the **stability** of the fp. The fp is called **stable** if the system returns to it after a small deviation, and **unstable** if the system moves away from it after a small deviation.

To determine whether a fp is stable, we use the following trick. Let x^* be a fp of the system (01) i.e. $f(x^*) = 0$. Then, assume that x is close to x^* i.e. $x = x^* + \Delta x$ where Δx is a small quantity, called the **perturbation** from the fp. Plugging into (01) we have

$$\frac{d}{dt}(x^* + \Delta x) = f(x^* + \Delta x) \quad . \tag{02}$$

On the LHS, x^* is a constant, a root of $f(x) = 0$, so it's d/dt is zero. On the RHS we can Taylor expand it as

$$f\left(x^*+\Delta x\right)=f\left(x^*\right)+\left.\frac{\mathrm{d}f}{\mathrm{d}x}\right|_{x=x^*}\Delta x+\left.\frac{1}{2}\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\right|_{x=x^*}\Delta x^2+\ldots\ldots\ ,$$

where the dots represent terms of increasingly higher power in Δx.

The first term on the RHS is $f(x^*)$; by definition of the fp, it evaluates to zero. What we now say is that Δx is sufficiently small so that **all terms of quadratic and higher power in Δx can be neglected** in the above; this reduces (02) to

$$\frac{\mathrm{d}}{\mathrm{d}t}\Delta x=\left.\frac{\mathrm{d}f}{\mathrm{d}x}\right|_{x^*}\Delta x\quad .$$

We are now on familiar ground; since $\mathrm{d}f/\mathrm{d}x|_{x^*}$ is a constant, this has the form of T005. Given an initial value of Δx, it decays to zero in time if $\mathrm{d}f/\mathrm{d}x<0$ at $x=x^*$ and it grows exponentially if $\mathrm{d}f/\mathrm{d}x>0$ at $x=x^*$. In the case of decay we have stability of the fp and in case of growth we have instability of the fp. The derivative $\mathrm{d}f/\mathrm{d}x$ at the fixed point is often called the **eigenvalue**, a terminology which comes from the theory of higher order systems (see L08). In case the derivative is zero, nothing can be said. We thus have the stability test for fp :

A fixed point x^* of $\dot{x}=f(x)$ is stable if the eigenvalue $\mathrm{d}f/\mathrm{d}x$ at the fp is negative, and is unstable if the eigenvalue $\mathrm{d}f/\mathrm{d}x$ at the fp is positive. If the eigenvalue $\mathrm{d}f/\mathrm{d}x$ at x^* is zero, then the test is inconclusive.

Note that the stability test is based on linearization of the original equation – given the original nonlinear equation (hard to solve) we converted it to a linear equation (easy to solve) and solved it. One issue which might arise is, in case of an unstable fp, we seem to have the $\dot{x}-\gamma x=0$ form of T005, which generally does not arise in real situations. Does that mean that unstable fps do not exist in reality ? The answer is no : unstable fps are very much real and the paradox (like any other paradox) is just a fallacy. The 'unphysical' equation is valid only in the immediate neighbourhood of the fp because it is obtained from a one-term Taylor expansion – the moment the system moves away from the fp, the linearization is no longer valid. The standalone linear equation $\dot{x}-\gamma x=0$ is unrealistic because its solution grows without bound; in the nonlinear case, the moment the system leaves the vicinity of the unstable fp, the equation ceases to hold and the nonlinearity can very well cause the solution to become bounded.

Finally, some useful fixed point facts :

- If the system $\dot{x} = f(x)$ has n fixed points x_1^*, x_2^*, ….., x_n^* arranged from left to right and if df/dx is not zero at any of them, then they **alternate in stability**. That is, if say x_3^* is stable, then automatically x_2^* and x_4^* are unstable, x_1^* and x_5^* are stable and so on. This often enables you to predict stabilities of multiple fixed points just by inspection, after calculating explicitly for only one of them.
- If the system $\dot{x} = f(x)$ is started with an initial condition $x = x_0$, then the asymptotic behaviour of x as $t \rightarrow \infty$ can be found by a very simple procedure. First calculate the two fixed points which are nearest to x_0 on the left and right side. Then see which of these two fps is stable. That is the desired answer.
- In continuation of the above process, if the IC x_0 lies either to the left of the leftmost fp or the right of the rightmost fp then you can treat $x = \infty$ or $x = -\infty$ as a de facto fixed point with stability opposite to that of its finite neighbour, and then apply the above procedure. Although this is mathematically sloppy, it is physically quite rigorous and produces the right answer. In practice, "∞" in this context means a very large but finite number, at which point the model which produced the equation ceases to hold, and some other behaviour takes place.
- If a fixed point has $df/dx = 0$ then what generally happens is that the system tends to approach it from one side and go away from it on the other side. Whenever we have a zero case, either here or in higher order systems, it is a failure of the linearization test and we need to analyse the properties of the nonlinear equation as a whole. Such cases are of great interest in the theory of nonlinear dynamics but outside the scope here.
- Finally, you might be wondering whether the result of linearization can be entirely wrong i.e. is it possible that linearization predicts a fp is stable while actually the nonlinearities are so strong that they cause the fp to become unstable. The answer fortunately is no. The **theorem of GROSSMANN and HARTMANN** (way too sophisticated for even a typical nonlinear dynamics course) guarantees that there is always a non-zero sized region around the fixed point where the result of linearization is strictly valid.

If you see a real system sitting quietly at any point then you immediately know that **the point is a fp AND that it is stable**. The first observation is obvious, the second maybe not so much but is just as true. Real systems are always subject to perturbations from external sources, and only stable fps resist these perturbations and come back. The slightest deviation from an unstable fp causes the system to permanently go away from it so you can NEVER EVER see an unstable fp in practice.

Example 1.4.1

Stability analysis verification

Consider the elementary nonlinear system

$$\frac{\mathrm{d}x}{\mathrm{d}t} = 1 - x^2 \quad .$$

(a) Find all the fixed points and determine their stabilities.

(b) Using stability concepts, determine the asymptotic value of x as $t \to \infty$ for the cases where the initial condition is $x(0) = -2$, $x(0) = 0$ and $x(0) = 2$. For this question, assign to "infinity" the value "100" i.e. assume that if the system reaches the value $x = \pm 100$ then it stays there while the equation ceases to hold any longer.

(c) Solve the equation exactly with the initial conditions of Part (b) and verify the predictions obtained there.

[Level : 2]

(a) The fixed points are clearly the roots of $1-x^2=0$ which are $x^*=\pm 1$. Drawing the graph of the function $f(x)$ (please draw this for yourself!) we can see that it is a parabola which opens downwards, so the slope $\mathrm{d}f/\mathrm{d}x$ is negative at $x=1$ and positive at $x=-1$. Hence $x^*=1$ is stable and $x^*=-1$ is unstable.

(b) For $x(0)=-2$, there is only one neighbouring fp which is $x^*=-1$ on the right. Hence we treat the domain boundary $x=-100$ as a virtual fp. Since $x^*=-1$ is unstable, the virtual one is stable and the system flies off to the distant point $x=-100$.

For $x(0)=0$, the two neighbouring fps are $x^*=-1$ and $x^*=1$. Since the latter is stable, the system asymptotically ends up at the point $x=1$.

For $x(0)=2$, there is only one neighbouring fp which is $x^*=1$ on the left, Hence we treat the domain boundary $x=100$ as a virtual fp. Since $x^*=1$ is stable, the virtual one is unstable and the system asymptotically ends up at the point $x=1$.

(c) The equation is separable so we solve by transferring variables. We have

$$\int_{x_0}^{x_1}\frac{\mathrm{d}x}{1-x^2}=\int_{t_0}^{t_1}\mathrm{d}t \quad .$$

The integral on the LHS can be cracked using partial fractions; the solution is

$$t_1-t_0=\frac{1}{2}\left(\log\left|\frac{1+x_1}{1-x_1}\right|-\log\left|\frac{1+x_0}{1-x_0}\right|\right) \quad .$$

The interpretation of this solution needs some care.

Consider the IC $(t_0,x_0)=(0,-2)$. In this case the log term involving x_0 evaluates to $\log(1/3)$ which is negative, and becomes positive if the minus outside is included. At $t_1=t_0=0$, the log term involving x_1 is also $\log(1/3)$. As t_1 starts increasing and the LHS turns positive, x_1 must move in such a direction that the log term involving x_1 becomes less negative than $\log(1/3)$. This means that $|(1+x_1)/(1-x_1)|$ must be greater than $1/3$, which happens if x_1 is more negative than x_0. Hence the system starts moving towards the negative side. At the time $t=\log(99/101)-\log(1/3)$ it reaches the value $x_1=-100$, which by the definition of the problem is the limit point.

With the IC $(t_0,x_0)=(0,0)$, the log term involving x_0 is zero. As t increases, x must move such that $\log|(1+x_1)/(1-x_1)|$ is positive. This happens if x_1 is positive so the system starts moving towards the positive side. At $x_1=1$, $|(1+x_1)/(1-x_1)|$ becomes plus infinity and so does its logarithm, hence t_1 also becomes infinity by the time the system reaches there. Thus we conclude that asymptotically the system arrives at the point $x=1$.

With the IC $(t_0,x_0)=(0,2)$, the log term involving x_0 is $\log 3$. As t increases, x must move such that $\log|(1+x_1)/(1-x_1)|$ is greater than $\log 3$ i.e. $|(1+x_1)/(1-x_1)|$ is greater than 3. This happens if x_1 decreases from the starting value 2 towards the value 1 where the function blows up. Hence in this case the system starts moving towards the negative side. As with the previous IC, the elapsed time becomes infinity when it reaches $x=1$, so that becomes the desired asymptotic state. This completes the verification of the predictions obtained in Part (b).

*

COMMENTS : First of all we see that the predictions of the linearized stability analysis are confirmed by the exact solution of the nonlinear equation. This validates the stability approach to nonlinear problems. We also see the vast difference in labour required to generate the answers by the two methods. In fact, the calculation of Part (c) was at all possible in this case because the system is sufficiently simple; in most cases the exact solution is impossible to find but the stability analysis always works because it is based on linearization. Of course, stability calculations don't recover ALL the features of the exact solution. For instance, how much time does the system take to go from $x=0.1$ to $x=0.9$ can never be answered by stability arguments. Nevertheless, when the system admits no other inroad, a stability calculation often goes a very long way.

With the first IC $x(0)=-2$, the system took a certain time to reach the 'defined infinity' $x=-100$. Note that if we redefine infinity as -1000, it will only take a slightly greater time to get there. In fact, whatever our definition of infinity, the time taken to reach it will always be finite, with the limiting value (when it's really infinite) being $-\log(1/3)$. This phenomenon is called finite time singularity.

Finally, now is a good time to go back to the train of T004 and its Comments. The equation of motion there was very similar to this – the force $f(x)$ [or rather $f(v)$] has a zero, and hence the dynamics a fixed point, at the MPS of 130 km/hr.

Since traction decreases with v and friction increases with v, the derivative $\partial f/\partial v$ is negative at the fp (it's negative everywhere else also), showing that the fp is stable. As we argued qualitatively then, perturbations from this fp decayed to zero in time. On the contrary, in the unphysical model where traction and friction reverse roles, the derivative $\partial f/\partial v$ is positive everywhere including at the fp, which thus becomes unstable.

Example 1.4.2

Nuclear reactor

In a nuclear reactor the power output is proportional to the number of neutrons N present inside. Neutrons are generated by the fission process which takes place at a rate proportional to N. Neutrons are absorbed by the control rods which absorb at a rate proportional to N and to the depth to which they are inserted into the reactor. In a particular reactor, the control algorithm used is that the insertion depth is proportional to N.

(a) Write an ODE for the number of neutrons inside the reactor as a function of time. Express things in terms of letter constants k_1, k_2 etc.

(b) Show that by regulating the proportionality constant between neutron number and control rod penetration depth, it is possible to maintain the reactor at any desired number of neutrons, and hence any desired power output.

(c) In the absence of the control rods, what would happen in theory and in practice ?

[Level : 2]

(a) The principle behind the equation is similar to that of tank problems – change in neutron number equals number generated minus number absorbed. The generation rate is proportional to N so it contributes a term $k_1 N$ to $\mathrm{d}N/\mathrm{d}t$. The absorption rate is proportional to N and also to h where h is the insertion depth so we can write its contribution to $\mathrm{d}N/\mathrm{d}t$ as $-CNh$. Now it is given that for the reactor in question, h is proportional to N. Hence the absorption rate is $-k_2 N^2$ for some k_2. Combining the two, we have

* $$\frac{\mathrm{d}N}{\mathrm{d}t} = k_1 N - k_2 N^2 \quad ,$$ *

which is the ODE asked for. Here k_1 and k_2 are both positive for the model to be meaningful. Again note that you must be very careful about the signs here – any error and you risk a zero score on the entire problem.

(b) We are told that k_1 is given. Variation of the proportionality constant means that k_2 can be allowed to vary. The fixed points of the system are $N^* = 0$ and $N^* = k_1/k_2$. The stability eigenvalue $\mathrm{d}f/\mathrm{d}N$ at $N = 0$ is k_1 which is positive hence $N = 0$ is unstable. At $N = k_1/k_2$ the eigenvalue is $-k_1$ which is negative so that fp is stable. Hence, starting from an arbitrary IC, the system will always go to the fp $N^* = -k_1/k_2$, and hover around it. By choosing k_2, it is possible to make N^* have any desired value. The stability of the fp guarantees that the reactor will sustain the level on its own. Note that the demonstration of stability is an essential component of the correct answer !

(c) In the absence of the control rods, the equation of the system is

$$\dot{N} - k_1 N = 0 \quad .$$

This is the 'bad' equation from T005 whose solution grows to infinity in time as $\exp k_1 t$.

In practice what happens in this case is a nuclear disaster. The reaction keeps on occurring, with escalating power output and neutron number, until the reactor explodes catastrophically. This was more or less what happened in the Chernobyl accident of 1986, where, after a series of mis-steps by the operators leading to instability in the reactor, the control rods were unable to be deployed to stop the reaction on an emergency basis. Things spiraled out of control and an explosion ensued. A nuclear fission bomb (like the infamous ones dropped on Hiroshima and Nagasaki) is a reactor designed without control rod – the value

of N at the moment of the explosion can be considered to be ‘infinity’ for this situation.

Example 1.4.3

Population dynamics

Part (a) is Read Only while Part (b) is suitable for practice.

The population of a specie satisfies the ODE "logistic equation"

$$\dot{x} = 3x\left(8 - x\right) \quad ,$$

with the IC $x(0) = 1$.

(a) Explain how the logistic equation represents a reasonable model of population dynamics.

(b) Using the substitution $u(t) = 1/x(t)$, recast the ODE in terms of u and hence determine $x(t)$.

[Level : 1]

(a) The logistic equation has the form

$$\frac{\mathrm{d}x}{\mathrm{d}t} = Ax - Bx^2 \quad ,$$

where A and B are positive. This is the same equation as we saw for the nuclear reactor in the previous tutorial. The A term represents growth of the population due to births; the direct dependence on x is reasonable because the more specimens there are, the greater the rate at which they produce babies (reproducing sexually or otherwise). The $-Bx^2$ term represents the decay of population due to deaths. Its origin is more empirical, and reflects the experimental observation that specie populations tend to saturate at a certain finite value, called the carrying capacity. As we saw for the reactor, this model yields a carrying capacity of A/B and indicates that the level is stable. Some motivation for the quadratic dependence of the death term comes from the fact that a fraction of the deaths are from contagious diseases, which spread from specimen to specimen. The spread of such diseases is proportional to the number of infected specimens as well as the number of uninfected specimens, and both scale as the total population x. Nevertheless, when modeling biological phenomena, market dynamics and other phenomena whose behaviour is not governed by laws of Nature, we should always keep in mind that the models are empirical and approximate and might or might not show good agreement with reality.

(b) We have $u(t) = 1/x(t)$ whereby (Chain Rule again)

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \frac{\mathrm{d}u}{\mathrm{d}x} \cdot \frac{\mathrm{d}x}{\mathrm{d}t} = -\frac{1}{x^2}\frac{\mathrm{d}x}{\mathrm{d}t} = -u^2\frac{\mathrm{d}x}{\mathrm{d}t} \quad \text{, or}$$

$$\dot{x} = -\frac{\dot{u}}{u^2} \quad .$$

The original ODE is

$$\dot{x} = 24x - 3x^2 \quad .$$

Introducing the substitution we have

$$-\frac{\dot{u}}{u^2} = \frac{24}{u} - \frac{3}{u^2} \quad ,$$

or multiplying throughout by u^2

$$\dot{u} + 24u = 3 \quad .$$

The IC is $x(0)=1$ so $u(0)$ must equal 1 also.

Now of course we have the familiar form T005; $SS=1/8$ and $IC=1$ so the solution is

$$u=\frac{1}{8}+\frac{7}{8}\mathrm{e}^{-24t} \quad .$$

Inverting the substitution, $x=1/u$ or

$$x=\frac{8}{1+7\mathrm{e}^{-24t}} \quad , \qquad *$$

which is the desired answer.

*

COMMENTS : Clearly, $x(t)$ asymptotically approaches the value 8, which is what would be predicted from a stability analysis (as we did on the reactor). Thus, this problem provides another example of a full nonlinear solution validating the result of the linearized analysis.

Here by the way are three kinds of first order nonlinear equations which have exact solutions :

- Bernoulli equation : $y'+P(x)y=Q(x)y^{\beta}$ where β is any real number. To solve it, introduce $u=y^{1-\beta}$. The problem we did here is evidently of this form.
- Ricatti equation : $y'=P(x)+Q(x)y+R(x)y^2$. Here, we must somehow (for example by guessing) know one special solution $y_1(x)$ which satisfies this equation. Then, the substitution $y=y_1+u(x)$ reduces the ODE to a simple equation for $u(x)$ and allows us to find all possible solutions. Exams by the way always provide the y_1 in the question, so if an exam problem has a first order equation which you can't recognize plus a given solution, that means the equation is Ricatti !
- Clairaut equation : $f(y')+xy'=y$. Here, if we take a derivative on both sides and factor, it reduces to simpler equations.

More detailed theory of these equations is not of much interest here. They don't arise much in applications but can occasionally make an appearance; in such an instance, you should be aware that they can be solved exactly. As is happening a lot in this Chapter, the possibility of exact solution arises only because these

equations are first order – there is no generalization of any of these equations to any higher order.

Example 1.4.4

Bifurcation

Consider a variation on the logistic equation

$$\dot{x} = rx - x^2 \quad ,$$

where r can be positive or negative.

Find all the fixed points and determine their stabilities for all values of r.

[Level : 1]

The fixed points clearly are $x^* = 0$ and $x^* = r$. The stability eigenvalue at $x^* = 0$ is r and that at $x^* = r$ is $-r$. What we see is that for positive r, the fp at zero is unstable while that at r is stable. For negative r on the other hand, zero becomes a stable fp while r becomes an unstable fp. Thus, at the point $r = 0$, the two fixed points 0 and r interchange stabilities. This is called a bifuraction. At the point $r = 0$ itself, the stability eigenvalue is zero and linearization test fails. We also note that by allowing r to become negative, the interpretation of the equation as a model of population growth is no longer valid.

Actually, this is an example of one type of bifurcation called “transcritical”. A detailed analysis of all the various cases is more appropriate for a nonlinear dynamics course than here; if you are interested, I will highly recommend “Nonlinear Dynamics and Chaos” by STEVEN STROGATZ or “Nonlinear Dynamics Primer with Applications to Magnetohydronamics” by JAYANTA BHATTACHARJEE, SAGAR CHAKRABORTY and ASOK MALLIK.

Example 1.4.5

Malaria epidemic

After a session of heavy rains, a malaria epidemic breaks out in a locality. The Municipal Corporation responds by initiating an insecticide spraying program. The variables used to track the progress of the epidemic are x, the number of mosquitoes and y, the number of sick people. You are given that the ODEs satisfied by them are

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \begin{cases} ax - b \text{ if } x > 0 \\ 0 \text{ if } x = 0 \end{cases},$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = cx - fy \quad ,$$

where a, b, c and f are positive constants. The initial conditions are $x(0) = x_0$ and $y(0) = 0$ i.e. there are some number of mosquitoes but no sick people to begin with.

(a) Interpret the model, explaining clearly what all the terms represent.

(b) Determine the minimum value of b for which local control of the epidemic is achieved.

[Level : 2]

(a) We are given that the mosquito population can satisfy two different equations depending on whether it is zero or non-zero. This duality arises from the spraying program of the Municipal Corporation; if there are mosquitoes present then the spray will kill some of them while if there are no mosquitoes but the authorities keep spraying, then the population will not be allowed to get off the ground. The term ax is the growth of the mosquitoes on account of reproduction, while the $-b$ represents the extermination by the Corporation. The question does not include the quadratic term to account for carrying capacity because it assumes that the mosquito population will remain far below the carrying capacity (when they will presumably take over the whole city).

The sick population has a growth term which is proportional to the number of mosquitoes. This is reasonable since the mosquitoes are spreading the disease. The sick population has a decay term because the patients recover from their illness; this model features the recovery rate as proportional to the number of sick people, which is again very plausible. (A more pessimistic interpretation of this decay term is that the sick people are dying rather than recovering; even though that wouldn't change the mathematical model, let's go with the brighter picture here.)

(b) Note that this part does NOT require the interpretation of the model ! So on an exam, even if you have no clue as to Part (a), it will be sheer folly to leave this part alone also and go on to the next question !

To solve the system, start from the mosquito equation which is standalone. It has the form T005 and admits the solution

$$x = \frac{b}{a} + \left(x_0 - \frac{b}{a} \right) \mathrm{e}^{at} \quad , \tag{01}$$

when $x > 0$ and the trivial solution $x = 0$ if x once becomes zero.

The sick people equation is

$$\dot{y} + fy = c \left[\frac{b}{a} + \left(x_0 - \frac{b}{a} \right) \mathrm{e}^{at} \right] \quad ,$$

when x is nonzero and

$$\dot{y} + fy = 0 \quad ,$$

when x is zero.

In the second case, the number of sick people clearly decreases exponentially. In the first case, it GROWS exponentially, a fact you can obtain either by solving for y using integrating factors, or taking a sneak peak ahead at undetermined coefficients (N^{o}. 09), or simply by observing that there is a dangerous e^{at} term on the RHS which must be balanced by a similar growth in y. Obviously, in this case, the Corporation fails to achieve local control of the epidemic.

So for local control of the malaria, what we want is that the mosquito population must become zero at some finite time. To achieve this, the thing on the RHS of (01) must attain the value zero. This will happen if $x_0 - b/a$ is negative. Since it is multiplied by positive exponential, its size will keep increasing until it balances the positive b/a and the mosquito population becomes zero. From then on, the sick population will all recover and the epidemic will be over. If $x_0 - b/a$ is positive however then x will never become zero and the disease will spiral out of control. The threshold separating the two cases is $x_0 - b/a = 0$. Hence the minimum value of b for local control of the epidemic is

$$b_{\min} = ax_0 \quad . \tag{*}$$

Physically this is a very reasonable result. It implies that the greater the initial mosquito population and the faster their intrinsic growth rate, the greater the rate at which the Corporation must kill the mosquitoes off.

Example 1.4.6

Fireball

This is a variation on a problem which appeared in the first mid-semester examination in 2011-12/II.

The size of a spherical fireball can be estimated from a simple model of oxygen balance in the ball. One says that the rate of growth of the radius is proportional to the difference between the rate at which oxygen is absorbed from the atmosphere and the rate at which it is consumed by the flames. Absorption of oxygen takes place from the surface of the ball and is proportional to the area, while consumption of oxygen occurs within the bulk of the ball and is proportional to the volume.

(a) Write an ODE for the radius of the ball as a function of time. Assume that all proportionality constants are such that the final equation has all coefficients numerically equal to unity (no floating "2", "π" etc.).

(b) Determine the radius attained asymptotically by the ball.

(c) Does the ball attain the above radius in a finite time ? Justify your answer.

[Level : 2]

(a) The rate of oxygen absorption is proportional to the surface area, which varies as the square of the radius (area of sphere is $4\pi R^2$, but even if you forget the numerical factor, area is always proportional to the square of the linear dimension). The consumption rate is proportional to volume which varies as the cube of the radius ($4\pi/3$ is the factor for sphere, and again the cubic dependence is general). Absorption of oxygen causes the flame size to increase, so it should carry a positive sign while consumption causes the size to decrease so it should carry negative sign. Hence we have

$$* \qquad \frac{\mathrm{d}R}{\mathrm{d}t} = R^2 - R^3 \quad , \qquad *$$

where we have chosen all coefficients to numerically equal unity, as the problem wanted. This is the ODE asked for.

(b) The fixed points of the system are $R^* = 0$ repeated twice and $R^* = 1$. The linearization fails at the former point due to zero eigenvalue but at the latter fp, $\mathrm{d}f/\mathrm{d}R$ equals -1 so it is stable. Hence R asymptotically approaches the value $\underline{R = 1}$.

(c) To handle this part, we must go to the definition of linearization. Whether the time will be finite or infinite will depend on the dynamics very near the fp and not far away. So, we let $R = R^* + \Delta R$ and substitute into the ODE to get

$$\frac{\mathrm{d}}{\mathrm{d}t}(\Delta R) = R^{*2} + 2R^* \Delta R - R^{*3} - 3R^{*2} \Delta R \quad ,$$

where higher order terms in ΔR are rejected. Since $R^*=1$, this reduces to

$$\Delta \dot{R} = -\Delta R \quad . \qquad (01)$$

Of course, if you really understand and remember the process which gave rise to the method of linearization, then you can bypass all this calculation and straightaway write that an eigenvalue of -1 at the fp necessarily means that perturbations from the fp satisfy (01).

Given this equation, we already saw in T010 that it does not attain steady state in a finite time, hence R also does <u>NOT</u> attain the value 1 in a finite time.

* * * * *

This page intentionally left blank

2

HIGHER ORDER HOMOGENEOUS INITIAL VALUE PROBLEMS

This Chapter examines homogeneous differential equations of second and higher order. Initial value problems only are considered i.e. problems where the function and its derivatives are all specified at a single value of the independent variable. Unless poorly defined, such problems always have a unique and well-behaved solution over their domain of validity.

2.1

LCCH EQUATIONS OF ANY ORDER

The most general form of n^{th} order LCCH (that's linear constant coefficient homogeneous) ODE is

$$\frac{\mathrm{d}^n x}{\mathrm{d}t^n} + A_{n-1}\frac{\mathrm{d}^{n-1} x}{\mathrm{d}t^{n-1}} + A_{n-2}\frac{\mathrm{d}^{n-2} x}{\mathrm{d}t^{n-2}} + \ldots\ldots + A_0 x = 0 \quad , \tag{01}$$

where A_0, A_1, ….., A_{n-1} are real constants which may or may not be zero. The leading coefficient i.e. that of $\mathrm{d}^n x/\mathrm{d}t^n$ cannot be zero for the equation to be of n^{th} order, hence we can always divide all the coefficients by it to make its value equal unity. In an initial value problem (IVP) the additional information apart from the equation itself are the **initial conditions (IC)** x, $\mathrm{d}x/\mathrm{d}t$, $\mathrm{d}^2x/\mathrm{d}t^2$, ….., $\mathrm{d}^{n-1}x/\mathrm{d}t^{n-1}$ all specified at **one particular value** of time, usually $t = 0$ (other types of problems have conditions specified at more than one value of the independent variable). A theorem whose proof is outside our scope assures us that a well-defined IVP always has a **unique solution** within its domain of validity.

Now we recall the solution properties of linear and LCCH ODEs from L02. **Property 1 :** An n^{th} order linear homogeneous ODE $\mathbf{L}x = 0$ has n fundamental solutions $x_1(t)$, $x_2(t)$, ….., $x_n(t)$ which are linearly independent of each other. Any arbitrary linear combination of these fundamental solutions is also a solution of the ODE. **Property 3 :** The solution of an LCCH ODE acquires the exponential form $x = \mathrm{e}^{\lambda t}$ where λ is unknown. Plugging the solution into the ODE yields the values of λ. (Property 2 deals with inhomogeneous equations which are outside the scope of this Chapter.)

To solve the ODE (01), we use Property 3 and plug $x = \mathrm{e}^{\lambda t}$ where λ is unknown. Since the k^{th} derivative of this function is $\lambda^k \mathrm{e}^{\lambda t}$, this substitution, followed by peeling off the common factor $\mathrm{e}^{\lambda t}$, reduces (01) to an algebraic equation for λ :

$$\lambda^n + A_{n-1}\lambda^{n-1} + A_{n-2}\lambda^{n-2} + \ldots\ldots + A_0 = 0 \quad . \tag{02}$$

This is called **characteristic equation or characteristic polynomial** of the ODE (01). The possible values of λ are the roots of the characteristic polynomial.

By the fundamental theorem of algebra, the polynomial (02) can be factorized as

$$(\lambda-\lambda_1)(\lambda-\lambda_2).....(\lambda-\lambda_n)=0 \quad ,$$

where the λ_i's are the roots of the polynomial, which may be real or complex and distinct or repeated. Since the polynomial is real, complex roots must come in conjugate pairs. Depending on the nature of each root, we can write a corresponding fundamental solution of the ODE as follows.

- If a root λ_i is **real and occurs only once**, then $\exp(\lambda_i t)$ is a fundamental solution of the ODE (01).
- If a root λ_i is **real and occurs *k* times**, then the corresponding fundamental solutions are $e^{(\lambda_i)t}$, $te^{(\lambda_i)t}$, $t^2e^{(\lambda_i)t}$, ..., $t^{k-1}e^{(\lambda_i)t}$. These k solutions are linearly independent of each other.
- If a **complex pair of roots** λ_i and λ_i* (* denotes complex conjugation) **occurs only once**, then the corresponding pair of fundamental solutions is exponential of the real part times cos and sin of the imaginary part. If $\lambda_i=\alpha+j\beta$ where $j^2=-1$ (the engineering convention, which is the one I will use here), then $\lambda_i^*=\alpha-j\beta$ and the corresponding solutions are $e^{\alpha t}\cos\beta t$ and $e^{\alpha t}\sin\beta t$.
- If a **complex pair of roots is repeated**, then the successive fundamental solutions are the above pair multiplied by t, t^2, etc. as in the repeated real root case.

Putting all the above together, we can account for a total of n linearly independent fundamental solutions of (01). By Property 1, the general solution of (01) is an arbitrary linear combination of the n fundamental solutions i.e. $x_G=C_1x_1(t)+C_2x_2(t)+.....+C_nx_n(t)$. To find the arbitrary constants C, we must use the initial conditions. Since there are n C's and n ICs there is a neat match here (it had better be so for the theory to be consistent).

Wherefrom did I get the four facts I quoted in the bullet points above ? The first is obvious, the explanation for the third is coming up right away while proving the second and fourth is not at all easy – we'll have to wait quite a bit

for that. But the most important thing to remember while solving ODEs is that they hold true.

We have thus solved every possible LCCH ODE in theory (hahahahaha) – now is the time to put theory to practice.

Example 2.1.1

Harmonic oscillator

Solve completely the equation

$$\ddot{x} + \omega^2 x = 0 \quad ,$$

which is called harmonic oscillator. Fit arbitrary initial conditions to the general solution. Discuss some salient features of the solution. Explain how a spring-mass system lends itself to such a model. Compare with the solutions of

$$\ddot{x} - \omega^2 x = 0 \quad ,$$

which is called a repeller. Explain this name. Express the general solution of $y''-k^2y = 0$ in terms of exponential and hyperbolic functions.

[Read Only]

The harmonic oscillator equation

$$\ddot{x} + \omega^2 x = 0 \quad , \tag{01}$$

is probably the single most important ODE in all of science and engineering, occurring in more places than one can possibly imagine. So it is essential that you acquire INTIMATE familiarity with this equation and its solution. Since this is our first higher order LCCH equation, I will solve it long hand, dotting all j's and crossing all t's; this will give you a concrete example of how to apply the abstract theories of the preceding Lecture.

Substituting $x = \mathrm{e}^{\lambda t}$ the characteristic equation is

$$\lambda^2 + \omega^2 = 0 \quad ,$$

which has the solutions $\lambda = \pm \mathrm{j}\omega$. The two fundamental solutions are thus $x_1 = \mathrm{e}^{\mathrm{j}\omega t}$ and $x_2 = \mathrm{e}^{-\mathrm{j}\omega t}$. The general solution is an arbitrary linear combination of these i.e.

$$x_G = \mathbb{C}_1 \mathrm{e}^{\mathrm{j}\omega t} + \mathbb{C}_2 \mathrm{e}^{-\mathrm{j}\omega t} \quad ,$$

where the fancy C's are complex because the exponentials are also complex.

BUT, x must be real – the ODE is perfectly real, often acting as a model for a real physical process, and it cannot be having a complex solution. Two complex numbers always add up to a real resultant if they are conjugates of each other; since $\mathrm{e}^{\mathrm{j}\omega t}$ and $\mathrm{e}^{-\mathrm{j}\omega t}$ are already conjugates, $\mathbb{C}_1$ and $\mathbb{C}_2$ must be conjugates also. Hence $\mathbb{C}_1 = \mathbb{C}_2{}^*$. Thus if $\mathbb{C}_1 = A + \mathrm{j}B$ for some real A and B, $\mathbb{C}_2 = A - \mathrm{j}B$, and

$$x_G = \left(A + \mathrm{j}B\right)\mathrm{e}^{\mathrm{j}\omega t} + \left(A - \mathrm{j}B\right)\mathrm{e}^{-\mathrm{j}\omega t} \quad . \tag{02}$$

Now we use the identities

$$\cos \omega t = \frac{\mathrm{e}^{\mathrm{j}\omega t} + \mathrm{e}^{-\mathrm{j}\omega t}}{2} \quad ,$$

$$\sin \omega t = \frac{\mathrm{e}^{\mathrm{j}\omega t} - \mathrm{e}^{-\mathrm{j}\omega t}}{\mathrm{j}2} \quad .$$

With this substitution, the solution (02) reduces to

$$x_G = 2A\cos \omega t - 2B \sin \omega t \quad .$$

But since A and B are completely arbitrary, so are $2A$ and $-2B$; hence we can rename them to C_1 and C_2 (the C's NOT fancy this time) and write x_G in real form as

$$x_G = C_1 \cos \omega t + C_2 \sin \omega t \quad . \tag{03}$$

This is the general solution of the harmonic oscillator equation.

So you have now seen the long way of going from complex root of the characteristic equation to cos and sin in the general solution. In practice, the moment you get the complex roots you can straightaway write down the corresponding trigonometric solutions bypassing the intermediate steps. Further, for the specific case of harmonic oscillator equation, you should soon get to a point where you can write down the solution without obtaining the characteristic equation at all. A couple more weeks later, you should be as likely to make a mistake in harmonic oscillator equation as to misspell your name.

The primary feature of the solution (03) is that cos and sin are both PERIODIC in t; they go up and down, up and down. This is of course why the equation is called "oscillator". The frequency of oscillation is ω (called the "natural frequency") and the period is $2\pi/\omega$. The SI Unit of ω is rad/s; a related unit is cycles/s or Hertz (Hz), defined as 1 Hz equals 2π rad/s. Since radian is dimensionless, ω has the dimension of 1/time. The solutions neither grow, nor decay in time. The size of the oscillations will be determined by the initial conditions – let us now see how that part works out.

In the most general case, you will have arbitrary ICs, $x(0) = x_0$ and $\dot{x}(0) = v_0$. Fitting the x IC first, set $t = 0$ into the general solution (03); $\cos 0 = 1$ and $\sin 0 = 0$ so we have $C_1 = x_0$. For the derivative IC, differentiate (03) to get

$$\dot{x}_G(t) = -\omega C_1 \sin \omega t + \omega C_2 \cos \omega t \quad ,$$

now set $t = 0$ here to get $\omega C_2 = v_0$ or $C_2 = v_0/\omega$. Thus the specific solution which fits these ICs is

$$x = x_0 \cos \omega t + \frac{v_0}{\omega} \sin \omega t \quad . \tag{04}$$

Clearly, the larger the values of x_0 and v_0, the larger are the resulting oscillations.

In case the ICs are not defined at $t = 0$ but at some other time, then we use the fact that the harmonic oscillator equation, and in fact ANY other ODE where time

does not appear explicitly, is invariant under the transformation $t \rightarrow t+\tau$ for all τ. Accordingly, if the ICs are given at say $t = T$, we define a new variable $t' = t - T$. Since the equation remains invariant, so does the general solution, which is $x_G = C_1 \cos \omega t' + C_2 \sin \omega t'$. The ICs are specified at $t = T$ which means at $t' = 0$, and we can use the coefficient formula (04) to write the specific solution as $x = x_0 \cos \omega t' + (v_0/\omega) \sin \omega t'$. Finally, substituting $t - T$ for t', we have

$$x = x_0 \cos \omega (t - T) + \frac{v_0}{\omega} \sin \omega (t - T) \quad .$$

This shifting of the time axis might appear obvious but it can save you a lot of calculation if the examiner gives you ICs at a time other than zero, or if a problem naturally throws up such ICs (for the record, it will at a later stage in this course).

A very simple but universal mechanical system acts as the archetype of the harmonic oscillator equation. This is a mass which is connected to a spring and subject to no other forces. The definition of a spring is that if it is displaced from its equilibrium, it exerts a restoring force proportional to the displacement. Let m be the mass and x its displacement from equilibrium. By definition, the spring force F has magnitude kx where k is called the spring constant. Restoring means that if the displacement is towards right, the force is towards left, hence F must be $-kx$. Take special note of the negative sign here. Now from NEWTON's Second Law,

$$m\ddot{x} = -kx \quad , \qquad (05)$$

or,

$$\ddot{x} + \frac{k}{m} x = 0 \quad ,$$

which is the harmonic oscillator equation if we make the obvious identification $\omega^2 = k/m$.

Let us see what would happen if the word "restoring" were absent from the definition of spring, or we were to make a sign error in formulating (05). This equation would be

$$\ddot{x} - \omega^2 x = 0 \quad ,$$

with the general solution

$$x_G = C_1 e^{\omega t} + C_2 e^{-\omega t} \quad .$$

Forthwith you should be seeing an array of red signals; here is the "bad guy" exponential growth from Chapter 1 ! It doesn't matter if ω is positive or negative; one of the terms is trouble either way.

If this reminds you of T005, then you are good at spotting patterns. There we had decaying motions with the correct sign, while here we have bounded motions with the correct sign. In both cases the wrong sign leads to growing solutions, and is a catastrophic error to make in a physical modeling problem. The reasoning behind this similar behaviour is also similar. We have already seen in T005 what happens in the first order case, so here I will only discuss the present case. For the oscillator, force always opposes displacement so bounded motions are physically very plausible. Decaying motion could also have been an option; it is ruled out because of the inertia of the mass – when the force has become zero, it is already going ahead with some velocity. On the other hand, if the sign on force were reversed, it would augment displacement rather than opposing it. If you displaced the mass to the right, it would go further and further rightwards while accelerating continuously. This would lead to the growing solution. This system is called a repeller because the mass is repelled away from the point $x = 0$; a small (infinitesimal) rightward displacement sends it to $+\infty$ while a small leftward displacement sends it to $-\infty$.

As with the first order equations, the oscillator equation is vastly more common in reality than the repeller equation. Other than spring, a system which satisfies oscillator equation in time is an LC circuit (coming in N°. 11). Many systems satisfy harmonic oscillator equations with the independent variable being space – some examples are electric potential and quantum mechanical wave function.

In mechanical systems, harmonic oscillator often arises in contraptions which at first sight have nothing to with springs. For example, the small oscillations of a pendulum near the vertical are harmonic oscillations. This happens because the result of linearizing a mechanical system about a fixed point is usually a harmonic oscillator (we'll see some examples as the course goes on), and contributes to the ubiquitousness of this equation. A physical example of a repeller is hard to find, except as a linearization of a mechanical system about an unstable fixed point. For example, the behaviour of a pendulum near the inverted position is governed by the repeller equation. True enough, howsoever

carefully you release a pencil standing on its tip, it falls over before you can say bingo.

Although dynamical repellers are unphysical, the repeller equation structure does have physical significance in some applications where the independent variable is space. So I will now spend some time on the equation

$$\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} - k^2 y = 0 \quad . \tag{06}$$

We have already seen that its general solution is

$$y_G = C_1 \mathrm{e}^{kx} + C_2 \mathrm{e}^{-kx} \quad . \tag{07}$$

Now the hyperbolic functions are defined as

$$\cosh kx = \frac{\mathrm{e}^{kx} + \mathrm{e}^{-kx}}{2} \quad \text{, and}$$

$$\sinh kx = \frac{\mathrm{e}^{kx} - \mathrm{e}^{-kx}}{2} \quad .$$

In terms of these we can write (07) as

$$y_G = (C_1 + C_2)\cosh kx + (C_1 - C_2)\sinh kx \quad .$$

But for arbitrary C_1 and C_2, C_1+C_2 and C_1-C_2 are two new arbitrary constants. So if we let C_1 and C_2 denote arbitrary constants in general, then we can rewrite this general solution as

$$y_G = C_1 \cosh kx + C_2 \sinh kx \quad . \tag{08}$$

Of course, C_1 and C_2 don't stand for the SAME two arbitrary constants in (07) and (08) – rather, they stand for generic arbitrary constants. The derivatives of the hyperbolic functions are

$$\frac{\mathrm{d}}{\mathrm{d}x}\cosh kx = k \sinh kx \quad ,$$

$$\frac{\mathrm{d}}{\mathrm{d}x}\sinh kx = k \cosh kx \quad ,$$

which you can easily verify from the definitions.

So we have two equivalent forms (07) and (08) of writing the general solution of (06). Both representations have their own advantages and disadvantages. For IVPs (which is the only thing we'll do in this and the next Chapter, and where the

repeller equation doesn't really arise), the exponential form is simple, and usually sufficient. For boundary value problems however the hyperbolic form is sometimes more convenient – although we won't get to these problems for another 150 pages, no harm if you already know about both forms of the general solution.

Example 2.1.2

Repeated root and reduction of order

(a) Write the second order LCCH ODE for the function $y(x)$ whose characteristic polynomial has the root $\lambda = -2$ repeated twice.

(b) Check that $y_1(x) = e^{-2x}$ is a solution.

(c) Try a second solution of the form $y_2(x) = v(x)y_1(x)$ where $v(x)$ is an unknown function. Substitute it into the ODE and hence obtain its general solution.

[Level : 2]

(a) This is an 'inverse problem' – the usual thing is, given the equation find the solution, but this says given the solution find the equation. We are given that the characteristic equation has the roots $-2, -2$. Hence it must be

$$(\lambda-(-2))(\lambda-(-2))=0 \quad .$$

Note the double negative on the '2's ! This simplifies to

$$\lambda^2+4\lambda+4=0 \quad .$$

Since λ in the characteristic equation means a derivative in the differential equation, and since the target function here is given to be $y(x)$ rather than $x(t)$ – this mistake, although trivial, is quite common in exams! – we have

* $$\frac{\mathrm{d}^2y}{\mathrm{d}x^2}+4\frac{\mathrm{d}y}{\mathrm{d}x}+4y=0 \quad , \qquad (01)^*$$

which is the ODE asked for. Note the 'y' attached to the last '4' : the required ODE is NOT $y''+4y'+4=0$! This is another very common error [and much more serious than writing $x(t)$ instead of $y(x)$] in the exam environment.

(b) The check of satisfaction is always by plugging : since $y' = -2\,\mathrm{e}^{-2x}$ and $y'' = 4\,\mathrm{e}^{-2x}$ we have

$$ODE \to 4\mathrm{e}^{-2x}-8\mathrm{e}^{-2x}+4\mathrm{e}^{-2x} \quad ,$$

which is clearly zero, showing that e^{-2x} is a solution.

(c) The question tells us how to start : try $y_2 = vy_1$. The derivatives are (Product Rule)

$$y_2' = v'y_1+vy_1' \quad ,$$
$$y_2'' = v''y_1+2v'y_1'+vy_1'' \quad ,$$

and now plugging into the ODE gives

$$v''y_1+2v'y_1'+vy_1''+4\left(v'y_1+vy_1'\right)+4vy_1=0 \quad .$$

Rearranging terms,

$$v''y_1+v'\left(2y_1'+4y_1\right)+v\left(y_1''+4y_1'+4y_1\right)=0 \quad ,$$

and we are ready for a big simplification.

Now is the key step in the process so watch carefully. The coefficient of v in the above equation is the ODE itself with y_1 in place of y. But since y_1 is a SOLUTION function of the ODE, that term is identically zero. I could have performed this

procedure by substituting $y_1 = \mathrm{e}^{-2x}$ right at the start, in which case the term would have just evaluated to an algebraic zero. However, this is the most general way of going about it. Reduction of order works for linear ODEs of any order, and not just constant coefficient ones; in all cases, the starting substitution is $y_2 = vy_1$ and the coefficient of v becomes identically zero through this argument.

So we have

$$v''y_1 + v'\left(2y_1' + 4y_1\right) = 0 \quad .$$

This is the equation of reduced order – if we let $w = v'$ then this is a FIRST order ODE for w. Hence the name of this procedure. Now substituting $y_1 = \mathrm{e}^{-2x}$ we see that in this case the coefficient on v' (which is w) also evaluates to zero. Unlike the zero on v, this is non-generic and only happened because of the specific nature of the ODE; in the general case this term does not vanish. Thus, the reduced equation is

$$v'' = 0 \quad .$$

This has the solution $v = Ax + B$ for arbitrary constants A and B, hence the second solution of (01) is

$$y_2 = vy_1 = \left(Ax + B\right)\mathrm{e}^{-2x} \quad .$$

But the term $B\,\mathrm{e}^{-2x}$ here is just a multiple of the first solution, so it can be dropped. Hence the second fundamental solution of the ODE (01) is $y_2 = x\,\mathrm{e}^{-2x}$ and the general solution is

$$y_G = C_1\mathrm{e}^{-2x} + C_2 x\,\mathrm{e}^{-2x} \quad , \qquad (*)$$

which is the desired answer.

*

COMMENTS : This example acts as a demonstration of the repeated real roots case we considered in L05. Modification of the coefficients 4 and 4 to letter values α and β would turn it into a proof rather than a demo for the second order case. However, we will do the proof later, for arbitrary order, using a different technique – see T042.

Example 2.1.3

Damped harmonic oscillator

Solve completely the equation

$$\ddot{x} + \gamma\dot{x} + \omega^2 x = 0 \quad ,$$

where γ is a positive parameter. This is called damped oscillator equation. Fit arbitrary initial conditions to the general solution. Discuss some salient features of the solution. Explain how the equation acts as a model for physical systems. Compare with the solution of

$$\ddot{x} - \gamma\dot{x} + \omega^2 x = 0 \quad ,$$

which is said to be negatively damped.

[Read Only]

The ODE is

$$\ddot{x} + \gamma\dot{x} + \omega^2 x = 0 \quad ,$$

where γ is positive (I could have written γ^2 but this is nearly universal convention). Note that γ and ω both have dimensions of 1/time, hence γ/ω is a ratio, a dimensionless number. The characteristic equation is

$$\lambda^2 + \gamma\lambda + \omega^2 = 0 \quad ,$$

which has the solutions (SRIDHARACHARYA's formula)

$$\lambda = \frac{-\gamma \pm \sqrt{\gamma^2 - 4\omega^2}}{2} \quad . \qquad (01)$$

Depending on whether γ is less than, greater than, or equal to 2ω, three cases arise here. We will consider each case separately.

CASE 1 : Underdamped

In this case $\gamma < 2\omega$. The thing inside the radical in (01) is negative so the surd is imaginary and the characteristic roots complex. We have

$$\lambda = -\frac{\gamma}{2} \pm \mathrm{j}\frac{\sqrt{4\omega^2 - \gamma^2}}{2} \quad .$$

Note that the radicand has reversed in sign because the j has come outside ! Taking out the j and not flipping the sign is a very common mistake on the course exams ! The corresponding solutions are exponential of the real part times cos and sin of imaginary part i.e.

$$x_G = C_1 \mathrm{e}^{-\gamma t/2}\cos\omega_d t + C_2 \mathrm{e}^{-\gamma t/2}\sin\omega_d t \quad \text{, where}$$

$$\omega_d = \frac{\sqrt{4\omega^2 - \gamma^2}}{2} \quad .$$

The frequency ω_d is called the damped frequency and the system is said to be underdamped.

Let us fit arbitrary initial conditions $x(0) = x_0,\ \dot{x}(0) = v_0$ to this solution. Plugging $t = 0$ into the general solution gives $C_1 = x_0$. The derivative is

$$\dot{x}_G = \mathrm{e}^{-\gamma t/2}\left(-\omega_d C_1 \sin\omega_d t + \omega_d C_2 \cos\omega_d t\right) - \frac{\gamma}{2}\mathrm{e}^{-\gamma t/2}\left(C_1\cos\omega_d t + C_2 \sin\omega_d t\right) \quad ,$$

and setting $t = 0$ here gives

$$\omega_d C_2 - \frac{\gamma}{2} C_1 = v_0 \quad .$$

Combining these we have the specific solution

$$x = \mathrm{e}^{-\gamma t/2}\left[x_0 \cos\omega_d t + \frac{1}{\omega_d}\left(v_0 + \frac{\gamma x_0}{2} \right)\sin\omega_d t \right] \quad . \qquad (02)$$

It is a ready check that in the limit $\gamma \to 0$ this expression reduces to the one we found in T020 for the undamped oscillator.

It is also noteworthy that in the limit of small damping, $\gamma << \omega$, the damped frequency is very nearly equal to the undamped or natural frequency; the difference is second order in γ/ω. Hence for lightly damped systems, you can treat the oscillation frequency as the natural frequency with very small error.

CASE 2 : Overdamped

In this case $\gamma > 2\omega$. Both roots are real, being given by the formula (01). Note that both the roots are negative – the one with the minus sign on the radical is obvious while the other is $-\gamma$ plus square root of something strictly less than γ. Hence, if we call the roots $-\alpha$ and $-\beta$, the general solution is

$$x_G = C_1 \mathrm{e}^{-\alpha t} + C_2 \mathrm{e}^{-\beta t} \quad .$$

To fit arbitrary ICs x_0 and v_0, we plug $t = 0$ into x_G and its derivative $\dot{x}_G = -\alpha C_1 \mathrm{e}^{-\alpha t} - \beta C_2 \mathrm{e}^{-\beta t}$ to get

$$C_1 + C_2 = x_0 \quad ,$$
$$-\alpha C_1 - \beta C_2 = v_0 \quad .$$

To solve these, we use KRAMER's rule. This states that the solution of

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} u \\ v \end{bmatrix} \quad ,$$

for unknowns x and y is

$$\begin{bmatrix} x \\ y \end{bmatrix} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}\begin{bmatrix} u \\ v \end{bmatrix} \quad ,$$

if $ad - bc$ does not equal zero. If it does equal zero then there is trouble; we need not go into this here, as it doesn't arise in solving IVPs. If you are unfamiliar with KRAMER's rule, now is a good time to become acquainted with it in the 2x2 case.

Although the proofs of it etc are not at all relevant, simultaneous equations come up often while fitting ICs and later while doing problems involving undetermined coefficients, and the Rule will help you to crack them formulaically. Of course, you can use other methods also to solve simultaneous equations, but I'm recommending this one because the process can be automated. Using this rule, we get the specific solution

$$x = \frac{-v_0 - \beta x_0}{\alpha - \beta} \mathrm{e}^{-\alpha t} + \frac{v_0 + \alpha x_0}{\alpha - \beta} \mathrm{e}^{-\beta t} \quad . \tag{03}$$

Of course, whatever method you use to solve the simultaneous equations for C_1 and C_2, your answer should be the same.

CASE 3 : Critically damped

In this case $\gamma = 2\omega$. The radicand in (01) is zero and the two roots are $-\gamma/2$ repeated twice. The general solution is

$$x_G = C_1 \mathrm{e}^{-\gamma t/2} + C_2 t \mathrm{e}^{-\gamma t/2} \quad .$$

For ICs x_0, v_0 we have $C_1 = x_0$. The derivative of x_G is

$$\dot{x}_G = C_2 \mathrm{e}^{-\gamma t/2} - \frac{\gamma}{2}\left(C_1 + C_2 t\right)\mathrm{e}^{-\gamma t/2} \quad .$$

Setting $t = 0$ here leads to the specific solution

$$x = \left[x_0 + \left(v_0 + \frac{\gamma x_0}{2} \right) t \right] \mathrm{e}^{-\gamma t/2} \quad . \tag{04}$$

This completes the three cases of the damped harmonic oscillator. The primary feature of the solution in all three cases is the EXPONENTIAL DECAY of x. If you have a good memory, you should mug up the final formulae (02-4) – they come up quite a lot and take a few minutes to derive in the exam hall. Again, like the undamped oscillator, the damped harmonic oscillator equation does not feature time explicitly, so if ICs are prescribed at $t = T$ then all formulae are the same with t being replaced by $t - T$.

The archetypal system governed by the damped harmonic oscillator equation is a mass and spring which is connected to a dashpot. A dashpot is a chamber filled with viscous fluid, which exerts a retarding force proportional to the velocity. So

the forces on the mass are the spring force $-kx$ (recall that x is displacement from the spring's equilibrium) and the damping force $-C\dot{x}$ where C is the damping constant (very importantly, the force is negative because it RETARDS motion). Writing NEWTON's Second Law, we have $m\ddot{x} = -kx - C\dot{x}$, which becomes the canonical damped harmonic oscillator equation if we make the identification $\omega = \sqrt{(k/m)}$ and $\gamma = C/m$. LC circuits also satisfy the damped harmonic oscillator equation if they have non-zero resistance (which they always do) – the resistance contributes a first order damping term. In some systems, the friction might have a different form in reality (say the internal friction of a supposedly undamped spring) and be impossible to handle analytically. In such cases also, a first order damping term with suitably chosen γ is often a good enough approximation which yields a reasonable answer.

When the sign on the damping is reversed,

$$\ddot{x} - \gamma\dot{x} + \omega^2 x = 0 \quad ,$$

we can again partion the situation into three cases like we did for the regular situation. The primary feature is the exponential growth of the solution in all three cases (verify this yourself or see the next Tute). Once again we have a scenario where a sign flip leads to bounded motions becoming unbounded, so this is another sign which you have to be very careful of while writing models. As always, the positive sign occurs far more commonly in reality than the negative – one example where we DO see the negative however is aeroelastic flutter, coming up in more detail in later Chapters.

In vibration theory, it is very common practice to specify the damping in terms of the ratio of γ to ω rather than the actual value of γ. This is especially so if the damping is light. Because of the special nature of $\gamma = 2\omega$, the ratio most commonly used is $\gamma/2\omega$; in words, the damping is expressed as a percentage of the critical. Sometimes, a percentage is quoted without mentioning relative to what; in such cases, you must assume that it is the critical. Thus, if for a machine part, ω is given to be 100 and damping is quoted as 2 percent, that means γ equals 4 and not 2.

Example 2.1.4

Bounded solutions

This is a variation on a problem which appeared in the final examination in 2011-12/II.

(a) Considering the second order LCCH ODE

$$\ddot{x} + a\dot{x} + bx = 0 \quad ,$$

where a and b are arbitrary real numbers, determine the most general conditions on a and b such that the solution starting from an arbitrary IC remains bounded in time.

(b) In which of the cases above do the solutions decay to zero in time ?

(c) Extend the reasoning to obtain the signs of all the coefficients on a third order ODE (leading coefficient unity) for its all solutions to decay in time.

[Level : 2]

(a) For boundedness of the solution starting from arbitrary IC, we need both the fundamental solutions to be bounded. Writing and solving the characteristic equation we see that :

- If $a=0$, $b>0$ we have the harmonic oscillator where both fundamental solutions are purely oscillatory.
- If $a=0$, $b<0$ we have the repeller and one solution grows.
- If $a=0$, $b=0$ then the solution is $x=C_1t+C_2$ which grows linearly.
- If $a<0$ i.e. $a=-\alpha$ where α positive then there is negative damping. Either it is negatively damped spring or negatively damped repeller. The roots of the characteristic equation are $\lambda=\dfrac{\alpha\pm\sqrt{\alpha^2-4b}}{2}$, which can be real or complex but either way both the roots represent growth.
- If $a>0$, i.e. $a=\alpha$ where α positive and $b>0$ i.e. $b=\beta$ where β positive then it is positively damped spring where we saw that all solutions decay.
- If $a>0$ i.e. $a=\alpha$ where α positive and $b<0$ i.e. $b=-\beta$ where β positive then it is positively damped repeller. The characteristic roots are $\lambda=\dfrac{-\alpha\pm\sqrt{\alpha^2+4\beta}}{2}$, where the root with the positive sign is positive and the corresponding solution grows.
- If $a>0$ i.e. $a=\alpha$ and $b=0$ then the roots are 0 and $-\alpha$. The first of these corresponds to a constant solution while the second shows exponential decay.

Thus, the conditions for bounded solutions are $a\geq 0$, $b\geq 0$ but not $a=b=0$. In words, the equation can contain NEITHER a negative damping term NOR a repeller term.

(b) The solution decays to zero in time if $a>0$ and $b>0$, positively damped spring.

(c) For a third order equation, it can either have all roots real or one complex conjugate pair and one real. Either way, one root must be real, and we start from there. Since the real root must be producing a decaying solution, it must be negative, say $\lambda=-c$. But the other two roots must correspond to a second order

equation with decaying solutions, as in Part (b) above. Hence the structure of the characteristic polynomial is

$$\left(\lambda^2 + a\lambda + b\right)\left(\lambda + c\right) = 0 \quad ,$$

and all its coefficients are positive. Hence the third order ODE must have all coefficients positive for the solutions to decay in time.

Example 2.1.5

Special IC

Consider the ODE

$$\ddot{x} + \dot{x} - 6x = 0 \quad .$$

(a) Obtain a special relation between initial position and initial velocity such that the resulting solution remains bounded in time.

(b) A laboratory apparatus is built where the position of a particle satisfies this ODE, and the particle is then released with the special IC. Does the particle's motion really remain bounded ? Justify your answer.

[Level : 1]

(a) The characteristic equation is

$$\lambda^2 + \lambda - 6 = 0 \quad ,$$

which has the roots $\lambda = -3$ and $\lambda = 2$. Hence the general solution is

$$x_G = C_1 \mathrm{e}^{-3t} + C_2 \mathrm{e}^{2t} \quad .$$

For a bounded solution, C_2 must equal zero, so this solution has the form $C\mathrm{e}^{-3t}$ where C is arbitrary. Its derivative is $-3C\mathrm{e}^{-3t}$, so at $t = 0$, $x_0 = C$ and $v_0 = -3C$, giving the relation $\underline{v_0 = -3x_0}$, which is the desired answer. Note that this way of doing it is less work than fitting x_0 and v_0 to x_G and then imposing that C_2 be zero.

(b) We have seen similar questions before in the context of unstable fixed points in first order systems. The answer is NO. In the lab there will always be error in setting the IC (v_0 will become either $-3.001x_0$ or $-2.999x_0$) and there will be disturbances from outside (vibrations from the experimenter walking around or something) which will cause C_2 and hence the growing solution to become small rather than zero. Thereafter, due to exponential nature of the growth, it will take an instant for 'small' to become 'large' and that will be it.

Example 2.1.6

Approximation analysis

Each of the equations in Parts (a) to (c) is second order LCCH and has the property that two coefficients equal 1 while the third equals 0.01.

(a) $\ddot{x}+0.01\dot{x}+x=0$

(b) $\ddot{x}+\dot{x}+0.01x=0$

(c) $0.01\ddot{x}+\dot{x}+x=0$

For each equation, perform an approximation by throwing away the term whose coefficient is 0.01. Solve both the actual and approximate equations with the ICs $x_0=1$ and $v_0=0$. Compare the two solutions and hence discuss the regime of validity of the approximation in each case. Clearly mention what happens if the initial velocity is given as some other value in Part (c).

[Level : 2]

EXTRA :

(d) Show how the solution of an equation like the one in Part (c) can feature a boundary layer.

(a) The full equation is the damped harmonic oscillator with very slight damping; the approximate equation is the undamped harmonic oscillator (γ=0). To fit the IC in the full system, use the formula (T02202); further use that since damping is ultra light, $\omega_d = \omega$. Then we have

$$x = \mathrm{e}^{-0.005t}\left(\cos t + 0.005 \sin t\right) \quad .$$

The undamped solution is of course

$$x = \cos t \quad .$$

Now to compare the two solutions.

The extra $0.005 \sin t$ in the actual solution is ignorable, but the $\mathrm{e}^{-0.005t}$ outside is certainly not. At small times when $\mathrm{e}^{-0.005t}$ is very close to unity, the approximation is excellent. But as t starts getting larger, the exponential keeps getting smaller, and the approximation worse. At large times of course, the approximation is pure garbage – the actual equation has a decaying solution while the approximate solution oscillates with unchanging amplitude. Hence this approximation scheme is good for small times.

(b) The general solution of the full system is

$$x_G = C_1 \mathrm{e}^{-0.99t} + C_2 \mathrm{e}^{-0.01t} \quad ,$$

and using the formula (T02203) with $\alpha = 0.99$ and $\beta = 0.01$ to fit the ICs we have

$$x = -\frac{0.01}{0.98}\mathrm{e}^{-0.99t} + \frac{0.99}{0.98}\mathrm{e}^{-0.01t} \quad .$$

The solution of the reduced system is

$$x_G = C_1 \mathrm{e}^{-t} + C_2 \quad ,$$

and an easy IC fit gives $C_1 = 0$ and $C_2 = 1$ whereby

$$x = 1 \quad .$$

Now to compare the two solutions.

The first term in the actual solution is $O(0.01)$ so it is ignorable, but again the second term has a decay which is not captured by the approximate solution. Hence this approximation, like the one in Part (a), is good for small times and deteriorates without bound as time goes on.

(c) The general solution of the full equation is

$$x_G = C_1 \mathrm{e}^{-98.99t} + C_2 \mathrm{e}^{-1.0102t} \quad ,$$

and (T02203) with $\alpha = 99$ and $\beta = 1.01$ gives

$$x = -\frac{1}{98}\mathrm{e}^{-99t} + \frac{99}{98}\mathrm{e}^{-1.01t} \quad .$$

The approximation here is the most audacious : the reduced system is

$$\dot{x} + x = 0 \quad ,$$

which is FIRST order and cannot even accommodate two initial conditions ! The general solution of this equation is of course

$$x_G = C\mathrm{e}^{-t} \quad ,$$

and to fit the IC we must go with x_0 alone, giving $C = 1$.

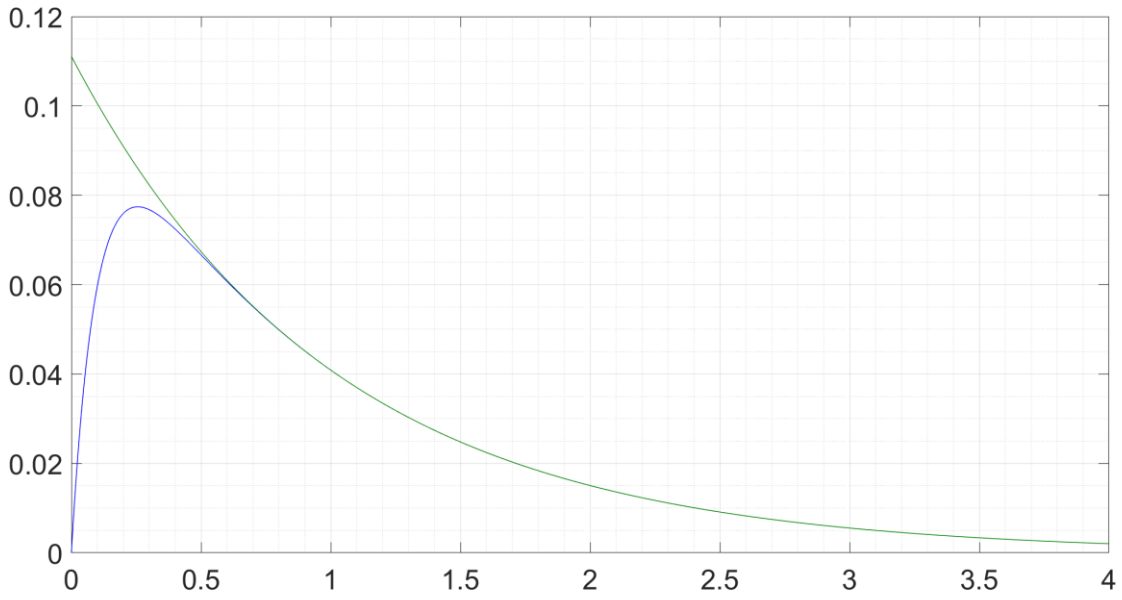
This approximation is an excellent fit to the actual solution at all times – the e^{-99t} component becomes dead as soon as the system starts to evolve, and what we are left with is $(99/98)\,\mathrm{e}^{-1.01t}$. This has a 1 percent error in the coefficient and 1 percent error in the exponent, both of which are insignificant. But of course, since we threw away an IC, we have to examine the possibility of the agreement being entirely fortuitious. If v_0 had some value other than zero, then the coefficient of e^{-t} in the general solution would change – the greater v_0, the greater the change. But since in the formula (T02203) v_0 is balanced by αx_0 with α being the large root, it would take a considerably large v_0 to make the error really gross. Hence the quality of the approximation in this case is governed by the size of v_0 compared to $99x_0$, i.e. the approximation is <u>good for certain initial conditions</u>.

*

COMMENTS : Summarizing the results, in Parts (a) and (b) we saw approximations which hold very good at small time and become worse and worse as time goes on. In Part (c) the error does not continue to grow in time, so with a tractable IC, the approximation is very good at all times. However, the tradeoff is that if v_0 happens to be really greater than $99x_0$ (which can easily happen if the IC features zero or small x) then we have a bad approximation at almost all times, with no region of good accuracy. Such analysis proves very useful when approximating complicated model equations with simpler ones – we'll see examples in later Chapters.

(d) Now let's do the extra bit. The e^{-99t} component in the solution to Part (c), if excited with a finite amplitude, contributes to an extremely rapid change of x in time at very small times, before it becomes practically zero. This sudden change over a very short time produces a unique phenomenon called BOUNDARY LAYER. The concept of boundary layer becomes more practical if the independent variable is space rather than time. For example, let's consider the equation $y''+11y'+10y=0$ with ICs $y(0)=0$ and $y'(0)=1$. The characteristic equation is $\lambda^2+11\lambda+10=0$ which has the two roots -1 and -10, and (T02203) gives the specific solution $y=(1/9)(e^{-x}-e^{-10x})$.

The below Figure shows the graph of this function, and also of the related function $(1/9)\,e^{-x}$. The first plot is in blue and the second one in green.



For $x>1$ the two curves are practically coincident since the e^{-10x} component is negligible. However, at small x this component is significant and ensures that the blue plot curves in and satisfies the condition that $y(0)=0$, which e^{-x} by itself does not satisfy. This region near $x=0$ is the boundary layer, where the function behaves differently from its asymptotic form so as to satisfy the condition at the origin.

In this example we generated the boundary layer by throwing a small coefficient onto the highest derivative – the original Part (c) had a leading coefficient 100 times smaller than the other two while the ODE whose solution we plotted

above has the leading coefficient about 10 times smaller. A salient feature of a boundary layer is that it must have a characteristic thickness. An exponential function does have it : e^{-10x} has a characteristic decay length of order 0.1. This is why exponentials make excellent boundary layers – power law functions like $1/x$ or $1/x^2$, although they too are decaying, don't have a characteristic length scale and don't make good boundary layers.

Finally, why am I spending so much time on this ? Because later, much later, we'll be seeing boundary layers again. Then we'll no longer be dealing with artificially constructed equations but real physical situations where the solutions have this special feature. And there it will be a help if you've already seen the phenomenon in an easy setting.

Example 2.1.7

Fourth order ODEs

Obtain the general solutions :

(a) $\dfrac{\mathrm{d}^4 y}{\mathrm{d}x^4} - y = 0$

(b) $\dfrac{\mathrm{d}^4 y}{\mathrm{d}x^4} + 4y = 0$

[Level : 2]

(a) The characteristic equation is

$$\lambda^4 - 1 = 0 \quad ,$$

i.e. the characteristic roots are the four fourth roots of unity (and they are NOT $\lambda = 1$ repeated four times, which is a regrettably common mistake). If you know how to find them, the problem is trivial (if you don't then see the next part of this question). But roots of unity concept is not essential for this problem since the polynomial can be factorized. We have

$$\begin{aligned} &\lambda^4 - 1 \\ &= \left(\lambda^2 + 1\right)\left(\lambda^2 - 1\right) \\ &= (\lambda + \mathrm{j})(\lambda - \mathrm{j})(\lambda + 1)(\lambda - 1) \end{aligned} \quad ,$$

implying that the four roots are 1, j, -1 and $-\mathrm{j}$.

Given the roots, the four corresponding fundamental solutions of the ODE are e^{x}, e^{-x}, $\cos x$ and $\sin x$ so we can write

* $$y_G = C_1 \mathrm{e}^{x} + C_2 \mathrm{e}^{-x} + C_3 \cos x + C_4 \sin x \quad ,$$ *

which is the desired answer. Again, since $\cosh x$ and $\sinh x$ are linear combinations of e^{x} and e^{-x} (see T020), we can also write this general solution as

* $$y_G = C_1 \cosh x + C_2 \sinh x + C_3 \cos x + C_4 \sin x \quad .$$ *

For the present, both alternatives are equivalent; later we will see that certain situations naturally favour one or the other.

(b) The characteristic equation now is

$$\lambda^4 + 4 = 0 \quad ,$$

i.e. the characteristic roots are the fourth roots of -4. This time there's no option but to do it via the complex numbers way. We first write -4 as a complex number in the polar form : $-4 = 4\,\mathrm{e}^{\mathrm{j}\pi}$. Then, since $\mathrm{e}^{\mathrm{j}2n\pi}$ is unity for any integer n, we add this on as a factor, writing $-4 = 4\,\mathrm{e}^{\mathrm{j}\pi}\,\mathrm{e}^{\mathrm{j}2n\pi}$. So far this factor is purely cosmetic. But now actually extract the fourth root and this cosmetic factor acquires great significance – we have

$$-4^{1/4} = \left(4\,\mathrm{e}^{\mathrm{j}\pi}\,\mathrm{e}^{\mathrm{j}2n\pi}\right)^{1/4} = 4^{1/4}\,\mathrm{e}^{\mathrm{j}\pi/4}\,\mathrm{e}^{\mathrm{j}2n\pi/4} \quad .$$

The first term in the RHS product is $\sqrt{2}$, the next one is $(1+\mathrm{j})/\sqrt{2}$ while the third one varies with n. For $n=0$, it is unity, for $n=1$ it is $\mathrm{e}^{\mathrm{j}\pi/2}$ which is j, for $n=2$ is $\mathrm{e}^{\mathrm{j}\pi}$ which is -1 and for $n=3$ it is $\mathrm{e}^{\mathrm{j}3\pi/2}$ which is $-\mathrm{j}$. Only then, for $n=4$, we get the same value as for $n=0$ and the cycle starts repeating itself. The $n=0$ root is thus $1+\mathrm{j}$ while the other three roots are $-1+\mathrm{j}$, $-1-\mathrm{j}$ and $1-\mathrm{j}$. These are the four fourth roots of -4. We can see that in the complex plane, all the roots are at the same radius and each root makes an angle of 90° with the next one. For arbitrary n^{th} root, this angle generalizes to $360^{\circ}/n$.

To create the real-valued fundamental solutions, we group the roots by conjugate pairs and have

$$y_G = \mathrm{e}^{x}\left(C_1\cos x + C_2\sin x\right) + \mathrm{e}^{-x}\left(C_3\cos x + C_4\sin x\right) \quad .$$

Again, for an alternative representation we expand out the above and collect terms involving $\cos x$ and $\sin x$ getting

$$\begin{aligned} y_G = {} & C_1\cos x\cosh x + C_2\cos x\sinh x \\ & + C_3\sin x\cosh x + C_4\sin x\sinh x \end{aligned} \quad .$$

Either of the above is the desired answer.

Example 2.1.8

Loco coupling to train

A locomotive of mass 123 tons approaches a static train at idle power and a speed of 1 m/s, with intention to couple to the train. The coupler is a spring of constant $k = 2.5\text{x}10^8\ \text{N/m}$, which is initially unstretched. As soon as the loco makes contact with it, the coupler engages and loco and spring move together. The rake is assumed to remain static throughout the coupling process. The damping in the coupler is 5 percent of the critical.

(a) Write an IVP for the displacement $x(t)$ of the spring from its equilibrium as a function of time.

(b) Determine $x(t)$.

[Level : 1]

(a) Since the rake is assumed to be static throughout, the coupler is a spring which is fixed at one end and connected to the loco at the other. This makes for usual mass-spring-damper equation – since the loco is idle during the coupling process, there is no external forcing on the system. So, letting x denote the displacement from the spring's equilibrium, we have

$$m\ddot{x} + C\dot{x} + kx = 0 \quad \text{, or}$$
$$\ddot{x} + \gamma\dot{x} + \omega^2 x = 0 \quad ,$$

where $\gamma = C/m$ and $\omega^2 = k/m$. Here $m = 123000$ (in kg) while $k = 2.5\text{x}10^8$, which gives $\omega = \sqrt{(2.5\text{x}10^8/123000)}$ which is 45 rad/s (approximately). Now the damping is given to be 5 percent of the critical. Recall from T022 that critical damping means $\gamma = 2\omega$; 5 percent of that means $\gamma = \omega/10 = 4.5$ rad/s. So the equation becomes

* $$\ddot{x} + 4.5\dot{x} + 45x = 0 \quad , \qquad (01)^*$$

which is the ODE asked for.

As for the IC, coupling takes place as soon as the loco touches the spring. Since it is in equilibrium at this time, $x(0) = 0$. The velocity at the instant of coupling is simply the approach velocity of the loco, which is 1 m/s, so $\dot{x}(0) = 1$. Combining the two we have $\underline{x_0 = 0 \text{ and } v_0 = 1}$. Note that the ICs are as much a part of the IVP as the ODE itself – if you write only the equation, you risk losing a huge chunk of marks !

(b) From T022, the general solution of (01) is

$$x_G = \text{e}^{-2.25t}\left(C_1 \cos\omega_d t + C_2 \sin\omega_d t\right) \quad \text{, where}$$
$$\omega_d = \frac{\sqrt{8100 - 20.25}}{2} \quad .$$

Since 20 is negligible compared to 8100, we let it go and treat the damped frequency as 45. Then using (T02202) to fit the IC we get the specific solution

* $$x = \frac{1}{45}\text{e}^{-2.25t}\sin 45t \quad , \qquad *$$

which is the desired answer.

*

COMMENTS : This is a very typical applied problem in the course exams so make sure you understand this completely and can do its entirety yourself (WITHOUT taking recourse to T022 – you won't be allowed to carry that into the exam hall!). As always, the signs on (01) are vital – they're plus and plus. Minus sign on either term is far too common – remember that the exponentially growing solutions are as undesirable for the loco pilot as they are for you. If you hack an incorrect equation and get such a solution, STOP at once, go back to the equation, correct the signs and solve again.

We also see here what a simple model of a real phenomenon can and cannot yield. The amplitude of the loco's motion has come out to 2 cm, the oscillation frequency is quite rapid (about 7 cycles per second) and the motion dies out in a couple of seconds, all of which are very plausible. The weakness in the model is the assumption that the train is static throughout. Although the rake's mass (typically 980 tons) is considerably higher than the loco's, the rake itself is also composed of springs and masses so the assumption of stationarity is not a very good one. Indeed, the act of coupling creates a jolt which can be felt all the way back down to the last coach – the more skilful the shunting loco pilot, the less is the jolt. If we want the motion of the train, the model as well as the solution techniques will become considerably more sophisticated – we'll get to that in due course of time.

2.2

LINEAR INDEPENDENCE AND WRONSKIAN

This Unit is primarily of theoretical utility and does not have a significant bearing on practical applications. Hence this material may be skipped in an application-centric course.

The basic results on linear equations say that the n fundamental solution of an n^{th} order linear homogeneous ODE should be linearly independent. Although in most cases, linear independence can be judged by inspection, some equations have such complicated solutions that inspection does not work. Fortunately, there is a systematic method of telling whether a set of n given functions is linearly independent. It consists of constructing their Wronskian determinant, defined as follows.

If $x_1(t)$, $x_2(t)$, ….., $x_n(t)$ be n functions of time then their **Wronskian** is defined as

$$W = \det\begin{bmatrix} x_1 & x_2 & ..... & x_n \\ \dot{x}_1 & \dot{x}_2 & ..... & \dot{x}_n \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ x_1^{(n-1)} & x_2^{(n-1)} & ..... & x_n^{(n-1)} \end{bmatrix},$$

where the superscript $(n-1)$ denotes the $n-1^{\text{th}}$ order derivative $\mathrm{d}^{n-1}/\mathrm{d}t^{n-1}$. In terms of the Wronskian, the functions are **linearly independent if and only if W does not equal zero**.

The Wronskian is named after Polish mad-scientist JOZEF WRONSKI (1776-1853) whose sole aim in life was to prove everyone wrong. His own "theory of everything" overthrew all the scientific findings of GALILEO, NEWTON, LAGRANGE, LAPLACE, (insert any great scientist of your choice who worked prior to 1800) and was met with the ridicule it deserved. The Wronskian was discovered by him in the context of this august pursuit –

he hoped to use it to find an error in an infinite series which LAGRANGE had written, and thus prove the Master false. Although the challenge was wrong as usual, the concept of the determinant was one instance where Pan WRONSKI was right !

Example 2.2.1

Basic Wronskians

Calculate the Wronskian of

(a) $\cos t$, $\sin t$

(b) e^{2t}, $2e^{2t}$

(c) $e^{-\alpha t}$, $e^{-\beta t}$

(d) t^m, t^n where $m \neq n$

(e) $e^{-\gamma t/2}$, $t e^{-\gamma t/2}$

and use it to determine whether the given functions are linearly independent or dependent in each case.

[Level : 1]

(a) We have

$$W = \det\begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix} = 1 \quad ,$$

so linearly independent.

(b) In this case

$$W = \det\begin{bmatrix} \mathrm{e}^{2t} & 2\mathrm{e}^{2t} \\ 2\mathrm{e}^{2t} & 4\mathrm{e}^{2t} \end{bmatrix} = 0 \quad ,$$

so the functions are linearly dependent, as was obvious.

(c) The relevant Wronskian is

$$W = \det\begin{bmatrix} \mathrm{e}^{-\alpha t} & \mathrm{e}^{-\beta t} \\ -\alpha\mathrm{e}^{-\alpha t} & -\beta\mathrm{e}^{-\beta t} \end{bmatrix} = (\alpha-\beta)\mathrm{e}^{-(\alpha+\beta)t} \quad ,$$

so linearly independent.

(d) For this problem we write

$$W = \det\begin{bmatrix} t^m & t^n \\ mt^{m-1} & nt^{n-1} \end{bmatrix} = (n-m)t^{m+n-1} \quad .$$

Since $m \neq n$ this is not zero so the functions are linearly independent (which is not a surprise).

(e) The calculation here is simplified if we pull out the common factor of $\mathrm{e}^{-\gamma t/2}$ from the two functions and then evaluate the Wronskian of the rest. This is OK because a common factor cannot affect the linear independence/dependence of a pair of functions. With this done, the problem reduces to finding Wronskian of 1 and t, which by the previous part ($m = 0$, $n = 1$) is 1. Finally, since the determinant is quadratic in $\mathrm{e}^{-\gamma t/2}$, it gets multiplied by TWO powers of that factor and not just one, so the overall Wronskian is

$$W = \mathrm{e}^{-\gamma t} \quad ,$$

and the functions are linearly independent. The trick of pulling out a common factor is worth noting for the exams – it can simplify your life considerably.

Example 2.2.2

High order Wronskian

It is desired to prove the linear independence of the functions 1, t, t^2, t^3, ..., t^n.

(a) A student presents the following argument : "*Instead of writing a gigantic Wronskian, we break it up into small parts. Considering any one pair (*t^m*,* t^n*), we know from the previous problem that the Wronskian is nonzero and hence they are linearly independent. Since this holds true for all* m *and* n*, i.e. for every possible pair, all the functions are in fact linearly independent of each other.*" Is this argument true or false ? Justify your answer.

(b) Irrespective of the result of the previous part, write the Wronskian of the given functions.

(c) Evaluate the said Wronskian.

[Level : 1]

(a) The argument is false since pairwise linear independence does NOT guarantee overall linear independence of the functions taken as a whole. The three functions $x_1 = 1$, $x_2 = t$ and $x_3 = 1+t$ should serve as sufficient proof of the argument's weakness. This pairwise proof of linear independence is a potential error in exams (I have seen more than one instances of this) so please be careful here.

(b) We can write the Wronskian mostly by inspection, noting that the size of the matrix is $(n+1)$ squared. We have

$$* \qquad W = \det \begin{bmatrix} 1 & t & t^2 & ..... & t^n \\ 0 & 1 & 2t & ..... & nt^{n-1} \\ 0 & 0 & 2 & ..... & n(n-1)t^{n-2} \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ 0 & 0 & 0 & ..... & \lfloor n \end{bmatrix} , \qquad *$$

which is the Wronskian asked for. Be careful not to write the bottom right corner element as zero – you will end up getting it !

(c) We can see that the determinant is lower triangular. Its value is the product of the elements on the diagonals (to see why, expand W along the first column, then the relevant minor along its first column and so on). Hence the value is

$$* \qquad W = 1 \cdot 1 \cdot \lfloor 2 \cdot \lfloor 3 \cdot ..... \cdot \lfloor n-1 \cdot \lfloor n \quad , \qquad *$$

proving that the functions are indeed linearly independent.

Example 2.2.3

ABEL's theorem

Consider the linear but not necessarily constant coefficient equation

$$y''+p(x)y'+q(x)y=0 \quad .$$

Let y_1 and y_2 be the two fundamental solutions and W their Wronskian.

(a) Check that W satisfies the ODE

$$W'+p(x)W=0 \quad .$$

(b) Hence show that if W is not zero at $x=0$ then it is not zero at any x.

[Level : 1]

(a) This is a plug and chug problem; plugging the definition of W into the ODE it is claimed to satisfy we have

$$\begin{aligned} W'+pW &= \frac{\mathrm{d}}{\mathrm{d}x}\left(y_1y_2'-y_2y_1'\right)+p\left(y_1y_2'-y_2y_1'\right) \\ &= y_1\left(y_2''+py_2\right)-y_2\left(y_1''+py_1\right) \end{aligned} .$$

Now use that y_1 and y_2 are SOLUTIONS of the starting second order ODE (till now we haven't used that, only the definition of Wronskian), and this becomes

$$\begin{aligned} W'+pW &= y_1\left(-qy_2\right)-y_2\left(-qy_1\right) \\ &= 0 \end{aligned} ,$$

and we have the desired result.

(b) The ODE for W is first order linear so we solve it using integrating factors (you haven't forgotten that, have you?) to get

$$W = C\exp\left(-\int_0^x p(s)\mathrm{d}s\right) .$$

If $W(0)$ is nonzero then so is C. Now, so long as p is well-behaved, the exponential of its integral cannot be zero, so W is not zero at any x, which was to be shown.

*

COMMENTS : As often with happens with simple and elegant results, the first discovery of this one was by a genius, NIELS ABEL. Thereafter, lakhs of undergraduates mechanically go through the derivation each year as homework and quizzes. With this problem, the boring Unit on Wronskians is finally over. The next Unit again features linear algebra, but this time in a context which has enormous theoretical and practical significance.

2.3

LINEAR ALGEBRA AND LCCH SYSTEMS

Computer linear algebra package software is very useful for this Unit. I have specified everything in terms of Matlab since that is the software I am familiar with. However the steps can be carried out in any software like Mathematica, Maple etc.

This Unit develops the 'LT' of DELTA. The aspects of linearity we have seen so far – namely the definition and three properties of linear equations in 1.2 – have been pretty elementary. As the course progresses however, we'll be encountering more linear algebra, especially in the PDE parts. A precursor to PDEs is the LCCH systems which are the focus of this Unit, so I might as well cover the fundamentals of linear algebra now itself. Note that this linear algebra is not absolutely necessary for a basic differential equations course (for instance, Cornell's MATH2930) – such a course can skip this Unit and also those sections of later material where hardcore linear algebra appears. For a more in-depth understanding however, familiarity with linear algebra is an invaluable asset. To facilitate such familiarity, I am including below a " capsule summary" of the linear algebra concepts we'll require in the rest of the book. This is obviously not a ground-up course in linear algebra but a memory refresher if you've last seen the stuffs some time before. After presenting this summary, I'll proceed to the theory of LCCH systems.

LINEAR ALGEBRA – A CRASH COURSE

Notation : vectors are boldface, upright and in Times New Roman, and usually denoted by a small letter, thus $\mathbf{x}$. Matrices are also boldface, upright and Times New Roman, but they are written in a larger font, and usually with a capital letter, thus $\mathbf{A}$. For column vectors, I will use the Matlab semicolon notation i.e. I will denote the column vector consisting of elements x_1, x_2 etc as $[x_1; x_2; \ldots..; x_n]$ instead of the more conventional but slightly confusing $[x_1 \; x_2 \; \ldots.. \; x_n]^T$.

We'll have to start off somewhere, so let's do it with the vector equation $\mathbf{Ax} = 0$. Here $\mathbf{A}$ is a square matrix of size nxn, $\mathbf{x} = [x_1; x_2; \ldots..; x_n]$ is the unknown vector of size nx1 and 0 denotes the zero vector of size nx1 (I won't be using boldface, enlargement etc for zero vector or zero matrix). This equation is entirely equivalent to the vector equation

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \ldots.. + x_n\mathbf{a}_n = 0 \quad ,$$

where $\mathbf{a}_k$ denotes the k^{th} column of $\mathbf{A}$. By analogy with ODEs, it is called a homogeneous equation. To solve this, we perform the following **elementary row operations** on $\mathbf{A}$: swap two rows, multiply a row by a (non-zero) scalar, and add one row to another row. We do the row ops in such a way as to reduce $\mathbf{A}$ to **echelon form** i.e. convert $\mathbf{A}$ to a matrix which has only zeroes below the main diagonal. Echelons are easy because they can be solved by inspection starting from the bottom.

There are two ways the thing can proceed. One is that the echelon of $\mathbf{A}$ has non-zero entry (called **pivot**) in every diagonal position. This would imply that $x_n = 0$, followed by $x_{n-1} = 0$ etc. until we come all the way back to $x_1 = 0$. In this case there is only one solution of $\mathbf{Ax} = 0$ which is $\mathbf{x} = 0$. This solution is the trivial solution. In this case, the matrix $\mathbf{A}$ is said to be **invertible**. Since the only solution of $\mathbf{Ax} = 0$ is the trivial one, it implies that no column of $\mathbf{A}$ is a linear combination of any other column/s i.e. the columns of $\mathbf{A}$ are all **linearly independent**. The second possibility which can occur is that some positions on the main diagonal are zero. Then, the corresponding unknowns can be arbitrary, and are called free variables. In terms of these free variables, we can solve the rest of the system to get a non-trivial solution. Non-trivial solutions always occur in families; if the vector $\mathbf{x}$ is a solution then $c\,\mathbf{x}$ is also a solution for every scalar c. In this case the matrix $\mathbf{A}$ is **non-invertible**. Since the non-trivial vector $\mathbf{x}$ satisfies $\mathbf{Ax} = 0$, that means that one or more columns of $\mathbf{A}$ can be written as linear combination of the other columns, and the columns of $\mathbf{A}$ are **linearly dependent**.

The next step is the inhomogeneous system $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{b}$ is an arbitrary vector. To solve it, we construct the augmented matrix $[\mathbf{A}\,|\,\mathbf{b}]$ and perform on this matrix the row operations which reduce $\mathbf{A}$ to echelon. If $\mathbf{A}$ is invertible, then this always has a **unique solution**, which we can construct explicitly by

solving the echelon. If $\mathbf{A}$ is non-invertible, then whether there is a solution depends on whether a zero row in the echelon corresponds to a zero in the augmented column. If yes, then the system is consistent. We can find a particular solution by setting the free variables to zero; then the **general solution** is the particular solution plus any of the homogeneous solutions of $\mathbf{Ax} = 0$. This is exactly like in ODEs. There is however a second case here. This happens if a zero row in the echelon corresponds to a non-zero entry in the augmented column. In this case, $\mathbf{Ax} = \mathbf{b}$ has **no solution** at all.

In case $\mathbf{A}$ is invertible, we just saw that $\mathbf{Ax} = \mathbf{b}$ has a unique solution for any and all vector $\mathbf{b}$. This means that there always exist constants x_1, x_2, ….., x_n such that $x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + ... + x_n\mathbf{a}_n$ add up to the vector $\mathbf{b}$. In other words, you can create an arbitrary vector as a linear combination of columns of $\mathbf{A}$. This leads to the definition of **span** : the span of a set of given vectors is the set of all possible vectors which can be expressed as linear combinations of the given vectors. Span leads to another definition : the **column space** of a matrix is the span of its columns. So in terms of these definitions we can say that if $\mathbf{A}$ is invertible, then its column space includes every possible nx1 vector.

Since I've mentioned the word column space, I must define what is a **vector space**. A set of vectors is called a vector space V if it satisfies the following three properties :

- The zero vector belongs to V
- If a vector $\mathbf{v}$ belongs to V then so does $c\mathbf{v}$ for every scalar c
- If two vectors $\mathbf{v}$ and $\mathbf{w}$ belong to V then so does $\mathbf{v}+\mathbf{w}$

These properties are basically the same as the second definition of linear operator I presented in1.2. Our everyday three-dimensional space is a vector space – every vector $[x; y; z]$ lives in it. The x-y plane is also a vector space – it houses all vectors $[x; y]$. The combination of the x-axis and y-axis however is not a vector space – it evidently fails to satisfy the third definitional property.

Given a vector space, now we come to **basis**. A basis for a vector space is a set of linearly independent vectors which span the space. Thus, the columns of an invertible matrix form a basis for the space of all nx1 vectors. The **dimension** of a vector space is the number of elements in its basis – thus, our everyday space is three-dimensional because the three vectors [1; 0; 0],

[0; 1; 0] and [0; 0; 1] form a basis for this space. The dimension of the column space of $\mathbf{A}$ is called the **rank** of $\mathbf{A}$. An invertible matrix $\mathbf{A}$ always has full rank n.

Like column space, another vector space associated with a matrix is the **null space**. This is the set of all vectors $\mathbf{x}$ such that $\mathbf{Ax} = 0$. If $\mathbf{A}$ is invertible then its null space consists of the single vector $\mathbf{x} = 0$. The dimension of the null space is called **nullity**; an important theorem (which we won't prove here) states that the rank of $\mathbf{A}$ added to the nullity of $\mathbf{A}$ always produces n.

Since we're talking about invertible, it makes sense to formally define the **inverse** of a matrix. The inverse of $\mathbf{A}$ is a matrix $\mathbf{A}^{-1}$ such that $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix of size n. The definition of inverse in terms of cofactors etc is quite cumbersome and we won't need it anywhere. But that definition shows that an invertible matrix must have a non-zero determinant while a non-invertible matrix has a zero determinant.

Given a matrix $\mathbf{A}$, suppose that there is a non-trivial vector $\mathbf{v}$ and a scalar λ such that $\mathbf{Av} = \lambda\mathbf{v}$. In such a case, λ is said to be an **eigenvalue** of $\mathbf{A}$ and $\mathbf{v}$ the corresponding **eigenvector**. If there are more than one linearly independent eigenvectors corresponding to a single eigenvalue, then the space spanned by all those vectors is called the eigenspace corresponding to the said eigenvalue. To find λ, we write the eigenvalue equation as $(\mathbf{A}-\lambda\mathbf{I})\mathbf{v} = 0$. Since $\mathbf{v}$ cannot be trivial by definition, $\mathbf{A}-\lambda\mathbf{I}$ must be non-invertible and its determinant must be zero. The condition $\det(\mathbf{A}-\lambda\mathbf{I}) = 0$ turns out to be an n^{th} order polynomial in λ which has a total of n roots, real or complex, distinct or repeated. These roots are the eigenvalues of $\mathbf{A}$. Given a λ, the corresponding eigenvector is a non-trivial solution of $(\mathbf{A}-\lambda\mathbf{I})\mathbf{v} = 0$. This means that the eigenvector can be defined only upto a multiplicative constant – for reasons we'll see shortly, the most conventional choice of this constant is the one that makes the eigenvector a unit vector (definition coming up). It also follows that a non-invertible matrix has zero as an eigenvalue – since $(\mathbf{A}-0\mathbf{I})\mathbf{x} = 0$ has a non-trivial solution for such an $\mathbf{A}$, that $\mathbf{x}$ becomes an eigenvector of $\mathbf{A}$ corresponding to the eigenvalue zero. An invertible $\mathbf{A}$ on the other hand does not have zero as an eigenvalue.

If $\mathbf{A}$ has all real entries, then its all eigenvalues may or may not be real. Complex λ's must come in conjugate pairs, and their corresponding eigenvectors are conjugates also. If all eigenvalues are real, then there can be several cases. The simplest case is when all the eigenvalues are distinct. For any matrix, eigenvectors corresponding to different eigenvalues are linearly independent. In this case, the eigenvectors of $\mathbf{A}$ form a basis of the n-dimensional space. The other case is when $\mathbf{A}$ has a repeated eigenvalue, say the eigenvalue λ_0 is repeated k times. Then, that eigenvalue may have an eigenspace of dimension k or dimension less than k, i.e. there may be k or less than k linearly independent vectors $\mathbf{v}$ such that $\mathbf{Av} = \lambda_0\mathbf{v}$. In the first case, the eigenvectors of $\mathbf{A}$ again form a basis of the total space, but in the second case they don't. In the latter case, the matrix $\mathbf{A}$ is said to be **defective**.

If $\mathbf{A}$ has all real eigenvalues, and if the corresponding eigenvectors span the entire space, then we say that $\mathbf{A}$ is **diagonalizable**. In this case, we can write $\mathbf{A} = \mathbf{PDP}^{-1}$ where $\mathbf{D}$ is a diagonal matrix consisting of the eigenvalues of $\mathbf{A}$ and $\mathbf{P}$ is a matrix whose columns are the corresponding eigenvectors. This is called **similarity transformation**. This transformation shows that, for diagonalizable $\mathbf{A}$, $\mathbf{A}^m = \mathbf{PD}^m\mathbf{P}^{-1}$, for any integer m. The advantage of this representation is that $\mathbf{D}^m$ is simply another diagonal matrix whose diagonal elements are the m^{th} powers of the corresponding entries of $\mathbf{D}$.

There is a special class of matrices which are always diagonalizable; before going to that however, I must introduce the **dot product** (also called scalar product or inner product in mathematically hifi courses). The dot product of two vectors $\mathbf{v}$ and $\mathbf{w}$ is $\mathbf{v}\cdot\mathbf{w} = v_1w_1 + v_2w_2 + \ldots.. + v_nw_n$, which can also be written as $\mathbf{v}^T\mathbf{w}$. Dot product is linear in either of the two elements, thus $(c\mathbf{v})\cdot\mathbf{w} = \mathbf{v}\cdot(c\mathbf{w}) = c(\mathbf{v}\cdot\mathbf{w})$ where c is a scalar, $(\mathbf{v}_1 + \mathbf{v}_2)\cdot\mathbf{w} = \mathbf{v}_1\cdot\mathbf{w} + \mathbf{v}_2\cdot\mathbf{w}$ and $\mathbf{v}\cdot(\mathbf{w}_1 + \mathbf{w}_2) = \mathbf{v}\cdot\mathbf{w}_1 + \mathbf{v}\cdot\mathbf{w}_2$. Dot product is commutative i.e. $\mathbf{v}\cdot\mathbf{w} = \mathbf{w}\cdot\mathbf{v}$ for all $\mathbf{v}$ and $\mathbf{w}$. The dot product of any vector with itself is $\mathbf{v}\cdot\mathbf{v} = v_1^2 + v_2^2 + \ldots.. + v_n^2$, which is positive if $\mathbf{v}$ is not the zero vector. This allows us to define the **length** of $\mathbf{v}$ as $(\mathbf{v}\cdot\mathbf{v})^{1/2}$. A vector whose length is unity is called a **unit vector**; every arbitrary non-zero vector can be rescaled to a unit vector by dividing it by its length. If two vectors $\mathbf{v}$ and $\mathbf{w}$ are such that $\mathbf{v}\cdot\mathbf{w} = 0$, and more generally if a set of vectors $\mathbf{v}_1$, $\mathbf{v}_2$, ….., $\mathbf{v}_n$ is such that

$\mathbf{v}_i \cdot \mathbf{v}_j = 0$ whenever $i \neq j$ then we say that the vectors are **orthogonal**. If further each of the $\mathbf{v}_i$'s be a unit vector, then the vectors are said to be **orthonormal**.

Now let's introduce the matrices which are the most well-behaved from an eigenvalue point of view. These are the **symmetric** matrices – all $\mathbf{A}$ such that $\mathbf{A}^T = \mathbf{A}$. The **spectral theorem** (one of the most important theorems in linear algebra as a whole) states that if a matrix $\mathbf{A}$ is symmetric then

- its all eigenvalues are real
- its eigenvectors span the *n*-dimensional space
- its eigenvectors are orthonormal

I wouldn't have proved this, except that we'll need some of the steps of the proof in a later problem. The primary operation which we need for the proof is the following **transposition trick** : for any (not necessarily symmetric) matrix $\mathbf{A}$ and two vectors $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{x} \cdot (\mathbf{A}\mathbf{y}) = (\mathbf{A}^T \mathbf{x}) \cdot \mathbf{y}$. This follows quickly from the definition of dot product and the fact that $(\mathbf{A}\mathbf{v})^T = \mathbf{v}^T \mathbf{A}^T$, and is a very important thing to remember and use whenever symmetric matrices enter the picture.

Given this, let's first prove that the eigenvalue of symmetric $\mathbf{A}$ is real. Suppose a complex λ is an eigenvalue of $\mathbf{A}$. Then, there is a complex vector $\mathbf{v}$ such that $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$. Since complex eigs occurs in conjugate pairs, there is also a vector $\mathbf{v}^*$ such that $\mathbf{A}\mathbf{v}^* = \lambda^*\mathbf{v}^*$. Now consider the product $\mathbf{v}^* \cdot (\mathbf{A}\mathbf{v})$. Since $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$, this is $\mathbf{v}^* \cdot (\lambda\mathbf{v})$ which is $\lambda(\mathbf{v}^* \cdot \mathbf{v})$. On the other hand, we can use the transposition trick to write $\mathbf{v}^* \cdot \mathbf{A}\mathbf{v} = (\mathbf{A}^T \mathbf{v}^*) \cdot \mathbf{v} = (\mathbf{A}\mathbf{v}^*) \cdot \mathbf{v}$, where the last step uses the fact that $\mathbf{A}^T = \mathbf{A}$ due to symmetric. Now by the eigenvalue condition, this equals $\lambda^*(\mathbf{v}^* \cdot \mathbf{v})$. We thus have the equality $\lambda(\mathbf{v}^* \cdot \mathbf{v}) = \lambda^*(\mathbf{v}^* \cdot \mathbf{v})$, so either $\lambda=\lambda^*$ or $\mathbf{v}^* \cdot \mathbf{v} = 0$. Now, $\mathbf{v}^* \cdot \mathbf{v} = v_1 v_1^* + v_2 v_2^* + \ldots + v_n v_n^*$, which is a sum of strictly positive real numbers unless $\mathbf{v}$ is zero. But an eigenvector cannot be trivial, so $\mathbf{v}^* \cdot \mathbf{v}$ cannot be zero, which means $\lambda = \lambda^*$ i.e. λ is real.

In the same way we'll prove that eigenvectors corresponding to different eigenvalues are orthogonal. Let λ_1 and λ_2 be such that $\mathbf{A}\mathbf{v}_1=\lambda_1\mathbf{v}_1$ and $\mathbf{A}\mathbf{v}_2=\lambda_2\mathbf{v}_2$.

Then consider the dot product $\mathbf{v}_1 \cdot \mathbf{A}\mathbf{v}_2$. By definition of eigenvalue, this is $\mathbf{v}_1 \cdot (\lambda_2 \mathbf{v}_2)$ which equals $\lambda_2 (\mathbf{v}_1 \cdot \mathbf{v}_2)$. Now use the transposition trick and $\mathbf{A}^T = \mathbf{A}$ to write the dot product as $(\mathbf{A}\mathbf{v}_1) \cdot \mathbf{v}_2$ which is $\lambda_1 (\mathbf{v}_1 \cdot \mathbf{v}_2)$. Thus we have got that $\lambda_1 (\mathbf{v}_1 \cdot \mathbf{v}_2) = \lambda_2 (\mathbf{v}_1 \cdot \mathbf{v}_2)$. Since $\lambda_1 \neq \lambda_2$ by definition, $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$, and this part of the proof is complete. To extend orthogonal to orthonormal, we note that every scalar multiple of an eigenvector is itself an eigenvector, so we can normalize the eigenvectors to unit vector with no loss of generality.

Before doing the rest of the proof of spectral theorem, let me digress a bit and briefly explain how we can always convert a set of linearly independent vectors into a set of orthonormal vectors which span the same space. This is done by **Gram Schmidt orthonormalization process**, whose concept is pretty simple, as follows (if only I'd realized this – or anyone taught me this – in my second year). Suppose you have *k* linearly independent vectors $\mathbf{v}_1$, $\mathbf{v}_2$, ….., $\mathbf{v}_k$. Then, you normalize $\mathbf{v}_1$ to a unit vector $\mathbf{e}_1$ by dividing by its length. Then, take $\mathbf{v}_2$ and subtract from it its component parallel to $\mathbf{e}_1$ i.e. do $\mathbf{v}_2 \rightarrow \mathbf{v}_2 - (\mathbf{v}_2 \cdot \mathbf{e}_1)\mathbf{e}_1$. This ensures that the transformed $\mathbf{v}_2$ is orthogonal to $\mathbf{e}_1$. Then normalize this new $\mathbf{v}_2$ to a unit vector $\mathbf{e}_2$. Next do the same thing with $\mathbf{v}_3$: first subtract off its components along the existing orthonormalized unit vectors $\mathbf{v}_3 \rightarrow \mathbf{v}_3 - (\mathbf{v}_3 \cdot \mathbf{e}_1)\mathbf{e}_1 - (\mathbf{v}_3 \cdot \mathbf{e}_2)\mathbf{e}_2$, and then normalize the transformed $\mathbf{v}_3$ to unit vector $\mathbf{e}_3$. This $\mathbf{e}_3$ is orthogonal to $\mathbf{v}_1$ and $\mathbf{v}_2$. You proceed similarly until all *k* vectors have been taken care of, thus converting the set of linearly independent **v**'s to a set of orthonormal **e**'s. Although we won't be doing a Gram Schmidt anywhere in this book, the process sometimes comes in handy because solving $(\mathbf{A}-\lambda\mathbf{I})\mathbf{x} = 0$ with symmetric $\mathbf{A}$ doesn't always yield orthonormal eigenvectors corresponding to a repeated λ. Then, we have to do Gram Schmidt to orthonormalize the linearly independent eigenvectors corresponding to that λ.

Now we come back to the rest of the proof of the spectral theorem. If you wish to skip it, you can go on to the next section of this Lecture with no loss of continuity. (This is why I put the Gram-Schmidt before this – that is practically important and not good to skip.)

Let's summarize where we stand so far as regards the spectral theorem. Its statement had three bullet points – we have shown the first one outright, while proving the second and third points partially, in the case that all eigenvalues

of $\mathbf{A}$ are distinct. In this case, we just showed the third point above; since n orthonormal vectors in n-dimensional space span it automatically, the second point also follows immediately. The remaining case is where $\mathbf{A}$ has a repeated eigenvalue – the spectral theorem states that in this case, a symmetric matrix can never be defective. Although the proof of this will involve a little fancy footwork, the way we'll do it here will be clever rather than profound – it will just be repeated applications of the transposition trick.

We shall prove the statement by a contradiction. Let the eigenvalues of $\mathbf{A}$ be $\lambda_1, \lambda_2, \ldots\ldots, \lambda_k$ which are all distinct with orthonormal eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots\ldots, \mathbf{v}_k$, and the repeated eigenvalue λ_0 occurring $n-k$ times. Let's claim that there is just a single linearly independent eigenvector $\mathbf{v}_0$ corresponding to λ_0. Evidently, $\mathbf{v}_0$ and $\mathbf{v}_1$ through $\mathbf{v}_k$ are all orthonormal; they thus form the basis of a $k+1$-dimensional subspace of the total n-dimensional space. Let $\mathbf{W}$ denote the $n-k-1$-dimensional subspace which is the "rest" of the total space i.e. the subspace which lies completely outside the span of $\mathbf{v}_0$ and $\mathbf{v}_1$ through $\mathbf{v}_k$. Evidently, every vector in $\mathbf{W}$ is orthogonal to all the $\mathbf{v}$'s and their linear combinations. Let us define the vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots\ldots, \mathbf{w}_{n-k-1}$ to be orthonormal basis vectors of $\mathbf{W}$.

Consider the dot product $\mathbf{w}_i \cdot \mathbf{A}\mathbf{v}_j$. By definition of eigenvalue this is $\lambda_j \mathbf{w}_i \cdot \mathbf{v}_j$ which is zero since the $\mathbf{w}_i$'s are orthogonal to the $\mathbf{v}_j$'s by definition. Now by transposition trick followed by symmetric, $\mathbf{w}_i \cdot \mathbf{A}\mathbf{v}_j = \mathbf{A}\mathbf{w}_i \cdot \mathbf{v}_j = 0$, which implies that $\mathbf{A}\mathbf{w}_i$ is also perpendicular to every $\mathbf{v}$ i.e. $\mathbf{A}\mathbf{w}_i$ also lies in the subspace $\mathbf{W}$ (we ignore the special case that $\mathbf{A}\mathbf{w}_i$ is zero – this case is relevant only if $\lambda_0 = 0$ and can be handled with little extra hassle once we figure out what to do in the general case). Of course this holds true for all i from 1 to $n-k-1$, and the conclusion here is that $\mathbf{A}$ acting on every basis vector of the subspace $\mathbf{W}$ generates another vector which lies inside $\mathbf{W}$.

Basis the above we can write

$$\mathbf{A}\mathbf{w}_1 = c_{1,1}\mathbf{w}_1 + c_{1,2}\mathbf{w}_2 + \ldots\ldots + c_{1,n-k-1}\mathbf{w}_{n-k-1} \quad ,$$

$$\mathbf{A}\mathbf{w}_2 = c_{2,1}\mathbf{w}_1 + c_{2,2}\mathbf{w}_2 + \ldots\ldots + c_{2,n-k-1}\mathbf{w}_{n-k-1} \quad ,$$

$$\vdots$$

$$\mathbf{A}\mathbf{w}_{n-k-1} = c_{n-k-1,1}\mathbf{w}_1 + c_{n-k-1,2}\mathbf{w}_2 + \ldots\ldots + c_{n-k-1,n-k-1}\mathbf{w}_{n-k-1} \quad ,$$

where the $c_{i,j}$'s are some constants. Evidently the $c_{i,j}$'s form the entries of a square matrix $\mathbf{C}$ of size $n–k–1$; we can say that this matrix is the representation of the operator $\mathbf{A}$ when expressed in the terms of the basis vectors of $\mathbf{W}$.

Now let's use the transposition trick and symmetric once more; we have $\mathbf{w}_i \cdot \mathbf{A}\mathbf{w}_j = \mathbf{A}\mathbf{w}_i \cdot \mathbf{w}_j$. For the LHS, dot $\mathbf{w}_i$ into the j^{th} row of the above set of equations; the result is $c_{j,i}$. For the RHS, dot $\mathbf{w}_j$ into the i^{th} row of the above; the result is $c_{i,j}$. So what this tells us is $c_{i,j} = c_{j,i}$. In other words the matrix $\mathbf{C}$ is itself symmetric.

And now we are almost done. Since $\mathbf{C}$ is a symmetric matrix which acts on vectors in $\mathbf{W}$ and returns vectors in $\mathbf{W}$, by the first bullet point of the spectral theorem (which is now rigorously proved), it must have at least one real eigenvalue λ^* with a corresponding eigenvector $\mathbf{w}^*$ in $\mathbf{W}$. But, since $\mathbf{C}$ is just the representation of $\mathbf{A}$ acting on vectors in $\mathbf{W}$, any eigenvalue and vector of $\mathbf{C}$ is also an eigenvalue and vector of $\mathbf{A}$. Evidently, λ^* can't equal any of the eigenvalues $\lambda_1, \lambda_2, \ldots\ldots, \lambda_k$ whose eigenvectors are all accounted for. The only eig which λ^* can be is the repeated one λ_0. But then, we have shown that there is some vector $\mathbf{w}^*$ in $\mathbf{W}$ which is an eigenvector of $\mathbf{A}$ corresponding to eigenvalue λ_0. This contradicts our initial claim that $\mathbf{v}_0$ is the only eigenvector corresponding to λ_0.

Proceeding in this way, we can then claim that $\mathbf{v}_0$ and $\mathbf{w}^*$ are the only two eigenvectors corresponding to λ_0, and thereafter show that claim to be false by ferreting out another eigenvector from the subspace orthogonal to the $\mathbf{v}_i$'s and $\mathbf{w}^*$. In similar fashion we can keep pulling out eigenvectors from the remaining subspaces until we have obtained $n–k$ linearly independent eigenvectors inside the subspace $\mathbf{W}$. The spectral theorem is now fully proved and the mathsiest section of this book concluded.

This completes our capsule summary of linear algebra, and we can now go on to the LCCH systems for which at least some of the above information is relevant (the rest will be useful as we progress through the course).

LCCH SYSTEMS

A system of coupled LCCH equations has the following structure :

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} = \mathbf{M}\mathbf{x} \quad , \tag{01}$$

where $\mathbf{x}$ is the vector of dependent variables $\mathbf{x} = [x_1(t); x_2(t); \ldots..; x_n(t)]$ and $\mathbf{M}$ is a matrix whose all entries are constants. In fact, a single higher order ODE for a scalar variable $x(t)$ can also be expressed in this form by defining the variables $x_1 = x,\ x_2 = \dot{x}$ etc. For example the harmonic oscillator equation

$$\ddot{x} = -\omega^2 x \quad ,$$

can be written as

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\omega^2 x_1 \end{aligned} \quad ,$$

which in the matrix form is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} = \begin{bmatrix} 0 & 1 \\ -\omega^2 & 0 \end{bmatrix}\mathbf{x} \quad .$$

To solve (01) we try a solution in the form

$$\mathbf{x} = \mathbf{v}_i \exp\left(\lambda_i t\right) \quad ,$$

where i is a number between 1 to n. Plugging it into the ODE system we have

$$\lambda_i \mathbf{v}_i \exp\left(\lambda_i t\right) = \mathbf{M}\mathbf{v}_i \exp\left(\lambda_i t\right) \quad .$$

Cancel the obvious common factor from both sides and we are left with the eigenvalue equation

$$\mathbf{M}\mathbf{v}_i = \lambda_i \mathbf{v}_i \quad , \tag{02}$$

i.e. $\mathbf{v}_i$ is an eigenvector of $\mathbf{M}$ with eigenvalue λ_i. Thus, we have a fundamental solution of the system,

$$\mathbf{x}_i = \mathbf{v}_i \exp\left(\lambda_i t\right) \quad . \tag{03}$$

If there is a complex λ and $\mathbf{v}$, then we can form real-valued combinations just as we did for the harmonic oscillator. If there is a repeated eigenvalue then there are two possibilities : either $\mathbf{M}$ has an n-dimensional eigenspace or $\mathbf{M}$ is defective. In the first case, (03) accounts for all fundamental solutions of

(01) if we consider k linearly independent eigenvectors $\mathbf{v}_i$ corresponding to a λ_i of multiplicity k. In the second case, we get fundamental solutions which are t, t^2 etc times the basic exponential solution. The complicated proof of this rather borderline case we won't go into here.

In fact, we'll completely ignore defective matrices in this book – they just aren't useful enough. Then the fundamental solutions $\mathbf{x}_i$ for i goes from 1 to n constitute all possible solutions of the system. The general solution is (no surprise) their arbitrary linear combination, which is

$$\mathbf{x}_G(t) = c_1\mathbf{v}_1 \exp(\lambda_1 t) + c_2\mathbf{v}_2 \exp(\lambda_2 t) + \ldots\ldots + c_n\mathbf{v}_n \exp(\lambda_n t) \quad ,$$

where the c_i's are arbitrary constants. If λ_i is real, c_i is real also, while if λ_i and λ_{i+1} are complex conjugate pairs, then so are c_i and c_{i+1}. In this case, a tedious manipulation enables us to extract real-valued solutions in terms of sines and cosines and the real and imaginary parts of $\mathbf{v}_i$. Since we won't need this calculation anywhere in this book, I'll let this case go also and assume henceforth that all eigenvalues are real and the eigenvectors span the space, i.e. that $\mathbf{M}$ is diagonalizable. In this case, we can say that the general solution is a linear combination of the eigenvectors of $\mathbf{M}$, each multiplied by exponential of the corresponding eigenvalue times t. The eigenvectors are also called the **normal modes**. If we know the constants c_i, then we can get the entire solution at time t.

How to find the c_i's ? As expected, those have to come from the IC. So suppose we have a generic IC, $t = 0$ $\mathbf{x} = \mathbf{x}_0$. In general this IC is not going to be an eigenvector of $\mathbf{M}$ so it is not going to evolve in time as a single exponential. Rather, the IC is going to be an **arbitrary linear combination** of the various eigenvectors of the system. Setting $t = 0$ into the $\mathbf{x}_G$ we found above, we have

$$\mathbf{x}_0 = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots\ldots + c_n\mathbf{v}_n \quad , \tag{04}$$

where the c_i's are all real. This is the matrix equation

$$\mathbf{V}\mathbf{c} = \mathbf{x}_0 \quad , \tag{05}$$

where $\mathbf{V}$ is the matrix whose columns are the eigenvectors of $\mathbf{M}$. We can easily solve this equation to determine the coefficients c_i explicitly from the IC. Since $\mathbf{M}$ is diagonalizable, the columns of $\mathbf{V}$ are linearly independent and (05) always has a unique solution.

A very important special case arises when $\mathbf{M}$ is symmetric. By the spectral theorem, the $\mathbf{v}$'s are orthonormal. Now we can bypass (05) and use the following **orthonormality trick**. Consider the relation (04) and dot with any particular $\mathbf{v}_k$ on the both sides. We have

$$\mathbf{x}_0 \cdot \mathbf{v}_k = c_1 \mathbf{v}_1 \cdot \mathbf{v}_k + c_2 \mathbf{v}_2 \cdot \mathbf{v}_k + \ldots\ldots + c_n \mathbf{v}_n \cdot \mathbf{v}_k \quad .$$

Now from orthonormality, the only term on the RHS which survives is $i = k$; that equals c_k. All other terms evaluate to zero. Hence we have

* $$c_k = \mathbf{x}_0 \cdot \mathbf{v}_k \quad , \qquad \textbf{(06)*}$$

in this case. Although a symmetric $\mathbf{M}$ might appear like a useless case, it turns out that the vast majority of $\mathbf{M}$'s which arise in real world applications are actually symmetric (in fact, the thought of what if that weren't true is enough to send chills down one's spine – you'll see what I mean when we come to Chapter 5). So the orthonormality trick is actually of paramount importance – take the trouble to become very familiar with it now, and I can guarantee that you'll reap the dividends later.

Once we have the c_i's, we have solved the system completely. The relation (04) tells us how much of each normal mode is present in the IC. Then, as the system evolves, each mode evolves with its own characteristic exponent, and the specific solution of (01) which satisfies the IC is

$$\mathbf{x}(t) = c_1 \mathbf{v}_1 \exp\left(\lambda_1 t\right) + c_2 \mathbf{v}_2 \exp\left(\lambda_2 t\right) + \ldots\ldots c_n \mathbf{v}_n \exp\left(\lambda_n t\right) \quad ,$$

and we are done.

This is the most general theory, now we'll see how to use Matlab to do matrix math for us. Finding eigenvalues and vectors is the following code (where $\mathbf{M}$ is of size 4) :

```
M = [m11 m12 m13 m14; m21 m22 m23 m24; m31 m32
m33 m34; m41 m42 m43 m44];
[V, D] = eig(M);
disp(D)
disp(V)
```

The output of this program is a diagonal matrix $\mathbf{D}$ whose elements are the eigenvalues of $\mathbf{M}$ and the matrix $\mathbf{V}$ whose columns are the corresponding eigenvectors of $\mathbf{M}$. For good measure, these eigenvectors are automatically

normalized to unity, and linearly independent eigenvectors of a single eigenvalue are Gram Schmidted to orthonormal.

The next step, determination of the coefficient vector **c** given the IC vector $\mathbf{x}_0$, is one more line of Matlab code after the previous steps. It is

```
c = V\x0;
```

This completes the computer solution. Of course, you should **NEVER** do $\mathbf{V}\backslash\mathbf{x}_0$ if $\mathbf{M}$ is symmetric; the equivalent code then is

```
c_k = dot(x0,v_k);
```

I am strongly discouraging the unnecessary $\mathbf{V}\backslash\mathbf{x}_0$ for two reasons : one is that it will not make you realize the importance of orthonormality and the other is that evaluating the dot product is far quicker and more accurate than inverting $\mathbf{V}$. Even though it won't matter right now because the matrices will be very small, the issue can become very significant when the matrices are a few hundred or few thousand elements in each direction. And yes, you won't be spared such examples later.

SECOND ORDER SYSTEMS

Suppose we are given a system which consists of all coupled springs and masses, with no damping. Its equation of motion works out to something like

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{x} = \mathbf{M}\mathbf{x} \quad . \tag{07}$$

Although the most general procedure would have been to convert everything to coupled first order form, we can exploit the physical origins of the system to make a short cut. We know that spring-mass systems will show oscillatory behaviour only, so every mode will evolve as $\cos \omega t$ or $\sin \omega t$ for some ω. Further, an IC consisting of displacement alone and no velocity will feature cos terms only while an IC consisting of velocity alone and no displacement will feature sin terms only. Supposing we have the first case, let's plug $\mathbf{x} = \mathbf{v}_i \cos \omega_i t$ into (07). This shows that shows that $\mathbf{v}_i$ is an eigenvector of $\mathbf{M}$ with eigenvalue $-\omega_i^2$. This gives the characteristic frequencies and the corresponding modes, and we can fit the IC as we did in the first order case to obtain the specific solution.

Now suppose we have a velocity only IC. Then, the trial solution is $\mathbf{v}_i \sin \omega_i t$ (so that $\mathbf{x}(0)$ can evaluate to zero). Plugging into (07) yields the same ω's and $\mathbf{v}$'s as in the cos case. This time, to fit the IC, note that

$$\dot{\mathbf{x}}(0) = \left[\frac{\mathrm{d}}{\mathrm{d}t} \left(\sum_i c_i \mathbf{v}_i \sin \omega_i t \right) \right]_{t=0} = \sum_i c_i \omega_i \mathbf{v}_i \quad .$$

Thus, the constants we'll get by doing $\mathbf{V} \setminus \dot{\mathbf{x}}_0$ (or $\dot{\mathbf{x}}_0 \cdot \mathbf{v}_k$ if $\mathbf{M}$ is symmetric) will actually be ω_i times the amount of the corresponding mode present in the solution. This is something we'll have to be careful of.

Finally, what if the IC has both position and velocity ? Then, we can exploit linearity to break it up into two problems, one which has the position IC alone and no velocity, and the other which has the velocity alone and no position. We can use the above process to solve the two problems, and then declare the combined solution to be the sum of the two. Again note that the short cut was possible only because we took the oscillatory behaviour of the system as a starting point, on physical grounds. For an arbitrary $\ddot{\mathbf{x}} = \mathbf{M}\mathbf{x}$, this trick is not going to work, and you'll have to reduce to coupled first order form and do the whole works on that.

Example 2.3.1

Easy eig

Drawing reference to 2.1.5, determine the eigenvalues of the matrix

$$\mathbf{M} = \begin{bmatrix} 0 & 1 \\ 6 & -1 \end{bmatrix},$$

with no calculation. Also determine without calculation the eigenvector corresponding to one eigenvalue.

[Level : 1]

We recognize that the second order equation in 2.1.5

$$\ddot{x} + \dot{x} - 6x = 0 \quad ,$$

when converted to first order form gives

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= 6x_1 - x_2 \end{aligned} \quad ,$$

or letting $\mathbf{x} = [x_1; x_2]$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} = \begin{bmatrix} 0 & 1 \\ 6 & -1 \end{bmatrix}\mathbf{x} \quad .$$

This of course is the matrix asked for in the present question.

The fundamental solutions of 2.1.5 are e^{-3t} and e^{2t}. Viewed as a matrix ODE system, the argument of the exponentials must equal the eigenvalues of $\mathbf{M}$, so they are -3 and $+2$.

For one eigenvector, we use the information that the special initial condition $v_0 = -3x_0$ excites only the decaying mode, evolving as e^{-3t}. Therefore, $[x_1; x_2] = [1; -3]$ must be the eigenvector corresponding to the eigenvalue -3. Of course, any multiple of this is also an eigenvector; the one of unit length is

* $$\mathbf{v} = \frac{1}{\sqrt{10}}\begin{bmatrix} 1 \\ -3 \end{bmatrix} \quad ,$$ *

which is the desired answer. For some soft matrix practice, you can check that this is indeed the eigenvector corresponding to the claimed eigenvalue.

Example 2.3.2

Self-closing door

A self-closing door consists of a mass (the door) attached to a spring and damper. The reference $x = 0$ is the closed position and there are no other forces on the door during its motion.

(a) Write its equation of motion in a matrix form.

It is desirable that the door close quickly starting from any arbitrary IC position and velocity.

(b) Expressing an arbitrary IC as an arbitrary linear combination of the eigenvectors of the matrix, show that, given the mass and the spring, the optimal choice of the damping constant is the critical value.

The door is now fitted with an automatic lock which snaps shut if the door coasts past $x = 0$ at a sufficiently low velocity.

(c) With the lock installed, is it preferable to have subcritical or supercritical damping ? Justify your answer.

[Level : 2]

(a) The door obviously satisfies the canonical mass spring damper equation $\ddot{x}+\gamma\dot{x}+\omega^2 x=0$. Defining $x_1=x$ and $x_2=\dot{x}$, we can write this in the matrix form as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}=\begin{bmatrix}0 & 1\\ -\omega^2 & -\gamma\end{bmatrix}\mathbf{x} \quad , \qquad *$$

which is the ODE system asked for.

(b) We refer to 2.1.3 for the solutions of the damped oscillator equation. With subcritical damping, the two characteristic roots (or equivalently the two eigenvalues) are

$$\lambda=\frac{-\gamma\pm \mathrm{j}\omega_d}{2} \quad .$$

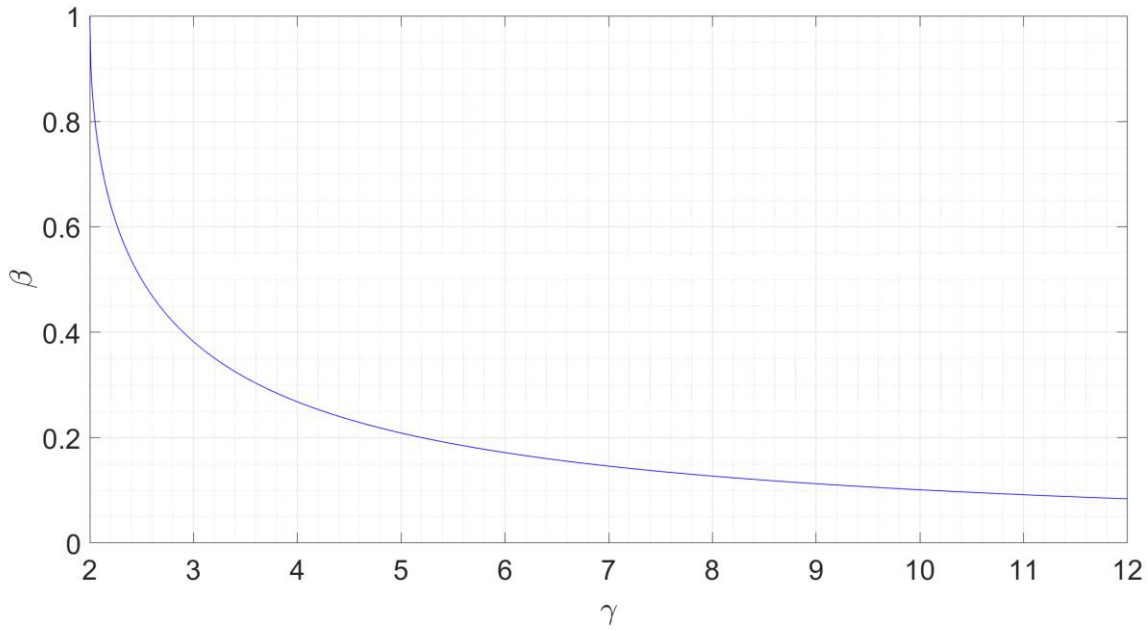
Each eigenvalue has its own eigenvector. For an IC consisting of a random linear combination of the two eigenvectors, each vector will evolve as per its own characteristic exponent. Since both of them have the same decay rate $\gamma/2$, it is clearly advantageous to increase γ to achieve faster damping.

With supercritical damping, the two characteristic roots are

$$-\alpha=\frac{-\gamma-\sqrt{\gamma^2-4\omega^2}}{2} \quad ,$$

$$-\beta=\frac{-\gamma+\sqrt{\gamma^2-4\omega^2}}{2} \quad .$$

Once again, each has its corresponding eigenvector and the IC is a random combination of these vectors. Now, for fastest damping of all motions, we want that both eigenvectors damp out as quickly as possible. Evidently, the large root $-\alpha$ increases in size without bound as γ increases. It's the small root $-\beta$ which is more interesting since that will be the limiting factor and its dependence on γ is not readily apparent. Below we present a plot of γ vs β for $\omega=1$ (obviously without loss of generality). We only consider $\gamma>2$ since the formula for β is valid there.



We can see that the maximum value of β occurs at $\gamma = 2$ which is the critical value. Hence, in this regime it is advantageous to decrease γ to achieve faster damping, a result which might appear paradoxical at first glance.

Summarizing the results, with both subcritical and supercritical damping, it is advantageous to drive the damping towards the critical value. Hence that is the optimal value.

(c) With the automatic lock in place, we want the door to arrive at $x = 0$ in a finite time and attempt to pass it with a small but finite velocity. This should happen irrespective of the ICs x_0 and v_0. Now we go back to 2.1.3 and look at the solution formulae. In the underdamped case, the solution has the structure

$$x = \mathrm{e}^{-\gamma t/2}\left(A\cos\omega_d t + B\sin\omega_d t\right) \quad .$$

Since we can write

$$A\cos\omega_d t + B\sin\omega_d t = R\cos\left(\omega_d t - \varphi\right) \quad \text{, where}$$

$$R = \sqrt{A^2 + B^2}, \qquad \varphi = \arctan\frac{B}{A} \quad ,$$

we see that x attains the value zero as soon as the argument of the trig becomes zero. Whatever the values of A and B, this is bound to happen within at most half a period of oscillation i.e. within $t = \pi/\omega_d$. So subcritical is a bonafide damping to use on the door.

For supercritical damping, the solution (2.1.3.03) is

$$x = \frac{-v_0 - \beta x_0}{\alpha - \beta} e^{-\alpha t} + \frac{v_0 + \alpha x_0}{\alpha - \beta} e^{-\beta t} \quad .$$

Suppose x_0 is positive and $v_0 = 0$. Then, at $t = 0$, the second term is greater in size (absolute value) than the first, since α is the large root. As time evolves, the second term decays more slowly than the first, so it remains greater in size at all time. This means that their resultant is always positive, so x never becomes zero (except at infinite time) and the lock does not engage.

The above arguments show that subcritical damping is the optimal choice to achieve automatic locking of the door.

Example 2.3.3

Baby heat equation

Four identical bodies, each holding heat, are connected together in a line (chain). Each body in the chain is connected to two neighbours, one on the left and one on the right. Bodies #1 and #4, the two ends of the chain, have as their missing neighbour the atmosphere which is maintained at temperature zero. The heat transfer between any two connected bodies (including atmosphere) follows NEWTON's rule i.e. each body gives heat to the other at a rate proportional to the difference between the temperatures of itself and the other one. Define the vector $\mathbf{u}=[u_1; u_2; u_3; u_4]$ to be the temperature of the bodies, taken as functions of time.

(a) Letting the proportionality constant in NEWTON's rule be equal to unity, write a coupled ODE system for $\mathbf{u}(t)$.

(b) If the initial temperature vector is [100; 50; 20; 0], determine the temperatures of all the bodies at any time t.

[Level : 2]

(a) We are given that body #i and #j exchange heat among each other as per NEWTON's rule. Thus, the effect of the exchange on body #i is

$$\frac{\mathrm{d}}{\mathrm{d}t}u_i = -\left(u_i - u_j\right) \quad ,$$

where we have taken the proportionality constant to be unity, as the problem said. Check the signs here : if $u_i > u_j$ then body #i LOSES heat as it should. Now, each body has two neighbours, one on left and one on right. So for body #i, the only relevant values of j are $i+1$ and $i-1$. Combining these, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}u_i &= -\left(u_i - u_{i-1}\right) - \left(u_i - u_{i+1}\right) \\ &= u_{i-1} + u_{i+1} - 2u_i\end{aligned} \quad .$$

For the ends of the chain, $i = 0$ and $i = 4$, we are given that the 'missing neighbour' is the atmosphere, $u = 0$. Hence for those two bodies we can substitute zero in place of u_0 and u_5, which do not exist. This leads to the matrix equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} \quad , \qquad *$$

which is the ODE system asked for.

(b) To solve the system we first calculate (or rather Matlab calculates) the eigenvalues of $\mathbf{M}$ and the corresponding eigenvectors. We get

$$\mathbf{D} = \mathrm{diag}\begin{pmatrix} -3.618 & -2.618 & -1.328 & -0.328 \end{pmatrix} \quad ,$$

$$\mathbf{V} = \begin{bmatrix} 0.3717 & -0.6015 & -0.6015 & -0.3717 \\ -0.6015 & 0.3717 & -0.3717 & -0.6015 \\ 0.6015 & 0.3717 & 0.3717 & -0.6015 \\ -0.3717 & -0.6015 & 0.6015 & -0.3717 \end{bmatrix} \quad .$$

The notation diag(.....) means a diagonal matrix whose entries are the items listed. Recall that the eigenvectors $\mathbf{v}$ of $\mathbf{M}$ are the columns of $\mathbf{V}$, and that Matlab automatically normalizes them to unity. Thus the general solution is

$$\mathbf{u}(t) = \sum_{i=1}^{4} c_i \mathbf{v}_i \exp(\lambda_i t) \quad , \tag{*}$$

where the λ_i's are of course the entries of $\mathbf{D}$.

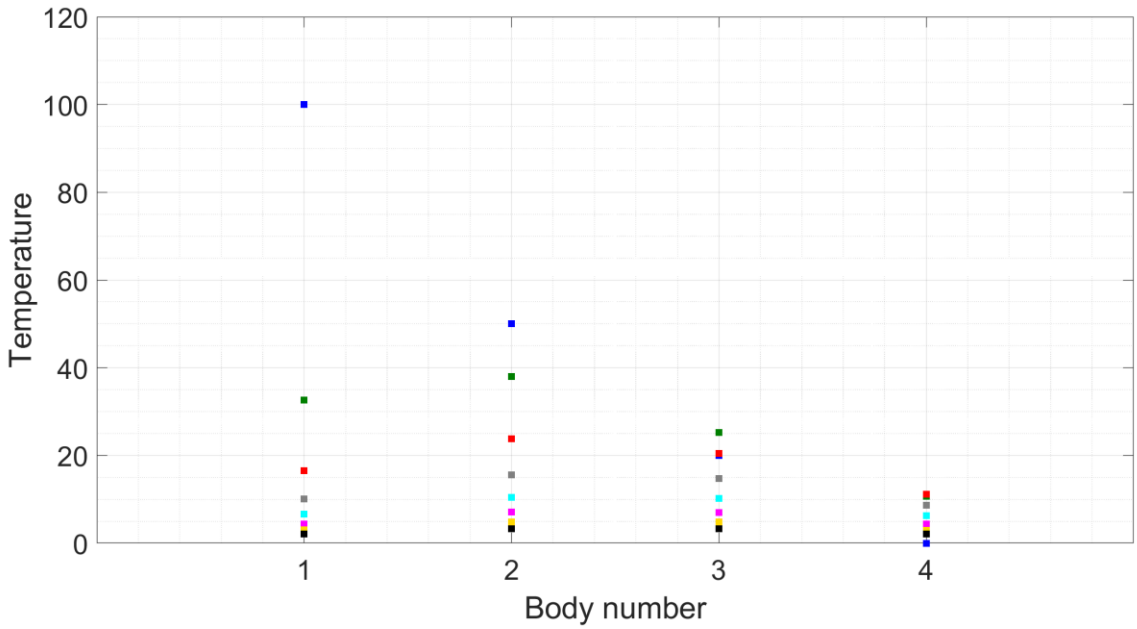
To determine the coefficients c_i, we use the IC vector $\mathbf{x}_0 = [100; 50; 20; 0]$. Since $\mathbf{M}$ is symmetric, its eigenvectors are orthonormal. So we use $c_i = \mathbf{x}_0 \cdot \mathbf{v}_i$, instead of $\mathbf{V} \backslash \mathbf{x}_0$ (this is important enough that I'll very likely impose a marks penalty for using $\mathbf{V} \backslash \mathbf{x}_0$ unnecessarily). Implementing this on Matlab gives

$$\mathbf{c} = [19.1298; \quad -34.1277; \quad -71.3025; \quad -79.2799] \quad . \tag{*}$$

Substituting this into the general solution completes the determination of $\mathbf{u}(t)$.

*

COMMENTS : Below is a plot of the four temperatures, at $t = 0$ to $t = 7$ in steps of 1.



The blue data points represent the IC; the subsequent time traces are shown as per our convention in green, red, gray, cyan, magenta, yellow and black. Two features of the solution are apparent. The first and most striking is the decay of all temperatures towards zero as time goes on. This happens physically because all bodies eventually give up their heat to the atmosphere, and mathematically

because all four eigenvalues of $\mathbf{M}$ are negative. The second feature is the smoothing out of the solution in time. The initial condition (blue squares) was quite jagged, featuring $u = 100$ on the first body and a sharp decrease to $u = 0$ on the fourth. The final state ($t = 7$ here, black squares) has a small value of u_1 and u_4 and a slightly larger value of u_2 and u_3. The spreading out of heat among the bodies occurs on account of the mutual exchange. To understand the long-time temperature profile, note that at large times, the dominant component of the solution will be the eigenvector which has the slowest decay rate. This is $\mathbf{v}_4$, which is approximately $[1;\ 1.6;\ 1.6;\ 1]$, and indeed, the black squares show a similar profile with u_2 and u_3 being about 50 percent greater than u_1 and u_4. In fact, the pattern is quite clear starting from the grey square only.

Finally, why baby heat equation ? As you have probably heard, heat equation is a PDE which describes the evolution of temperature in a body as a function of space and time. You can imagine this ODE system as a very crude mock-up of the PDE with the space divided into only four discrete points. When the time comes, I will let four become 10, then 100 etc. and finally infinity, and you will see how the ODE systems and their solutions converge towards the PDE.

Example 2.3.4

Coupled tanks

Three tanks are connected to each other using pipes, and water is circulated among the tanks using pumps. There is no external source or drain. If salt is introduced into the tanks then you are given that the quantity of salt in the three tanks, $\mathbf{q} = [q_1;\ q_2;\ q_3]$ follows the ODE system

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} = \begin{bmatrix} -2 & 3 & 1 \\ 1 & -5 & 2 \\ 1 & 2 & -3 \end{bmatrix}\begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} \quad .$$

Initially, the second tank contains 50 kg of salt while the first and third contain pure water.

(a) Show that the total quantity of salt inside the three tanks remains constant in time.

(b) Does the system converge to a steady state ? If yes, determine the steady state vector **q**. If no, justify your answer.

[Level : 2]

(a) The total quantity is of course $q_1+q_2+q_3$. To show that it is a constant, we must prove that its time derivative is zero. We have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(q_1+q_2+q_3) &= \dot{q}_1+\dot{q}_2+\dot{q}_3 \\ &= (-2q_1+3q_2+q_3)+(q_1-5q_2+2q_3)+(q_1+2q_2-3q_3) \quad , \\ &= q_1(-2+1+1)+q_2(3-5+2)+q_3(1+2-3) \\ &= 0\end{aligned}$$

which proves the assertion.

(b) As usual we feed the system to Matlab and it cranks out the following :

$$\mathbf{D} = \mathrm{diag}\begin{pmatrix}0 & -3.5858 & -6.4142\end{pmatrix} \quad ,$$

$$\mathbf{V} = \begin{bmatrix}-0.7877 & -0.7941 & 0.4760 \\ -0.3581 & 0.2326 & -0.8125 \\ -0.5013 & 0.5615 & 0.3366\end{bmatrix} \quad .$$

The IC is given as $\mathbf{x}_0 = [0;\ 50;\ 0]$. This time the matrix $\mathbf{M}$ is NOT symmetric so the eigenvectors are NOT expected to be orthogonal. To determine the coefficient vector we use

$$\mathbf{c} = \mathbf{V} \setminus \mathbf{x}_0 \quad ,$$

which yields the answer $\mathbf{c} = \begin{bmatrix}-30.3570; & 1.5064; & -47.7282\end{bmatrix}$. Again, be careful about this step – don't get influenced by my reservations about $\mathbf{V}\setminus\mathbf{x}_0$ and try taking dot products here ! There's no reason why that will work for this $\mathbf{M}$! The mechanical part of the process complete, now we interpret the results.

A steady state solution of an LCCH system means that $\mathrm{d}\mathbf{x}/\mathrm{d}t = 0$ hence $\mathbf{M}\mathbf{x} = 0$. This has a non-trivial solution if and only if $\mathbf{M}$ is non-invertible, and it has an eigenvalue of zero. For the present problem, we see that this is indeed the case. The eigenvector corresponding to the zero is the first column of $\mathbf{V}$. Thus, a non-trivial steady state solution exists here.

Existence of steady state however is only half the story. To claim that the system actually ends up at the steady state, we must also show that ALL other eigenvectors go to zero in time. This will happen if the corresponding eigenvalues have strictly negative real parts so that their exponentials go to zero in time. Even if there is one eigenvector attached to a positive eigenvalue, then

that mode will grow unboundedly as time goes on and you will never see the steady state in practice even if it exists. (This is the same as what happens with the fixed points in the equations $\dot{x}+\gamma x=b$ vis-a-vis $\dot{x}-\gamma x=b$ and what happened to the special IC in 2.1.5). This requirement – that the real parts of all the eigenvalues other than the zero be negative – is the test of STABILITY of the steady state solution.

With the tank problem, we see that the other two eigenvalues are negative so the steady state is stable. Hence as time goes on, the salt quantities in the three tanks indeed converge to their steady state values. These values are given by

$$\begin{aligned}\mathbf{q} &= c_1\mathbf{v}_1 \\ &= -30.3570\left[-0.7877;\quad -0.3581;\quad -0.5013\right] \\ &= \left[23.9122;\quad 10.8708;\quad 15.2180\right]\end{aligned}$$

It is a ready check that all three quantities are positive (no negative salt) and that they add up to 50, consistent with the conservation of salt mentioned in the problem and verified in Part (a). I cannot overestimate the importance of such checks both in the exam hall and in research work – error-free calculation is impossible unless you are a genius but with good fault diagnostics, you can become aware of mistakes and correct them before they really cause harm.

*

COMMENTS : Compare this with the previous problem – there $\mathbf{M}$ was invertible, all eigenvalues were nonzero, and the only possible steady state was the trivial solution. Since all the eigenvalues were negative, that state was stable and the system collapsed into it as time went on.

Example 2.3.5

Baby wave equation

Five identical masses are connected by four identical springs in the form of a line (chain). Every mass in the chain is connected to two neighbours, one on the left and one on the right. Define the variable vector **x** to be the displacements x_i of all the bodies from their static equilibrium positions. Let all masses and spring constants be unity.

(a) Assume that the masses at the two ends of the chain are connected by an extra pair of identical springs to a further two identical masses #0 and #6, which are held fixed at their equilibrium positions for all time. Write a coupled ODE system for $\mathbf{x}(t)$.

(b) The system is released from rest with the initial conditions (i) [1; 2; 4; 2; 4] and (ii) [2; –2; 0; 2; –2]. Will it come back to its initial configuration at any later time in either case ? Justify your answer. You may use the fact that $A\cos\omega_1 t + B\cos\omega_2 t$ where A and B are constants is periodic if and only if the ratio ω_1/ω_2 is rational.

(c) Modify the system by deleting the extra masses and springs in Part (a) and assuming that the masses #1 and #5 are free. Again write a coupled ODE system for $\mathbf{x}(t)$.

(d) You are told that the matrix in Part (c) has an eigenvalue of zero. In one step, and without using the computer, determine the corresponding eigenvector.

[Level : 3]

(a) We note that x_i is the displacement of each mass from its static equilibrium, and NOT its displacement along some absolute scale. As in 2.3.3 (baby heat), each mass is connected only to its immediate neighbours, hence it feels a force only if it or its neighbours are displaced. Considering the left neighbour alone (letting x and i increase rightwards as is conventional), we have

$$m\ddot{x}_i = -k\left(x_i - x_{i-1}\right) \quad . \tag{01}$$

Check the sign : if mass #i is pulled rightwards i.e. away from its neighbour, the force it feels is towards the left i.e. towards the neighbour. Likewise, considering the right neighbour alone,

$$m\ddot{x}_i = -k\left(x_i - x_{i+1}\right) \quad . \tag{02}$$

Adding the two and setting $m = k = 1$,

$$\ddot{x}_i = x_{i-1} + x_{i+1} - 2x_i \quad . \tag{03}$$

For the masses at the ends, #1 and #5, they are also connected by springs to two further masses which are held static. Hence (03) holds for these also if we take x_0 and x_6 to be zero. Combining this we have the matrix equation

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 1 & -2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} \quad , \tag{*}$$

which is the ODE system asked for. The static masses at the ends can also be replaced by walls – the effect is the same.

(b) Here we use the short cut I showed in the Lecture for handling coupled spring mass systems. Taking the oscillatory nature of solutions for granted, since the system is released with position-only IC, we try $\mathbf{x} = \mathbf{v}_i \cos\omega_i t$, with $-\omega_i^2$ an eigenvalue of $\mathbf{M}$ and $\mathbf{v}$ the corresponding eigenvector (or normal mode). Cranking the Matlab, we have for the given $\mathbf{M}$,

$$\mathbf{D} = \mathrm{diag}\left(-3.7321 \quad -3 \quad -2 \quad -1 \quad -0.2679\right) \quad ,$$

$$\mathbf{V} = \begin{bmatrix} -0.2887 & -0.5 & 0.5774 & 0.5 & 0.2887 \\ 0.5 & 0.5 & 0 & 0.5 & 0.5 \\ -0.5774 & 0 & -0.5774 & 0 & 0.5774 \\ 0.5 & -0.5 & 0 & -0.5 & 0.5 \\ -0.2887 & 0.5 & 0.5774 & -0.5 & 0.2887 \end{bmatrix} .$$

Since the problem does not say "determine X(t) given the ICs (i) and (ii)" we refrain from blindly calculating the coefficient vectors in each case.

What we are asked for is whether the solution starting from some given IC is periodic. We have been told that $A\cos\omega_1 t + B\cos\omega_2 t$ is periodic if and only if ω_1 and ω_2 are commensurate i.e. the ratio ω_1/ω_2 is rational. (To see why, suppose the ratio is p/q where p and q are integers, and that both waveforms start evolving at $t=0$. Then, at $t=pq$, the first function has completed q whole periods and the second has completed p whole periods, so both waveforms start again from what they were like at $t=0$. Hence the function is periodic with period pq. On the other hand, if the ratio of the two frequencies is irrational, then the function is not periodic because the two waveforms never begin a period together at any time. In this case, the function is said to be QUASIPERIODIC.)

In the present case, we note that the frequencies of the various modes are the square roots of the negatives of the corresponding entries of $\mathbf{D}$. Now, we can see that two of the entries, λ_3 and λ_4 equal -1 and -2; the corresponding frequencies are obviously related in an irrational ratio. For the IC (i), to prove that the solution is not periodic, it is sufficient to show that the coefficients associated with these two modes are not zero. Since $\mathbf{M}$ is symmetric, the eigenvectors are orthonormal, and $c_k = \mathbf{x}_0 \cdot \mathbf{v}_k$. Evidently, $[1; 2; 4; 2; 4]$ dotted with neither $\mathbf{v}_3$ nor $\mathbf{v}_4$ gives zero. Hence both these modes are present in the IC and the solution is NOT periodic. For the IC (ii) on the other hand, we see that $[2; -2; 0; 2; -2]$ is a direct scalar multiple of the second normal mode $\mathbf{v}_2$. Hence, that IC evolves in time as $\cos\sqrt{3}t$ and the solution IS periodic with time period $2\pi/\sqrt{3}$.

(c) The general relation (03) for the i^{th} mass holds true in this case also; all that changes is the treatment of the masses at the two ends, #1 and #5. Now, mass #1 only has a right neighbour and no left neighbour so for $i=1$ we have the force (02) only with no (01). Likewise, mass #5 only has a left neighbour and no right

neighbour so for $i=5$ we have the force (01) only with no (02). Combining these new end point results with Part (a), we have the matrix equation

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} \quad , \qquad *$$

which is the ODE system asked for.

(d) We can do this two ways, mathematically and physically. In the math way, we observe that the sum of the non-trivial entries in every row of $\mathbf{M}$ is zero. Hence, if we multiply it into the vector $\underline{[1;\ 1;\ 1;\ 1;\ 1]}$, the result will be zero in every row. This is basically "seeing the solution," an art worthy of practice !

Physically, a zero eigenvalue means a steady state solution – if we start off the masses with the corresponding eigenvector, then they will sit quietly without oscillating at all. Such behaviour is highly untypical for a spring chain, and must be a very special state. Clearly, one displacement profile which will evoke no response from the system is where all the masses are displaced through the same amount. This amounts to moving the chain as a whole in absolute space, while keeping the relative positions of the masses invariant. Physically, there is no difference in the state of the chain before and after the displacement. Hence, the eigenvector of the zero eigenvalue must be $\underline{[1;\ 1;\ 1;\ 1;\ 1]}$.

The normalization constant to make the eigenvector a unit vector is of course $1/\sqrt{5}$; throwing in this factor, the '1' becomes '0.4472'.

*

COMMENTS : Just as 2.3.3 was a mock-up of the heat equation , this is a discrete version of the wave equation . When the time comes , I will let the number of masses increase to infinity, and then the system will represent a bar which is subject to longitudinal expansion and compression.

If we hack the Matlab on Part (c), we in fact get the following eigenvalues and vectors of $\mathbf{M}$:

$$\mathbf{D} = \mathrm{diag}\begin{pmatrix} -3.6 & -2.6 & -1.4 & -0.38 & 0 \end{pmatrix} \quad ,$$

$$\mathbf{V} = \begin{bmatrix} -0.20 & -0.37 & -0.51 & -0.60 & 0.45 \\ 0.51 & 0.60 & 0.20 & -0.37 & 0.45 \\ -0.63 & 0 & 0.63 & 0 & 0.45 \\ 0.51 & -0.60 & 0.20 & 0.37 & 0.45 \\ -0.20 & 0.37 & -0.51 & 0.60 & 0.45 \end{bmatrix},$$

which are completely different from the $\mathbf{D}$ and $\mathbf{V}$ of Part (a). This might be a surprise considering that the only difference between the two $\mathbf{M}$'s is that a '2' becomes a '1' in the top left and bottom right corners ! What these corner elements represent are the boundary conditions on the PDE. The first case is that the ends of the bar are clamped and the second is that the ends are strain-free. When we come to PDEs, we will see that the slightest change in the boundary condition can indeed cause a drastic change in the solution.

2.4

MULTIVARIATE STABILITY ANALYSIS

In N^{o}. 04 we covered stability analysis for univariate, first order systems. Here we will look at the case where the system is higher order or equivalently features more than one variable in a coupled first order form. Since the latter representation is the most general, we will go with that picture only. This is a direct application of the LCCH systems theory we did in the last Unit.

Consider a nonlinear system in n variables x_1, x_2, ….., x_n written in coupled first order form which has the general structure

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, x_2, ...x_n) \\ \dot{x}_2 &= f_2(x_1, x_2, ...x_n) \\ &\vdots \\ \dot{x}_n &= f_n(x_1, x_2, ...x_n) \end{aligned} \quad , \qquad (01)$$

where f_1, f_2, ….. f_n are arbitrary functions of the n variables (NOT of time though – if there's that too then all hell breaks loose). As with univariate systems, a fixed point (fp) of this multivariate system, $\mathbf{x}^* = [x_1^*;\ x_2^*;\ \ldots..\ x_n^*]$, is a point where all the derivatives evaluate to zero. If the system is started off at a fp, it remains there for all time. Again as in L04, the question to ask is, if you start the system from a nearby point then will it approach the fp or go away from it. No surprise that the fp is called stable in the former case and unstable in the latter.

To obtain the criterion for stability, the process is the same as before : Taylor expand the nonlinear system (01) about the fp $\mathbf{x}^*$. This time, there can be perturbations in any and all variables, so in the most general case we must write $x_1 = x_1^* + \Delta x_1$, $x_2 = x_2^* + \Delta x_2$, ….., $x_n = x_n^* + \Delta x_n$. Now substitute this into (01); we have

$$
\begin{aligned}
&\dot{x}_1{}^* + \Delta\dot{x}_1 \\
&= f_1\left(x_1{}^*, x_2{}^*, \ldots . x_n{}^*\right) + \left.\frac{\partial f_1}{\partial x_1}\right|_{\mathbf{x}^*} \Delta x_1 + \left.\frac{\partial f_1}{\partial x_2}\right|_{\mathbf{x}^*} \Delta x_2 + \ldots . + \left.\frac{\partial f_1}{\partial x_n}\right|_{\mathbf{x}^*} \Delta x_n \\
&\dot{x}_2{}^* + \Delta\dot{x}_2 \\
&= f_2\left(x_1{}^*, x_2{}^*, \ldots . x_n{}^*\right) + \left.\frac{\partial f_2}{\partial x_1}\right|_{\mathbf{x}^*} \Delta x_1 + \left.\frac{\partial f_2}{\partial x_2}\right|_{\mathbf{x}^*} \Delta x_2 + \ldots . + \left.\frac{\partial f_2}{\partial x_n}\right|_{\mathbf{x}^*} \Delta x_n \\
&. \\
&. \\
&. \\
&\dot{x}_n{}^* + \Delta\dot{x}_n \\
&= f_n\left(x_1{}^*, x_2{}^*, \ldots . x_n{}^*\right) + \left.\frac{\partial f_n}{\partial x_1}\right|_{\mathbf{x}^*} \Delta x_1 + \left.\frac{\partial f_n}{\partial x_2}\right|_{\mathbf{x}^*} \Delta x_2 + \ldots . + \left.\frac{\partial f_n}{\partial x_n}\right|_{\mathbf{x}^*} \Delta x_n
\end{aligned}
\quad ,
$$

plus higher order terms in the perturbations, which we neglect. As before, on the LHS, the derivatives of starred quantities are zero; on the RHS, all the f's evaluated at $x_1{}^*$, $x_2{}^*$, ….., $x_n{}^*$ are also zero by definition of the fp. This leaves us with a linear system

$$
\frac{\mathrm{d}}{\mathrm{d}t}\Delta\mathbf{x} = \mathbf{J}\Delta\mathbf{x} \quad \text{, where}
$$

$$
\mathbf{J} = \begin{bmatrix}
\dfrac{\partial f_1}{\partial x_1} & \dfrac{\partial f_1}{\partial x_2} & \ldots . & \dfrac{\partial f_1}{\partial x_n} \\
\dfrac{\partial f_2}{\partial x_1} & \dfrac{\partial f_2}{\partial x_2} & \ldots . & \dfrac{\partial f_2}{\partial x_n} \\
. & . & & . \\
. & . & & . \\
. & . & & . \\
\dfrac{\partial f_n}{\partial x_1} & \dfrac{\partial f_n}{\partial x_2} & \ldots . & \dfrac{\partial f_n}{\partial x_n}
\end{bmatrix} \quad .
$$

Normally I would have used for **M** for the dynamic matrix but **J** is the near-universal convention in this context as it called the Jacobian, after CARL JACOBI who invented it. Note that all the derivatives in the Jacobian are evaluated at the fp itself, hence **J is a constant matrix**.

Now of course we have the situation we saw in T034 – the fp will be stable if **all** perturbations decay in time i.e. **all** eigenvalues of the Jacobian have negative real parts. Conversely the fp will be unstable if **even one** combination of perturbations grows in time i.e. **even one** eigenvalue has a positive real part. Because of the all-or-one dependence of stability on the eigenvalues, it makes sense to recast the criterion in terms of the eigenvalue of **J** which has the **largest real part** (including the sign!). By abuse of language we will call this the 'largest eigenvalue' of **J** and denote it by λ_{max}. In terms of this, we have the formal enunciation of the stability condition below :

The fixed point $\mathbf{x}^* = [x_1^*; x_2^*; \ldots.; x_n^*]$ of the nonlinear system (01) is stable if the real part of the largest eigenvalue of the Jacobian matrix evaluated at $\mathbf{x}^*$ is negative. The fp is unstable if the said real part is positive. If the real part is identically zero, then this test fails to yield the stability of the fixed point.

As with the univariate case, the Grossmann-Hartmann theorem guarantees the accuracy of the results of the linear stability test in a non-zero sized region around the fp. The case $\mathrm{Re}\ \lambda_{max} = 0$ (i.e. the largest eigenvalue is either identically zero or purely imaginary) is a failure of the test because in that case, the stability is determined entirely by the nonlinear terms – depending on their nature and signs, the system can be stable or unstable. See the book by STROGATZ for elaboration of this topic.

One case where a purely imaginary λ_{max} implies stability rather than indeterminacy however is in **undamped mechanical systems** i.e. mechanical systems (more realistically, mechanical models) where no damping is present. Because NEWTON's Law is second order, the linearization of undamped mechanical systems about stable fps generally gives rise to harmonic oscillator equations, whose eigenvalues are $\pm j\omega$. So if an undamped mechanical system throws up all purely imaginary eigenvalues upon linearization, then you very likely know that the fp is stable with the nonlinearities included also. If you add damping to the system (in the real world, a bit of damping exists anyway) then the real part of λ_{max} will in fact become negative, achieving conformity to the linearization test.

In this context, let me mention an extremely plausible but completely incorrect way of determining stability of a (primarily mechanical) system of coupled second order equations. Suppose that in the LHS of (01), the dots on

x_1, x_2 etc. were replaced by double dots, and the RHS were kept the same. Then you would linearize about the fp $\mathbf{x}^*$ and arrive at

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Delta\mathbf{x} = \mathbf{J}\Delta\mathbf{x} \quad .$$

Next you would seek the eigenvalues of $\mathbf{J}$; if all were negative, then you would claim that every mode is a harmonic oscillator therefore the system is stable. This shortcut appears to bypass the reduction to first order form and deals with eigenvalues of a matrix of order n as against $2n$ which you would get if you brought the system down to first order. It was so tempting that I once used it in the first version of a paper which I had written, only to have the error pointed out by the Reviewer.

Although the reason is very involved and I won't get into it here, the second order trick works only if $\mathbf{J}$ is symmetric. In all other cases it can yield a wrong result, so in general please always go with the first order form, which can sometimes be more work but will always give the correct answer.

Example 2.4.1

Simple pendulum

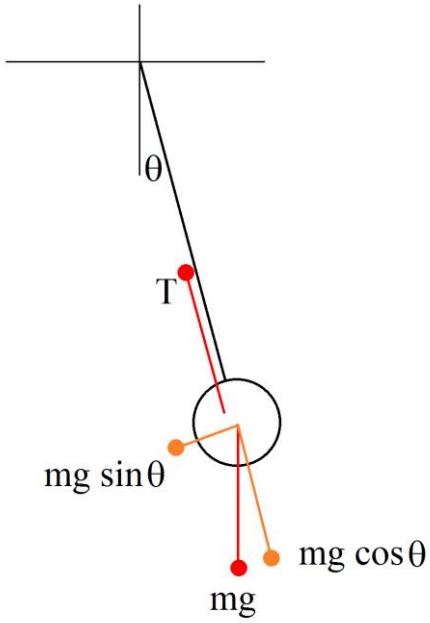
A simple pendulum consists of a point bob of mass m attached to one end of an inextensible rod of negligible mass, having length l. The rod is pivoted at the other end and the contraption allowed to hang in the vertical plane. Let θ be the angle made by the pendulum measured counter-clockwise from the downward vertical.

(a) Write an ODE satisfied by the pendulum.

(b) Find all the fixed points of the system and determine their stabilities.

[Level : 1]

(a) The free body diagram of the pendulum is shown in the Figure, where T denotes the tension in the rod.



Since the rod is inextensible, there is no motion in the radial direction. The tangential velocity of the bob is given by $l\dot{\theta}$ while the tangential acceleration is $l\ddot{\theta}$. The tangential force is clearly $-mg\sin\theta$ (it is NEGATIVE because θ increases counter-clockwise). Writing NEWTON's Law we have

$$ml\ddot{\theta} = -mg\sin\theta \quad ,$$

or,

$$\ddot{\theta} + \frac{g}{l}\sin\theta = 0 \quad , \qquad (01)^*$$

which is the ODE asked for.

(b) For the stability analysis we express it in coupled first order form. Letting x_1 denote θ and x_2 denote $\dot{\theta}$, we can write the pendulum equation as

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = -\frac{g}{l}\sin x_1 \quad .$$

The first RHS is zero if x_1 is a constant; the second is zero if $\sin x_1 = 0$ i.e. $x_1 = 0$ or $x_1 = \pi$. The Jacobian matrix at an arbitrary point $(x_1^*, 0)$ is

$$\mathbf{J} = \begin{bmatrix} 0 & 1 \\ -\frac{g}{l}\cos x_1 * & 0 \end{bmatrix} .$$

We have to evaluate this matrix at $x_1* = 0$ and $x_1* = \pi$ respectively, giving the two cases

$$\mathbf{J}(0) = \begin{bmatrix} 0 & 1 \\ -g/l & 0 \end{bmatrix} ,$$

$$\mathbf{J}(\pi) = \begin{bmatrix} 0 & 1 \\ g/l & 0 \end{bmatrix} .$$

The eigenvalues in the first case are $\pm \mathrm{j}\sqrt{g/l}$ while in the second case they are $\pm\sqrt{g/l}$. In the first case, $\lambda_{\max}$ has a real part of exactly zero, which is the indeterminate situation. However, since the pendulum is a mechanical system with no damping, we can say that the fp $x_1 = 0$ i.e. θ=0 is actually stable. In the second case, $\lambda_{\max} = +\sqrt{g/l}$ which clearly indicates instability hence $\theta=\pi$ is unstable. These results are of course in agreement with our intuition – a pendulum is very much stable while hanging near its vertically downward equilibrium position and is unstable near the vertically upward equilibrium position.

*

COMMENTS : In this particular case, the Jacobian formulation is overkill; I only did it to show how the process works in a very simple setting. Since the pendulum is governed by a single equation of a single variable (01), we can do this almost by inspection. Evidently, the fp has to satisfy $\dot{\theta}* = 0$ (else how will the pendulum remain 'fixed') and $\sin\theta* = 0$ so the two possibilities are $\theta* = 0$ and $\theta* = \pi$. Letting $\theta = \theta* + \Delta\theta$, we substitute into (01) and keep only the linear term; about the zero point it yields $\Delta\ddot{\theta} + (g/l)\Delta\theta = 0$ which is an oscillator equation so stable while about the π point it yields $\Delta\ddot{\theta} - (g/l)\Delta\theta = 0$ which is a repeller equation so unstable. Also note how the repeller equation arose as the linearization of a mechanical system about an unstable fp, as I had mentioned in T020.

Example 2.4.2

Lorenz model

The Lorenz model refers to the following system in three variables x, y and z :

$$\begin{aligned} \dot{x} &= -\sigma(x-y) \\ \dot{y} &= rx - y - xz \\ \dot{z} &= b(xy - z) \end{aligned} \quad ,$$

where σ, b and r are parameters. Take $\sigma = 10$ and $b = 8/3$ for this problem and let r be a variable, positive quantity.

(a) Find all fixed points of the Lorenz system as r is varied.

(b) Determine the stabilities of all of them. You may use the computer to evaluate the eigenvalues of a non-trivial third order Jacobian; for this routine, try $r = 1,3,5\ldots..$ until you get a transition from stable to unstable. Report the value of r at which you find the transition.

[Level : 2]

(a) The fixed point must satisfy

$$\begin{aligned} -\sigma\left(x^*-y^*\right)&=0 \\ rx^*-y^*-x^*z^*&=0 \\ b\left(x^*y^*-z^*\right)&=0 \end{aligned} \quad .$$

An obvious fp is the trivial one, the origin $x^*=y^*=z^*=0$.

To find other possible fp's we proceed systematically. Clearly, x^* must equal y^*; substituting this into the third equation gives $z^*=x^{*2}$, and plugging both of these into the second equation result in a cubic

$$\left(r-1\right)x^*-x^{*3}=0 \quad .$$

The root $x^*=0$ corresponds to the origin; the other roots are $x^*=\pm\left(r-1\right)^{1/2}$. For the roots to be real, r must be greater than 1. Hence the fps of the Lorenz system are

$$O: x^*=y^*=z^*=0 \quad ,$$

valid for all r, and

$$C^+: x^*=\left(r-1\right)^{1/2} \quad y^*=\left(r-1\right)^{1/2} \quad z^*=r-1 \quad ,$$

$$C^-: x^*=-\left(r-1\right)^{1/2} \quad y^*=-\left(r-1\right)^{1/2} \quad z^*=r-1 \quad ,$$

valid for $r>1$. These are the fixed points asked for. The names C^+ and C^-, looking suspiciously like grades which you should try to avoid in this course, are just conventionally assigned to the latter pair of points, and are not important for the analysis.

(b) We first construct the Jacobian of the Lorentz system at an arbitrary point (x^*,y^*,z^*) before fitting specific values. We have

$$\mathbf{J}=\begin{bmatrix} -\sigma & \sigma & 0 \\ r-x^* & -1 & -x^* \\ by^* & bx^* & -b \end{bmatrix} \quad .$$

At the origin O, this matrix acquires a really simple form. It is

$$\mathbf{J}(O)=\begin{bmatrix} -\sigma & \sigma & 0 \\ r & -1 & 0 \\ 0 & 0 & -b \end{bmatrix} \quad .$$

We resist the temptation to blindly evaluate eigenvalues of this using the computer !

We can see that this $\mathbf{J}$ is partially uncoupled – the z equation is standalone and the other two form a coupled two-equation system. The standalone equation contributes an eigenvalue of $-b$, which is $-8/3$ which is negative. The eigenvalues of the remaining 2 x 2 matrix can be evaluated using

$$\det\left(\mathbf{A}-\lambda\mathbf{I}\right)=0 \quad \text{, or}$$

$$\det\begin{bmatrix}-\sigma-\lambda & \sigma \\ r & -1-\lambda\end{bmatrix}=0 \quad .$$

This determinant yields the quadratic equation

$$\lambda^2+\left(1+\sigma\right)\lambda+\sigma\left(1-r\right)=0 \quad .$$

Although we can do it the long way and find the roots, a shorter method is to recognize that the above equation is nothing but the characteristic equation of the second order LCCH ODE $\ddot{x}+\left(1+\sigma\right)\dot{x}+\sigma\left(1-r\right)x=0$. Now we can apply the results of T023. There we saw that the solutions are bounded if and only if the coefficients of the ODE are both positive. Since coefficients of the ODE and of the characteristic equation are the same, the above will produce a decaying motion and hence a stable fixed point if and only if $1+\sigma$ and $\sigma(1-r)$ are both positive. By the problem, $1+\sigma=11$ and does not pose trouble; the other one changes from positive at $r<1$ to negative at $r>1$. Hence the origin O is stable for $r<1$ and unstable for $r>1$. The case $r=1$ is the borderline and nothing can be said.

As for C^+ and C^-, we analyse their stabilities by plugging them into the Jacobian. Taking C^+, the Jacobian there is

$$\mathbf{J}(C^+)=\begin{bmatrix}-\sigma & \sigma & 0 \\ 1 & -1 & -\left(r-1\right)^{1/2} \\ b\left(r-1\right)^{1/2} & b\left(r-1\right)^{1/2} & -b\end{bmatrix} \quad .$$

This is the matrix whose eigenvalues we evaluate numerically by trying $r=1,3,5$ etc. We see that one eigenvalue of $\mathbf{J}$ is negative real while the other two are a complex pair. At low r, their real parts are negative but as r increases they

become less negative and ultimately go over to positive, at which point the fp loses stability. The transition takes place at $r = 25$. For C^-, the transition occurs at the exact same value. Hence C^+ and C^- are stable for $r < 25$ and unstable for $r > 25$ (approximately).

*

COMMENTS : The Lorenz model, first discovered by EDWARD LORENZ (not to be confused with HENDRIK LORENTZ of the eponymous transformation in relativity theory) in 1963, is one of the most famous equations in nonlinear dynamics. It is one of the simplest systems (and the first to be discovered) which exhibits chaos. The chaos occurs beyond $r = 25$, when O, C^+ and C^- are all unstable. In fact, the critical value of r at which C^+ and C^- lose stability can be calculated analytically, by forming the eigenvalue equation and seeking a solution having one negative real root and one purely imaginary pair. The answer is $r = \sigma \dfrac{\sigma + b + 3}{\sigma - b - 1} = 24.74$, a derivation which I did not do in detail as it is just messy algebra, and the exact value of the transition is not the most important or interesting feature of the Lorentz model. A discussion of chaos would be out of scope here, but please consult any textbook on chaos if you are interested.

Example 2.4.3

Brouwer saddle

A certain class of two-dimensional undamped mechanical systems possesses an equation of motion having the following structure :

$$\begin{aligned} \ddot{x} &= \alpha x + \beta \dot{y} \quad , \\ \ddot{y} &= \gamma y - \beta \dot{x} \quad , \end{aligned} \tag{*}$$

where α, β and γ are some parameters of arbitrary sign and size.

(a) Write the characteristic polynomial for the eigenvalues.

(b) The AM-GM inequality states that for any two real numbers p and q, $p^2 + q^2 > 2pq$. Using this inequality, determine the conditions which α, β and γ must satisfy for the system (*) to be stable.

A Brouwer saddle consists of a particle of unit mass constrained to move in the x-y plane and mounted in a saddle potential of strength A^2 which rotates with angular velocity $\Omega\hat{\mathbf{z}}$ relative to the laboratory frame. It can be shown that the displacements x and y from the centre of the saddle in the rotating frame satisfy the equation (*) with $\alpha = -A^2 + \Omega^2$, $\gamma = A^2 + \Omega^2$ and $\beta = 2\Omega$.

(c) Determine the relation between A and Ω for which the particle remains stable at the centre of the saddle.

In a practical application, a Brouwer saddle is used to overcome the effect of a repelling potential of strength B^2. In the presence of this potential, the system equation is (*) with $\alpha = -A^2 + B^2 + \Omega^2$, $\gamma = A^2 + B^2 + \Omega^2$ and $\beta = 2\Omega$.

(d) Given A and Ω, determine the maximum value of B for which the particle can still remain stable at the centre of the saddle.

If the particle is replaced by a rigid spherical shell which rolls without slipping on the plane, it can be shown that the system equation is again (*) with $\alpha = 3\left(-A^2 + B^2 + \Omega^2\right)/5$, $\gamma = 3\left(A^2 + B^2 + \Omega^2\right)/5$ and $\beta = 8\Omega/5$.

(e) Given A and Ω significantly greater than A, determine the maximum value of B for which the ball remains stable at the centre of the saddle.

[Level : 3]

Another page-long problem like T012 but again the bark is worse than the bite because the question guides us step by step to the final results. So let's just take on the various parts systematically.

(a) We convert the system to coupled first order form. Letting the variables $[x_1; x_2; x_3; x_4]$ be $[x; \quad \dot{x}; \quad y; \quad \dot{y}]$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x} = \mathbf{M}\mathbf{x} \quad \text{, where}$$

$$\mathbf{M} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \alpha & 0 & 0 & \beta \\ 0 & 0 & 0 & 1 \\ 0 & -\beta & \gamma & 0 \end{bmatrix} .$$

Note that it's $\mathbf{M}$ here and not $\mathbf{J}$ – this is not a Jacobian of anything but the full system equation ! Since there is no option to evaluate the eigenvalues by computer, we must go for the characteristic equation

$$\det(\mathbf{M} - \lambda\mathbf{I}) = 0 \quad \text{, or}$$

$$\det\begin{bmatrix} -\lambda & 1 & 0 & 0 \\ \alpha & -\lambda & 0 & \beta \\ 0 & 0 & -\lambda & 1 \\ 0 & -\beta & \gamma & -\lambda \end{bmatrix} = 0 \quad .$$

We expand the determinant along the first row and then hack at some algebra to get the polynomial

$$\lambda^4 + \lambda^2\left(\beta^2 - \alpha - \gamma\right) + \alpha\gamma = 0 \quad , \qquad (01)^*$$

which is the characteristic polynomial asked for.

(b) Since we are told that the system is mechanical and undamped, we expect that in the regions of stability we will have all eigenvalues purely imaginary – no growing solutions but no decay either. Then, the eigenvalues will be $\lambda = \pm\mathrm{j}\omega_1, \pm\mathrm{j}\omega_2$ for some real ω_1 and ω_2. The characteristic polynomial will then have the form

$$(\lambda - j\omega_1)(\lambda + j\omega_1)(\lambda - j\omega_2)(\lambda + j\omega_2) = 0 \quad \text{, or}$$

$$(\lambda^2 + \omega_1^2)(\lambda^2 + \omega_2^2) = 0 \quad \text{, or}$$

$$\lambda^4 + (\omega_1^2 + \omega_2^2)\lambda^2 + \omega_1^2\omega_2^2 = 0 \quad . \qquad (02)$$

For stability of the system (*) we must force the polynomial (01) to have the form (02). This implies that for stability, all the coefficients in (01) must be positive. Moreover, by the AM-GM inequality,

$$\frac{\omega_1^2 + \omega_2^2}{2} \geq (\omega_1^2\omega_2^2)^{1/2} \quad ,$$

for all ω_1 and ω_2. This relation too must be satisfied by (01), i.e. half of the coefficient of λ^2 must be greater than the square root of the constant term. Putting all these together, we have the three conditions

* $$\beta^2 - \alpha - \gamma > 0 \quad , \qquad (03a)^*$$

* $$\alpha\gamma > 0 \quad , \qquad (03b)^*$$

* $$\frac{\beta^2 - \alpha - \gamma}{2} > \sqrt{\alpha\gamma} \quad , \qquad (03c)^*$$

required for stability of the system (*).

(c) From this point onwards it's basically plugging expressions into (03) and interpreting the results. For the basic Brouwer saddle, (03a) implies $2\Omega^2 > 0$ which is trivially true. Equation (03b) implies $\Omega^4 - A^4 > 0$ which holds true if $\Omega > A$ (we can take A to be positive without loss of generality). Finally, (03c) implies $\Omega^2 > (\Omega^4 - A^4)^{1/2}$ which is again satisfied trivially. So only (03b) imposes a constraint on Ω and it says that the desired condition is $\underline{\Omega > A}$.

(d) In this case, (03a) implies $2\Omega^2 - 2B^2 > 0$ or $B < \Omega$ (if we take B to be positive). Equation (03b) implies $(\Omega^2 + B^2)^2 - A^4 > 0$, which is trivially true since Ω must be greater than A for a Brouwer saddle. Finally, (03c) says

$$\Omega^2 - B^2 \geq \left[(\Omega^2 + B^2)^2 - A^4 \right]^{1/2} \quad .$$

Since both sides are positive, the inequality is preserved by squaring; we then open out and simplify to get

$$-4\Omega^2 B^2 > -A^4 \quad \text{, or}$$

$$B \le \frac{A^2}{2\Omega} \quad .$$

Thus, (03a) and (03c) throw up two different upper bounds for B; the actual bound must be the one which is more restrictive. Since $\Omega > A$, $A^2/2\Omega$ is clearly less than Ω and the second is the more stringent bound. The maximum permissible B occurs at the boundary of this inequality, i.e. $\underline{B_{\max} = A^2/2\Omega}$.

(e) The same procedure again, with the numbers a little more complicated : (03a) now says

$$\frac{34\Omega^2}{25} - \frac{6B^2}{5} > 0 \quad ,$$

or $B^2 < (17/15)\,\Omega^2$. Equation (03b) implies

$$\frac{9}{25}\left[\left(B^2 + \Omega^2\right)^2 - A^4\right] > 0 \quad ,$$

which is trivially true since $\Omega > A$. Finally, (03c) leads to

$$\frac{64}{625}\Omega^4 - \frac{192}{125}\Omega^2 B^2 + \frac{9}{25}A^4 > 0 \quad .$$

Simplifying this we have

$$B^2 < \frac{\Omega^2}{15} + \frac{15A^2}{64\Omega^2} \quad .$$

When Ω is significantly larger than A as the problem says, the first term above is dominant. Evidently, this is a more stringent inequality than what (03a) gave. So the maximum B occurs at the boundary of this inequality and evaluates to $\underline{B_{\max} = \Omega/\sqrt{15}}$.

*

COMMENTS : The Brouwer saddle is named after the Dutch mathematician BERTUS BROUWER who first discovered this result in 1918. It turns out that its equations are simple enough in the rotating frame but look hideous in the laboratory frame. If X and Y denote the particle's displacements from the centre in the inertial frame then they satisfy the system

$$\ddot{X} = -\left(A^2 \cos 2\Omega t\right) X - \left(A^2 \sin 2\Omega t\right) Y \quad ,$$

$$\ddot{Y} = -\left(A^2 \sin 2\Omega t\right) X + \left(A^2 \cos 2\Omega t\right) Y \quad .$$

Although this is linear (check against the definition!), it looks impossible to solve; recall from T003 that Mathieu equation $\ddot{x} + \omega^2 \left(a + b\cos\Omega t\right) x = 0$ is unsolvable. However, because of the wonderful properties of rotations, a frame transformation tames this monster completely.

To derive the equation in Part (c), we basically add the centrifugal and Coriolis forces to the potential force; the derivation of Part (e) requires us to also calculate the centrifugal and Coriolis torques on the rigid body and becomes outside our scope. So let's just take the equations for granted. The rather surprising stability in fact comes from the Coriolis force – the β term in (*).

It is also interesting to note that replacing the point mass by a rigid body significantly enhances the system stability. $B_{\max}$ varies as $1/\Omega$ in Part (d) but as Ω in Part (e). That is why we did a nearly identical calculation twice – in general, this book has little space for such redundancy but in this case the two final answers are sufficiently disparate to warrant the repeat.

* * * * *

3

INHOMOGENEOUS INITIAL VALUE PROBLEMS

This Chapter examines inhomogeneous differential equations and presents the solution techniques of undetermined coefficients and variation of parameters. A theoretical discussion is followed by applications to vibration theory, circuit theory and other physical problems.

3.1

LINEAR INHOMOGENEOUS EQUATIONS

A linear inhomogeneous equation has the form $\mathbf{L}x = f(t)$ where $f(t)$ does not equal zero. Recall from N°. 02 that the solution of this equation is 'homogeneous plus particular' i.e. $x = x_h + x_p$ where x_p satisfies the complete ODE but not the IC and x_h satisfies the homogeneous ODE $\mathbf{L}x = 0$. Whereas in the first order case, the integrating factor formula yields both solutions at one go, in higher order systems we need to find x_h and x_p separately. We have already seen how to find x_h for many equations; we will now look at x_p. There are two primary methods for finding x_p; these I describe below.

VARIATION OF PARAMETERS

Consider the second order equation

$$\ddot{x} + p(t)\dot{x} + q(t)x = f(t) \quad , \tag{01}$$

where p and q are not necessarily constants. Assume that the homogeneous solutions $x_1(t)$ and $x_2(t)$ are known, and that W is their Wronskian (if you skipped that inessential Unit, then $W = x_1\dot{x}_2 - x_2\dot{x}_1$). Then, the var of par formula is

* $$x_p = -x_1 \int \frac{x_2 f}{W}\,\mathrm{d}t + x_2 \int \frac{x_1 f}{W}\,\mathrm{d}t \quad . \tag{02}*$$

We can also write this in a more pedantically correct manner by defining the integrations in terms of a dummy variable s and letting t be the upper limit, thus

$$x_p = -x_1 \int^t \frac{x_2(s) f(s)}{W(s)}\,\mathrm{d}s + x_2 \int^t \frac{x_1(s) f(s)}{W(s)}\,\mathrm{d}s \quad .$$

This notation is more highbrow than (02) but doesn't serve any additional purpose. On a later occasion however, we'll have to deal with a situation where the limits and variables of integration will become very tricky; so it's a

good idea for you to remain familiar with the more mathsy notation even at this stage.

The way (02) comes about is you start from a trial solution $x_p = u_1(t)x_1 + u_2(t)x_2$, where u_1 and u_2 are unknown functions, plug it into the ODE, then impose the 'constraint condition' (more or less taken out of the air) that $\dot{u}_1 x_1 + \dot{u}_2 x_2 = 0$, plug that in and then hack at the algebra. In practice, the formula (02) itself is more important than its derivation. In any case, variation of parameters works primarily in situations which can also be cracked using undetermined coefficients (below), and in an isolated handful of other problems, whose primary utility is in exams and quizzes. A variation of this technique, the Bogoliubov-Krylov method, on the other hand has great practical significance in the analysis of nonlinear systems, but I can't go into a discussion of that here. The biggest source of error in var of par setup is that people apply (02) without ensuring that their ODE is in the form (01) with the coefficient of $\ddot{x}$ normalized to unity. **If you try applying (02) to the ODE $a(t)\ddot{x} + b(t)\dot{x} + c(t)x = d(t)$ with *a*(*t*) not equal to unity then you will get a plausible but completely garbage answer.** Before using (02) you must define $p = b/a$, $q = c/a$ and $f = d/a$.

An extension of this formula to higher order equations also exists, but it is too cumbersome and has too little practical significance to be quoted here.

UNDETERMINED COEFFICIENTS

This is the second method for finding particular solutions. It works for ODEs of any order, not just second order, provided that the equation satisfies certain conditions. These conditions are that :

- All coefficients of the ODE are **constants**
- The inhomogeneity or excitation function $f(t)$ has certain **special forms**, namely polynomials, exponentials, trigs and their sums and products.

Although the conditions might appear restrictive, it so happens that almost all practical manifestations of linear ODEs feature problems which can be cracked using undet coeff, so we look at this method in far more detail than var of par.

The basic principle is : when your ODE satisfies the criteria for using the method, you pick a **trial function** to be the particular solution. This trial function has a specific form depending on the ODE (details coming) and features certain **unknown constants** (the undetermined coefficients). You **plug** the trial into the ODE and obtain the coefficients by demanding that it be a solution.

How to select the trials ? Let the ODE be

$$x^{(n)} + A_{n-1}x^{(n-1)} + \ldots\ldots + A_1\dot{x} + A_0 x = f(t) \quad ,$$

where the A_k's are all constants and we recall that $x^{(k)}$ means $\mathrm{d}^k x/\mathrm{d}t^k$. Given the ODE, we start off by obtaining the homogeneous solutions. Since the homogeneous ODE is LCCH, it will have the usual exponential form solutions, real or complex, with or without repeats. It might also have polynomial solutions, which correspond to a characteristic root λ=0 with repeats as appropriate. Then, we break up the function $f(t)$ into a sum of its constituent parts and apply the following procedure to each part, which we call $g(t)$. [What do I mean by a "part" ? It's easiest for me to define this indirectly, by asking you to look at the Table coming shortly.] First, we see if $g(t)$ is identical to any of the homogeneous solutions or contains a homogeneous solution as a factor. The method proceeds differently according to this test is true or false so we consider case by case.

CASE 1 : No coincidence

In this case, the part of the excitation under consideration does not equal any of the homogeneous solutions and does not contain any homogeneous solution as a factor (for example, something like $g(t) = (12t+30)\,\mathrm{e}^{6t}$ where e^{6t} is a homogeneous solution is not allowed). This way of putting it is intuitive but a little sloppy as it gives a wrong result in some cases. So let's be absolutely precise, at the risk of sounding pedantic. In this case the function $g(t)$ has the form of a polynomial in t, of degree equal to or greater than zero, multiplying an exponential function $\mathrm{e}^{\alpha t}$ with α real including zero, or a real-valued linear combination of exponentials $\mathrm{e}^{\alpha t}$ and $\mathrm{e}^{\alpha^* t}$ with α complex, where α (and α^* if appropriate) does not equal a characteristic root of the homogeneous equation.

Then, corresponding to the function $g(t)$ we pick the trial functions $x_p(t)$ as in the below Table.

$g(t)$	Trial $x_p(t)$
Polynomial of degree k	A different polynomial of degree k, all coefficients unknown
$A\,e^{\gamma t}$, γ is real	$Ce^{\gamma t}$ where C is unknown
$A\cos\omega t + B\sin\omega t$	$C_1\cos\omega t + C_2\sin\omega t$, C_1 and C_2 unknown
Product of two or more of the above	The corresponding product

Note that if $g(t)$ is of a form not included in the above table, then undet coeff is not applicable at all and you have to do var of par to find the particular solution. Examples of tempting but unsuitable g's are fractional and negative powers ($t^{1/2}$, $1/t$), quotient, reciprocal and inverse trigs ($\tan t$, $\sec t$, $\arcsin t$), the corresponding hyperbolics ($\tanh t$, $\operatorname{sech} t$, $\operatorname{arcsinh} t$), logarithms etc. I said this once already – the number of unsuitable g's vastly exceeds that of the applicable g's but it's the latter case which happens to be practically the most relevant.

CASE 2 : Coincidence

In this case, the function $g(t)$ has the form of a polynomial in t, of degree equal to or greater than zero, multiplying an exponential function $e^{\alpha t}$ with α real including zero, or a real-valued linear combination of exponentials $e^{\alpha t}$ and $e^{\alpha^* t}$ with α complex, where α (and α^* if appropriate) is a characteristic root of the homogeneous equation, in which it appears with a multiplicity of k (i.e. the root $\lambda = \alpha$ is repeated k times in the characteristic equation of the homogeneous ODE.)

Then, to form the trial function, first take a **bump-up factor of t^k** and then multiply this bump-up factor with the trial function corresponding to the given g in the no coincidence Case 1. To get the bump-up factor correct, recall from L05 that if a root λ is repeated k times then the corresponding homogeneous solutions are $e^{\lambda t}$, $t\,e^{\lambda t}$, $t^2\,e^{\lambda t}$, ….., $t^{k-1}\,e^{\lambda t}$; the bump-up factor features the next highest power of t.

After the trial functions for each constituent $g(t)$ of the excitation are out, we add them up to form the overall trial corresponding to the excitation $f(t)$. All

this procedural detail perhaps sounds like Greek or worse legalese; so if your head is spinning, take a five minutes break and then go on to the Tutes where I have included lots of examples and an explanation for the appearance of the bump-up factors.

Example 3.1.1

Undet coeff – uncomplicated case

This is taken from the (erstwhile?) course textbook BOYCE and DIPRIMA and is assigned as homework in every semester that the course is taught.

Consider the ODE

$$\dot{x} - x = 1 + 3\sin t \quad ,$$

with initial condition $x(0) = x_0$.

(a) Obtain the particular solution x_p.

(b) Determine the value of x_0 such that the solution remains bounded in time.

[Level : 1]

(a) As the first step we calculate the homogeneous solution. By this point it has been implanted into our heads that the homogeneous equation $\dot{x}_h - x_h = 0$ has the solution $x_h = \mathrm{e}^t$ (you didn't get it reflexively? go back to T005 again.) Now we calculate the particular solution using undetermined coefficients. The excitation function has two parts – the '1' and the sine term, so we must calculate the particular solution separately for each. The first one is a polynomial of degree zero, like the first row of the trial table while the second part is obviously taken from the third row of the trial table.

To solve x_p corresponding to the '1', we first note that a constant (i.e. e^0) is not a homogeneous solution so it comes under the no-coincidence Case 1. Since 1 is a polynomial of degree zero, we try a different polynomial of degree zero, i.e. an unknown constant $x_p = C$. Plugging this into the ODE, we have

$$\dot{C} - C = 1 \quad .$$

Obviously, C does not have a time derivative, so this means $C = -1$, and x_p for this part is -1.

We now solve x_p corresponding to $3\sin t$. Again, $\sin t$ is not a homogeneous solution and does not contain it as a factor so this corresponds to Case 1. The trial function we take is $x_p = C_1 \cos t + C_2 \sin t$. Note that just because the RHS lacks the cos term, we CANNOT ignore it in the trial ! A trial of $C_2 \sin t$ alone is not going to work in this case, as you can see for yourself. Plugging this in, we have

$$\begin{aligned} ODE \rightarrow \dot{x}_p - x_p \\ = -C_1 \sin t + C_2 \cos t - C_1 \cos t - C_2 \sin t = 3\sin t \end{aligned} \quad .$$

Now, because $\cos t$ and $\sin t$ are linearly independent, their coefficients must be equal each to each. This gives

$$\begin{aligned} -C_1 + C_2 &= 0 \\ -C_1 - C_2 &= 3 \end{aligned} \quad ,$$

or $C_1 = C_2 = -3/2$. Hence x_p for this part is $-(3/2)\cos t - (3/2)\sin t$. The overall particular solution is the sum of both parts :

$$* \qquad x_p = -1 - \frac{3}{2}\cos t - \frac{3}{2}\sin t \quad , \qquad *$$

which is the desired answer.

(b) To answer the question regarding bounded solutions, we note that x_p is bounded while x_h is not. Hence, the solution starting from the IC x_0 will remain bounded if and only if the homogeneous solution does not get excited at all i.e. if $x(t)$ equals $x_p(t)$. This means $x(0) = x_p(0)$. Plugging $t = 0$ into x_p, we find it equals $-5/2$ hence the desired initial condition is $\underline{x_0 = -5/2}$.

*

COMMENTS : While assigning this problem for homework, the instructor generally expects a solution using integrating factors rather than undet coeff. The book also poses it much before undet coeff comes up for discussion. Just for comparison, you can try doing it that way – the integral involved is no picnic. It is in fact the same integral as the one in T011, which problem you can now do using undet coeff as a practice exercise. Hence, other than acting as a simple undet coeff example, this problem also shows how the method can be an easy way to get to the answer when it is applicable. Finally, the concept of special IC leading to a bounded solution is more theoretical than practical; see T024 as well as any problem on stability analysis.

Example 3.1.2

Another easy example

Considering the equation

$$\ddot{x} + \omega^2 x = A\cos\omega t \quad ,$$

obtain the particular solution x_p using

(a) variation of parameters,

(b) undetermined coefficients.

[Level : 2]

(a) The homogeneous solutions are of course $x_1 = \cos \omega t$ and $x_2 = \sin \omega t$. Their Wronskian is

$$\begin{aligned} W &= \det \begin{bmatrix} \cos \omega t & \sin \omega t \\ -\omega \sin \omega t & \omega \cos \omega t \end{bmatrix} \\ &= \cos \omega t \left(\omega \cos \omega t \right) - \sin \omega t \left(-\omega \sin \omega t \right) \quad , \\ &= \omega \end{aligned}$$

which shows both the formal determinant definition and the 2x2 'operational definition' of W. Then we have

$$\begin{aligned} x_p &= -x_1 \int \frac{x_2 f}{W} \mathrm{d}t + x_2 \int \frac{x_1 f}{W} \mathrm{d}t \\ &= -\cos \omega t \int \frac{A \sin \omega t \cos \omega t}{\omega} \mathrm{d}t + \sin \omega t \int \frac{A \cos^2 \omega t}{\omega} \mathrm{d}t \end{aligned} \quad .$$

Now invoke the trigonometric identities

$$\begin{aligned} \cos 2\theta &= 2 \cos^2 \theta - 1 \quad , \\ \sin 2\theta &= 2 \sin \theta \cos \theta \quad , \end{aligned}$$

to get

$$\begin{aligned} x_p &= \frac{A}{\omega} \left[-\cos \omega t \left(-\frac{1}{2\omega} \cos 2\omega t \right) + \sin \omega t \left(\frac{t}{2} + \frac{1}{2\omega} \sin 2\omega t \right) \right] \\ &= \frac{A}{2\omega^2} \left(\cos 2\omega t \cos \omega t + \sin 2\omega t \sin \omega t \right) + \frac{At}{2\omega} \sin 2\omega t \quad , \qquad (01) \\ &= \frac{A}{2\omega^2} \cos \omega t + \frac{A}{2\omega} t \sin \omega t \end{aligned}$$

where the last step uses the trig identity $\cos(a–b) = \cos a \cos b + \sin a \sin b$. These and other trig identities crop up occasionally in this part of the course so it's good to be familiar with them. They will start appearing all the time in the next Chapter when we do boundary value problems and Fourier analysis, so I will include a brief review between this Chapter and the next.

Finally, note that the $\cos \omega t$ term in the last line of (01) is just a multiple of one of the homogeneous solutions, so it is extraneous and can be dropped. Hence the actual particular solution is

* $$x_p = \frac{A}{2\omega} t \sin \omega t \quad ,$$ *

which is the desired answer.

(b) We start from the homogeneous solutions which are $\cos \omega t$ and $\sin \omega t$. Evidently, the excitation function is one of these two only, so we have Case 2 of the Lecture. Now is also a good time to translate some of that language into ordinary sentences. The excitation $\cos \omega t$ is an exponential function corresponding to $\alpha = \pm \mathrm{j}\omega$; moreover it is a real-valued linear combination of these two exponentials. The roots $\pm \mathrm{j}\omega$ are solutions of the characteristic equation of the homogeneous ODE, and they occur there with multiplicity $k = 1$. The constant A multiplying the cos is a polynomial of degree zero.

Since there is a coincidence, the trial solution has to start off with a bump-up factor. This factor is t^k which equals t in this case. Then tack on the trial function corresponding to the given excitation i.e. $A \cos \omega t$ in the no coincidence case. This trial is $C_1 \cos\omega t + C_2 \sin\omega t$. Putting these together, our required trial function is

$$x_p = t\left(C_1 \cos \omega t + C_2 \sin \omega t\right) \quad .$$

Due to a notational convenience, we rename it as

$$x_p = Dt \cos \omega t + Et \sin \omega t \quad .$$

This enables us to use the shorthand C for $\cos \omega t$ and S for $\sin \omega t$ – a shorthand which cleans up the inevitably messy calculation of derivatives to a great extent. We then have

$$\begin{aligned} ODE \rightarrow \ddot{x}_p + \omega^2 x_p & \\ &= -2\omega D\,\mathrm{S} - \omega^2 Dt\,\mathrm{C} + 2\omega E\,\mathrm{C} - \omega^2 Et\,\mathrm{S} + \omega^2\left(Dt\,\mathrm{C} + Et\,\mathrm{S}\right) \quad . \\ &= -2\omega D\,\mathrm{S} + 2\omega E\,\mathrm{C} = A\,\mathrm{C} + 0\,\mathrm{S} \end{aligned}$$

Note that the terms $t\mathrm{C}$ and $t\mathrm{S}$ vanished automatically on substitution into the ODE. Finally, equating the coefficients of C and S each to each in the last line, we have $D = 0$, $E = A/2\omega$, giving the result

$$x_p = \frac{A}{2\omega} t \sin \omega t \quad ,$$

which is the desired answer.

*

COMMENTS : Of course, the two methods yielded the same answer, which they had better. Variation of parameters was more mechanical as it is just plug and chug, but the calculation needed evaluation of integrals and several trigonometric identities. Undetermined coefficients on the other hand needed a well-informed starting point, but after that it was evaluation of derivatives rather than integrals. The phenomenon being studied here is LINEAR RESONANCE. It refers to the linear growth of the particular solution when the frequency of excitation is exactly equal to the natural frequency of the harmonic oscillator. I will have much more to say on this topic when we come to the applications of LCC equations.

Example 3.1.3

‘Homogeneous resonance’

(a) Write the LCCH ODE of lowest possible order which has $t\sin\omega t$ as a fundamental solution.

(b) A laboratory apparatus is built where the position of a particle satisfies the above equation. The particle is released with a carefully chosen IC which excludes all the growing modes of motion. Will the motion of the particle really be bounded in time ? Justify your answer.

[Level : 2]

(a) We are given that the equation being sought is LCCH. There, a solution of $t\sin\omega t$ means that $\sin\omega t$ is also a solution, and the characteristic root is repeated at least once. The root corresponding to $\sin\omega t$ is of course $+\mathrm{j}\omega$ or $-\mathrm{j}\omega$; for the equation to be real, BOTH of them must always occur together. Taking the repeat into account, we have at least the following factors of the characteristic polynomial :

$$(\lambda-\mathrm{j}\omega)(\lambda+\mathrm{j}\omega)(\lambda-\mathrm{j}\omega)(\lambda+\mathrm{j}\omega)(.....)=0 \quad .$$

The dots represent additional roots which might exist – the moment we have the $\pm\mathrm{j}\omega$ coming twice, we are guaranteed the $\sin\omega t$ and $t\sin\omega t$ solution, the latter being the one given in the question.

But we are also told to write the ODE of lowest possible order; if we have additional roots in place of the dots in the characteristic polynomial, then we will be unnecessarily raising the order of the ODE (an n^{th} order ODE always has n characteristic roots). Hence the required characteristic polynomial is

$$\begin{aligned}&(\lambda-\mathrm{j}\omega)(\lambda+\mathrm{j}\omega)(\lambda-\mathrm{j}\omega)(\lambda+\mathrm{j}\omega)\\&=\left(\lambda^2+\omega^2\right)^2 \qquad\qquad ,\\&=\lambda^4+2\omega^2\lambda^2+\omega^4=0\end{aligned}$$

and the corresponding differential equation is

$$x^{(4)}+2\omega^2\ddot{x}+\omega^4 x=0 \quad ,$$

which is the ODE asked for.

(b) By now you are probably ready to scream NO by instinct, following T024 and various discussions of stability. This time however the answer is not quite so clear. Of course, you can never manage an exact IC, there will be perturbations etc which will excite the growing modes with small, non-zero amplitudes. But the catch is that this time the growth is linear instead of exponential. With exponential growth, if you improve your precision by a factor of thousand i.e. reduce the amplitude of growing modes by thousand, then it will take only thrice – or five or some other $O(1)$ number times – as much time for the modes to grow to unit size and the extra precision to be lost. But with linear growth, it will take a thousand times longer. So if you manage to run the experiment with sufficient precision and for a short enough duration then you probably CAN get to see bounded motions to a good degree of accuracy – the experiment will be

over before the growing modes have got a chance to really become big. Thus although both linear and exponential modes are growing, the similarity between the two more or less ends there – if one is a sheep, the other is a wolf; both are animals but that's about it.

*

COMMENTS : The problem title 'homogeneous resonance' is more glitzy than substantial – a more factually accurate version would have been 'LCCH equation with linearly growing oscillatory solution' which very likely does not make you want to try doing or reading it. While this ODE does have the resonant solution of the previous problem, this is not really an example of resonance because it lacks the important feature of a match between an internal frequency and an external excitation frequency. In T054 we will see a set of examples clarifying what really makes a resonance and what does not.

Example 3.1.4

Why undet coeff works

Recall from L02 the D notation for differential operator. Expressions involving D and constants can be factorized like algebraic polynomials, thus $\ddot{x}+5\dot{x}+6x=0$ can be written as $(D+2)(D+3)x=0$ or equivalently $(D+3)(D+2)x=0$. Consider a LCC n^{th} order ODE $\mathbf{L}x=f(t)$ where the characteristic polynomial of $\mathbf{L}x=0$ has the real root $\lambda=\lambda_0$ repeated k times and none of the other roots equals λ_0.

(a) Express $\mathbf{L}$ in factorized differential operator form.

(b) Use the above representation of $\mathbf{L}$ to show that the homogeneous ODE $\mathbf{L}x=0$ has the solutions $e^{(\lambda_0)t}$, $t\,e^{(\lambda_0)t}$, $t^2\,e^{(\lambda_0)t}$, ..., $t^{k-1}\,e^{(\lambda_0)t}$ corresponding to the k repetitions of the root λ_0.

(c) Extend the previous argument to show that $\mathbf{L}x_p=e^{\alpha t}$ is satisfied by $x_p=Ce^{\alpha t}$ (C is some constant) when α does not equal any of the characteristic roots λ, and by $x_p=Ct^k e^{(\lambda_0)t}$ when α equals the repeated root λ_0.

[Level : 3]

(a) The characteristic polynomial of the homogeneous equation must be

$$(\lambda-\lambda_1)(\lambda-\lambda_2).....(\lambda-\lambda_{n-k})(\lambda-\lambda_0)^k=0 \quad ,$$

where the λ_1 to λ_{n-k} denote any values, real or complex, distinct or repeated, other than λ_0. Now, λ in the characteristic equation is equivalent to D in the ODE hence the differential operator $\mathbf{L}$ is

* $$\mathbf{L}=(D-\lambda_1)(D-\lambda_2).....(D-\lambda_{n-k})(D-\lambda_0)^k \quad , \qquad (01)^*$$

which is the factorized representation asked for.

(b) We note that when $\mathbf{L}$ is written in the factorized form and made to act on a function $g(t)$, the successive factors of $\mathbf{L}$ will operate on g in the order right to left. Thus, $\mathbf{L}g$ with $\mathbf{L}$ as written in (01) will be $(D–\lambda_0)$ acting on g, followed by $(D–\lambda_0)$ acting on that etc. k times and then $(D–\lambda_{n-k})$, $(D–\lambda_{n-k-1})$ and all the way backwards to $(D–\lambda_1)$ acting on the successive resultants, thus

$$\begin{aligned}&\mathbf{L}g(t)\\&=(D-\lambda_1)\circ(D-\lambda_2)\circ.....\circ(D-\lambda_{n-k})\circ \quad .\\&(D-\lambda_0)\circ(D-\lambda_0)\circ.....\circ(D-\lambda_0)g(t)\end{aligned}$$

Since we want to show that powers of t multiplied by $e^{(\lambda_0)t}$ make possible solutions of $\mathbf{L}x=0$, we choose as g the "monomial of degree m" i.e. the function $t^m e^{(\lambda_0)t}$ for some integer m. Since this function will "react with" the $(D–\lambda_0)$ components of $\mathbf{L}$ and not the other ones, the arrangement (01) is most convenient as it has all the $(D–\lambda_0)$'s bunched up on the right. We have

$$\begin{aligned}&(D-\lambda_0)t^m e^{(\lambda_0)t}\\&=\lambda_0 t^m e^{(\lambda_0)t}+mt^{m-1}e^{(\lambda_0)t}-\lambda_0 t^m e^{(\lambda_0)t} \quad ,\\&=mt^{m-1}e^{(\lambda_0)t}\end{aligned}$$

which is a monomial of degree $m–1$.

Thus we see that $(D–\lambda_0)$ acting on a monomial REDUCES ITS DEGREE by one. Hence, if we keep hitting the monomial with $(D–\lambda_0)$'s then its degree will keep reducing. But this process cannot go on for ever. For when we reach the monomial of degree 0 i.e. just $e^{(\lambda_0)t}$, then that is a solution of $(D–\lambda_0)g=0$, so one

more application of $(D–\lambda_0)$ will convert it to zero. And once it reaches zero, it of course stays there for ever. For a monomial of degree m, we need to hit it with a minimum of $m+1$ copies of $(D–\lambda_0)$ to reach the zero level. Now coming back to the original equation $\mathbf{L}x=0$, there are k copies of $(D–\lambda_0)$ available in $\mathbf{L}$. So $\mathbf{L}$ acting on any monomial upto and including degree $k–1$ i.e. $t^{k-1}e^{(\lambda_0)t}$ will convert it to zero. In other words, the monomials $e^{(\lambda_0)t}$, $t\,e^{(\lambda_0)t}$ etc until $t^{k-1}\,e^{(\lambda_0)t}$ will be solutions of the homogeneous ODE, which was to be shown.

(c) We first consider the case when α is not equal to any of the roots of the homogeneous equation. In this case we have

$$(D-\lambda)Ce^{\alpha t}=(\alpha-\lambda)Ce^{\alpha t} \quad ,$$

where λ is any of the roots. In words, each of the factors of $\mathbf{L}$, acting on the trial function, multiplies it by a constant without changing the nature of the function. Clearly, with a suitable choice of C (which happens to equal $\left[(\alpha-\lambda_1)(\alpha-\lambda_2).....(\alpha-\lambda_{n-k})(\alpha-\lambda_0)^k\right]^{-1}$), we can get the overall constant in the resultant to equal unity, and thus generate a particular solution of $\mathbf{L}x=e^{\alpha t}$. Since none of the λ's equals α, the expression for C is not singular and the method works fine, and we have shown the desired result.

If $\alpha=\lambda_0$, then the action of $(D–\lambda_0)$ gives zero (or equivalently the expression for C becomes garbage), but in this case we know that (and are also told in the problem that) we should try it with the bump-up factor t^k. The calculation for this has already been done in Part (b); this is just a conceptual extension of that. We saw that k applications of $(D–\lambda_0)$ reduce the degree of the monomial by k. Hence the last k terms of $\mathbf{L}$ (the k repeated factors) acting on the bumped-up monomial $t^k\,e^{(\lambda_0)t}$ will bring it down to a monomial of degree zero i.e. $e^{(\lambda_0)t}$ with some constant in the front. The remaining terms of $\mathbf{L}$ acting on this will only change the value of the constant, as happened in the previous case ($\alpha\neq\lambda$). Hence $\mathbf{L}$ acting on $Ct^k\,e^{(\lambda_0)t}$ will produce $e^{(\lambda_0)t}$ times some constant, which can be made to equal unity with a suitable choice of C, thus proving the given proposition.

*

COMMENTS : In this example we took the repeating root λ_0 to be real. The argument works identically if we have it complex and the pair λ_0 and λ_0* repeat k times – there is more mess but no new mathematical content. Part (b) is the proof behind the appearance of powers of t in the solutions of LCCH equations with repeated roots. Part (c) is the proof behind the method of undetermined coefficients. As you can see, this proof is not particularly easy. There are two ways you can react to this – if you are mathematically minded, then the method by itself will appear like witchcraft and start making sense only now, while if you are engineering minded, then you will prefer to focus on the operational aspects rather than the underlying mathematics. Both approaches are fine, so long as you get the answers right in the exams and in your research.

Example 3.1.5

Undet coeff – tricky cases

For each of the following six inhomogeneous ODEs, the particular solution is sought. The solutions of the homogeneous equation are given for your convenience. Write the best possible starting guess for the trial function which you would use in each case. You do NOT need to evaluate the unknown coefficients.

(a) $\ddot{x}+\omega_0^2 x=\cos\omega t:\quad x_h=C_1\cos\omega_0 t+C_2\sin\omega_0 t$

(b) $\ddot{x}-4\dot{x}+4x=\cosh 2t:\quad x_h=C_1\mathrm{e}^{2t}+C_2 t\mathrm{e}^{2t}$

(c) $\ddot{x}+\gamma\dot{x}+\omega^2 x=\cos\omega_d t:\quad x_h=\mathrm{e}^{-\gamma t/2}\left(C_1\cos\omega_d t+C_2\sin\omega_d t\right)$. Note that ω_d is the damped frequency

(d) $\ddot{x}-3\dot{x}+2x=\left(2t^3+3t^2+2\right)\mathrm{e}^{2t}:\quad x_h=C_1\mathrm{e}^{t}+C_2\mathrm{e}^{2t}$

(e) $\ddot{x}+5\dot{x}=t^2+3t\mathrm{e}^{-5t}:\quad x_h=C_1+C_2\mathrm{e}^{-5t}$

(f) $\ddot{x}+2\dot{x}+x=\mathrm{e}^{-t}\sin 3t:\quad x_h=C_1\mathrm{e}^{-t}+C_2 t\mathrm{e}^{-t}$

[Level : 2]

(a) We clearly have the no coincidence Case 1 of the Lecture, so use the third row of the table to pick the trial function $\underline{A\cos\omega t + B\sin\omega t}$ where A and B are unknown (the table uses C_1 and C_2 but any notation does).

Although the problem does not ask for this, let us solve the system also. Plugging the trial into this ODE we get $A = 1/\left(\omega_0^2 - \omega^2\right)$, $B = 0$. Thus the particular solution is

$$x_p = \frac{1}{\omega_0^2 - \omega^2}\cos\omega t \quad .$$

One thing we note from this example is that the coefficient B came out to zero. This happened because the ODE LHS $\mathbf{L}$ featured only EVEN ORDERS of derivative, which transform cosines to cosines and sines to sines. Since there was no sine in the excitation, the sine didn't appear in the particular solution also. This observation often simplifies your life considerably when you are analysing undamped spring systems, so remember it. Of course, this is not a tricky example; the ones coming next are.

(b) For $\cosh 2t$ we first write it as $\left(\mathrm{e}^{2t} + \mathrm{e}^{-2t}\right)/2$. We then see that the first component here is equal to one of the homogeneous solutions while the second component is not so. Hence we need to treat these two components separately.

For the first component, e^{2t} corresponds to λ=2 which in fact is a double root of $\mathbf{L}$. So we insert the bump-up factor of t^2 and then go to the second row of the table, getting the trial $At^2\,\mathrm{e}^{2t}$. For the second component, we have a straight-forward no coincidence case, so we try $B\,\mathrm{e}^{-2t}$. Putting them both together, the overall trial function is $\underline{x_p = At^2\,\mathrm{e}^{2t} + B\,\mathrm{e}^{-2t}}$.

Be aware of a potential fallacy here : drawing analogies between trigs and hyperbolics, you might be tempted to try $A\cosh 2t + B\sinh 2t$ with or without some overall bump-up factor outside. This trial is NOT going to work, as you can see by looking at the actual trial. A constant combination of cosh and sinh is equivalent to a constant combination of positive and negative exponentials, but cannot account for the t^2 factor preferentially attatching to one of them. This difference between trigs and hyperbolics occurs because a trig is a linear combination of exponentials associated with two roots which NECESSARILY

occur together (to ensure that the ODE is real) while a hyperbolic is a linear combination of exponentials which are not bound by any such constraint.

(c) With the $\cos \omega_d t$ term appearing in both the excitation and the homogeneous solution, it is tempting to insert a bump-up factor of t. WRONG ! The reason for this is that the homogeneous solutions correspond to a complex pair of roots $\lambda = -\gamma / 2 \pm \mathrm{j}\omega_d$, while the RHS corresponds to the exponential of $\pm \mathrm{j}\omega_d$. The two are completely different – there is NO coincidence here ! Hence this is a Case 1 situation and the trial function is $\underline{A \cos \omega_d t + B \sin \omega_d t}$. This example shows why I had to write the rules for Case 1 and Case 2 in a possibly obscure but pedantically correct language – if I'd gone with the intuitive phrasing "the excitation equals a homogeneous solution or contains it as a factor" then the rule would have produced an incorrect answer to this question.

(d) This time we have a situation straight out of Case 2 : the excitation has the form of a polynomial (of degree 3) multiplying an exponential which is a homogeneous solution. This solution occurs only once, no repeats, so the bump-up factor is t. Then look at the table of Case 1. Corresponding to the cubic polynomial, the first row asks for a different cubic polynomial while corresponding to the exponential, the second row asks for the same exponential. The fourth row tells us that to account for the product of the two, we should also try the product of the two. Putting all this together, we have the trial function $x_p = t\,(At^3+Bt^2+Ct+D)\,(E\,\mathrm{e}^{2t})$. The constant E is an extra since it multiplies all the other constants, so we can absorb it into those and simply write $\underline{x_p = (At^4+Bt^3+Ct^2+Dt)\,\mathrm{e}^{2t}}$. (Technically I should have written this expression with coefficients A' standing for AE, B' standing for BE etc. but since they are all unknown anyway, I can be lax about notation here.) Note that for this problem, it wasn't necessary to break up the polynomial into monomials and calculate x_p for each one separately. This is in fact a general rule – a polynomial can always be treated as a single entity.

(e) We consider separately the two functions t^2 and $3t\,\mathrm{e}^{-5t}$. The homogeneous ODE evidently has $\lambda = 0$ as a characteristic root, with the corresponding solution being $x_h = 1$. Now t^2 is a polynomial (of degree 2) multiplying e^0 so we have a coincidence case ! (In fact, this is why I had to explicitly mention the case $\alpha = 0$ in the Rules for coincidence.) Thus, the bump-up factor is t. Then, the trial table gives us a quadratic polynomial At^2+Bt+C and the trial function corresponding

to the t^2 becomes At^3+Bt^2+Ct. As regards the part $3te^{-5t}$, that has the same form as Part (d) above – a polynomial multiplying an exponential of a characteristic root. So we have a bump-up factor of t and a trial of $t(Dt+E)e^{-5t}$ corresponding to this part. Combining the two trial functions gives us the overall trial function $\underline{At^3+Bt^2+Ct+(Dt^2+Et)e^{-5t}}$.

Since the polynomial t^2 needed a bump-up, it is tempting to write a double bump-up (i.e. t^2) for the term $3te^{-5t}$, once for the polynomial t and once more for the exponential e^{-5t}. That however is wrong – only the exponential counts ! Look at the Rule for Case 2 – the excitation has the form of a polynomial (of degree one in this case) multiplying an exponential which corresponds to a homogeneous solution. Bump-up is required only for the exponential – since $\lambda = -5$ occurs only once, the bump-up is t. Bump-up for the polynomial t^2 occurred only because it multiplied e^0 with zero being a root – there also it was the exponential and not the polynomial which necessitated bumping up.

(f) Identical to Part (c), a t^2 bump-up is attractive but all that glisters is not gold. No bump-up here, use second and third rows to get the pieces Ae^{-t} and $B\cos 3t+C\sin 3t$, then use the fourth row to multiply them together, and finally absorb the outside factor A into B and C to get the trial function $\underline{e^{-t}(A\cos 3t+B\sin 3t)}$ where A and B, like chameleons, change their meaning from line to line !

*

COMMENTS : The two Rules and these six examples should prepare you for every possible undet coeff problem which you can face. The underlying concepts are all covered in the last Tute. If you internalize that, then your trial functions will be correct automatically. If on the other hand you find the concept too difficult then I'll recommend that you learn the Rules by heart, and understand thoroughly how those Rules led to the six answers here. That too will ensure that given a new problem, you'll be able to recall the Rules and apply them step by step to form the desired trials.

Just for curiosity, suppose you start hacking the algebra on an undet coeff procedure with the wrong trial function – then what happens ? What happens is that you eventually reach meaningless equations for the coefficients. For

example, if the trial should have been e^{2t} and you try with e^{3t} then your equation will have e^{2t} on one side and e^{3t} on the other (an obvious mismatch). In a more plausible error, suppose you have got the bump-up factor wrong (underbumped or overbumped). Then, if you have underbumped, your coefficient will come out to infinity (1/0 form). If you have overbumped, your coefficient will come out to zero. Thus, the method is to some extent self-correcting – you can adjust your trial functions if you can recognize and interpret the error messages correctly. Nevertheless, do not use this property to avoid learning how to select the trials. Evaluating the coefficients is time-consuming, so an erroneous starting point means a lot of time wasted in the exam hall. And on questions like this, where the coefficients aren't even asked for, a wrong trial will serve no purpose at all.

As an extension of this, you might devise the exam strategy that, when unsure of the bump-up, write a plethora of terms in the hope that some of them will stick. For example for Part (a) you can write a trial like $A_0 \cos \omega t + B_0 \sin \omega t + A_1 t \cos \omega t + B_1 t \sin \omega t + A_2 t^2 \cos \omega t + B_2 t^2 \sin \omega t + \ldots\ldots$ upto as many terms as you like, reasoning that some of the extra A_n's and B_n's will just turn out to be zero. The phrase "best possible starting guess" in the question rules this out. In fact, whatever the wording of the question, these smart-alec trial functions are always rejected and they cast a negative impression on the examiner. So don't do it. If you can't find the trial function, leaving it blank or writing what you think should be correct is much better than trying to cover your tracks in this manner.

The intricate cases of bump-up, no bump-up etc are fascinating from a mathematical point of view, and very useful for the exams. Operationally, more bump-ups mean lots more calculation to actually find the unknown coefficients. But we must not lose sight of the fact that the primary strength and utility of the method of undetermined coefficients lies in the quick derivation of simple results like the one in Part (a), and it is these results rather than the highly degenerate cases which have the most profound applications in science and technology. This fact you will see repeatedly in the examples T045-056, and again in the examples of the next Chapter.

3.2

EULER CAUCHY EQUATIONS

LEONHARD EULER (pronounced "oiler" – normally people don't care about pronunciation but saying "ewwler" is universally considered vulgar) was probably the second greatest mathematical physicist of all time after Sir ISAAC NEWTON (we'll encounter EULER's contributions several times in this book). AUGUSTIN-LOUIS CAUCHY also made groundbreaking contributions to the theory of partial differential equations. Indeed, the equation named after them is probably among the more insignificant of their differential equations achievements. It takes the form

$$x^n y^{(n)} + A_{n-1}x^{n-1}y^{(n-1)} + \ldots\ldots + A_2 x^2 y'' + A_1 x y' + A_0 y = f(x) \quad ,$$

where the A_i's are all constants. It is a linear equation (check against the definition if there's doubt). Because the number of derivatives in each term is equal to the number of powers of x multiplying it, it is called **equidimensional equation**.

There are two ways of solving this type of equation :

In the first way, for the homogeneous equation, try the **power substitution** $\boldsymbol{y = x^r}$ to obtain a characteristic equation for r. As in LCCH equations, there are four cases here.

- If the root r is real and distinct then of course the corresponding solution is x^r.
- If the root r is real and appears k times, then the corresponding solutions are x^r, $x^r \log x$, $x^r (\log x)^2$, ….., $x^r (\log x)^{k-1}$.
- If the root r is complex and has the value $\alpha + j\beta$, then of course its conjugate $\alpha - j\beta$ is another root. If the root is distinct, then the corresponding real-valued fundamental solutions are $x^\alpha \cos(\beta \log x)$ and $x^\alpha \sin(\beta \log x)$.
- If a complex root is repeated, then factors of $\log x$ start attaching to the above fundamental solutions.

After the homogeneous solutions are out, use variation of parameters to obtain the particular solution.

In the second way of solving Euler Cauchy equation, perform the **transformation $w = \log x$** to define a new independent variable w. This reduces the system to a LCC equation (homogeneous or inhomogeneous) for $y(w)$. Solve this equation and transform back to x from w.

The single example coming up should cover everything you can possibly need to know about this topic.

Example 3.2.1

Euler Cauchy equation

Considering the ODE

$$x^2y''+xy'-9y=2x^3 \quad ,$$

obtain the general solution using

(a) $y=x^r$ and variation of parameters,

(b) the transformation $w=\log x$.

[Level : 2]

EXTRA :

(c) Show that the substitution $w=\log x$ always converts a homogeneous Euler Cauchy equation into a LCCH equation.

(d) Basis the above result, show that the coefficients of the resulting LCCH equation are the same as those of the characteristic polynomial obtained by substituting $y=x^r$ into the Euler Cauchy equation.

(e) Considering the ODE

$$x^4y^{(4)}+6x^3y'''+9x^2y''+3xy'+y=2\cos\log x \quad ,$$

write the best possible starting guess for the trial function y_p.

(a) We plug $y = x^r$ into the homogeneous equation to get

$$\begin{aligned} ODE &\to x^2 r(r-1)x^{r-2} + xrx^{r-1} - 9x^r \\ &= x^r\left[r(r-1)+r-9\right] \\ &= x^r\left(r^2-9\right)=0 \end{aligned} \quad .$$

Note that x^r came out of the equation as a factor, which is why the substitution is effective. Now of course, x^r cannot be zero, so the thing multiplying it must be zero and we get our characteristic equation for r :

$$r^2 - 9 = 0 \quad ,$$

which we can solve for r.

This is very similar to the characteristic equation for a second order LCCH equation, but there are a couple of important differences also, which is why I have used r here rather than λ. Firstly, the substitution being tried is $y = x^r$ and NOT $y = e^{\lambda x}$ as in the LCCH case. An avoidable but common error in the exams is to obtain the characteristic equation, get the values of r correctly and then write the solutions as e^{rx}. This error is very costly as it implies either a careless attitude or blind memorization. Secondly, for LCCH, the characteristic polynomial has the same coefficients as the ODE but here that is not true (the characteristic equation for $y''+y'-9y=0$ would have been $\lambda^2+\lambda-9=0$).

The characteristic equation of course has the solutions $r = 3$ and $r = -3$, so the homogeneous solution is

$$y_h = C_1 x^3 + \frac{C_2}{x^3} \quad .$$

For the particular solution, we use variation of parameters. The excitation function is identical to one of the homogeneous solutions but that is an issue with undet coeff and not var of par. Here we just have a straightforward application of the var of par formula, so I won't cover it in detail. BUT, the thing to be careful of is, the question does not give you the ODE with the leading coefficient unity ! So before applying the formula, you MUST convert it to

$$x^2 y'' + xy' - 9y = 2x^3 \to y'' + \frac{1}{x}y' - \frac{9}{x^2}y = 2x \quad .$$

Of course this does not change the homogeneous solutions as you are multiplying the homogeneous ODE throughout by a factor. But, as per the original ODE, the inhomogeneity or excitation $f(x)$ is $2x^3$ while after the conversion, $f(x)$ is $2x$. The var of par formula (L1102) will give the right answer only if you take $f(x) = 2x$ – with $2x^3$ it will just give a wrong answer which you wouldn't even recognize as wrong unless you are really alert.

Thereafter, things are straightforward. The Wronskian is $W = -6/x$ and a messy calculation leads to the particular solution $y_p = \frac{1}{3}x^3 \log x$. Combining the two, we have the general solution

$$y_G = C_1 x^3 + \frac{C_2}{x^3} + \frac{1}{3}x^3 \log x \quad , \qquad *$$

which is the desired answer.

An extra factor of $\log x$ appeared in the particular solution when the excitation overlapped a homogeneous solution, and the same $\log x$ is the multiplier when the characteristic equation features root repetition; if this reminds you of the bump-up factors of t in LCC equations, then you are good at spotting patterns. The reason for this 'coincidence' will become apparent when we do the next part of this question.

(b) The problem tells us to define the new independent variable $w = \log x$ i.e. $x = \mathrm{e}^w$. Note that the derivative $\mathrm{d}x/\mathrm{d}w$ evaluates to x, a fact which simplifies the computations involving the Chain Rule (which thus makes another appearance in this course). Taking the derivatives we have

$$\frac{\mathrm{d}y}{\mathrm{d}w} = \frac{\mathrm{d}x}{\mathrm{d}w}\frac{\mathrm{d}y}{\mathrm{d}x} = x\frac{\mathrm{d}y}{\mathrm{d}x} \quad , \text{ and}$$

$$\frac{\mathrm{d}^2y}{\mathrm{d}w^2} = \frac{\mathrm{d}x}{\mathrm{d}w}\frac{\mathrm{d}}{\mathrm{d}x}\left(x\frac{\mathrm{d}y}{\mathrm{d}x}\right) = x\left(\frac{\mathrm{d}y}{\mathrm{d}x} + x\frac{\mathrm{d}^2y}{\mathrm{d}x^2}\right) = x^2\frac{\mathrm{d}^2y}{\mathrm{d}x^2} + x\frac{\mathrm{d}y}{\mathrm{d}x} \quad .$$

The above two relations yield that

$$x^2\frac{\mathrm{d}^2y}{\mathrm{d}x^2} = \frac{\mathrm{d}^2y}{\mathrm{d}w^2} - \frac{\mathrm{d}y}{\mathrm{d}w} \quad ,$$

and of course we have $x(\mathrm{d}y/\mathrm{d}x) = \mathrm{d}y/\mathrm{d}w$.

Plugging all these into the original ODE, the LHS reduces to

$$\frac{\mathrm{d}^2 y}{\mathrm{d}w^2} - 9y \quad ,$$

while the RHS is $2\mathrm{e}^{3w}$. Thus we have the ODE for $y(w)$

$$\frac{\mathrm{d}^2 y}{\mathrm{d}w^2} - 9w = 2\mathrm{e}^{3w} \quad ,$$

which is LCC inhomogeneous. Homogeneous solutions are e^{3w} and e^{-3w}, then undet coeff, trial function has one bump-up factor of w and particular solution is $(1/3)w\mathrm{e}^{3w}$. Transforming back to x from w gives the general solution

* $$y_G = C_1 x^3 + \frac{C_2}{x^3} + \frac{1}{3}x^3 \log x \quad ,$$ *

which is the desired answer. As we saw in T003, $\mathrm{e}^{3\log x}$ is x^3 and $\mathrm{e}^{-3\log x}$ is x^{-3}; they are NOT $3x$ and $-3x$ respectively !

This example shows why $\log x$ plays the same 'bump-up' role in Euler Cauchy equations as t in LCC equations, and also explains why there is really not much additional content in the theory of Euler Cauchy equations over what we have seen in the preceding pages.

*

COMMENTS : These extra Parts are more mathematical and are relevant only if the course emphasizes Euler Cauchy (EC) equations, whose applicational value is after all quite limited (there's only one instance where EC equation will come up later, and there it will be a simple example). But in case you are caught in a maths-heavy course, here goes.

(c) We want to show that the generic homogeneous EC equation for $y(x)$

$$x^n y^{(n)} + A_{n-1}x^{n-1}y^{(n-1)} + \ldots\ldots + A_1 xy' + A_0 y = 0 \quad ,$$

reduces to an LCCH equation for $y(w)$ if we perform the transformation $w = \log x$. The proof is not easy because of the generality involved here. The RHS of the equation is trivial so we focus on the LHS alone, which we now view as a differential operator acting on the function y. Evidently, the transformation will reduce this operator to a polynomial in $\mathrm{d}/\mathrm{d}w$ if and only if it converts every constituent term $x^k y^{(k)}$ into a polynomial in $\mathrm{d}/\mathrm{d}w$. Hence it is sufficient to show that $w = \log x$ converts $x^n(\mathrm{d}^n/\mathrm{d}x^n)$ to a n^{th} degree polynomial in $\mathrm{d}/\mathrm{d}w$ for all n (we already showed this for the case $n = 2$ in Part (b) above).

We prove by induction on n. Suppose it is true that $x^n(\mathrm{d}^n/\mathrm{d}x^n)$ is a polynomial in $\mathrm{d}/\mathrm{d}w$ of degree n with its leading coefficient unity (called a MONIC polynomial). Then we want to show that $x^{n+1}(\mathrm{d}^{n+1}/\mathrm{d}x^{n+1})$ is a monic polynomial in $\mathrm{d}/\mathrm{d}w$ of degree $n+1$. So let

$$x^n \frac{\mathrm{d}^n}{\mathrm{d}x^n} = \frac{\mathrm{d}^n}{\mathrm{d}w^n} + \alpha_{n-1}\frac{\mathrm{d}^{n-1}}{\mathrm{d}w^{n-1}} + \ldots\ldots + \alpha_0 \quad , \tag{01}$$

where the α's are constants. Since we are only supposing this at this time, it's called the induction hypothesis. To prove the hypothesis true, let's take a $\mathrm{d}/\mathrm{d}x$ on the above LHS :

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(x^n \frac{\mathrm{d}^n}{\mathrm{d}x^n} \right) = x^n \frac{\mathrm{d}^{n+1}}{\mathrm{d}x^{n+1}} + nx^{n-1}\frac{\mathrm{d}^n}{\mathrm{d}x^n} \quad .$$

Then, we multiply by x on the left. We have to be careful about the ordering of x and $\mathrm{d}/\mathrm{d}x$ here because x times $\mathrm{d}/\mathrm{d}x$ of a function is NOT the same as $\mathrm{d}/\mathrm{d}x$ of x times that function ! In fancier language, which actually comes up all the time in quantum mechanics, x and $\mathrm{d}/\mathrm{d}x$ (position and momentum if you're in the business) do not commute with each other. Anyway, doing the multiplication by x, we get

$$x\frac{\mathrm{d}}{\mathrm{d}x}\left(x^n \frac{\mathrm{d}^n}{\mathrm{d}x^n} \right) = x^{n+1}\frac{\mathrm{d}^{n+1}}{\mathrm{d}x^{n+1}} + nx^n \frac{\mathrm{d}^n}{\mathrm{d}x^n} \quad , \text{ or}$$

$$x^{n+1}\frac{\mathrm{d}^{n+1}}{\mathrm{d}x^{n+1}} = x\frac{\mathrm{d}}{\mathrm{d}x}\left(x^n \frac{\mathrm{d}^n}{\mathrm{d}x^n} \right) - nx^n \frac{\mathrm{d}^n}{\mathrm{d}x^n} \quad . \tag{02}$$

This completes the manipulations with the LHS of (01).

Now we have to work the RHS of (01). By Chain Rule,

$$\frac{\mathrm{d}}{\mathrm{d}x} = \frac{\mathrm{d}w}{\mathrm{d}x}\frac{\mathrm{d}}{\mathrm{d}w} = \frac{1}{x}\frac{\mathrm{d}}{\mathrm{d}w} = \mathrm{e}^{-w}\frac{\mathrm{d}}{\mathrm{d}w} \quad ,$$

so $x(\mathrm{d}/\mathrm{d}x) = \mathrm{d}/\mathrm{d}w$. Hence $x(\mathrm{d}/\mathrm{d}x)$ of the RHS of (01) is

$$\frac{\mathrm{d}^{n+1}}{\mathrm{d}w^{n+1}} + \alpha_{n-1}\frac{\mathrm{d}^n}{\mathrm{d}w^n} + \ldots\ldots + \alpha_0\frac{\mathrm{d}}{\mathrm{d}w} \quad ,$$

which is a polynomial of degree $n+1$ in $\mathrm{d}/\mathrm{d}w$.

Now invoke (02). We have, $x^{n+1}(\mathrm{d}^{n+1}/\mathrm{d}x^{n+1})$ equals a $n+1^{\text{th}}$ degree polynomial minus n times a n^{th} degree polynomial in $\mathrm{d}/\mathrm{d}w$, which is a $n+1^{\text{th}}$ degree

polynomial in $\mathrm{d}/\mathrm{d}w$. Further, the leading term of this polynomial, $\mathrm{d}^{n+1}/\mathrm{d}x^{n+1}$, occurs only in the first term of the RHS of (02), where it has a coefficient of unity. Hence this polynomial is monic, and the induction hypothesis has been proved. Since we have already shown the base case of $n=2$ to be true, the hypothesis indeed holds true for all n, which was to be shown.

(d) We now know that an EC equation for $y(x)$ always turns into LCCH equation for $y(w)$ on introducing $w=\log x$. Then, trying $y=x^r$ in the EC equation is identical to trying $y=\mathrm{e}^{rw}$ in its LCCH form. Both must yield the same values of r, and hence the same characteristic equation. But in an LCCH equation, the coefficients of the characteristic polynomial are identical to those of the ODE. Hence the coefficients of the transformed LCCH equation must be equal to those of the characteristic equation found by plugging x^r into the Euler Cauchy equation. In fact, you should know this trick even if you find the proof in Part (c) too difficult – it often simplifies life, as in the following example.

(e) Fourth order equation, asking for a trial function, and $\cos\log x$ on the RHS all point to an undet coeff calculation. So we introduce the substitution $w=\log x$. The RHS of the new equation is simple – it's $2\cos w$. And now, we use the result of Part (d) to bypass some calculation in finding the LHS. Obtaining the w equivalent of $x^4y^{(4)}$ using Chain Rule is no easy task, so instead we plug x^r into the homogeneous equation. The four derivatives of x^r are rx^{r-1}, $r\,(r-1)\,x^{r-2}$, $r\,(r-1)\,(r-2)\,x^{r-3}$ and $r\,(r-1)\,(r-2)\,(r-3)\,x^{r-4}$ respectively. It's still a lot of algebra but less than in the absence of the trick – if everything goes correctly you should end up at

$$\frac{\mathrm{d}^4y}{\mathrm{d}w^4}+2\frac{\mathrm{d}^2y}{\mathrm{d}w^2}+y=2\cos w \quad .$$

We've already seen the homogeneous solutions in T041 – they are $\cos w$, $\sin w$, $w\cos w$ and $w\sin w$. Since the excitation features $\cos w$ again, we must have bump-up factor of w^2, giving a trial of $w^2\,(A\cos w+B\sin w)$. Finally, converting from w to x, the trial function should be $\underline{(\log x)^2\,(A\cos\log x+B\sin\log x)}$.

3.3

MODELING WITH LCC EQUATIONS

Here we amplify upon the 'A' of DELTA – although we have seen some applications of first order equations in Chapter 1, they're hardly enough to warrant a new book. Second order equations with initial conditions specified arise most frequently in various kinds of dynamical systems. Coupled dynamical systems lead to higher order equations. Fourth order systems also arise in elasticity theory, but there, the problem typically comes with end point or boundary conditions, which we'll see in the next Chapter. Hence the focus here will be on dynamical systems.

Possibly the most common dynamical system to be governed by a linear ODE is the **spring mass damper system with forcing**

$$\ddot{x} + \gamma\dot{x} + \omega_0^2 x = f(t) \quad ,$$

where the **forcing function** or **excitation function** $f(t)$ is either a constant or a sinusoidal function.

With the damping not equal to zero, the homogeneous solutions decay exponentially so at large times we only get to see the particular solution. Since the ICs affect only the homogeneous solution, the system ends up at the same final state x_p irrespective of its initial condition. When f is a constant, the response is a constant also (the steady state). When f is oscillatory and has the form $F\cos\omega t$, the particular solution is always periodic at the same frequency ω. By a slight abuse of language, this state is called the steady state also – although the system is not "steady" in the puritanical sense of the word, the terminology doesn't really cause confusion. The amplitude R of this particular solution is $R = G(\omega)F$ where $G(\omega)$, the **gain function** (to use a term from electrical engineering), is

* $$G(\omega) = \frac{1}{\left[\left(\omega^2 - \omega_0^2\right)^2 + \gamma^2\omega^2\right]^{1/2}} \quad . \qquad \textbf{(01)*}$$

I'll derive this formula in an example. Note that G is not a function of t; the actual motion of the particle is $G(\omega)F\cos(\omega t-\varphi)$ for some phase angle φ, and that is the function of t. If the damping is light i.e. if $\gamma << \omega_0$ then G is sharply peaked at ω very close to ω_0, and we have a **resonance**. If $\omega >> \gamma, \omega_0$ then we have a decrease of the response amplitude with increasing ω. Since the IC affects only the homogeneous solutions which are decaying , the system settles into this periodic state whatever the initial conditions.

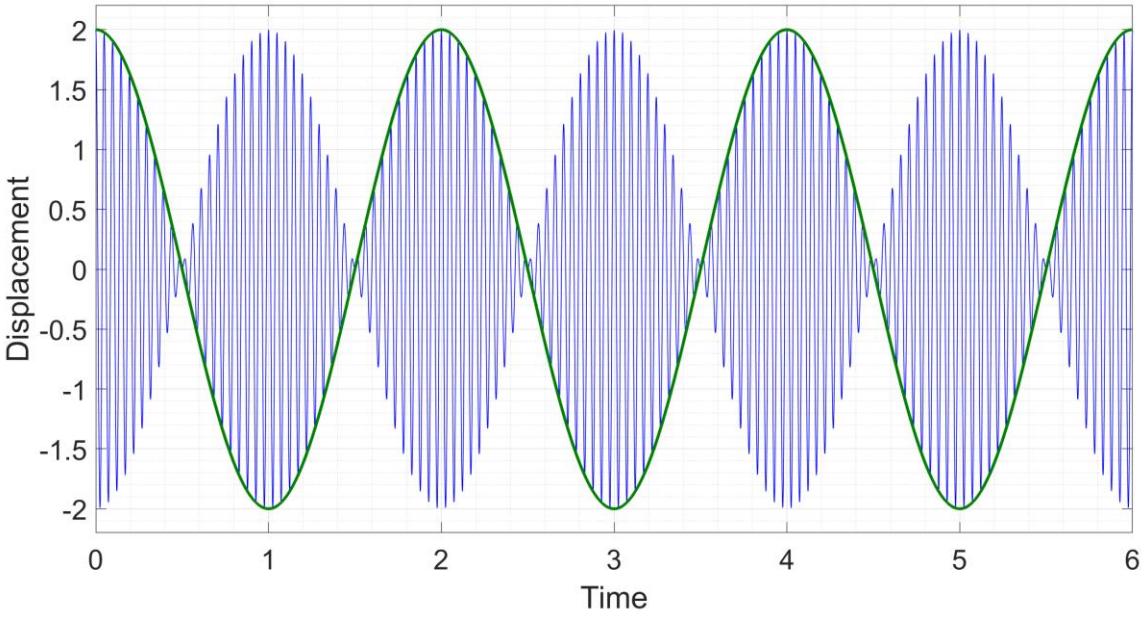
With exactly zero damping (which can be achieved using a controller so it's not unrealistic) and oscillatory forcing, we have some new phenomena. At $\omega = \omega_0$ the formula is divergent, and we actually have linear resonance which we saw in 3.1.2. Near the resonance , we have the phenomenon of **beats** which carries greater practical interest . These occur as a result of superposition of the homogeneous solution $\cos \omega_0 t$ and the particular solution $\cos \omega t$. The trig identity

$$\cos \omega t + \cos \omega_0 t = 2\cos\left(\frac{\omega - \omega_0}{2} t\right)\cos\left(\frac{\omega + \omega_0}{2} t\right) \quad ,$$

shows that the sum of these solutions can also be written as a product. Now when ω is very close to ω_0, the sum term $(\omega_0+\omega)/2$ is like ω but the difference term is something which oscillates at a much slower frequency $(\omega-\omega_0)/2$. The effect is that of oscillations at frequency ω being modulated by oscillations at a much lower frequency – this low frequency is called the beats frequency. Note that the beat frequency is **half of the difference** between the natural and excitation frequencies.

Below we show a graph of the function $\cos 40\pi t + \cos 42\pi t$ in blue.

(An unfortunate location for a page break but I'd rather have the wasted space than a diminutive Figure)



The beat frequency should be $2\pi(1/2)$ and indeed it is. The envelope of the oscillations (green) clearly shows oscillation with period 2 while the curve itself oscillates within this envelope at a rapid rate. The utility of beats is that it can bring down a high frequency (that of the actual oscillation) to a lower frequency (that of the envelope) which can be measured easily. The details of this, we'll see in the Tutes.

Another system which lends itself to a similar analysis is the circuit consisting of an inductor L, capacitor C and a resistor R connected in series, and called a **series LCR circuit**. The equation for this circuit is

$$Li''+Ri'+\frac{1}{C}i=\frac{\mathrm{d}}{\mathrm{d}t}V(t) \quad ,$$

where $i(t)$ is the current in the circuit, prime means $\mathrm{d}/\mathrm{d}t$ in this case as in 1.2 .7, and $V(t)$ is the voltage applied across the circuit . I won't go into the details of the derivation , but will merely mention that it comes from KIRCHHOFF's voltage and current laws and the definitions of L, C and R. The equation is identical to the one for the mass spring damper system with forcing.

That's enough of general theory, now for some juicy examples.

Example 3.3.1

Simplified induction motor

Following the two-reaction theory, an induction motor can be modelled as two coupled windings, direct and quadrature, each of which is like an LR circuit. The model equations for the windings are

$$\begin{aligned} Li_d' + Ri_d + \Omega L i_q &= V_d \\ Li_q' + Ri_q - \Omega L i_d &= V_q \end{aligned} \quad ,$$

where i_d and i_q are the direct and quadrature currents and V_d and V_q the corresponding source voltages. All applied voltages are on the direct winding alone.

(a) Uncouple the system to write a single second order ODE for i_q.

In a particular apparatus, the motor resistance is extremely small. The direct winding is connected to a dc voltage source which produces a constant voltage $V_d = V_0$ when it is ON and zero voltage when it is OFF. In an experiment, the source is OFF and the motor is quiescent ($i_d = i_q = 0$) when at $t = 0$ the source is suddenly turned ON.

(b) The maximum instantaneous current which the motor winding (either direct or quadrature) can withstand is $3V_0/2\Omega L$. Will the experiment run or will it destroy the motor ? Justify your answer.

[Level : 2]

(a) Since all voltage is on the direct winding alone, $V_q = 0$. Then, since we want an equation for i_q alone, we wish to eliminate i_d from the coupled system. Rearranging the second equation we have

$$i_d = \frac{Li_q' + Ri_q}{\Omega L} \quad .$$

Taking the derivative,

$$i_d' = \frac{Li_q'' + Ri_q'}{\Omega L} \quad .$$

Now we plug these two into the first equation to get

$$\frac{Li_q'' + Ri_q'}{\Omega} + R\frac{Li_q' + Ri_q}{\Omega L} + \Omega L i_q = V_d \quad \text{, or}$$

$$* \qquad i_q'' + \frac{2R}{L} i_q' + \Omega^2 i_q = \frac{\Omega V_d}{L} \quad , \qquad *$$

which is the ODE asked for. We note that this is a harmonic oscillator equation with damping and external excitation.

(b) When V_d is a constant V_0, the trial function for the particular solution i.e. steady state current $i_{q,p}$ is also a constant C. Plugging $i_{q,p} = C$ into the ODE gives $C = V_0/\Omega L$. This is clearly less than the tolerance limit of the windings. But don't jump at this as the answer ! Although the motor does attain this state eventually, we are given an INSTANTANEOUS current tolerance, so we also have to determine whether the tolerance is crossed at any time during the approach to the steady state. In other words, we need to account for the transient currents also. Since the resistance is very small, we start by taking it to be zero. Under this assumption, the equation for the current is

$$i_q'' + \Omega^2 i_q = \frac{\Omega V_0}{L} \quad ,$$

and the ICs are $i_q = 0$, $i_q' = 0$.

The trial function for particular solution is again a constant and again it evaluates to $i_{q,p} = V_0/\Omega L$ and adding the homogeneous solutions we have the general solution

$$i_{q,G} = C_1 \cos\Omega t + C_2 \sin\Omega t + \frac{V_0}{\Omega L} \quad .$$

The first IC gives $C_1 = -V_0/\Omega L$ while the second gives $C_2 = 0$ hence the specific solution is

$$i_q = \frac{V_0}{\Omega L}\left(1 - \cos\Omega t\right) \quad .$$

Within half a cycle, this attains the value $2V_0/\Omega L$ which is significantly greater than the tolerance value, so we suspect that the motor is destroyed by the experiment.

To make the answer water tight, we examine the approximation involved (that the motor has zero resistance). The problem mentioned that the resistance is very small, and 2.1.6 showed that dropping a small damping produces an excellent approximation at small times, which are of interest in this problem. The current attained in the winding is way above the threshold so the approximation does not affect the result in any way. Hence the motor indeed is destroyed by the experiment.

*

COMMENTS : The motor model I have presented here is grossly simplified (obviously, since that word is in the title). Firstly, the direct and quadrature windings are actually those of the rotor, and the model assumes that the stator is controlled by a current source. V_d and V_q here are actually the externally controlled direct and quadrature stator currents times some proportionality constants. The linearized model of a motor with stator connected to voltage source (which is more common in reality) is fourth order so it is far more difficult to handle. The parameter Ω is the rotation speed of the motor. Here I have assumed it to be constant – in reality it is also time-varying, with the equation for $\mathrm{d}\Omega/\mathrm{d}t$ featuring the stator and rotor currents. If we account for that variation, then we have a nonlinear system and analytical results in that case are few and far between.

Yet, despite all the simplifications of the model, the problem makes a real point. This is that the transient currents flowing in the windings are often significantly greater than the steady state values. Here we saw a factor of two amplification; with voltage source supply and the nonlinearities included, the factor is often five or more. Such transients arise whenever there is a sudden change in the applied voltage or current such as motor being started by direct connection to

supply line. Multiple times the rated resistive losses occur during such line starting, which is why the number of starts per hour is limited for large motors, such as the one considered here (very low resistance is a hallmark of a large machine). In sophisticated applications, elaborate control arrangements are used to ensure conformity of the currents to their tolerance values.

We will have more to say about induction motors – in the next Chapter we'll see the voltage source supplied motor model as well as a cutting-edge control technology used in industry. In the PDE part of the book, we'll see how to obtain the dynamic motor model from first principles.

Example 3.3.2

The Gain function

Derive the formula for gain function $G(\omega)$ of the damped driven harmonic oscillator

$$\ddot{x} + \gamma\dot{x} + \omega_0^2 x = F\cos\omega t \quad .$$

Discuss the salient features of this function, considering various limiting situations.

[Read Only]

First we recall the definition of the gain $G(\omega)$. Considering the ODE

$$\ddot{x}+\gamma\dot{x}+\omega_0^2 x = F\cos\omega t \quad ,$$

the gain is the ratio of the amplitude R of the solution function to the amplitude F of the excitation. With damping present i.e. γ non-zero, the homogeneous solutions decay in time so they are transient, and only the particular solution is of interest. I would have skipped the bulk of the steps leading to the particular solution and thence to the gain, but it so happens that they assign this as homework every single semester. So here's a homework solution on a platter.

To find particular solution we use undetermined coefficients; the trial function is $x_p = A\cos\omega t + B\sin\omega t$ (3.1.5 Part (c) assures us that there are no bump-ups even in the 'worst case' $\omega = \omega_d$). As in 3.1.2, we write this trial in shorthand as $x_p = A\mathrm{C}+B\mathrm{S}$. Taking the derivatives and plugging into the ODE, we find

$$\begin{aligned} ODE \to\ & \ddot{x}_p+\gamma\dot{x}_p+\omega_0^2 x_p \\ &= -\omega^2 A\mathrm{C}-\omega^2 B\mathrm{S}+\gamma\left(-\omega A\mathrm{S}+\omega B\mathrm{C}\right)+\omega_0^2\left(A\mathrm{C}+B\mathrm{S}\right) \quad . \\ &= F\mathrm{C}+0\mathrm{S} \end{aligned}$$

Equating the coefficients of C and S yields the simultaneous equations

$$\begin{aligned} \left(\omega_0^2-\omega^2\right)A+\gamma\omega B &= F \\ -\gamma\omega A+\left(\omega_0^2-\omega^2\right)B &= 0 \end{aligned} \quad ,$$

which we solve using KRAMER's rule. This rule yields

$$\begin{bmatrix} A \\ B \end{bmatrix} = \frac{1}{\left(\omega_0^2-\omega^2\right)^2+\gamma^2\omega^2}\begin{bmatrix} \omega_0^2-\omega^2 & -\gamma\omega \\ \gamma\omega & \omega_0^2-\omega^2 \end{bmatrix}\begin{bmatrix} F \\ 0 \end{bmatrix} \quad ,$$

i.e.

$$A = \frac{\left(\omega_0^2-\omega^2\right)F}{\left(\omega_0^2-\omega^2\right)^2+\gamma^2\omega^2} \qquad B = \frac{\gamma\omega F}{\left(\omega_0^2-\omega^2\right)^2+\gamma^2\omega^2} \quad .$$

So the particular solution is out; now we have to find its amplitude R.

We use the fact (mentioned once already in 2.3.2) that

$$A\cos\omega t + B\sin\omega t = R\cos\left(\omega t-\varphi\right) \quad \text{, where}$$

$$R = \left(A^2 + B^2 \right)^{1/2}, \quad \varphi = \arctan \frac{B}{A} \quad .$$

The phase is not of interest, and, substituting the A and B we found above, we get the amplitude of x_p as

$$R = \left[F^2 \left(\frac{\left(\omega_0^2 - \omega^2 \right)^2 + \gamma^2 \omega^2}{\left(\left(\omega_0^2 - \omega^2 \right)^2 + \gamma^2 \omega^2 \right)^2} \right) \right]^{1/2} \quad .$$

We're in luck – although the expressions could have become very messy here, the numerator in this case happens to be just the square root of the denominator. Thus, we get

$$R = \frac{F}{\left[\left(\omega_0^2 - \omega^2 \right)^2 + \gamma^2 \omega^2 \right]^{1/2}} \quad .$$

We can just as well write the term in the denominator as $(\omega^2 - \omega_0^2)^2$; doing so and recognizing that the gain $G(\omega)$ is defined as $G = R/F$, we have our formula for the gain

$$G(\omega) = \frac{1}{\left[\left(\omega^2 - \omega_0^2 \right)^2 + \gamma^2 \omega^2 \right]^{1/2}} \quad ,$$

which is the desired answer.

For exams, you might or might not be required to memorize this formula – some instructors allow or provide a formula sheet while others are less friendly. Nevertheless, if you are a mechanical or electrical engineer, you will need this formula so often that you would do well do learn it. Note that the dimension of the gain (at least when defined this way) is 1/frequency^2; $\omega_0^2 x$ matches F, and so does R (amplitude – same dimension as x) divided by G. Dimensionally incorrect formulae for G are avoidable, but I have seen lots of them, and not all from students.

Now we turn to the properties of G. Since $\cos(-\omega t) = \cos \omega t$, the relevant frequency range is ω going from zero to infinity. Evidently, G is positive everywhere, as it should be (it is a ratio of amplitudes which are always positive).

G always has one maximum over its domain of definition. To find the max given γ and ω_0, we can either use calculus, setting $\mathrm{d}G/\mathrm{d}\omega = 0$, or algebra, completing the square in the denominator. Since we'll use the calculus approach in the very next problem, I will do the algebraic method here.

The maximum of G occurs when its denominator is minimum; since square root is a monotonic function, this is equivalent to saying that $\left(\omega^2 - \omega_0^2\right)^2 + \gamma^2\omega^2$ must be a minimum. Opening out, we can write it as $\omega^4 - \left(2\omega_0^2 - \gamma^2\right)\omega^2 + \omega_0^4$; now complete the square to get

$$\left[\omega^2 - \left(\omega_0^2 - \frac{\gamma^2}{2}\right)\right]^2 + \omega_0^4 - \left(\omega_0^2 - \frac{\gamma^2}{2}\right)^2 \quad .$$

This expression consists of a perfect square involving ω inside the box bracket plus some terms independent of ω; the maximum value is clearly when the square is zero i.e. at

$$\omega = \left(\omega_0^2 - \frac{\gamma^2}{2}\right)^{1/2} \quad .$$

When γ is small, the above ω is very close to ω_0, and we have an important property of G: for light damping, the maximum gain occurs at excitation frequency almost equal to natural frequency.

The squaring argument also gives us the height of the maximum : it is

$$\frac{1}{\left[\omega_0^4 - \left(\omega_0^2 - \frac{\gamma^2}{2}\right)^2\right]^{1/2}} \quad ,$$

which for small γ reduces to $1/\gamma\omega_0$. In the limit $\gamma \to 0$, the frequency of maximum response becomes ω_0 and the amplitude becomes infinite, which is consistent with the breakdown of the solution and its replacement by the linearly growing thing from 3.1.2. Thus we have a second important property of G: for light damping, the maximum gain is the inverse of the damping constant and the natural frequency. If we increase γ starting from a small value, then the maximum starts to shift leftward and its height starts to reduce. At $\gamma = \sqrt{2}\omega_0$

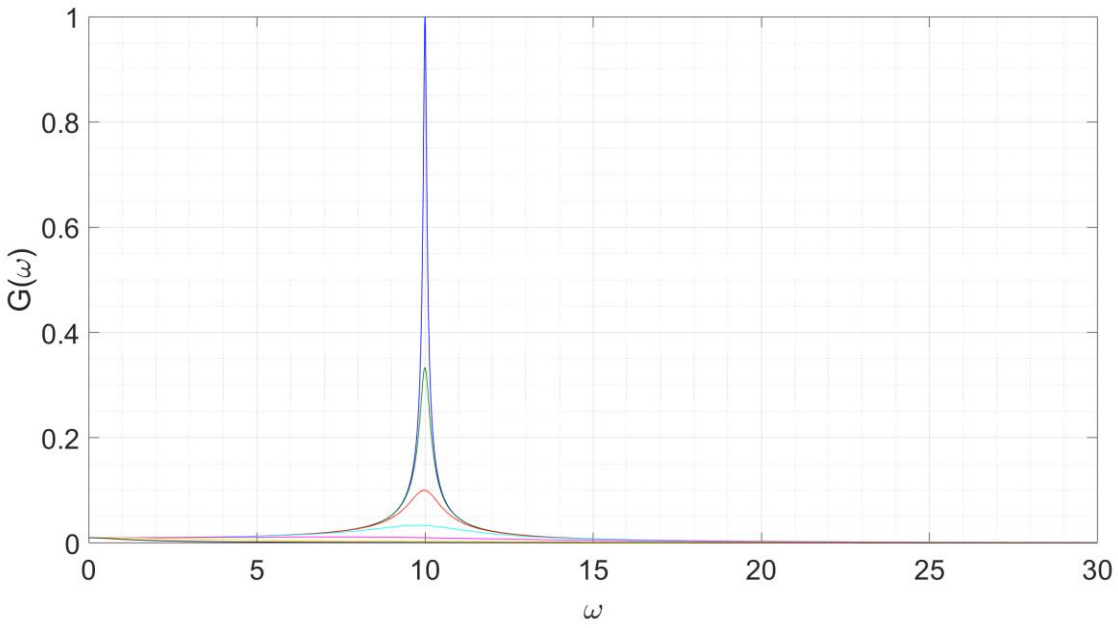
the maximum response occurs at $\omega=0$ and $G(\omega)$ becomes monotonically decreasing with the max occurring at the boundary of the domain.

A different limit is when the excitation frequency is extremely high i.e. $\omega>>\omega_0,\gamma$. In this case, everything inside the radicand in the denominator can be dropped in comparison with ω^4 and we have $G(\omega)=1/\omega^2$. Thus a third important property of G is that it at high frequencies the gain is the inverse square of the excitation frequency. Note that there are no proportionality constants here.

Here is a plot of $G(\omega)$ for the equation

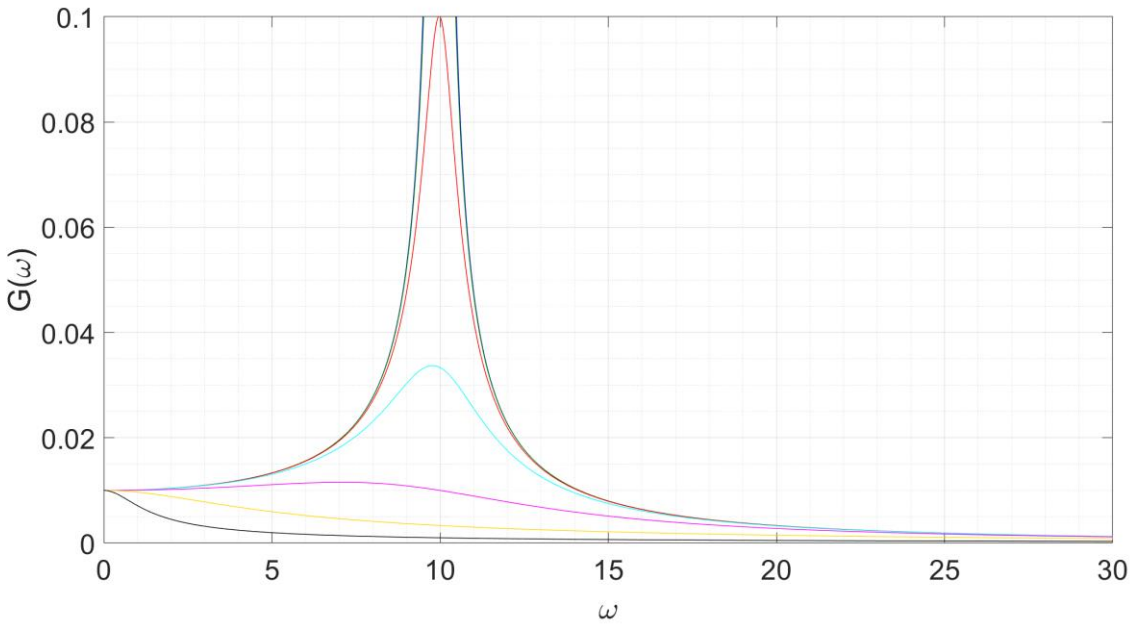
$$\ddot{x}+\gamma\dot{x}+100x=F\cos\omega t \quad ,$$

with γ taking the successive values 0.1, 0.3, 1, 3, 10, 30, 100. The corresponding curves are shown in blue, green, red, cyan, magenta, yellow and black respectively. Because of the huge maximum attained by the first one, the others are scarcely visible (don't worry, we'll zoom in).



Clearly, the maximum response on the blue curve occurs at frequency exactly 10 for all practical purposes (10 is the natural frequency). As per the limiting formula, its height should be $1/\gamma\omega_0$, which is 1, and indeed it is. We can also see that the blue curve is a narrow spike, while the green and red ones are more rounded. Thus we have a fourth important property of G : the width of the peak decreases sharply as the damping decreases. This fact can be shown

rigorously from a calculation (by solving for the frequency at which G is half the value at the peak itself) but I'll let it go. We now look at a zoomed up version of the plot, showing the other lines.



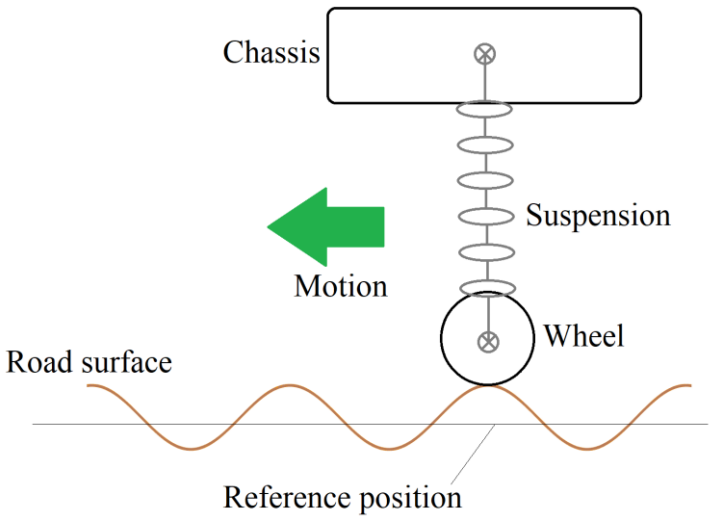
Note that the red line ($\gamma = 1$) is visible in both pictures, so it can act as a scale reference between the two. Even at this rather high damping (5 percent of critical) the location and height of the maximum are for all practical purposes exactly 10 and 0.1, which are obtained from the crudest approximation. With still further damping, the maximum really starts to shift leftward and the height becomes small. As per our predictions, the curve should be monotonic for $\gamma > 14$ i.e. in the last two plots (yellow and black, γ equals 30 and 100). Indeed, this is so.

These general properties of the gain function are extremely useful in understanding and calculating the responses of damped driven systems, a couple of which we will see in the examples which are coming up. If you get stuck at anything there, please refer back here.

Example 3.3.3

Vehicle on a bumpy road – unicycle model

A vehicle can crudely be modelled as a single wheel of mass m connected to a chassis of mass M via a spring of natural length L and constant k and a damper of constant C. The wheel travels with constant horizontal velocity v along a bumpy road whose vertical displacement from the reference position is given by $w = A\cos\lambda y$ where w is vertical displacement and y the horizontal distance travelled. The setup is shown in the Figure below.



Assume that the wheel remains attached to the road at all times.

(a) Write an ODE for the displacement of the chassis M as a function of time.

(b) Given k, is it possible to arrange the damping such that the amplitude of oscillation of the chassis decreases monotonically as v increases, for all v ? If yes, determine the minimum C required for the same. If no, justify your answer.

[Level : 2]

(a) Let z be the displacement of the chassis M from the horizontal reference plane. To write an equation of motion for M, we need to consider all the forces acting on it. These are the force from the spring and damper and the force of gravity. The force of gravity is just mg downward, which is $-mg$ if we let z be positive upwards.

For calculating spring and damper force we start from the profile of the road. That is $w=A\cos\lambda y$. Since the car travels at constant speed v, $y = vt$; moreover since the wheel always follows the road, its z-coordinate is $w = A\cos\omega t$ where $\omega = \lambda v$. The length of the spring at any instant is $z-w$. The restoring force exerted by the spring on the mass is k times the length of the spring minus the natural length i.e.

$$F_{sp} = -k\left(z - w - L\right) \quad .$$

Check the signs : if you raise the chassis relative to the wheel, then $z-w$ becomes greater than L, and the force is negative i.e. the chassis tends to come down, which is consistent. Likewise, the damper opposes relative motion between the wheel and the chassis; the damping force is

$$F_d = -C\left(\dot{z} - \dot{w}\right) \quad ,$$

and all forces on the chassis are out.

Writing NEWTON's Second,

$$\begin{aligned} M\ddot{z} &= -Mg + F_{sp} + F_d \\ &= -Mg - k\left(z - A\cos\omega t - L\right) - C\left(\dot{z} + \omega A\sin\omega t\right) \quad \text{, or} \\ M\ddot{z} + C\dot{z} + kz &= kL - Mg + kA\cos\omega t - C\omega A\sin\omega t \quad . \end{aligned}$$

Dividing through by the chassis mass M we have

$$\ddot{z} + \gamma\dot{z} + \omega_0^2 z = \omega_0^2 L - g + \omega_0^2 A\cos\omega t - \gamma\omega A\sin\omega t \quad , \qquad (*)$$

which is the ODE asked for.

(b) The excitation function consists of two parts – a dc term and an oscillatory term. Since the problem deals with the oscillations of the vehicle, we begin by throwing the dc terms out of consideration. Then of course we want to use the gain function. Since the excitation here consists of both cos and sin but the gain deals with amplitudes, we first convert the forcing to amplitude-phase representation i.e. we write

$$f(t)=\left(\omega_0^4A^2+\gamma^2\omega^2A^2\right)^{1/2}\cos\left(\omega t-\varphi\right)\quad,$$

where φ is not of interest. The important feature here is that the excitation amplitude is also a function of ω.

Now, before using the gain formula, we note that the question asks if we can get a response function which is monotonically decreasing with ω. This will be true irrespective of the value of the amplitude A multiplying the overall excitation, so we can take that out of the game. Then, the response amplitude is

$$\begin{aligned}R(\omega)&=G(\omega)\left(\omega_0^4+\gamma^2\omega^2\right)^{1/2}\\&=\frac{\left(\omega_0^4+\gamma^2\omega^2\right)^{1/2}}{\left[\left(\omega^2-\omega_0^2\right)^2+\gamma^2\omega^2\right]^{1/2}}\end{aligned}\quad.$$

We now need to analyse its properties.

Below, I will present two methods of answering the question; the first is systematic while the second involves seeing the solution. While doing it myself, I had ploddingly followed the first approach; only after getting the answer, I devised the second one as a test of my algebra.

The standard procedure to find monotonicity is to locate the maxima and minima of the function. If R is to be monotonically decreasing in ν, then it cannot have any maxima or minima for strictly positive ω. Accordingly, we must find the zeros of $\mathrm{d}R/\mathrm{d}\omega$. But, since R is positive, R^2 will have the same maxima and minima as R itself, and we can save a bit of algebra by hunting for max and min of R^2 rather than R. From the Quotient Rule we have

$$\frac{\mathrm{d}}{\mathrm{d}\omega}R^2=\frac{\gamma^2\omega\left[\left(\omega^2-\omega_0^2\right)^2+\gamma^2\omega^2\right]-\left(\omega_0^4+\gamma^2\omega^2\right)\left(2\omega^3-2\omega_0^2\omega+\gamma^2\omega\right)}{\mathrm{XXX}}\quad,$$

where XXX is something we need not bother to evaluate since that can't drive the expression to zero anyway. Hacking at the numerator and setting it to zero yields

$$\omega\left(\gamma^2\omega^4+2\omega_0^4\omega^2-2\omega_0^6\right)=0\quad.$$

This shows that $\omega = 0$ is always a critical point. The function will be monotonic if there are no other critical points for positive ω.

However, solving the quadratic for ω^2 gives

$$\omega^2 = \frac{-\frac{2\omega_0^4}{\gamma^2} \pm \sqrt{\frac{4\omega_0^8}{\gamma^4} + \frac{8\omega_0^6}{\gamma^2}}}{2} \quad ,$$

which always has one positive root for ω^2, which corresponds to a real and positive value of ω. Thus there is always a value of ω greater than zero where R has a critical point, and it is impossible to generate a monotonic response.

Now for the quick way to get to the answer. We know that at large ω, the response function must be decreasing – the numerator goes as ω but the denominator goes as ω^2. But we suspect that the behaviour near zero will pose a problem because of the ω in the numerator. Accordingly, we write R as

$$R = \frac{\left(\omega_0^4 + \gamma^2\omega^2\right)^{1/2}}{\left[\left(\omega_0^2 - \omega^2\right)^2 + \gamma^2\omega^2\right]^{1/2}} \quad ,$$

same as what it was before except for an inconsequential sign flip in the denominator. At ω=0, R=1; at some small ω, if the first term in the denominator had been ω_0^4 then R would again have remained 1. However, this term is actually less than ω_0^4 so R is greater than 1. This proves that $R(\omega)$ increases initially with ω and generating a monotonic response is impossible.

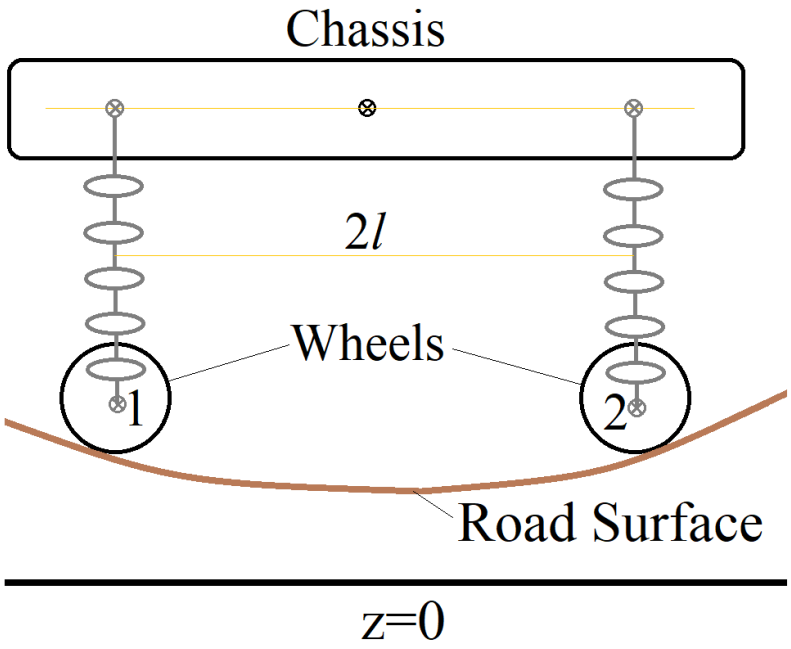
*

COMMENTS : Why ‘y’ and not ‘x’ for the direction of motion ? This is just a convention of vehicle dynamics – it is typical to have z as the vertical, y as the direction of travel and x as the line pointing directly to starboard. I have used the same convention in the next two problems, which are also on this topic.

Example 3.3.4

Biaxial bus

A bus having two axles can be modelled as shown in the Figure below.



Let the left wheel be #1 and the right one be #2. Let z_1 and z_2 denote the z-coordinates of the tops of the respective springs and let w_1 and w_2 denote those of the respective wheels measured relative to the baseline $z=0$. Assume that the two springs (L, k) and dampers (C) are identical and positioned symmetrically, and that the wheels are massless so that the centre of mass (CM) of the bus is in dead centre of the chassis. Let M be the chassis mass, $2l$ the wheelbase (shown) and I the moment of inertia of the chassis about an axis passing through the CM and perpendicular to the plane of the page.

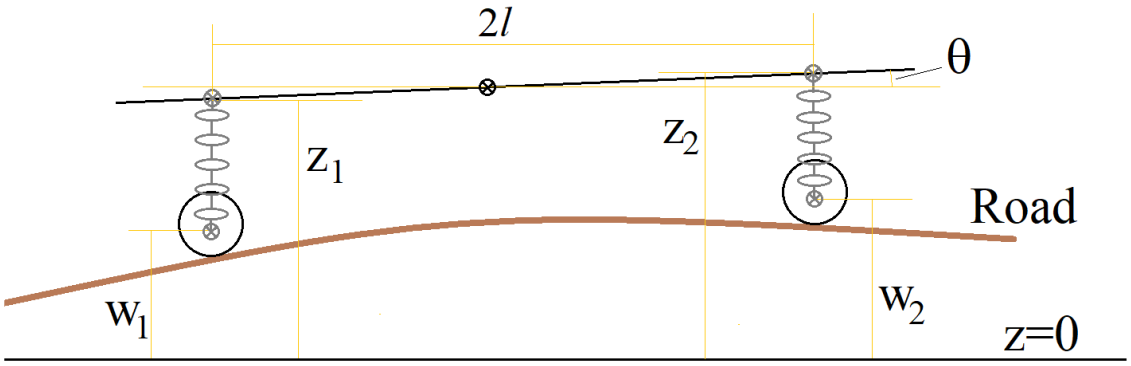
(a) Write a pair of coupled ODEs for z_1 and z_2 as functions of time, assuming that the tilt of the bus relative to the horizontal is small. Your ODEs should feature the w's as excitations.

The bus now travels on a bumpy road $w = A\cos\lambda y$ with constant velocity v.

(b) Ignoring the homogeneous solutions, determine the conditions on λ and v for which the bus performs up and down translational motion alone, and tilting motion alone.

[Level : 3]

(a) We define z_1, z_2, w_1, w_2 and θ as mentioned in the problem and/or shown in the Figure below. Let F_1 and F_2 be the forces exerted on the chassis by the two springs.



From the last example we know that

$$F_1 = -k(z_1 - w_1 - L) - C(\dot{z}_1 - \dot{w}_1) \quad ,$$
$$F_2 = -k(z_2 - w_2 - L) - C(\dot{z}_2 - \dot{w}_2) \quad .$$

In addition, gravity acts on the chassis, exerting a force $-mg$ passing through the CM.

Now we have to formulate the equations of motion of chassis. One is the translational equation, $M\ddot{z}_{CM} = F_z$. But since the CM lies halfway between the two wheels, the z-coordinate of the chassis is $(z_1+z_2)/2$. This gives

$$* \quad \frac{M}{2}(\ddot{z}_1 + \ddot{z}_2) = F_1 + F_2 - Mg$$
$$= -k(z_1 - w_1 - L) - k(z_2 - w_2 - L) - C(\dot{z}_1 - \dot{w}_1) - C(\dot{z}_2 - \dot{w}_2) - Mg \quad . \quad (01)^*$$

This is one of the two required equations.

The other comes from balancing the torque on the chassis. We have $I\ddot{\theta} = l(F_2 - F_1)$. But from the geometry, $z_2 - z_1 = 2l\sin\theta$ and for small θ (which is the only case that makes sense without either deforming the springs or elongating the bus) we have $z_2 - z_1 = 2l\theta$ i.e. $\theta = (z_2 - z_1)/2l$. Substituting this into the torque balance equation gives

$$* \quad \frac{I}{2l}(\ddot{z}_2 - \ddot{z}_1)$$
$$= l\left[-k(z_2 - w_2 - L) + k(z_1 - w_1 - L) - C(\dot{z}_2 - \dot{w}_2) + C(\dot{z}_1 - \dot{w}_1)\right] \quad . \quad (02)^*$$

Equations (01) and (02) make up the coupled ODE system asked for.

(b) We have been asked for conditions under which only translation or only rotation takes place. Clearly, pure translation means that $z_1 = z_2$ throughout i.e. their sum is something finite but their difference is identically zero. Likewise, a pure tilting motion means that $z_1 = -z_2$ throughout i.e. their difference is something but the sum identically zero. These observations, coupled with the fact that sum and difference arose naturally in the formulation of the ODE system, cause us to officially define coordinates $z_s = z_1 + z_2$ and $z_d = z_1 - z_2$.

In terms of these new coordinates we can recast the ODEs as

$$\frac{M}{2}\ddot{z}_s + C\dot{z}_s + kz_s = 2kL - Mg + k\left(w_1 + w_2\right) + C\left(\dot{w}_1 + \dot{w}_2\right) \quad , \qquad \text{(03a)}$$

$$\frac{I}{2l^2}\ddot{z}_d + C\dot{z}_d + kz_d = k\left(w_2 - w_1\right) + C\left(\dot{w}_2 - \dot{w}_1\right) \quad . \qquad \text{(03b)}$$

This simple transformation causes the two ODEs to completely decouple in this case.

Now we need to find w_1 and w_2. The road profile is $A\cos\lambda y$, $y = vt$ and without loss of generality we can let $w_1 = A\cos\lambda vt$ with no extra phase shift. The second wheel however is a constant distance $2l$ ahead of the first wheel. Hence, given w_1, w_2 must be equal to $A\cos\lambda(y+2l)$ which is $A\cos(\lambda vt + 2\lambda l)$. Thus we have the relations

$$\begin{aligned} w_1 &= A\cos\left(\lambda vt\right) \\ w_2 &= A\cos\left(\lambda vt + 2\lambda l\right) \end{aligned} \quad ,$$

which we now use to obtain the desired results.

For oscillatory translational motions alone, the difference mode z_d must be zero. Since we have been told to focus on the particular solutions alone, it will be sufficient if the excitation function on the RHS of (03b) is zero. Clearly, this implies that $w_2 - w_1$ must be zero. Since cos repeats with period 2π, this happens if and only if $2\lambda l$ is a multiple of 2π i.e. $\underline{\lambda l = n\pi}$ where $n = 1,2,3\ldots..$ is a natural number. For tilting motions alone, the sum mode cannot be excited (the dc terms there do not contribute to any motion, and are neglected) and the RHS of (03a) must vanish. Hence $w_1 + w_2$ must be zero. Since $\cos(\theta + m\pi) = -\cos\theta$ if m is odd, the condition is that $2\lambda l$ must be an odd multiple of π, or $\underline{\lambda l = (2n-1)\pi/2}$

where n is again a natural number. Note that the speed v does not enter the condition at all !

*

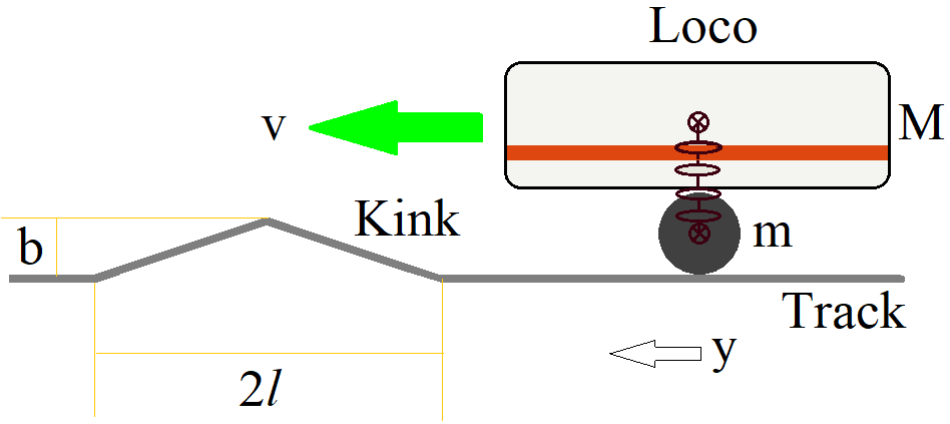
COMMENTS : The unicycle model of a vehicle we considered in the previous problem is undoubtedly a bit crude. This example shows the primary new behaviour which can occur when the effect of more axles is taken into account. Of course, things worked out so neatly because the vehicle was symmetric, the springs identical etc. but the bottom line is, whatever the parameters, the oscillations always get split between the translational and tilting modes.

You might object that an infinitely long sinusoidal road is unrealistic – what happens if the bus encounters a single bump in an otherwise level road ? Basically what happens is an algebraic mess with no new conceptual subtleties. You can model the bump as some function $w = f(y)$ between $y = 0$ and $y = L$ (where L is the length of the bump) which merges smoothly into the flat road on both sides [for example, $w = \sin^2(\pi y / L)$], and calculate w_1 and w_2 and their derivatives. Then you plug them into the ODEs for z_s and z_d and solve these ODEs while imposing continuity of position and velocity at every instant when the excitations change character. This calculation is straightforward but tedious and, since this not a book on vehicular dynamics, not worth going into here. A similar problem which makes a different point however is coming up next.

Example 3.3.5

Unsprung mass

Referring to 3.3.3, let the unicycle now model an axle of a railway locomotive. Here the mass m represents the axle and other bogie components, and it is connected by spring (L, k) and damper (C) to the mass M which represents the loco body (more accurately, $1/n$ of the loco body if there are n axles). m is called the unsprung mass and M the sprung mass. A surface roughness in the track can be modelled as kink as shown in the Figure : $w = by$ from $y = 0$ to $y = l$, $w = b(2l–y)$ from $y = l$ to $y = 2l$ and $w = 0$ everywhere else. The dimensions of the kink are grossly exaggerated in Figure. The loco approaches the kink from the negative y (right) side with constant velocity v. Assume that the mass m follows the contours of the kink throughout.



(a) Write an ODE describing the motion of M as the loco rides over the kink.

An impulsive force is defined as a 'hammer blow' i.e. a force which has the form $F = p\,\delta(t)$ where $\delta(\ldots)$ is the Dirac delta function from 1.3.

(b) Determine the impulsive forces exerted on the track by the loco at the kink.

The locomotive WAP5 has an axle load ($M+m$) of 19.5 tons and an unsprung mass per axle (m) of 2.6 tons. WAP7 has axle load 20.5 tons and unsprung mass per axle 3.9 tons.

(c) If track wear and tear were the only factor in selecting the loco link for a high speed train, would you choose WAP5 or WAP7 ? Justify your answer.

[Level : 3]

(a) The ODE has the same form as in the last two problems,

$$M\ddot{z} = -k\left(z - w - L\right) - C\left(\dot{z} - \dot{w}\right) - Mg \quad \text{, or}$$

$$M\ddot{z} + C\dot{z} + kz = kL - Mg + kw + C\dot{w} \quad .$$

Here w is defined in terms of y; since the velocity v is constant, we can write $y = vt$, which sets $t = 0$ as the time when the loco just hits the kink. Then we have

$$w = \begin{cases} bvt & \text{if } 0 < t < \dfrac{l}{v} \\ b\left(2L - vt\right) & \text{if } \dfrac{l}{v} \le t \le \dfrac{2l}{v} \\ 0 & \text{otherwise} \end{cases} \quad .$$

The derivative $\mathrm{d}w/\mathrm{d}t$ evaluates to

$$\dot{w} = \begin{cases} bv & \text{if } 0 < t < \dfrac{l}{v} \\ -bv & \text{if } \dfrac{l}{v} \le t \le \dfrac{2l}{v} \\ 0 & \text{otherwise} \end{cases} \quad .$$

Plugging these into the original equation gives the ODE asked for.

(b) We start by noting carefully what the problem does and does not require. YES : impulsive forces on the track. NO : trajectory of the sprung mass M as a function of time. An impulsive force means $F = p\,\delta(t)$, or $\mu\ddot{z} = p\delta(t)$, where μ is mass (typically m but that's already doing duty here). Integrating both sides of this equation across an infinitesimal interval straddling the delta function gives $\dot{z}\big|_{0+} - \dot{z}\big|_{0-} = p/\mu$ i.e. there is a discontinuous change in velocity at the instant of the impulse.

Now consider the first part of the kink i.e. the transition from the straight track to the up-sloping segment. A careful analysis here will make the subsequent parts of the motion obvious. We have already defined $t = 0$ to be the instant when the loco hits the transition. While the loco is approaching the kink, the big mass M is in equilibrium with the compression in the spring balancing gravity. Letting the reference level $z = 0$ be this equilibrium position, we can say that z

and $\mathrm{d}z/\mathrm{d}t$ equal zero at all times prior to the arrival of the kink, and hence at $t = 0^-$.

Now from $t = 0$ onwards, the big mass satisfies the ODE

$$M\ddot{z} + C\dot{z} + kz = kbvt + Cbv \quad , \qquad (01)$$

for the initial phase of the motion. This is a standard LCC inhomogeneous ODE and its solution features nice continuous functions for x and $\mathrm{d}x/\mathrm{d}t$, so taking another derivative will not throw up an impulsive force. So for $t > 0$ there are no impulsive forces on M. But what about the initial time $t = 0$? There also, things are tractable because there is no delta function on the RHS. Before $t = 0$, the forcing was zero; after $t = 0$, the forcing is finite – integrate (01) across an infinitesimal interval spanning zero and there is no impulsive change in velocity. At $t = 0$, a finite forcing was turned on – that alone is not sufficient to produce impulsive response in the system. Hence we can say that M is not subjected to impulsive forces during the transition from the straight to the kink.

But now consider the motion of m. Its vertical velocity is zero just before $t = 0$ while it is bv just after $t = 0$; clearly there is an impulsive force on it at $t = 0$. The impulse cannot have come from gravity. And we just showed that the spring exerts no impulsive force on M, therefore it cannot be exerting one on m either. Hence the only source of this impulse is the track. Now from Sir ISAAC's Third, m exerts the equal and opposite impulsive force on the track. The impulse on m is clearly $mbv\,\delta(t)$ so the impulse on the track must be $-mbv\,\delta(t)$.

We can readily extend this logic to find the other impulses. There will be one at the top of the kink and a third at the end of the kink, both places where m changes velocity abruptly. The spring shields M from impulse each time so it does not contribute. The top of the kink is reached at $t = l/v$ while the end is reached at $t = 2l/v$. At the top, the change in vertical velocity of m is from $+bv$ to $-bv$ while at the end it is $-bv$ to zero. Combining these, we have the impulsive force exerted by the loco on the track

$$* \qquad F = mbv\left[-\delta(t) + 2\delta\left(t - \frac{l}{v}\right) - \delta\left(t - \frac{2l}{v}\right)\right] \quad , \qquad (02)*$$

which is the desired answer.

(c) This last part is an application of the previous result. To minimize track wear and tear, we must of course try to subject it to as small impulsive forces as possible. Equation (02) clearly says that the size of the impulsive forces at a roughness on the track is proportional to the unsprung mass alone. The total mass does not even enter the picture. Since WAP5 has 33 percent lower unsprung mass than WAP7, and we need not bother about performance etc. we should clearly use WAP5 rather than WAP7 for the desired haulage.

*

COMMENTS : Like every model, this too has a lot of simplification. The shape of the kink is one, and the constraint that the wheel follow the kink precisely is another. But whatever the shape of the roughness, it will always cause sudden changes in vertical velocity of the wheel and then the impulses received on the track will be proportional to unsprung mass alone. For this reason, unsprung mass per axle is one of the very important parameters to be considered when allotting a loco link for a train.

Example 3.3.6

Lollipop LCR circuit

A series LCR circuit is excited with a voltage source which provides sinusoidal voltage of constant amplitude and frequency ω. For very large ω, the size of the response current goes as

A. Independent of ω

B. $1/\omega$

C. $1/\omega^2$

D. $1/\omega^4$

[Level : 1]

The correct answer is B.

If you are an electrical engineer then of course you will know this off the top of your head. Else, you might be tempted to go for choice C because of the asymptotic properties of $G(\omega)$. Note however that the LCR equation is like the damped driven harmonic oscillator but the RHS features $\mathrm{d}V/\mathrm{d}t$ and not just V. So the harmonic oscillator structure brings the gain function $G(\omega)$ which goes as $1/\omega^2$ for large ω, but the derivative on V throws in one power of ω for a resultant response of $1/\omega$.

This example is primarily intended as a break from mechanical vibrations.
Break over, here they come again.

Example 3.3.7

Vibration isolation

Two masses m_1 and m_2 are coupled using a spring (L, k) and damper (C) and an external force $F\cos\omega t$ acts on m_1. Let x_1 and x_2 ($x_2 > x_1$) be the coordinates of the two masses.

(a) Write a pair of coupled ODEs for x_1 and x_2 as functions of time.

(b) Assuming the damping to be zero, determine the value of k for which the particular solution of the system features m_1 not vibrating at all.

[Level : 2]

(a) Writing the force balance as we did in all the examples so far, we have

$$m_1\ddot{x}_1 = -k\left(x_1 - x_2 + L\right) - C\left(\dot{x}_1 - \dot{x}_2\right) + F\cos\omega t \quad , \qquad *$$

$$m_2\ddot{x}_2 = -k\left(x_2 - x_1 - L\right) - C\left(\dot{x}_2 - \dot{x}_1\right) \quad , \qquad *$$

which is the coupled ODE system asked for. As quick checks on such ODEs always note that (i) the extension of the spring is $x_2 - x_1 - L$ so this combination must appear as a unit, with either positive or negative sign (terms like $x_1 - x_2 - L$ are not allowed), and (ii) the ODE for each x_i generally features the corresponding variable and its derivative with negative sign on the RHS so that the homogeneous solutions are oscillatory and damped.

(b) In the absence of damping the system simplifies to

$$m_1\ddot{x}_1 = -k\left(x_1 - x_2 + L\right) + F\cos\omega t \quad ,$$

$$m_2\ddot{x}_2 = -k\left(x_2 - x_1 - L\right) \quad .$$

Since we are interested in the vibrations of the system, we can further drop the L terms from both equations.

To find the particular solution we use undet coeff; since the system features only even order derivatives and pure cos excitation, the trial functions should include cos only and not sin. Thus the trials are

$$x_1 = A\cos\omega t$$

$$x_2 = B\cos\omega t \quad .$$

Plugging these into the system (with L thrown away) gives

$$\left(k - \omega^2 m_1\right)A - kB = F$$

$$-kA + \left(k - \omega^2 m_2\right)B = 0 \quad ,$$

a pair of simultaneous equations for A and B.

These can be solved in any number of ways; here we use KRAMER's Rule. Application of the Rule gives

$$\begin{bmatrix} A \\ B \end{bmatrix} = \frac{1}{\left(k - \omega^2 m_1\right)\left(k - \omega^2 m_2\right) - k^2} \begin{bmatrix} k - \omega^2 m_2 & k \\ k & k - \omega^2 m_1 \end{bmatrix} \begin{bmatrix} F \\ 0 \end{bmatrix} \quad .$$

For m_1 to not vibrate, A must be zero. Clearly, $A=0$ if $\underline{k = \omega^2 m_2}$. Under this condition, the denominator is non-zero so the solution is valid, hence this is the desired answer.

*

COMMENTS : The way this result is used practically is as follows. Suppose you have a machine with a lot of vibrating components, one of which you really want to shield from vibration. As happens with most machines, suppose that your one has a more or less fixed speed of operation, which translates to a more or less constant frequency of vibration of the component which requires protection. This component is m_1 of the model; indeed it is being acted on by an oscillatory force, as a vibrating component in a machine would be. Then, to achieve the shielding, you connect to it an extra mass m_2 via a spring k with the values chosen such that $k - \omega^2 m_2 = 0$. The vibration disappears even though m_1 is still being forced !

To be more accurate, the vibration does not disappear altogether, but diminishes greatly in amplitude. The exact zero was a consequence of taking the damping to be zero. In reality, damping always exists, and a bit of it is needed to kill off the homogeneous solutions. But with the damping in place, the amplitude comes out to something of order γ/ω in place of zero. If γ is small (and there is no need to make it large), we still have a great reduction over order unity. This is surely an amazing result if there ever was one.

The calculation with the damping present has been done in the classic text on vibrations by JACOB DEN HARTOG. I didn't present it here because the algebra is very cumbersome, and the undamped analysis is sufficient to capture the physics behind vibration isolation.

Example 3.3.8

Beats and frequency measurement

A device to measure the vibration frequency of a rapidly vibrating machine consists of a spring (k) and an adjustable mass (m). The spring is perfectly undamped (achieved using a controller). The mass is adjusted to bring the natural frequency close to that of the machine, and then the device is hooked up to the machine. The equation of motion of the device is

$$m\ddot{x} + kx = F\cos\omega t \quad ,$$

where ω is the machine frequency which is to be determined. The experimenter measures the beats frequency.

In a particular apparatus, $m = 1$ kg and $k = 400$ N/s and the beats frequency is measured to be 0.25 rad/s. The mass is then reduced by 2 percent and the new beats frequency is 0.15 rad/s.

(a) Determine the frequency ω coming from the machine.

Since the measuring apparatus is undamped, the homogeneous solutions are not killed off in time and play a significant role. For an accurate measurement, it is desirable that all motions other than the beating one should be as small as possible. This happens if the oscillation amplitudes of the homogeneous and particular solutions are nearly identical.

(b) Show that the above holds true if the initial position and initial velocity of the mass m satisfy a certain relation which is dependent on the amplitude and frequency of the excitation.

(c) One particular initial condition is optimal irrespective of the excitation parameters. Determine that one.

[Level : 2]

(a) The beats formula says that the beating frequency is half of the difference between the natural and excitation frequencies. For the first measurement, the natural frequency is $\sqrt{(k/m)}$ which is 20 rad/s. Since the beat frequency is 0.25 rad/s and double that is 0.5 rad/s, the machine can be outputting either 19.5 or 20.5 rad/s. There is nothing in the formula to predict which way the defect lies – it can be either positive or negative. Also, in exams, please be very careful of the factor of half in the beat formula – forgetting it is a common but completely avoidable way of throwing the bulk of your marks into the water.

The second measurement resolves the ambiguity. Reducing the mass by 2 percent increases the natural frequency. Since this causes the beat frequency to reduce, the natural frequency must earlier have been LESS than the excitation frequency. This picks out the value of 20.5 rad/s for the excitation.

If you are not satisfied with the primarily qualitative treatment of the last para, you can check that reducing the mass by 2 percent increases natural frequency by 1 percent, so new ω_0 is 20.2, beats frequency is 0.15, double that is 0.3 so excitation can either be 20.5 or 19.9; the first value is consistent with the first measurement and thus is the actual answer.

(b) As an aside, let's see how the question arrived at the condition that the amplitudes of the homogeneous and particular solutions must be identical for a good beats phenomenon. The beats occur as a result of interference between the particular and homogeneous solutions, arising from the trig formula

$$\cos C + \cos D = 2\cos\left(\frac{C-D}{2}\right)\cos\left(\frac{C+D}{2}\right) \quad ,$$

with $C=\omega_0 t$ and $D=\omega t$.

However, $\cos \omega_0 t$ and $\cos \omega t$ are certainly not the most general form of the homogeneous and particular solutions. There can be a phase shift and a difference in amplitude between the two components. The phase part is easy – $\cos C + \cos(D+\varphi)$ only throws in extra φ terms into the product but does not change its basic character of a fast vibration modulated by a slow oscillation (which is visible to the naked eye). Different amplitude however cannot be spirited away like this – $10\cos C + \cos D$ is the beating motion plus $9\cos C$ which swamps it out and results in a lousy measurement. Hence to produce a good

measurement, we need to ensure that the amplitudes of the homogeneous and particular solutions are identical.

The particular solution of $\ddot{x}+\omega_0^2 x = F\cos\omega t$ we have already calculated in 3.1.5 Part (a); it is

$$x_p = \frac{F}{\omega_0^2-\omega^2}\cos\omega t \quad .$$

Letting $A=F/(\omega_0^2-\omega^2)$, the beats will come out best if and only if the homogeneous solution has the same amplitude A i.e.

$$x_h = A\cos\left(\omega_0 t+\varphi\right) \quad ,$$

where φ, following the phase shifting logic given above, can be arbitrary.

Under these conditions the general solution is

$$x_G = A\cos\left(\omega_0 t+\varphi\right)+A\cos\omega t \quad .$$

The time derivative is

$$\dot{x}_G = -\omega A\sin\omega t-\omega_0 A\sin\left(\omega_0 t+\varphi\right) .$$

The ICs x_0 and v_0 are of course x_G and its derivative at $t=0$; setting this into the two expressions above gives

* $$x_0 = A\left(1+\cos\varphi\right) \quad ,$$ *

* $$v_0 = -\omega_0 A\sin\varphi \quad ,$$ *

which expresses the relation between x_0 and v_0 in a parametric form, in terms of the parameter φ. This relation is evidently dependent on A – if x_0 is given, then the value of φ (provided it exists) gets determined in terms of A and then this φ goes into fixing v_0. Since A is dependent on the properties of the excitation, which we do not know before starting the experiment, this is not a pleasant situation as it appears to imply that we cannot control whether we will have good beats and hence an accurate measurement.

(c) Inspection reveals the IC which holds for all A. If we choose $\varphi=\pi$ then we get $x_0 = 0$ and $v_0 = 0$ to be a good IC, irrespective of the value of A. Hence the required IC is that $\underline{x_0 = 0 \text{ and } v_0 = 0}$. In other words, the measuring device should be quiescent when the excitation is turned on. Now that's a simple IC all right ! *****

COMMENTS : A common use of beats is in tuning pianos – the tuner strikes a tuning fork tuned to a particular frequency and adjusts the tension in the wire to try and achieve equal pitch between the note and the fork. The tuner senses differences in pitch by way of beats – this phenomenon brings down the enormously high frequencies (400 Hz or so) down to a few Hz which is on an everyday human scale. The slower the beats, the better the tuning. Although those beats make use of the same underlying trig as this one, modeling a tuning fork as a driven harmonic oscillator would not be physically quite sound, so I desisted from posing a problem on it. One of the factors limiting the accuracy of the tuning is in fact the damping of the sound of both the fork and piano.

A second limitation is the tuner's ability to ensure an exactly same volume of sound from the fork and the piano, the relevance of which becomes clear from the analysis of Part (b). This detailed calculation might appear like beating the phenomenon to death, but I included it because it is not at all an obvious conclusion that perfect beats will occur whenever there is a slight difference in frequencies – the initial conditions do indeed play a very important role. This fact doesn't get appropriate attention in Literature because of the universal nature of the IC which always does the job. But there is no harm in seeing that this IC in fact arises as a special (and extremely fortunate) case of a far more general class of ICs which can produce pure beats unalloyed with other extraneous motions.

Now another break from vibrations; this time we dive into the microscopic scale.

Example 3.3.9

Rabi Oscillations

This problem is taken from "Introduction to Quantum Mechanics" by DAVID GRIFFITHS. The solution follows the solution manual of the said book.

A qubit is a quantum mechanical system which can be in either of two states, called UP and DOWN. The general wave function of the system is a linear superposition of the two states, c_1<UP>+c_2<DOWN> where c_1 and c_2 are complex quantities constrained by the relation $|c_1|^2+|c_2|^2=1$. When an oscillating potential of strength V is turned on at frequency ω, c_1 and c_2 satisfy the ODE system

$$\dot{c}_1=-\frac{\mathrm{j}V}{2\hbar}\mathrm{e}^{\mathrm{j}(\omega-\omega_0)t}c_2$$
$$\dot{c}_2=-\frac{\mathrm{j}V}{2\hbar}\mathrm{e}^{-\mathrm{j}(\omega-\omega_0)t}c_1 \quad ,$$

where ω_0 is a characteristic frequency of the system.

(a) Eliminate c_1 from the given system to write an LCCH ODE for c_2.

(b) Use exponential form of solution to obtain the general solution. Express your answer in terms of the Rabi frequency, defined as $\omega_R=\left[\left(\omega_0-\omega\right)^2+V^2/\hbar^2\right]^{1/2}$.

(c) If the IC is $c_1(0)=1$ and $c_2(0)=0$, obtain the specific solution.

(d) Determine $|c_2|^2$ as a function of t.

[Level : 2]

(a) To eliminate c_1 we differentiate the second equation and substitute in the first. Performing the differentiation we have

$$\ddot{c}_2 = -\frac{\mathrm{j}V}{2\hbar}\left[\mathrm{j}(\omega_0 - \omega)\mathrm{e}^{\mathrm{j}(\omega_0-\omega)t}\, c_1 + \mathrm{e}^{\mathrm{j}(\omega_0-\omega)t}\, \dot{c}_1\right] \quad .$$

Now substitute the second equation of the given system in place of c_1 and the first equation in place of $\dot{c}_1$ to get

$$\begin{aligned}\ddot{c}_2 &= \mathrm{j}(\omega_0 - \omega)\left[-\frac{\mathrm{j}V}{2\hbar}\mathrm{e}^{\mathrm{j}(\omega_0-\omega)t}\, c_1\right] - \frac{\mathrm{j}V}{2\hbar}\mathrm{e}^{\mathrm{j}(\omega_0-\omega)t}\left[-\frac{\mathrm{j}V}{2\hbar}\mathrm{e}^{-\mathrm{j}(\omega_0-\omega)t}\, c_2\right] \\ &= \mathrm{j}(\omega_0 - \omega)\dot{c}_2 - \frac{V^2}{4\hbar^2}c_2\end{aligned} \quad .$$

Rearranging terms we have

$$\ddot{c}_2 - \mathrm{j}(\omega_0 - \omega)\dot{c}_2 + \frac{V^2}{4\hbar^2}c_2 = 0 \quad , \qquad *$$

which is the LCCH ODE asked for.

(b) Using exponential form, the characteristic equation is

$$\lambda^2 - \mathrm{j}(\omega_0 - \omega)\lambda + \frac{V^2}{4\hbar^2} = 0 \quad ,$$

and SRIDHARACHARYA's formula gives the two roots as

$$\mathrm{j}\frac{(\omega_0 - \omega)}{2} \pm \mathrm{j}\omega_R \quad ,$$

where ω_R is the Rabi frequency given in the question. The general solution is the complex linear combination of these two fundamental solutions i.e.

$$c_{2,G} = \exp \mathrm{j}\frac{\omega_0 - \omega}{2}t\left(A\mathrm{e}^{\mathrm{j}(\omega_R)t} + B\mathrm{e}^{-\mathrm{j}(\omega_R)t}\right) \quad , \qquad *$$

which is the desired answer. We note that since the ODE is complex, A and B are arbitrary complex constants which are not bound by any conjugation constraint as they are for real ODEs.

(c) More algebra here, this time I will skip the details. After slogging it, what we end up with is

$$c_2 = -\frac{\mathrm{j}V}{2\hbar\omega_R}\left(\exp \mathrm{j}\frac{\omega_0 - \omega}{2}t\right)\sin\omega_R t \quad , \qquad *$$

$$c_1 = \exp \mathrm{j}\frac{\omega_0 - \omega}{2} t \left[\cos \omega_R t + \mathrm{j}\frac{\omega_0 - \omega}{2\omega_R} \sin \omega_R t \right] \quad , \qquad (*)$$

which is the desired answer.

(d) Taking the mod squared we have

$$|c_2|^2 = \frac{V^2}{4\hbar^2 \omega_R^2} \sin^2 \omega_R t \quad , \qquad (*)$$

which is the desired answer.

*

COMMENTS : This problem has become a bit of a rote exercise unlike the other problems in this Unit. It happened because the full physics going into the ODE would be out of scope for this book. The given ODE is almost a direct statement of the SCHROEDINGER equation; there is only one non-trivial step where one has to approximate an oscillatory potential $\cos \omega t$ as $\mathrm{e}^{\mathrm{j}\omega t}/2$ while dropping the $\mathrm{e}^{-\mathrm{j}\omega t}/2$ part on physical grounds. Nevertheless, this example shows you how second order LCCH ODEs can arise in a context which is as far removed from classical mechanics and springs as it could possibly be. The interpretation of c_1 and c_2 is that, at any time t, $|c_1(t)|^2$ is the probability of finding the system in the state <UP> while $|c_2|^2$ is the probability of finding the system in <DOWN>. Note that although $|c_2|^2$ starts from zero, it becomes equal to unity at $t = \pi/2\omega_R$, indicating that the system has performed a transition from the UP to the DOWN state. Also, as happened with ABEL's theorem and similar results, this derivation is now a trivial exercise but when the oscillations were first discovered, it was very significant; so much so that this discovery and its ramifications earned ISIDOR (or ISRAEL as he was born – his school changed it to ISIDOR for him) ISAAC RABI the Nobel Prize in Physics.

Example 3.3.10

Resonance or no resonance

This question gives you five situations, each of which features a large-amplitude oscillation. For each one, determine whether the oscillation is being driven by a linear resonance effect i.e. $\ddot{x}+\gamma\dot{x}+\omega_0^2 x = F\cos\omega t$ type of equation with small γ and ω close to ω_0, or by some other effect (which you do not need to explain). You may focus both on the nature of the equation and on the characteristics of linear resonance to help make your choices. First write your choice (YES or NO); then explain your answer as completely and cogently as possible.

(a) An aircraft is starting its engine at the parking bay and the rpm of the fan is increasing slowly. At a particular point during this acceleration, the tray tables attached to the seats suddenly start vibrating and making a clattering noise, before again subsiding to quiescence as the engine attains the idle speed of 500 rpm.

(b) The same aircraft is now on the runway, spooling up from 30 to 100 percent rpm for the takeoff run. As the engine crosses 80 percent, the volume of sound emitted by it increases tremendously, and remains high as it stabilizes at full throttle.

(c) A child sitting on a swing is pumping her legs up and down to increase the amplitude of her oscillations. The system can be modelled as pendulum having periodically time-varying length. At a particular frequency of pumping, the amplitude of the child's oscillation increases significantly.

(d) Ten pendula of different lengths are hanging from a beam. An oscillatory force of a given frequency and lying in the horizontal plane is applied to each of them. One pendulum out of the ten starts swinging with a large amplitude.

(e) Ten clothes of different shapes and sizes are hanging from a line and being dried by the breeze from a fan running at constant speed

of 1900 rpm. Among them, a shirt is flopping around violently in a to and fro motion.

[Level : 2]

(a) YES. The tray table can indeed be modelled as an oscillator, with the table being connected to the seat back using spring and damper (the clamp). The rotating fan is likely to transmit some oscillatory force to the seats through the wings and fuselage, and this force will have the same frequency as that of the fan. These two facts suggest that the damped driven harmonic oscillator equation is valid in this case. Further confirmation comes from the fact that the oscillation amplitude is given to increase sharply at a particular value of engine rpm (i.e. excitation frequency), and fall away on either side, which is exactly what happens in a linear resonance situation. Finally, since the idle speed is given as 500 rpm, the suspected resonance occurs at say 200 or 300 rpm, i.e. an excitation frequency of about 3-5 Hz. If we push the tray table relative to the seat back and release it, we would expect to see it oscillate at approximately this frequency (as against say one period in 5 seconds or 50 periods per second), showing that we have a match between excitation and natural frequencies and making the case for linear resonance water tight.

(b) NO. For a linear resonance we need a match between two frequencies. One frequency here is of course the rotation of the fan, which we did in the previous example. However, there is no evident second frequency involved in the system. The air which transmits the sound does not have a natural frequency unless we confine it inside a flute or something – the airport apron is not a confined space. So a linear resonance between the jet and the air is unlikely. Secondly, the noise is given to be high at all rpms above 80 percent, so we do not see the narrow peak which is a fundamental characteristic of a linear resonance. (OPTIONAL) The actual phenomenon which is occurring here is that at 80 percent the tips of the fan become supersonic and then start producing 'sonic booms', which, despite all the design features to mitigate their effect, are still extremely loud.

(c) NO. The problem actually tells us the model we need to write. In 2.4.1 we obtained the equation of a pendulum . If the length is a periodic function of time then the equation is

$$\ddot{\theta} + \frac{g}{l_0 + \Delta l \cos \omega t} \sin \theta = 0 \quad ,$$

where l_0 is the basic length of the pendulum and Δl the time-varying part on account of the child's pumping her legs. This certainly does not look like the

linear resonance equation. But it is nonlinear, so let's linearize assuming θ close to zero and Δl is much smaller than l_0. This gives

$$\ddot{\theta} + \frac{g}{l_0}\left(1 - \frac{\Delta l}{l_0}\cos\omega t\right)\theta = 0 \quad .$$

This ODE is linear but not constant coefficient; at the same time it is homogeneous not inhomogeneous. Therefore the increase in amplitude occurring here is not a case of linear resonance in a forced damped harmonic oscillator.

(d) YES. Once again a pendulum situation; since they are all hanging near vertical, the linearization $\sin\theta \rightarrow \theta$ is appropriate so we indeed have a harmonic oscillator LHS. An oscillatory horizontal force on the bob is clearly an outside excitation on the RHS so we get the full equation of the linear oscillator, and a likely case for linear resonance. Moreover, ten pendula of different lengths mean ten different natural frequencies; since only one is responding enthusiastically to the external excitation, we have the frequency selectivity inherent in the phenomenon and the defence walks home with the trophy.

(e) NO. Although a shirt is somewhat like a pendulum and we can plausibly have an oscillator LHS with a natural frequency, where is the other frequency ? Air flowing past the shirt at a constant rate does not have a periodic component. Even a convoluted and contrived reasoning that the rotation of the fan imparts a periodic component to the airflow is rendered void, if not by the fact that such effects smoothe out over the distance from the fan to the clothe, by the given information that the fan rpm is 1900. So the frequency of the periodic components in the airflow is 1900 rpm or 30+ Hz and its multiples, which are orders of magnitude faster than the rate at which the shirt is flopping about. Hence we do not have a linear resonance here. (OPTIONAL) The phenomenon taking place here is aeroelastic flutter, which is not a linear resonance, but positive feedback in which the oscillator experiences a force proportional to the velocity and PARALLEL to it in sign.

*

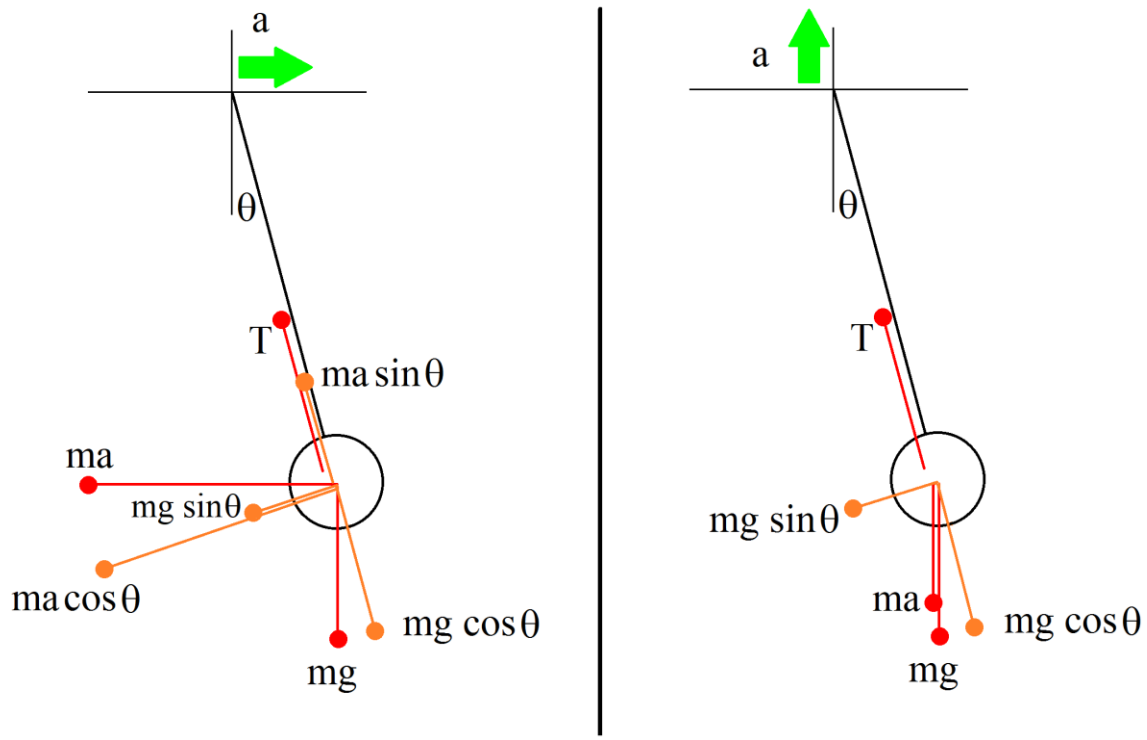
COMMENTS : There are some comments regarding almost every part of this question, so here they go.

(a) The phenomenon of vibrating tray tables, armrests etc. indeed occurs on aircraft during starting, especially on large aircraft where the big engines produce a lot of sound and vibration during startup.

(c) The equation of the child on the swing is a Mathieu equation, which we've seen on a couple of earlier occasions. The amplitude increase is called a PARAMETRIC resonance because the excitation multiplies the dependent variable rather than appearing independently. The biggest parametric resonance occurs when $\omega = 2(g/l_0)^{1/2}$ i.e. if the excitation frequency is twice the natural frequency; smaller ones occur at $\omega = (g/l_0)^{1/2}$ and other values also. Unlike the 'weak resonance' of the externally forced oscillator, where the growth is linear, parametric resonance is a 'strong resonance' as the growth (when the resonance occurs) is exponential. Why this is so, we'll see (at least approximately) in the next Chapter.

(d) This is relevant only if you are interested in classical mechanics and/or studying for a mechanics exam. Here we have not mentioned how the force is applied on the pendulum bobs. One way of doing it is to move the beam back and forth in a horizontal vibration. To see how this leads to the driven harmonic oscillator equation, we consider a more general case where the support of the pendulum is externally given an acceleration $\mathbf{a}$ in the vertical plane. We'll analyse the motion from a reference frame moving with the support point. In this frame, every body will feel an extra non-inertial force equal to its mass times $-\mathbf{a}$. One such body is the pendulum bob. We now consider the free body diagram of the bob, as in 2.4.1, but with the extra force added on. The case of horizontal force is shown in the left panel of the Figure. Here I have assumed that the pendulum support is given an acceleration a directed towards the right.

(Another ill-timed page break)



The radial force balance is trivial since the motion is constrained in that direction; the tangential balance yields

$$ml\ddot{\theta} = -mg\sin\theta - ma\cos\theta \quad .$$

Linearizing about θ=0, $\sin\theta$ becomes θ while $\cos\theta$ becomes 1; if the acceleration a is oscillatory i.e. $a = a_0\cos\omega t$, we get the driven harmonic oscillator equation.

The right panel of the Figure shows the case where the acceleration **a** of the support point is vertically upwards instead of horizontal rightwards. A trivial change, you think. Think again. In this case, the non-inertial force adds onto gravity, and the total vertically downward force is $m(g+a)$. Now, we can carry through the steps of 2.4.1 verbatim; the inevitable result is

$$ml\ddot{\theta} = -m(g+a)\sin\theta \quad .$$

Take $a = a_0\cos\omega t$ and linearize about θ=0; we have

$$\ddot{\theta} + \frac{1}{l}(g + a_0\cos\omega t)\theta = 0 \quad .$$

This is the Mathieu equation from Part (c) !

This fundamental difference between a pendulum with horizontally and vertically accelerated bob is rather surprising, and almost impossible to guess if blindsided with it in an exam. So I will strongly encourage you to save the bottom line to your memory; if you know the end product you can derive the rest easily.

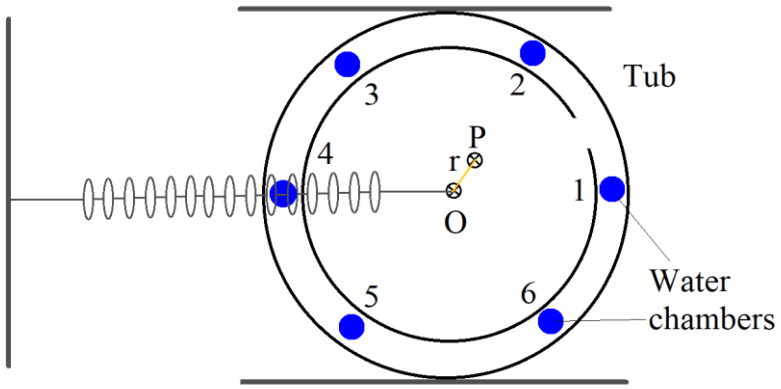
(e) Aeroelastic flutter is an example of the negatively damped oscillator $\ddot{x} - \gamma\dot{x} + \omega^2 x = 0$. In a later Chapter I have included a basic model which shows how the negative damping crops up. Like the parametric case, the growth here is exponential and not linear. It is a dangerous phenomenon, having been responsible for the 1944 failure of Tacoma Narrows bridge, and, combined with a parametric resonance, for repeated crashes of the shoddily designed Lockheed Electra aircraft. (It is extremely unfortunate that those of us who thought that air crashes due to design defects were history were found to be living in a fool's paradise.)

Indeed, whenever there is a case of large amplitude oscillations leading to a structural failure, you can be sure that the 'wolf' exponential growth is involved rather than the 'sheep' linear (see 3.1.3). Linear resonance can always be killed off by adding sufficient damping (the amplitude blows up for γ identically zero and is large only when γ is small). Structures generally tend to have a lot of inherent damping in the first place (a bridge often vibrates when a heavy truck passes – does it continue oscillating after the truck is gone ?), so they are extremely unlikely to reach failing amplitude just from a linear excitation. But parametric resonance, aeroelastic flutter and other mechanisms which can excite exponential growth can do so even in the presence of damping, so that becomes a redundant factor and the system spirals out of control until it fails.

Example 3.3.11

Washing machine

The tub of a front loading washing machine, loaded with clothes, is shown in front view in the Figure below. It consists of a drum of mass m with centre O, which rotates about O with constant angular velocity ω. Due to eccentric distribution of clothes inside the tub, the CM of the tub is not at O, but at P, which is fixed in the frame of the tub and is a distance r away from O. The tub is constrained to remain immobile in the y-direction by smooth walls and in the x-direction it is connected to the wall (the machine frame) via a spring and damper (representing the bearing). The water chambers (fixed relative to the tub) are empty and can be ignored for now.



It can be shown that the displacement x of O, measured from equilibrium, satisfies the ODE

$$\ddot{x} + \gamma\dot{x} + \omega_0^2 x = \omega^2 r \cos\omega t \quad .$$

In a particular washing machine, the spin cycle is being commenced with the water chambers empty. It is known that the damping constant is $\gamma = 1$ rad/s. As the tub accelerates slowly, a maximum vibration amplitude equal to 0.03 m (3 cm) is observed at a rotational speed of 10 rad/s (96 rpm).

(a) Determine the approximate amplitude of vibration at the rated operating speed of 120 rad/s (1150 rpm).

(b) The machine tries to reduce the amplitude by filling the water chambers with water. If O and P and the labelled chambers are as shown in the Figure, which chamber should be filled first ? Justify your answer.

[Level : 2]

EXTRA :

(c) Derive the ODE given in the question.

(a) The structure of the governing ODE is given on a platter, but some information is irritatingly withheld. In particular, we are told γ directly but not given values for ω_0 and r, hence we must use the information provided in the problem to deduce these closely guarded secrets. Slow acceleration of tub means that at every instant, the steady state response formula can be assumed to hold true. The first thing is that maximum response amplitude occurs at $\omega = 10$ rad/s. Suppose the oscillator were undamped. Then, maximum response at frequency 10 would imply that the natural frequency of the oscillator is 10 – although the excitation has an ω^2 term, the gain is infinite at $\omega = 10$ and finite everywhere else. Now, the actual damping is 1, which is 1/10 of the hypothetical undamped natural frequency, or 5 percent of critical. Since this implies LIGHT damping, we can continue to treat the oscillator as undamped, and write its natural frequency as 10. This is all the more permissible since the question asks for an approximate amplitude. This gives us the equation

$$\ddot{x} + \dot{x} + 100x = \omega^2 r \cos \omega t \quad ,$$

and part of the missing information has been unearthed.

There is a second tidbit thrown our way – the maximum vibration amplitude is 0.03 m. At the resonance, the function $G(\omega)$ has the value $1/\gamma\omega$. So this tells us that

$$\frac{10^2 r}{1 \cdot 10} = 0.03 \quad ,$$

or $r = 0.003$. Now we have a full equation

$$\ddot{x} + \dot{x} + 100x = 0.003\omega^2 \cos \omega t \quad ,$$

and the problem has reduced to a routine calculation.

At $\omega = 120$ rad/s, the response amplitude is

$$A = G(120) \cdot 120^2 \cdot 0.003 \quad . \qquad (01)$$

But $\omega = 120$ corresponds to a very large frequency, bigger than all other players in town. So we can use the asymptotic formula $G(120) = 1/120^2$. Plugging this into (02) gives $A = 0.003$ m i.e. the amplitude of oscillation of the tub is 3 mm at the rated operating speed. This is 1/10 of the maximum amplitude !

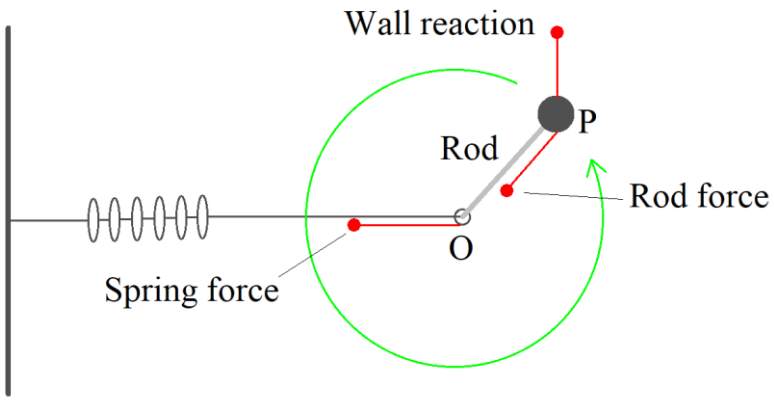
(b) Here is the time to put theory to practice. The vibration of the tub is occurring because of the eccentricity so to reduce the amplitude, we must reduce the

eccentricity. This will happen if we load the tub in the direction opposite to the radius vector OP. Water chamber 5 is located in this direction so that must be the first one to get filled with water.

*

COMMENTS : In Part (a) we exploited the word "approximate" present in the question to solve it using physical argument as opposed to rigorous mathematical analysis. The latter technique would have featured writing the given ODE with unknown ω_0 and r, solving for the position and height of the maximum response, and then imposing that they equal 10 and 0.03 respectively. This would have become a tedious exercise in algebra. It turns out that the error involved is negligible – with our chosen values of ω_0 and r, the actual position and height of the maximum turn out to be 10.01 and 0.03003 respectively ! In many situations, the approximation of throwing the damping from a lightly damped driven harmonic oscillator produces results which are better than we would normally expect – another example is the upcoming one.

(c) Now let's do the extra question. This is extra because it's more mechanics than differential equations. The forces on the tub are the spring force in the x-direction and the normal reactions from the walls in the y-direction (no x-force from walls since they are smooth). Given these, we now replace the tub by a point mass m at the CM of the tub, connected to O by a massless rod. The free body diagram of the modified setup is shown below.



The acceleration of P relative to the lab frame equals the acceleration of O relative to the lab plus the acceleration of P relative to O. The first is $\ddot{x}\hat{\mathbf{x}}$ by definition while the second is $-\omega^2 r\hat{\mathbf{r}}$, where $\mathbf{r}$ is the vector from O to P. Hence

the acceleration of P is $\ddot{x}\hat{\mathbf{x}} - \omega^2 r\hat{\mathbf{r}}$. Now of course we have to do the force balance. The y-direction balancing is not of interest since the motion is constrained – the reactions enforce the constraint and that's it. Because the rod is massless, the x-force it exerts on the mass P must be exactly equal to what it gets from the spring, F_{sp}. Then, taking the x-component of the acceleration, we have

$$m\left(\ddot{x} - \omega^2 r\cos\theta\right) = F_{sp} \quad , \tag{02}$$

where θ is the angle between OP and the horizontal. Since the rotation speed is constant, $\theta=\omega t$. The spring force F_{sp} evidently has the standard form $F_{sp} = -kx - C\dot{x}$ (recall that the spring is damped); putting all this into (02) gives

$$m\ddot{x} + C\dot{x} + kx = m\omega^2 r\cos\omega t \quad .$$

Dividing through by m leads to the ODE given in the question.

To see this problem in action, observe a reasonably modern front loading washing machine during the spin cycle at the end of the wash. Almost all front loaders with a minimum level of sophistication have tubs with water chambers and sensors which measure vibration. During the spin cycle, you will often see the tub accelerating and decelerating/stopping in the low speed regime many times before launching into the final run which takes it to the maximum permissible speed. Every time it slows/stops, it measures the vibration at low speed, performs a calculation just like Part (a) to determine whether the vibration at rated speed will be within tolerable levels, and if not, empties and fills the water chambers to reduce the eccentricity. Only when sufficiently well balanced does it embark on the acceleration run to the maximum. Sometimes you can clearly see with the naked eye that the vibration level at the same speed is reducing after several bursts of acceleration and deceleration. The peaking of the amplitude at low speed and the subsequent reduction as the tub really takes off are of course visible clearly.

Finally, here is a chance to explain why front loading washing machines generally have higher spin speeds (1000-1400 rpm) than top loaders (usually 600 rpm max). In a front loader, the rotation is in the vertical plane so an eccentricity results in fluctuating torque and speed on account of gravity; this fluctuation can be measured by the sensors to determine the offset. Once located, the offset can of course be corrected, like we saw here. In top loading machine on the

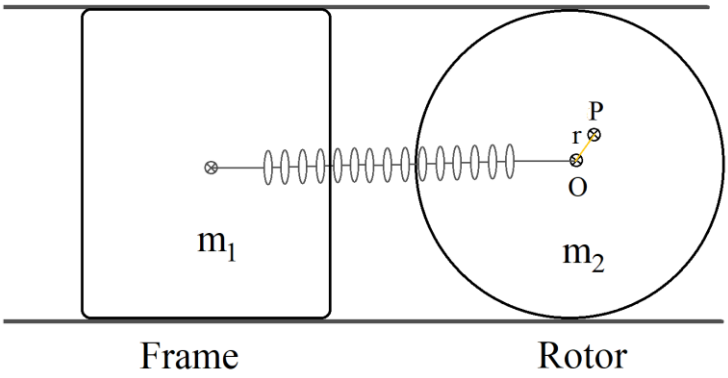
other hand, the tub rotates in the horizontal plane, eccentricity cannot be detected very accurately, so vibration reduction is also inaccurate; the end result is that the maximum permissible speed is limited to avoid excessive stresses on the bearings.

Example 3.3.12

Sommerfeld effect

Easier form : A version of this question with the mass m_1 *replaced by a wall.*

An eccentric rotor of mass m_2 and eccentricity radius r is connected to a frame of mass m_1 via spring (L, k) and damper (C) as shown in the Figure below. The speed of rotation is constant ω.



(a) Write a pair of coupled ODEs for the positions x_1 and x_2 of the two masses as functions of time.

For the subsequent parts, you might want to use a computer to solve the relevant equations. If so, use the values $m_1 = 50$ kg, $m_2 = 10$ kg, $k = 1000$ N/m, $C = 10$ kg/s and $r = 0.06$ m. If not, plug these values into an analytical formula. Recall that power of a force is the (dot) product of the force and the velocity of the point of application of the force.

(b) Plot the steady state time-averaged power developed by the applied force i.e. the external excitation, as a function of ω.

The device is being driven by a motor which has a finite and constant power output P_0. The power required for rotating the rotor is $0.02\omega^2$ W while that for performing the vibrations was obtained by you above. If at any value of ω the motor has power left over after driving the rotation and vibration, then the reserve (or suitable part thereof) goes into increasing the rotation speed.

Assume that the acceleration is slow enough for the spring to be in steady state oscillation at all times.

(c) If the rotor starts from rest, determine the ultimate operating speed of the machine if the power of the motor is P_0=100 W, and $P_0 = 200$ W.

(d) Explain why the above is an unexpected result for anyone who is not familiar with the Sommerfeld effect from beforehand.

[Level : 3]

(a) We put the answer together by a mix-and-match of the solutions of previous examples . 3.3.7 gives the equation of motion of a nearly identical system , the only difference being that the excitation there is applied on m_1 rather than m_2. The form of the excitation was determined in the last example only – it is $-m_2\omega^2 r\cos\omega t$. Putting all this together we have the two equations

$$m_1\ddot{x}_1 = -k\left(x_1 - x_2 + L\right) - C\left(\dot{x}_1 - \dot{x}_2\right) \quad , \qquad (01a)^*$$

$$m_2\ddot{x}_2 = -k\left(x_2 - x_1 - L\right) - C\left(\dot{x}_2 - \dot{x}_1\right) + m_2\omega^2 r\cos\omega t \quad , \qquad (01b)^*$$

which is the ODE system asked for.

(b) To find the power, we need the velocity of the point of application of the excitation i.e. of the mass m_2. To get the velocity we need the position, i.e. the solution of the above system. Since damping is present and the steady state power is asked for, the homogeneous solutions can be neglected from consideration. To find the particular solution we use undetermined coefficients (no surprise here). To focus on the oscillatory solutions we throw out the kL terms from the RHS. Because the system is damped, the homogeneous solutions are decaying and there cannot be a coincidence with the purely oscillatory excitation function – so no bump-up factors. Since there are first derivatives also in the ODEs, the trial cannot be cosines only as in 3.3.7 but must have both cos and sin. Accordingly, we choose the trial function

$$x_1 = A_1\cos\omega t + B_1\sin\omega t$$
$$x_2 = A_2\cos\omega t + B_2\sin\omega t \quad .$$

Because the manipulations are longish, we will use the C and S shorthand for cos and sin; this interferes with the C of the damper so we rename that to D. Substituting this into the system (01) gives

$$\begin{aligned} &m_1\left(-\omega^2 A_1\,\mathrm{C} - \omega^2 B_1\,\mathrm{S}\right) + D\left(-\omega A_1\,\mathrm{S} + \omega B_1\,\mathrm{C}\right) + k\left(A_1\,\mathrm{C} + B_1\,\mathrm{S}\right) \\ &-D\left(-\omega A_2\,\mathrm{S} + \omega B_2\,\mathrm{C}\right) - k\left(A_2\,\mathrm{C} + B_2\,\mathrm{S}\right) = 0 \quad , \\ &m_2\left(-\omega^2 A_2\,\mathrm{C} - \omega^2 B_3\,\mathrm{S}\right) + D\left(-\omega A_2\,\mathrm{S} + \omega B_2\,\mathrm{C}\right) + k\left(A_2\,\mathrm{C} + B_2\,\mathrm{S}\right) \\ &-D\left(-\omega A_1\,\mathrm{S} + \omega B_1\,\mathrm{C}\right) - k\left(A_1\,\mathrm{C} + B_1\,\mathrm{S}\right) = m_2\omega^2 r\,\mathrm{C} \quad . \end{aligned}$$

Equating C and S terms each to each gives four linear simultaneous equations for A_1, A_2, B_1 and B_2, which can be written in a matrix form as

$$\begin{bmatrix} k-\omega^2 m_1 & -k & D & -D \\ -D & D & k-\omega^2 m_1 & -k \\ -k & k-\omega^2 m_2 & -D & D \\ D & -D & -k & k-\omega^2 m_2 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ m_2\omega^2 r \\ 0 \end{bmatrix} .$$

Note that this matrix has several 'symmetries' in the arrangement of its various terms – such symmetries often arise when a longish calculation is going correctly and act as an unofficial test of its correctness.

Since the question allows it, we will solve the system by computer rather than by hand. But merely solving it won't lead to the marks allotted for this part. To get the prize we must calculate power of the force. Now power is force times velocity of the application point, which is x_2 here; since $x_2 = A_2\cos\omega t + B_2\sin\omega t$, the velocity is the derivative $v_2 = -\omega A_2\sin\omega t + \omega B_2\cos\omega t$, and the power is

$$P = \left(m_2\omega^2 F\cos\omega t\right)\left(-\omega A_2\sin\omega t + \omega B_2\cos\omega t\right) .$$

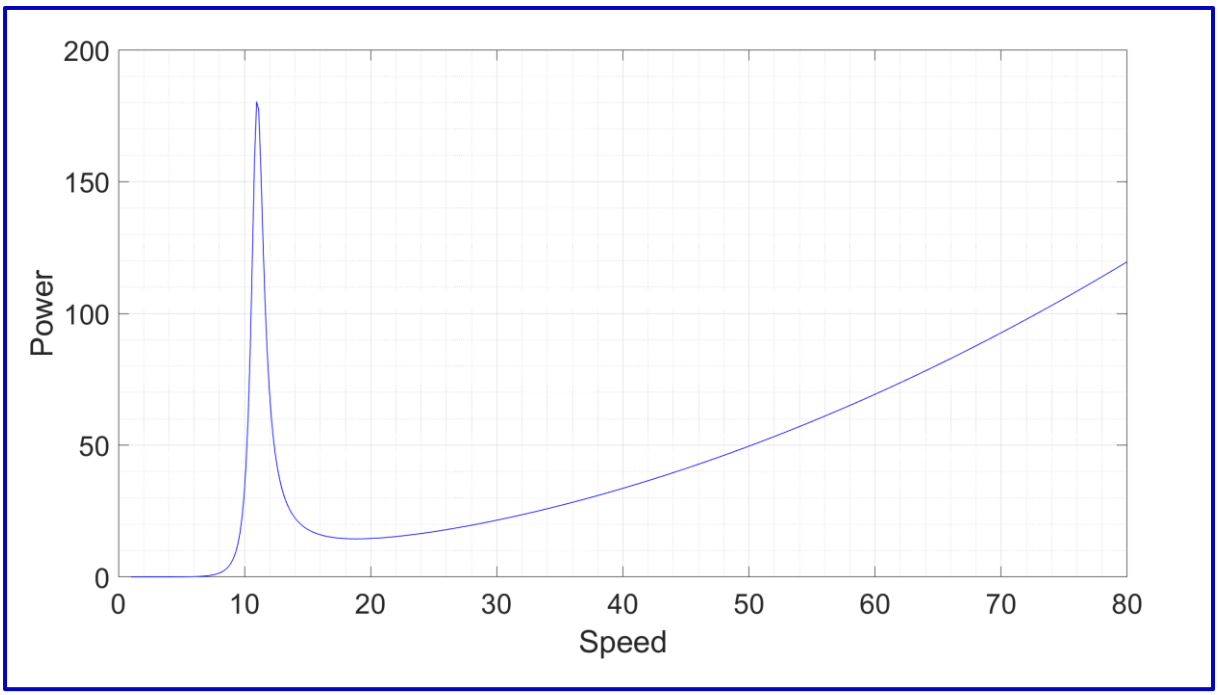
This however is the instantaneous power – the question asks for the time average. Some of you will know that the average of $\cos^2$ is 1/2 while that of cos.sin is zero. If you do, then skip to the end of the page; else read on. Since the functions are periodic with period $T = 2\pi/\omega$, the most natural time to average over is one period. Letting τ denote some arbitrary time, this average is

$$\begin{aligned} P_{\text{avg}} &= \frac{1}{T}\int_{\tau}^{\tau+T} m_2\omega^3 r\left(-A_2\cos\omega t\sin\omega t + B_2\cos^2\omega t\right)\mathrm{d}t \\ &= \frac{m_2\omega^3 r}{T}\int_{\tau}^{\tau+T}\left[-\frac{A_2}{2}\sin 2\omega t + \frac{B_2}{2}\left(1+\cos 2\omega t\right)\right]\mathrm{d}t \\ &= \frac{m_2\omega^3 r}{T}\left[\frac{A_2}{4}\cos 2\omega t + \frac{B_2}{2}t + \frac{B_2}{4}\sin 2\omega t\right]_{\tau}^{\tau+T} \end{aligned} .$$

But now, $\cos 2\omega t$ and $\sin 2\omega t$ are periodic with period π/ω which is $T/2$. Hence their values of course repeat at intervals of T, i.e. they are equal at $t = \tau$ and $t = \tau + T$ for ANY τ. Therefore the trig terms in the last line are both zero and the only term which survives is $(B_2/2)t$. Plugging in the limits,

$$P_{\text{avg}} = \frac{m_2\omega^3 F B_2}{2} ,$$

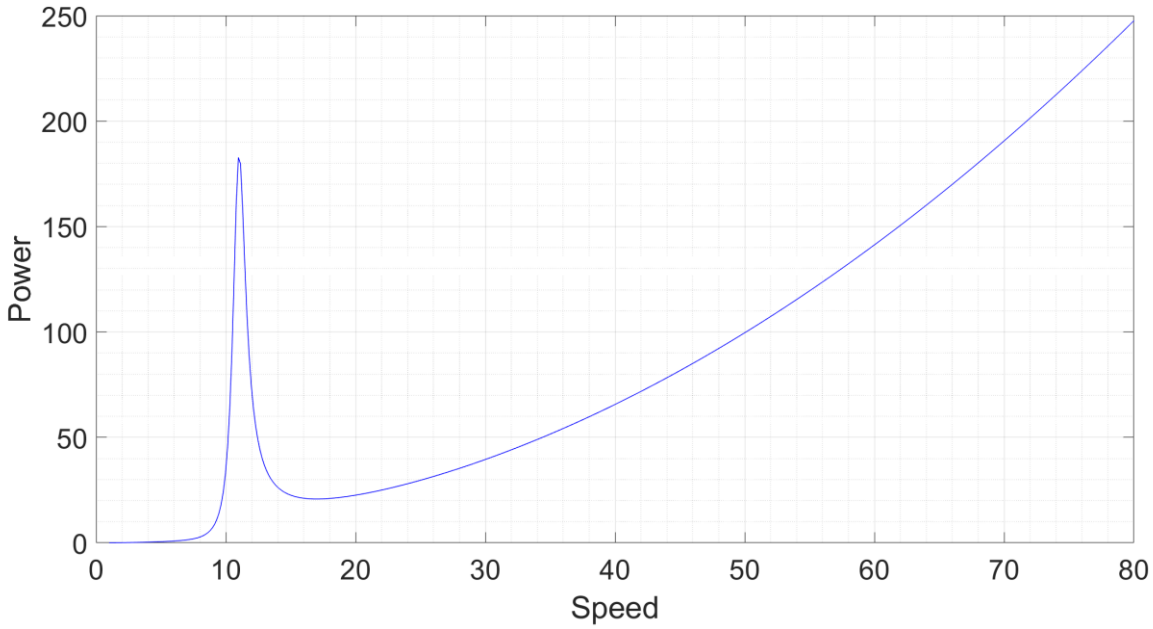
and now we can substitute the numerically determined B_2 to find P (now I'll let go of the subscript avg) as a function of ω. The numerical values given in the problem are substituted and B_2 and hence P determined using the computer. The plot of $P(\omega)$ is given in the Figure below.



Here, speed denotes ω in rad/s and power is in Watts. The noteworthy feature of this plot is of course the enormous spike at approximately $\omega = 10$.

(c) For this part, we add the power of rotation $0.02\omega^2$ to the vibration power and plot the result again. This is shown in the Figure below.

(Yet another unavoidable page break !)



Now we interpret this Figure. At zero speed, the power required is of course zero so the motor has the entire power P_0 as reserve and the rotor starts to accelerate. With $P_0 = 100$ W, the system reaches the level for the first time at ω just above 10 rad/s. To further increase speed, it would have needed more power which it does not have, so it settles into operation at this speed, which is approximately 100 rpm. With $P_0 = 200$ W, the motor has a reserve of about 20 W even at the top of the spike. So it rides clean over the spike and continues to accelerate. The motor runs out of reserve power only at ω just above 70 rad/s. So the speed of operation is approximately 700 rpm.

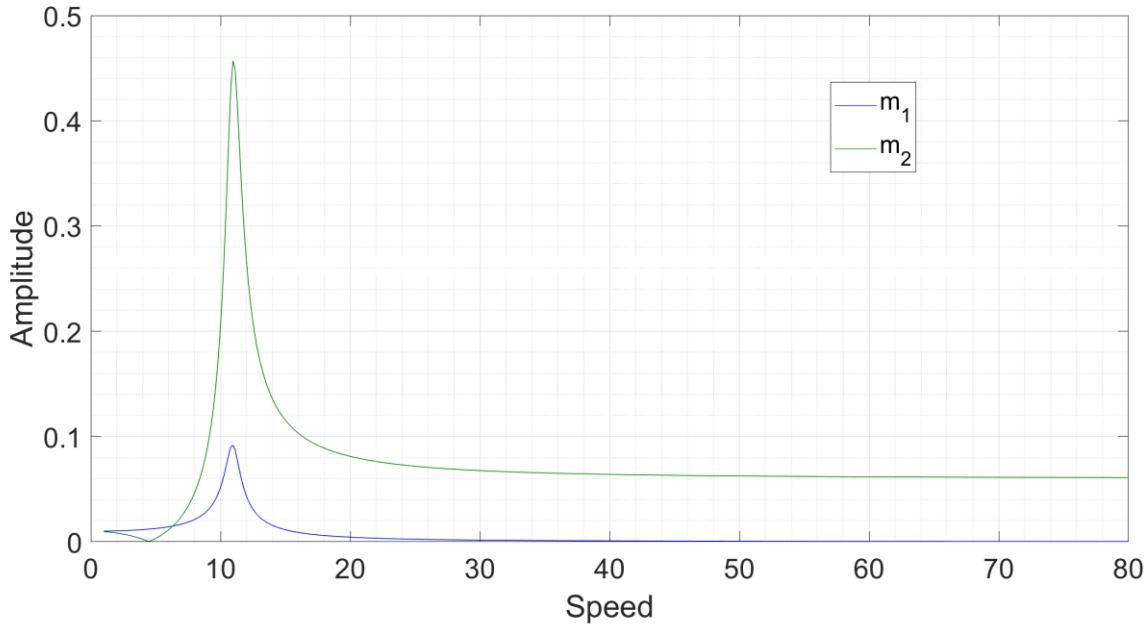
(d) This is a surprising result (to one who doesn't know it already) because we would hardly expect the speed to increase by a factor of seven for a mere doubling of the input power. Since the rotation power is given to vary as ω^2, we would think that the rotation speed will increase by 1.4 times if the power is doubled. However, because of the huge spike in the vibration power, the actual increase in speed is five times of the expected value !

*

COMMENTS : This is the simplest possible model of the phenomenon called Sommerfeld effect (after ARNOLD SOMMERFELD – an excellent physicist as well as a supreme teacher who has written a book on The Partial Differential Equations of Mathematical Physics) or resonant capture. What happens in this

case is that the motor tries to accelerate the rotor to rated speed but falls into the trap and ends up spending all its energy in driving large amplitude vibrations. The trap is nothing but linear resonance – if we drop the inhomogeneous terms from the ODE system and set $C = 0$ then a normal mode analysis yields the natural frequency to be $\omega_0 = \sqrt{120}$, which is almost 11 rad/s. And the spike indeed occurs at 11 rad/s.

To better understand the phenomenon, I will now show a plot of the amplitudes of vibration of the two masses as functions of ω.



Both amplitudes increase sharply at $\omega = 11$; in particular that of m_2 (the rotor itself) increases drastically. The dynamics of the frame m_1 also is quite interesting – it has a medium response at very low frequencies, rises to a maximum at the resonance and then falls off. This transfer of energy from the rotor to the frame at the resonance is also one of the salient features of the Sommerfeld effect. It was in fact to show this feature that I set up the model this way rather than going for the even simpler case where the rotor was connected to a wall, as in the last Tute. While the calculation for that system would have been really easy, it would have failed to show the transfer of vibrational energy at the resonance to other bodies in the system. As an aside, we also see the nearly zero amplitude of m_2 at ω just greater than 4 – this is the vibration isolation we saw in 3.3.7. As per our prediction there, it should occur at $\omega^2 = 1000/50$, which indeed it does; the amplitude also is very close to zero even

with 5 percent damping ($\gamma = \omega_0/10$) included. This is another demonstration of the accuracy of the results which we can obtain even after removing the damping from a lightly damped driven oscillator system.

Resonance capture occurs in mechanical as well as astronomical systems and is a phenomenon of considerable research interest. A practical demonstrator was the top loading washing machine in the laundry room of my rented house – due to an intrinsic eccentricity in the tub, the spin cycle always ended up at large amplitude vibrations of tub and body at a speed of about 150 rpm (the defect has since been repaired). Resonance capture is usually modelled using nonlinear equations but it seems that this linear system with its analysis suitable for our course is sufficient to get to the heart of the physics behind the phenomenon.

* * * * *

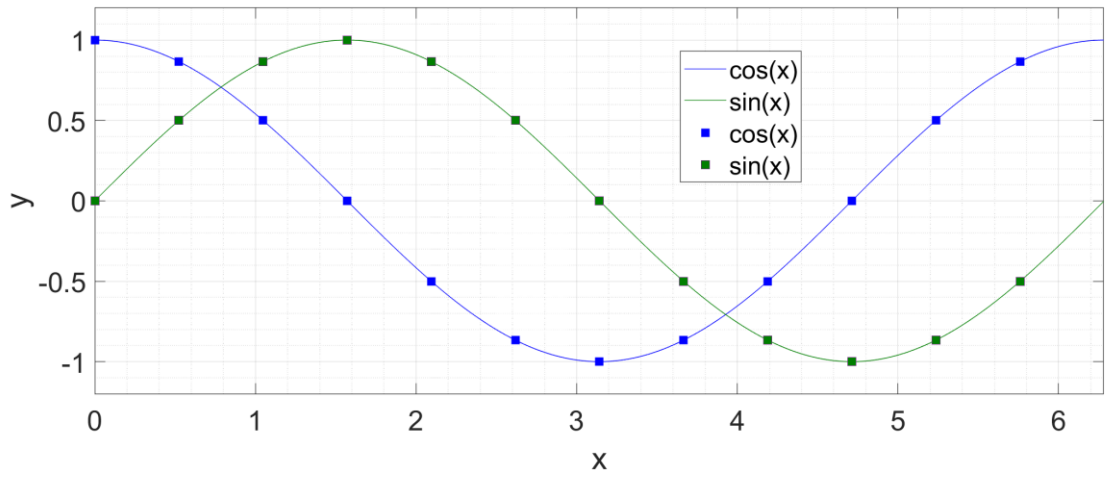
This page intentionally left blank

Review

TRIGONOMETRIC IDENTITIES

This is a review of the trigonometric (and hyperbolic) functions – definitions, basic properties and identities. The topics in Chapter 4 need a lot of trig because the solutions of boundary value problems are often trig functions and Fourier series are serieses of trigs. This material might appear irrelevant or trivial, but many times I have seen large penalties accruing from simple and avoidable trig errors, both in midsems and in the Final.

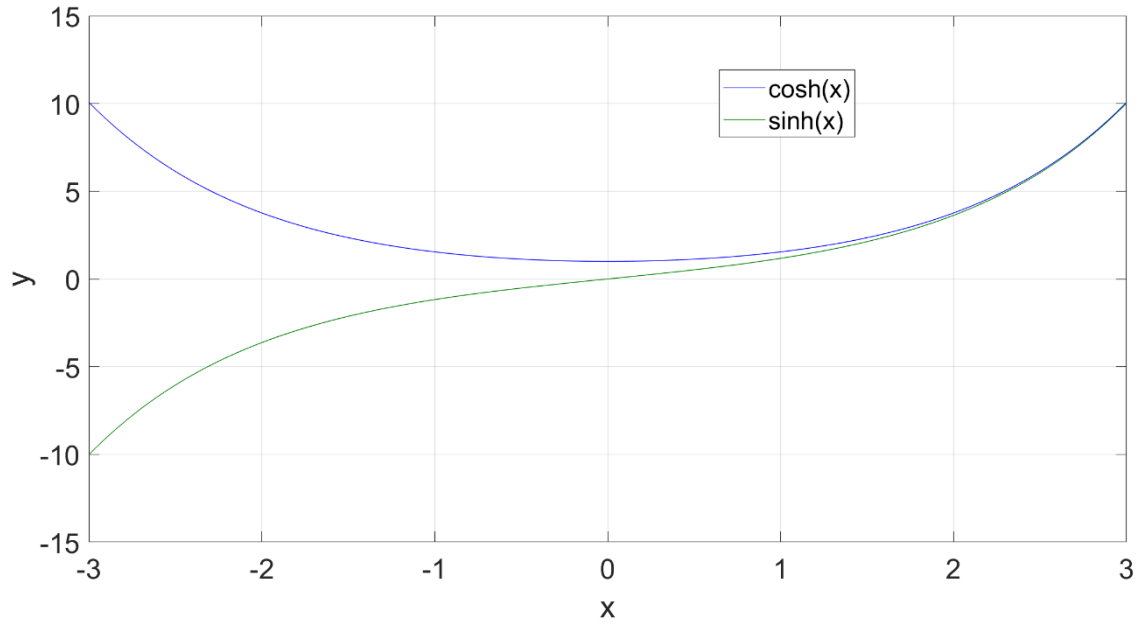
Here are the graphs of $\cos x$ and $\sin x$, shown in the interval x goes from 0 to 2π.



These functions are of course periodic with period 2π which is 360 degrees. Remember that π radians is 180°; we'll use the radian measures here because that is mathematically sound, but to generate values you can always think in terms of the degree equivalent if that is more convenient (for example, you might be unfamiliar with sin $\pi/2$ but might be fully aware that sin 90° is 1). Here is a table of the typical values of cos and sin. The special values below are also highlighted on the graph with large square markers.

Radian angle	Degree angle	cos	sin
0	0	1	0
$\pi/6$	30	$\sqrt{3}/2$	$1/2$
$\pi/3$	60	$1/2$	$\sqrt{3}/2$
$\pi/2$	90	0	1
$2\pi/3$	120	$-1/2$	$\sqrt{3}/2$
$5\pi/6$	150	$-\sqrt{3}/2$	$1/2$
π	180	-1	0
$7\pi/6$	210	$-\sqrt{3}/2$	$-1/2$
$4\pi/3$	240	$-1/2$	$-\sqrt{3}/2$
$3\pi/2$	270	0	-1
$5\pi/3$	300	$1/2$	$-\sqrt{3}/2$
$11\pi/6$	330	$\sqrt{3}/2$	$-1/2$

The hyperbolic functions are not periodic, and their graphs are shown below.



At large positive x, both go like e^x and the two curves overlap for all practical purposes. At large negative x, cosh is again positive exponential while sinh is negative exponential. The only special values you need to know are $\cosh 0 = 1$ and $\sinh 0 = 0$. That's the only zero which sinh has, while cosh has no zero at all, as you can see.

The trigs satisfy lots of identities, all of which you can find on suitable websites, but the main ones I am reproducing below for the sake of completeness.

* $$\sin^2\theta+\cos^2\theta=1$$ **(01)***

* $$\sin(a+b)=\sin a\cos b+\cos a\sin b$$ **(02)***

* $$\sin(a-b)=\sin a\cos b-\cos a\sin b$$ **(03)***

* $$\cos(a+b)=\cos a\cos b-\sin a\sin b$$ **(04)***

* $$\cos(a-b)=\cos a\cos b+\sin a\sin b$$ **(05)***

$$\sin 2\theta=2\sin\theta\cos\theta$$ (06)

$$\cos 2\theta=2\cos^2\theta-1$$ (07)

$$\sin C+\sin D=2\sin\frac{C+D}{2}\cos\frac{C-D}{2}$$ (08)

$$\sin C-\sin D=2\cos\frac{C+D}{2}\sin\frac{C-D}{2}$$ (09)

$$\cos C+\cos D=2\cos\frac{C+D}{2}\cos\frac{C-D}{2}$$ (10)

$$\cos C-\cos D=2\sin\frac{C+D}{2}\sin\frac{D-C}{2}$$ (11)

$$\sin C\sin D=\frac{1}{2}\left[-\cos(C+D)+\cos(C-D)\right]$$ (12)

$$\sin C\cos D=\frac{1}{2}\left[\sin(C+D)+\sin(C-D)\right]$$ (13)

$$\cos C\sin D=\frac{1}{2}\left[\sin(C+D)-\sin(C-D)\right]$$ (14)

$$\cos C\cos D=\frac{1}{2}\left[\cos(C+D)+\cos(C-D)\right]$$ (15)

Long list.

Successful IIT-JEE candidates often have all these and more (3θ, $\theta/2$ etc.) at their fingertips. But suppose you aren't a maggu (I myself am not – whenever I need trigs for my work, I either refer to SOSMath.com or worse, make mistakes) then what ? Perhaps the examiner will provide the necessary trig on the question paper or allow a formula sheet in the exam hall. What if that's not the case ?

Then my advice is, **memorize the first five** of these and then derive everything else based on those. For example, to find $\sin C+\sin D=?$, use (02)

and (03) with $a+b=C$ and $a-b=D$. That tells you, $a=(C+D)/2$, $b=(C-D)/2$ and then adding the RHSes gives (08). But if you don't know something within the first five also, then there can be trouble because deriving them in the exam hall is not particularly easy.

*

4

BOUNDARY VALUE PROBLEMS, FOURIER SERIES AND PERIODIC EXCITATION

This Chapter introduces boundary value problems and explores the connections between linear ODEs and matrix equations. An introduction to Fourier series is given, and the method of matched boundary conditions is presented for solving dynamical systems with periodic excitation.

4.1

BVPS, DISCRETE DERIVATIVES, DOT PRODUCTS AND THE MATRIX METHOD

This Unit, the one with the longest title in the book, is a lead-up to PDEs like 2.3. So if you just want a quick ODE review before going on to PDEs, this is one of the essential components. In the latter half (4.1.4 onwards), it becomes a bit heavy on linear algebra so you can skip those parts if your course doesn't require this aspect at all.

On the face of it, a boundary value problem (BVP) is little different from an initial value problem (IVP) – it's just an ODE (for us, i.e. throughout this book, linear) with the accompanying information being specified at two (or more) **different** values of the independent variable instead of everything at the same value of the independent variable. These specifications are called **boundary conditions (BC)**. The problem is generally understood to be confined within these boundaries. Just as with IVPs, the number of boundary conditions needed to be specified is the same as the order of the ODE. All these conditions can involve the value of the function or of a derivative of order strictly less than that of the ODE. Since BVP typically arises with the independent variable being space, we will usually denote the target function as $y(x)$ rather than $x(t)$.

So, a BVP looks like say $y''+y=0$, $y(0)=y_0$, $y(1)=y_1$ whereas an IVP looks like $\ddot{x}+x=0$, $x(0)=x_0$, $\dot{x}(0)=v_0$. Big deal, might be your initial reaction.

But it turns out, it is a really big deal. Let me just quote the results before going into theory. A LCC IVP we know always has a unique solution – given x_0 and v_0 we can always crank the machinery of Chapters 1 to 3 and find the solution, whether the problem is homogeneous or inhomogeneous. For BVP on the other hand things are not so straightforward.

If the BVP is homogeneous (the RHS of the ODE **and** of all BCs are zero), it can EITHER have only the trivial solution $y = 0$ OR infinitely many non-trivial solutions. If the BVP is inhomogeneous, the result depends on whether the homogeneous problem had only the trivial solution or non-trivial ones. In the former case, the inhomogeneous problem always has one unique solution. In the latter case, the problem has EITHER infinitely many solutions, OR no solution at all. These results are parts of **FREDHOLM's alternative theorem**, named after IVAR FREDHOLM.

Some BVPs come with a free parameter λ in the equation, say $y''+\lambda y = 0$, $y(0) = y(1) = 0$. Such a problem can be written as $\mathbf{L}_\lambda y = f(x)$. For homogeneous problems i.e. $f(x)=0$ and all BCs are zero, it often happens that for some special values of λ the problem admits non-trivial solutions while for all other λ's it only has the boring old $y = 0$. The special values of λ are called **eigenvalues** of the problem, and the corresponding non-trivial solutions the **eigenfunctions**. From this terminology, BVPs with an unknown λ in them are called eigenvalue problems.

The choice of word for the special value is surely a surprise. In Chapter 2, eigenvalue seemed to mean something completely different – some special value for a matrix. Now the same word for a special value in a differential equation ??!! To a greater or lesser extent, the pedagogical literature generally adopts the view that the word does double duty. But that's as far from reality as one can imagine. A linear differential equation is just an infinite matrix equation and eigenvalue means the same thing in both situations.

DISCRETE DERIVATIVE OPERATORS AND DOT PRODUCT

To see the connection between matrix algebra and differential equations, we start from a simple definition, which at its inception was probably the most epoch-making in all of maths and science and has since reduced to a trifle – the definition of derivative. ISAAC NEWTON and GOTTFRIED LEIBNITZ (and I won't go into the issue of who first) defined this as : $\mathrm{d}y/\mathrm{d}x$ is the limit of $[y(x+h)-y(x)]/h$ as h tends to zero. Here, we go backwards – starting from the continuous problem, we discretize it. Suppose we have a BVP on the interval [0, 1]. Then, we partion the continuous space into a number N of discrete points, each separated by equal distance h. In this discrete framework, we label the lattice points by the (integer) index i; then

the continuous function $y(x)$ becomes a discrete set y_i (y_1, y_2, y_3…..). An array of entries y_1, y_2 etc. suggests a **vector** – indeed, continuous functions are the limit of discretized vectors and share the same properties as vectors.

The discrete derivative at the i^{th} site is

$$y'_i = \frac{y_{i+1} - y_i}{h} \quad . \qquad \textbf{(01R)*}$$

We can equivalently write it as

$$y'_i = \frac{y_i - y_{i-1}}{h} \quad . \qquad \textbf{(01L)*}$$

These two forms are called the right and left side derivatives. They are equivalent in meaning, but the situation at hand generally forces us to use either one or the other, as appropriate. To get the discrete second derivative, we take one more discrete derivative of y'. To keep things centred on y_i, we either take a left side derivative on (01R) or a right side derivative on (01L). The result either way is

$$y''_i = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} \quad , \qquad \textbf{(02)*}$$

and this does not have right and left side forms.

The discrete derivative operators are **matrices** which act on the discretized function-vectors y. Thus, discretization transforms the functional equation $\mathbf{L}y = f(x)$ to the matrix equation $\mathbf{Ly} = \mathbf{f}$. No wonder that linear ODEs and matrix equations have extremely similar properties. The details of the discretization, and the matrix method of solving linear BVPs, we'll leave for the Tutes.

The vector space in which functions live is called **Hilbert space** (after DAVID HILBERT). While we won't be using this jargon much, you'll see it all the time in quantum mechanics books. This space consists of all sufficiently well-behaved functions (including a few 'rogues' like DIRAC's delta). The most important thing about Hilbert space is that it permits a dot product. So what is the definition of dot product of two functions ?

To get the answer, we think in terms of their discretized vectors. Suppose we have an array of N lattice sites in the interval x goes from 0 to 1, and suppose we have two functions f and g defined in this interval. The dot product of their discretized vectors **f** and **g** is of course the sum $f_1 g_1 + f_2 g_2 + \ldots\ldots + f_N g_N$. There's

only one problem with taking the limit $N \to \infty$ directly – as N increases, the number of terms in this sum, and hence its size, also increase. To keep the size tractable, we must multiply the sum by $1/N$. But, $1/N$ is precisely h, the spacing between successive lattice sites (or maybe h is $1/(N+1)$ or $1/(N-1)$ or something – it doesn't matter in the limit). Now, $h(f_1 g_1 + f_2 g_2 + \ldots + f_N g_N)$ in the limit $h \to 0$ and $N \to \infty$ is the integral $\int_0^1 f(x)g(x)\,dx$ evaluated by the rectangle method. Thus we have the following extremely important definition : **the dot product of two functions over an interval is the integral of the product of the functions evaluated over the interval.** Mathematically, for two functions $f(x)$ and $g(x)$ defined in the interval $[a, b]$,

* $$"f(x)\cdot g(x)" = \int_a^b f(x)g(x)\,dx \quad . \qquad \textbf{(03)*}$$

Two functions are **orthogonal** over an interval if their dot product is zero. The phenomenal utility of the concepts of dot product and orthogonality won't be immediately apparent, but it will become clear in the next Chapter onwards. So, now only, make sure to understand these things well.

A modification to (03) can arise if the function-vectors come about from discretization in a non-Cartesian geometry – in such a case, the dot product takes the form $"f\cdot g" = \int_a^b f(x)g(x)w(x)\,dx$, where $w(x)$ is called the weighting function. For example, the Bessel functions, which arise in polar coordinates, need a weighting function of x. Dot products featuring any weighting function other than the trivial value (unity) are outside the scope of this book (yaaay!).

EULER'S METHOD

While we're at discrete derivatives, let me at least motivate why an IVP always has a unique solution while BVP can have such variety in its solution structure. The motivation comes from the simplest numerical method used to solve ODEs – EULER's method, which is basically direct integration of the ODE. Suppose we have the simple ODE $y''+y=0$. First we write it in coupled first order form as $y_1 = y$, $y_2 = y'$ (note that subscript does not mean lattice site

in this paragraph) so that $y_1' = y_2$, $y_2' = -y_1$. Let's say we have ICs defined at $x = 0$. Then the discrete derivative implies that $y_1(0+h) = y_1(0) + hy_1'(0)$ and $y_2(0+h) = y_2(0) + hy_2'(0)$. Next we plug in the ODE to eliminate the derivatives, $y_1(h) = y_1(0) + hy_2(0)$, $y_2(h) = y_2(0) - hy_1(0)$. But $y_1(0)$ and $y_2(0)$ are just the ICs which by definition are known. Thus, using the ICs we have generated y_1 and y_2 (and hence y and y') at $x = h$. Continuing this process, we can generate y and y' at $2h$, $3h$ etc. and hence at any arbitrary value of x. The entire process uses bootstrapping from the initial values; the ICs are given, there is no indeterminacy anywhere, and the solution is unique. The example of harmonic oscillator equation was the simplest possible, but the argument we used was quite general, and unless you are a hardcore mathematician you should be reasonably convinced that a well-defined IVP always has a unique solution. (After reading 4.1.4 you will object that EULER's method doesn't even work for the harmonic oscillator equation. Here we are interested in the concepts rather than the mechanics of the method – if you aren't satisfied you can assume that we work only at such small x and h values where the numerical solution remains accurate.)

With BVP, things are not so straightforward. Let's say we have the BCs $y(0) = 0$ and $y(1) = 0$. The absence of $y'(0)$ prevents a direct bootstrap from $x = 0$; the max we can do is assume some unknown value and go ahead, best foot forward. With some assumed value of $y'(0)$, it is not at all clear where we will end up when we reach $x = 1$; in general $y(1)$ can become any random real number. But that won't do – to have a solution we need $y(1)$ to equal zero. So we certainly have iffiness about our solution. One solution is always there however – if we start with $y'(0) = 0$ then we will just crawl along the x-axis until we reach $y = 1$, and automatically our second BC will be satisfied. This of course is the trivial solution. But whether we'll have more solutions or not, that's anybody's guess.

Once again, this is motivation rather than proof. A more formal treatment will come in the Tutorials. Before going over to those, where do BVPs occur in reality ? Eigenvalue problems invariably arise after separation of variables on PDE problems. Standalone ODE BVPs occur in the deflection of beams, of which we will be seeing a couple of examples. Finally, dynamical systems with periodic excitations can sometimes be solved exactly by reduction to BVPs in the time domain and imposing periodic boundary conditions.

Example 4.1.1

Basic BVP

Consider the BVP

$$y'' + \lambda y = 0 \quad ,$$

defined in the interval $[0, L]$ with the BCs $y(0) = 0$, $y(L) = 0$, where λ can be any real number, positive, negative or zero.

Determine all eigenvalues and corresponding eigenfunctions.

[Read Only]

This is one of the most fundamental BVPs in the business, so watch carefully how we go about solving it.

We consider separately the cases (1) $\lambda > 0$, (2) $\lambda < 0$ and (3) $\lambda = 0$.

CASE 1

Let $\lambda = k^2$ where k is any real number other than zero. Then the ODE is $y'' + k^2 y = 0$, whose general solution is

$$y_G = C_1 \cos kx + C_2 \sin kx \quad ,$$

where C_1 and C_2 are arbitrary constants. (From this point onwards I will assume that you can solve such simple ODEs easily, preferably in your head or else in under half a minute on a rough sheet. Nowhere will I elaborate the details behind finding the general solution of an ODE like this – if you are having a confusion, go back to Chapters 2-3, read them as many times as required, work the problems over and over again, work out drills from homework assignments….. in summary do whatever it takes to become absolutely comfortable with these things before you proceed with the material here. This Chapter will rapidly become quite difficult as it is, and if you aren't familiar with the basics, you will find the going impossible.)

Coming back to where we were, we have got the general solution and now we want to fit the BCs. Setting $x = 0$, $C_1(1) + C_2(0) = 0$ so $C_1 = 0$ while C_2 can be anything. Then setting $x = L$, $C_2 \sin kL = 0$. This has two possibilities, either $C_2 = 0$ or $\sin kL = 0$. The first one means $y = 0$ i.e. the trivial solution. The second one is more interesting as it implies non-trivial solutions. The zeroes of sin are at $kL = n\pi$ where n is an integer (different from zero, so that k does not become zero and contradict the starting premise of this case). Then, the permissible values of k are $n\pi/L$; since $\lambda = k^2$, the eigenvalues are $\lambda = n^2\pi^2/L^2$. The corresponding eigenfunctions are $y = \sin(n\pi x/L)$; since any eigenfunction can be multiplied by an arbitrary constant to produce another eigenfunction, we generally factor it out when reporting the solution. Further, since $\sin(-n\pi x/L)$ is just -1 times $\sin(n\pi x/L)$, the relevant values of n here are 1,2,3….. i.e. n is a natural number.

CASE 2

Let $\lambda = -k^2$ where k is any non-zero real number. The ODE is $y'' - k^2 y = 0$ and general solution is $y_G = C_1 e^{kx} + C_2 e^{-kx}$. Since this is a BVP, x doesn't run to infinity

and the exponentially growing solution is not a problem. Here however, it is more advantageous to write the solution in the hyperbolic form (see T020) as

$$y_G = C_1 \cosh kx + C_2 \sinh kx \quad ,$$

for reasons we'll see in a moment.

Setting $x = 0$, $C_1(1) + C_2(0) = 0$ so $C_1 = 0$ while C_2 can be anything. Then setting $x = L$, $C_2 \sinh kL = 0$. But sinh does not have a zero at any argument other than zero, so the only option here is $C_2 = 0$. This means that the BVP has only the trivial solution in this case.

CASE 3

The ODE here is $y'' = 0$ with the general solution

$$y_G = C_1 x + C_2 \quad .$$

Setting $x = 0$, $C_2 = 0$ while C_1 can be anything. Then setting $x = 1$, C_1 must be zero also, and once again we have only the trivial solution.

Hence the BVP possesses eigenvalues $\underline{\lambda = n^2\pi^2/L^2}$ where n is a natural number with the corresponding eigenfunctions $\underline{y = \sin(n\pi x/L)}$.

*

COMMENTS : If there is even one BVP whose result you should know by heart, it is this, it is this, it is this. As we go on to PDEs, this BVP will crop up more times than you can count. Then, it will be an enormous time-saver if you can go on to the solution immediately. Instead of blindly memorizing the answer however, make sure that you understand the process which led to it – that will preclude any memory lapses such as including zero as an eigenvalue or writing cos instead of sin.

Sometimes, the examiner might pose this question with the phrase "defined in the interval $[0, L]$" omitted. In such a case, s/he assumes that the moment the BVP nature of the problem and the two endpoints are specified, the interval is obvious. Anyone who raises queries about it risks being thought of as either a dunce or a pedant, so please take the interval for granted even if it isn't specified explicitly.

Example 4.1.2

Miscellaneous BCs

Each of the three BVPs given below features the ODE

$$y'' + \lambda y = 0 \quad ,$$

with the interval and BCs being different. Determine all eigenvalues λ and the corresponding eigenfunctions in each case.

(a) Interval $[0, L]$; BC $y'(0) = y'(L) = 0$

(b) Interval $[-\infty, \infty]$; BC y is periodic with period 2π

(c) Interval $[0, \infty]$; BC $y(0)$ is finite and $y \to 0$ as $x \to \infty$

[Level : 2]

(a) We split into the three cases of the last example. Case 1 $\lambda = k^2$ gives general solution $C_1 \cos kx + C_2 \sin kx$. Since the BC is on its derivative, we compute y_G'; it is $-kC_1 \sin kx + kC_2 \cos kx$. Fit the $x = 0$ BC and C_2 must be zero. Then go to $x = L$; either $C_1 = 0$ which is trivial or $kL = n\pi$. So eigenvalues here are $n^2\pi^2/L^2$ with eigenfunctions $\cos(n\pi x/L)$. Again, only positive n makes sense since $\cos(-n\pi x/L)$ is the same as $\cos(n\pi x/L)$. Case 2 $\lambda = -k^2$ gives general solution $C_1 \cosh kx + C_2 \sinh kx$; its derivative is $kC_1 \sinh kx + kC_2 \cosh kx$. Fit the $x = 0$ BC and $C_2 = 0$, then fit $x = L$ and $kC_1 \sinh kL = 0$. Since sinh doesn't have a zero, and k is non-zero by definition, C_1 must be zero and we have only the trivial solution. Case 3 $\lambda = 0$ gives general solution $C_1 x + C_2$; derivative is C_1 so that must be zero. C_2 however can be anything (evidently; a constant DOES satisfy the BC that $y' = 0$ at the two ends). Hence zero is an eigenvalue with 1 (since we can factor out the arbitrary constant) being the corresponding eigenfunction. Putting all this together, the eigenvalues are $\lambda = 0$ and $\lambda = n^2\pi^2/L^2$ with corresponding eigenfunctions $y = 1$ and $y = \cos(n\pi x/L)$ respectively.

(b) Again we partion into the three standard cases. Case 1 $\lambda = k^2$ gives the fundamental solutions $\cos kx$ and $\sin kx$. These are certainly periodic; to make the periodicity 2π, we need to have k as an integer n. Hence the eigenvalues are natural numbers $n = 1, 2, 3 \ldots$ and the corresponding eigenfunctions are $\cos nx$ and $\sin nx$. Case 2 $\lambda = -k^2$ gives the hyperbolic solutions which are obviously non-periodic and should be rejected. Case 3 $\lambda = 0$ gives the fundamental solutions 1 and x. The first one is periodic with any period while the second is not, so $y = 1$ is an eigenfunction corresponding to eigenvalue 0. Thus the eigenvalues are $\lambda = 0$ and $\lambda = n^2$ with eigenfunctions $y = 1$ and $y = \{\cos nx, \sin nx\}$ respectively.

(c) The starting step is the same as usual. Case 1 has the fundamental solutions $\cos kx$ and $\sin kx$ which are periodic, so if they are finite at $x = 0$, they cannot tend to zero as $x \to \infty$. Case 2 has the fundamental solutions e^{kx} and e^{-kx}. The second of these has the desired property of being finite at $x = 0$ and going to zero as $x \to \infty$ so it is an eigenfunction of the problem. Clearly, e^{-kx} satisfies the given BCs for any k. Case 3 has fundamental solutions 1 and x, neither of which tends to zero at large x, so we again do not have a solution which satisfies the

BCs. Thus the eigenvalues are all $\lambda < 0$ and the corresponding eigenfunctions are $y = e^{-kx}$ where $k = \sqrt{-\lambda}$.

*

COMMENTS : In Parts (a,b), note that $\lambda = 0$ is an eigenvalue here, while it wasn't one in the last Tute. Sometimes I have seen confusion regarding $\lambda = 0$; potential errors are saying that it is always an eigenvalue, it can never be an eigenvalue, and that the corresponding eigenfunction is always the trivial solution. As regards eigenvalue, zero may or may not be one as you can see; as regards eigenfunction, the trivial solution can NEVER be one.

For Part (b), note that this is one instance where the exponential form of fundamental solutions is preferable to the hyperbolic form as a starting point. Of course, taking hyperbolics isn't wrong; if you start from hyperbolics, you can always say that the relevant solution is $[\cosh kx - \sinh kx]/2$. But many students start from hyperbolics and then, since both of them grow unboundedly, conclude that there is no solution of the ODE satisfying the BCs, which becomes wrong. After some practice you will automatically know which representation is best for which problem. While starting out, I will strongly recommend that you write the solution in both forms and see which one is more suitable to fit the BCs which you have on hand.

In all the problems so far, we split the cases at $\lambda = 0$. So is zero always the magic number for the casewise split ? No. The splitting occurs at that value of λ where the form of the solution changes. For instance, an exam once posed the BVP $y'' + y' + \lambda y = 0$ with BCs at $x = 0$ and $x = 1$. In this case, the critical value of λ is 1/4 since that's where the solutions change – for $\lambda > 1/4$ the characteristic equation has a pair of complex roots, for $\lambda < 1/4$ a pair of real roots and for $\lambda = 1/4$ a repeated real root. However, it so happens that BVPs like this hardly ever arise in practical applications – the usual ones are what we did in this and the last Tute and a few others which we'll see in the subsequent Chapters.

Finally, wherefrom did I get these problems ? Don't get TOO inquisitive – you'll see that in due course of time.

Example 4.1.3

Inhomogeneous BVP

Each of the two BVPs given below is defined in the interval $[0, \pi]$ and the BCs are given as $y(0) = 0$ and $y(\pi) = 0$. The ODE is different in the two problems. Obtain the homogeneous solution as well as the specific solution in each case.

(a) $y'' + 2y = 1$

(b) $y'' + y = 1$

[Level : 2]

(a) The general solution of the ODE is $y_G = C_1 \cos\sqrt{2}x + C_2 \sin\sqrt{2}x + \frac{1}{2}$, and the homogeneous solution y_h doesn't have the '1/2'. Now we fit the boundary conditions, where the first BC is $y(0) = 0$ and the second BC is $y(\pi) = 0$. In the homogeneous problem, we get

$$C_1 + 0 = 0$$
$$C_1 \cos\sqrt{2}\pi + C_2 \sin\sqrt{2}\pi = 0 \quad .$$

The first equation implies $C_1 = 0$; then since $\sin\sqrt{2}\pi$ is not zero, C_2 is zero also. Hence the homogeneous problem has only the trivial solution.

For the inhomogeneous problem we have

$$C_1 + \frac{1}{2} = 0$$
$$\frac{1}{2} + C_1 \cos\sqrt{2}\pi + C_2 \sin\sqrt{2}\pi = 0 \quad ,$$

which has the solution $C_1 = -1/2$, $C_2 = \frac{1}{2}\frac{\cos\sqrt{2}\pi - 1}{\sin\sqrt{2}\pi}$, yielding the specific solution

$$* \qquad y = -\frac{1}{2}\cos\sqrt{2}x + \frac{1}{2}\frac{\cos\sqrt{2}\pi - 1}{\sin\sqrt{2}\pi}\sin\sqrt{2}x + \frac{1}{2} \quad , \qquad *$$

which is the desired answer.

(b) Here the general solution is $y_G = C_1 \cos x + C_2 \sin x + 1$ and the homogeneous solution y_h doesn't have the '1'. Fitting BCs,

$$C_1 + 0 = 0$$
$$C_1(-1) + C_2(0) = 0 \quad ,$$

showing that $C_1 = 0$ and C_2 can be anything. Hence the homogeneous problem has the solution $y_h = C\sin x$ where C is an arbitrary constant.

For the inhomogeneous problem we have

$$C_1 + 1 = 0$$
$$C_1(-1) + C_2(0) + 1 = 0 \quad .$$

The first one says $C_1 = -1$ but then the second one implies $C_2(0) = -2$ so C_2 does not exist. Hence the inhomogeneous problem has NO solution.

*

COMMENTS : Part (b) here is a very popular question for oral examinations. An unprepared candidate arriving at the contradiction for C_2 becomes flustered, swears under his/her breath, does a da capo (obviously with no change in result), and in summary provides a lot of entertainment to the interrogators. You can avoid this clownish display if you learn right now, for the rest of your life, that a BVP can have the possibility of no solution and that an attempt to solve such a BVP leads to a contradiction like this one.

A similar thing happens if the inhomogeneity lies in the BCs instead of the ODE RHS. For example, $y''+2y = 0$ with $y(0) = 1$ and $y(\pi) = 0$ has the solution $y = \cos\sqrt{2}x + (\cot\sqrt{2}\pi)\sin\sqrt{2}x$. On the other hand, $y''+y = 0$ with these BCs has no solution. And yet, $y''+y = 0$ with $y(0) = 1$ and $y(\pi) = -1$ has the solutions $\cos x + C_2 \sin x$ for arbitrary C_2. We'll demystify all this in example 4.1.7.

Example 4.1.4

Matrix method preliminaries

Note : Knowledge of EULER's method is not essential to make progress with the subsequent material so Parts (a) and (b) may be skipped if the course does not cover this topic.

(a) Solve the IVP $\dot{x}+x=0$, $x(0)=1$ using EULER's method with step size h. Show that the accuracy of the solution increases without bound as h decreases.

(b) Solve the IVP $\ddot{x}+x=0$, $x(0)=1$ and $\dot{x}(0)=0$ using EULER's method with step size h. Show that the accuracy of the solution deteriorates without bound as t increases.

(c) Consider the matrices

$$\mathbf{A}=\begin{bmatrix}1 & 0 & 3\\ 2 & 1 & 0\\ 3 & 2 & 2\end{bmatrix} \text{ and } \mathbf{F}=\begin{bmatrix}1 & 0 & 4\\ 2 & 1 & 3\\ 3 & 2 & 2\end{bmatrix} \text{ , and}$$

and the vectors

$$\mathbf{b}_1=0,\ \mathbf{b}_2=\begin{bmatrix}5; & 5; & 5\end{bmatrix} \text{ and } \mathbf{b}_3=\begin{bmatrix}5; & 5; & 6\end{bmatrix} \ .$$

Solve the systems $*\mathbf{x}=\#$ where $\mathbf{x}=[x_1; x_2; x_3]$ is the unknown vector, * denotes any of the two matrices and # denotes any of the three vectors.

[Read Only]

(a) The ODE is $\mathrm{d}x/\mathrm{d}t = -x$ with IC $x(0) = 1$. We can find the analytical solution even after 20 beers : it's $x = \mathrm{e}^{-t}$. EULER's method gives us the approximate solutions of the IVP at times h, $2h$, $3h$, …..; let the solutions at these times be denoted by X_1, X_2, X_3, ….. i.e. let X_i denote the approximate solution at time ih. Here we are using X for the approximate solution to distinguish it from the true solution x. By definition of the discrete derivative, $\dot{X}$ at the i^{th} time step is

$$\dot{X}_i = \frac{X_{i+1} - X_i}{h} \quad ,$$

which yields

$$X_{i+1} = X_i + h\dot{X}_i \quad .$$

Now, from the ODE, $\dot{X}_i$ is simply $-X_i$, so that we can write the above as

$$X_{i+1} = (1-h)X_i \quad .$$

This is a recursion relation for the sequence X_i; in this case it is simple enough for us to solve it by inspection. We have

$$X_i = (1-h)^i X_0 = (1-h)^i \quad ,$$

the latter relation holding because the IC is $x(0) = X(0) = 1$.

Now we want to compare the sequence X_i with the analytical solution $x(t)$. Recalling that $t = ih$, the Eulerian solution at time t is

$$X(t) = \left(1 - \frac{t}{i}\right)^i \quad . \qquad (01)$$

If the step size h decreases, then for a given t, the value of i increases. As this happens, the expression on the RHS of (01) evidently converges to e^{-t}, which it equals in the limit $i \to \infty$. Thus, the Eulerian solution approaches closer and closer to the true solution as the step size is reduced, which was to be shown.

(b) Harmonic oscillator equation with unit initial displacement and zero initial velocity; if it takes you longer than a tenth of a period (SI Units!) to find $x = \cos t$, then you slide down a snake back to 2.1.1. As in Part (a), we let the sequence X_i denote the approximate solution at time $t = ih$. We've already started the solution process for harmonic oscillator in 4.1; we write it as $\dot{x} = y$, $\dot{y} = -x$, whose discrete form is

$$X_{i+1} = X_i + h\dot{X}_i = X_i + hY_i \quad ,$$
$$Y_{i+1} = Y_i + h\dot{Y}_i = Y_i - hX_i \quad .$$

This suggests a matrix relationship between $[X_{i+1};\ Y_{i+1}]$ and $[X_i;\ Y_i]$ – we have

$$\begin{bmatrix} X_{i+1} \\ Y_{i+1} \end{bmatrix} = \begin{bmatrix} 1 & h \\ -h & 1 \end{bmatrix} \begin{bmatrix} X_i \\ Y_i \end{bmatrix} \quad .$$

Once again, the recursion for the sequences X_i and Y_i can be solved by inspection; we have

$$\begin{bmatrix} X_i \\ Y_i \end{bmatrix} = \begin{bmatrix} 1 & h \\ -h & 1 \end{bmatrix}^i \begin{bmatrix} X_0 \\ Y_0 \end{bmatrix} \quad . \qquad (02)$$

Since the question mentions trouble as t increases, we must increase i here while keeping h fixed.

Now, we have to find the powers of the matrix $\mathbf{A}=[1\ h;\ -h\ 1]$. Recall from 2.3 that powers of a matrix $\mathbf{A}$ can be found easily by writing $\mathbf{A}=\mathbf{P}^{-1}\mathbf{D}\mathbf{P}$; the i_{th} power is $\mathbf{A}^i=\mathbf{P}^{-1}\mathbf{D}^i\mathbf{P}$. We take for granted that this formula holds even if the eigenvalues and vectors of $\mathbf{A}$ are complex. Let's find the eigenvalues of $\mathbf{A}$ in (02). The characteristic equation is

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0 \Rightarrow \lambda^2 - 2\lambda + 1 + h^2 = 0 \quad ,$$

which has the solutions $\lambda = 1 \pm \mathrm{j}h$. If we write either λ in polar form, the amplitude R works out to $(1+h^2)^{1/2}$; the angles are immaterial. Now, raise this λ to the i^{th} power; this is a complex number of amplitude $(1+h^2)^{i/2}$. As i increases, this amplitude increases without bound. Hence, the entries of $\mathbf{D}^i$ become larger with increasing i and so do the entries in $\mathbf{A}^i$ since $\mathbf{P}^{-1}$ and $\mathbf{P}$ are constant matrices. Equation (02) thus implies that as i increases, $[X_i;\ Y_i]$ becomes a larger and larger matrix times $[X_0;\ Y_0]$ i.e. the Eulerian solutions X and Y keep increasing in time. The actual solution x and its derivative are of course oscillatory. Hence, we conclude that the Eulerian solution represents a worse and worse approximation of the true solution as time increases, which was to be shown.

(c) We start from $\mathbf{A}\mathbf{x} = \mathbf{b}_1$, noting that '0' means $[0;\ 0;\ 0]$. The fourth column of the augmented matrix is just zero so we ignore it and perform elementary row

operations on $\mathbf{A}$. The operations R2 → R2−2R1 and R3 → R3−3R1, followed by R3 → R3−2R2 on the resulting matrix reduce $\mathbf{A}$ to

$$\tilde{\mathbf{A}} = \begin{bmatrix} 1 & 0 & 3 \\ 0 & 1 & -6 \\ 0 & 0 & 5 \end{bmatrix} .$$

The third row implies $x_3 = 0$, the second row gives $x_2 = 0$ and the first row then produces $x_1 = 0$, so the only solution of the homogeneous system is $\mathbf{x} = 0$, the trivial solution. This implies that $\mathbf{A}$ is invertible, which then means that $\mathbf{Ax} = \mathbf{b}_2$ and $\mathbf{Ax} = \mathbf{b}_3$ must both have unique solutions.

$\mathbf{Ax} = \mathbf{b}_2$ gives the augmented matrix in echelon form as

$$\left[\tilde{\mathbf{A}} \,|\, \tilde{\mathbf{b}}_2\right] = \left[\begin{array}{ccc|c} 1 & 0 & 3 & 5 \\ 0 & 1 & 6 & -5 \\ 0 & 0 & 5 & 0 \end{array}\right] ,$$

with the solution $\mathbf{x} = [5; -5; 0]$. $\mathbf{Ax} = \mathbf{b}_3$ gives the augmented matrix in echelon form as

$$\left[\tilde{\mathbf{A}} \,|\, \tilde{\mathbf{b}}_3\right] = \left[\begin{array}{ccc|c} 1 & 0 & 3 & 5 \\ 0 & 1 & 6 & -5 \\ 0 & 0 & 5 & 1 \end{array}\right] ,$$

with the solution $\mathbf{x} = [22/5; -19/5; 1/5]$.

Now consider the other matrix $\mathbf{F}$. For $\mathbf{Fx} = \mathbf{b}_1$, the elementary row operations are R2 → R2−2R1 and R3 → R3−3R1 followed by R3 → R3−2R2 on the resultant. This produces the echelon

$$\tilde{\mathbf{F}} = \begin{bmatrix} 1 & 0 & 4 \\ 0 & 1 & -5 \\ 0 & 0 & 0 \end{bmatrix} ,$$

showing that x_3 is a free variable. Hence the homogeneous system has one family of non-trivial solutions. To find that family, we set $x_3 = 1$, getting $x_1 = -4$ and $x_2 = 5$. Thus, $\mathbf{x} = [-4; 5; 1]$ is a solution of the homogeneous system. Since any multiple of that is also a homogeneous solution, we can write the solution

as $\underline{\mathbf{x} = c_1[-4;\ 5;\ 1]}$ where c_1 is an arbitrary constant. At once, we see that $\mathbf{F}$ is non-invertible, so there can be issues with the solutions of $\mathbf{Fx} = \mathbf{b}_2$ and $\mathbf{Fx} = \mathbf{b}_3$.

$\mathbf{Fx} = \mathbf{b}_2$ gives the augmented matrix in echelon form as

$$\left[\tilde{\mathbf{F}} \,|\, \tilde{\mathbf{b}}_2\right] = \left[\begin{array}{ccc|c} 1 & 0 & 4 & 5 \\ 0 & 1 & -5 & -5 \\ 0 & 0 & 0 & 0 \end{array}\right] .$$

The last row shows that the system is consistent, with x_3 a free variable as before. To find a particular solution, we set the free variable to zero, giving $x_1 = 5$ and $x_2 = -5$. So a particular solution is $[5;\ -5;\ 0]$, and the general solution is the particular solution plus any multiple of the homogeneous solution i.e. $\underline{\mathbf{x} = [5;\ -5;\ 0] + c_1[-4;\ 5;\ 1]}$.

Finally, $\mathbf{Fx} = \mathbf{b}_3$ gives the augmented matrix in echelon form as

$$\left[\tilde{\mathbf{F}} \,|\, \tilde{\mathbf{b}}_3\right] = \left[\begin{array}{ccc|c} 1 & 0 & 4 & 5 \\ 0 & 1 & -5 & -5 \\ 0 & 0 & 0 & 1 \end{array}\right] .$$

The last row shows that the system is inconsistent so a solution $\mathbf{x}$ <u>does not exist</u>.

*

COMMENTS : EULER's method and the simple matrix equation are warm-ups which get you familiar with discrete derivatives, and recap the properties of matrix equations. They lay the groundwork for the matrix method of discretizing BVPs, which we are going to see from the next Tute onwards.

Parts (a) and (b) here are about the only problems on EULER's method which can be solved by pen and paper rather than computer, so they make great questions for the course exams. Indeed, Part (a) or trivial variants thereof has appeared multiple times in the midsems and Final in the last few years. Part (b) also shows a very significant drawback of EULER's method – it generates a lousy solution to one of the most fundamental equations. For this reason, the method has near-zero credibility as a numerical technique for solving ODEs. It can be remedied by several procedures such as CRAMER's correction and Runge Kutta

method but numerical methods are absolutely NOT within the scope of this book.

The matrix equation of Part (c) reminds you of the similarities between BVPs and matrix equations. This Tute was essentially the same as the last one except that the nature of the equations was different in the two cases. In both cases, when the homogeneous problem had only the trivial solution, the inhomogeneous one had a unique solution. In both cases, when the homogeneous problem had non-trivial solutions, the inhomogeneous one had either infinitely many solutions or none at all. Of course, this is anything but a coincidence as we'll soon see.

Example 4.1.5

First order BVP, matrix method

Consider the very simple ODE

$$\frac{\mathrm{d}y}{\mathrm{d}x} + y = 0 \quad ,$$

defined in the interval [0, 1] with the BC $y(0) = b$.

(a) Solve the problem analytically.

The steps towards solution of this problem by matrix method are outlined below. Partion the interval into four points (lattice sites) x_1, x_2, x_3 and x_4 equal to 1/4, 2/4, 3/4 and 1 respectively. Let x_0 denote the boundary point $x = 0$. Define y_0 to be y at x_0 and similarly y_1, y_2, y_3 and y_4. Let $\mathbf{y} = [y_1; y_2; y_3; y_4]$.

(b) Using the concept of discrete derivative, write the vector $\mathbf{y}' = [y'_1; y'_2; y'_3; y'_4]$ in the form $\mathbf{P}_+\mathbf{y}_+$, where $\mathbf{P}_+$ is the derivative matrix and $\mathbf{y}_+$ is the vector $[y_0; y_1; y_2; y_3; y_4]$. The $\mathbf{y}'$ vector should be defined at the four points x_1, x_2, x_3 and x_4 and not at x_0 since that is the boundary point.

(c) Impose the BC on the above result to write the BVP as $\mathbf{L}\mathbf{y} = \mathbf{b}$ where $\mathbf{L}$ is square.

(d) Solve the matrix equation on a computer with $b = 1$.

(e) Instead of four points, split the interval into a variable number N of equispaced discrete points. Compare the analytical and matrix solutions as N increases.

[Read Only]

(a) This is a feel-good problem : $\underline{y = b\,e^{-x}}$.

(b) The discrete derivative formula is $y'_i = (y_{i+1} - y_i)/h$ or $y'_i = (y_i - y_{i-1})/h$, accordingly as we take the right or left side form. Here, h is of course 1/4, the spacing between two successive lattice points. The raw first derivative matrix $\mathbf{P}_+$ accepts a 5-vector $\mathbf{y}_+$ and returns a 4-vector $\mathbf{y}'$, so its size must be 4 x 5 (we'll use the letters $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{R}$, $\mathbf{S}$ etc to denote derivative matrices of order 1, 2, 3, 4 etc). Some thought now reveals the form (left side or right side) of the derivative we want. To define y' at site 1, the left side form says $y'_1 = (y_1 - y_0)/h$ while the right side form says $y'_1 = (y_2 - y_1)/h$. Evidently, since y_0 is known (it's the BC), it makes sense to bring that into the definition of y'_1 rather than using the unknown y_2. Hence we go with the left side derivative here.

The formula for left side derivative is

$$y'_i = \frac{1}{h}\left(y_i - y_{i-1}\right) \quad . \tag{01}$$

Then by inspection, we get the following matrix representation of the derivative $\mathrm{d}/\mathrm{d}x$:

$$* \qquad \frac{\mathrm{d}y}{\mathrm{d}x} \to \mathbf{y}' = \begin{bmatrix} y'_1 \\ y'_2 \\ y'_3 \\ y'_4 \end{bmatrix} = \mathbf{P}_+\mathbf{y}_+ = \frac{1}{h}\begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix}\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \quad . \qquad *$$

Check this : the first row says $y'_1 = (y_1 - y_0)/h$, the second row says $y'_2 = (y_2 - y_1)/h$ and so on. Evidently, each row of $\mathbf{P}_+$ contains the coefficients -1 and 1 of the discrete derivative formula (01) arranged in a stepwise manner, and there's an overall $1/h$ outside.

(c) The ODE now states $y' + y = 0$ or

$$\mathbf{P}_+\mathbf{y}_+ + \mathbf{y} = 0 \quad . \tag{02}$$

This is almost OK except for the fact that the first term features $\mathbf{y}_+$ rather than $\mathbf{y}$. So now we have to use the BC, which is that $y_0 = b$. Using this in the first row of the above equation gives $(y_1 - b)/h + y_1 = 0$, or, $y_1/h + y_1 = b/h$. All other rows of (02) are not affected by the value of y_0 since they feature 0 in the first column

of $\mathbf{P}_+$. We can now use this to eliminate the first column of $\mathbf{P}_+$ and the first element of $\mathbf{y}_+$. Carrying the b/h over to the RHS we have

$$\frac{1}{h}\begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} b/h \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{, or} \qquad (03a)^*$$

$$\mathbf{P}\mathbf{y} + \mathbf{I}\mathbf{y} = [b/h;\ \ 0;\ \ 0;\ \ 0] \ , \qquad (03b)^*$$

where $\mathbf{P}$ is the matrix in the LHS of (03a) and $\mathbf{I}$ is the identity matrix. We can mechanically verify that (03) is identical to (02) together with the given BC. Comparing with the notation of the question, we see that $\mathbf{L}=\mathbf{P}+\mathbf{I}$ and $\mathbf{b}=[b/h;\ 0;\ 0;\ 0]$. Thus, (03) is the desired answer.

(d) Matlab returns the solution $\underline{\mathbf{x} = [0.80;\ 0.64;\ 0.51;\ 0.41]}$.

(e) We note that since the total number of points including the boundary point is $N+1$, the inter-site spacing should be $h = 1/N$ (as it was with $N = 4$). Given this, it is a simple task to write the code to solve the matrix system of variable size N. As a check of accuracy, we compare the value of y_N i.e. $y(1)$ with the analytical value of $1/e = 0.368$. For $N = 4$, we had 0.41, for $N = 20$ we have 0.377 and for $N = 500$ we have 0.368. This shows that the numerical solution converges to the analytical solution as the grid refinement increases.

*

COMMENTS : This "MATRIX METHOD," numerical solution of the ODE together with BC by representing it as a matrix equation, is the simplest version of the finite element method used in engineering. It is also called finite difference method. Let's again summarize the steps involved in the procedure.

- Step 1. We first define the problem in lattice sites $\#1$ to $\#N$ (called "internal" or "significant" points) and then add on extra sites corresponding to boundaries to create the vector $\mathbf{y}_+$. The number of extra sites is k where k is the order of the derivative involved in the problem. We define the raw discrete derivative matrix (of order k) to be just the coefficients of the discrete derivative formula written in the successive

rows, arranged in a stepwise manner. It is not square, having k more columns than rows. It accepts $\mathbf{y}_+$ to produce a derivative vector of size N.

- Step 2. We impose the BCs to create a square matrix which accepts $\mathbf{y}$ and produces the derivative vector. Forcing the BCs on the non-square system allows us to eliminate extra columns from the raw derivative matrix until it becomes square, of size N. The square matrix thus represents the derivative together with the BCs.
- Point to note. If the operator $\mathbf{L}$ features terms with different orders of derivative then we must handle each term separately using Steps 1 and 2. For each derivative we create $\mathbf{y}_+$ using only as many boundary points as required. Suppose we want to discretize $y''+y'$. We write $\mathbf{y}''=\mathbf{Q}_+\mathbf{y}_+$ where $\mathbf{Q}_+$ is of size N x $N+2$ and $\mathbf{y}_+$ also has two boundary points whereas we write $\mathbf{y}' = \mathbf{P}_+\mathbf{y}_+$ with $\mathbf{P}_+$ of size N x $N+1$ and $\mathbf{y}_+$ now has only one boundary point. Two BCs are needed to convert $\mathbf{Q}_+$ to $\mathbf{Q}$ while only one of these BCs suffices to turn $\mathbf{P}_+$ into $\mathbf{P}$. Only after that can we say $\mathbf{L} = \mathbf{Q}+\mathbf{P}$.

As the course progresses we'll see that the matrix method is in fact a powerful tool – a hammer at one end and a knife at the other. If a BVP is resistant to all other approaches, we can just discretize it into a matrix and smash it open. On the other hand, we can use the matrix representations of ODEs (and later PDEs) to cut through layers of formalism and obtain subtle theoretical properties of their solutions, which are quite difficult to get otherwise.

Example 4.1.6

Second order BVP, matrix method

Consider the ODE

$$y'' = \lambda y \quad ,$$

defined in the interval [0, 1] with the BCs $y(0) = y(1) = 0$.

(a) Write the matrix method formulation of the BVP and explain why the term “eigenvalue” is appropriate for λ.

(b) If x_i are the lattice sites and n is an integer from 1 to N, check that the vector $\mathbf{v} = [\sin n\pi x_1;\ \sin n\pi x_2;\ .....;\ \sin n\pi x_N]$ is an eigenvector of $\mathbf{Q}$ and determine the corresponding eigenvalue.

[Level : 2]

(a) We follow the algorithm of the last example. For Step 1 we let x_1, x_2, ….., x_N be the internal points and let x_0 and x_{N+1} be the boundaries. The vector $\mathbf{y}$ is of course $[y_1; y_2; \ldots..; y_N]$, while $\mathbf{y}_+$ is $[y_0; y_1; y_2; \ldots..; y_N; y_{N+1}]$. We let $\mathbf{y}'' = \mathbf{Q}_+\mathbf{y}_+$ where $\mathbf{Q}_+$ is the raw second derivative matrix – its size is evidently N x $N+2$. To construct $\mathbf{Q}_+$, we note that the discrete second derivative is

$$y_i'' = \frac{y_{i+1} + y_{i-1} - 2y_i}{h^2} \quad , \tag{01}$$

which is centred and doesn't raise any issue of right or left side. Since site #0 corresponds to $x = 0$ while #N+1 corresponds to $x = 1$, $h = 1/(N+1)$ rather than $1/N$! Each row of $\mathbf{Q}_+$ must contain the formula (01) written in a stepwise manner. Hence we have, taking $N = 4$ for display purposes,

$$\frac{\mathrm{d}^2 y}{\mathrm{d}x^2} \to \mathbf{y}'' = \mathbf{Q}_+\mathbf{y}_+ = \frac{1}{h^2}\begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix}\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} \quad .$$

This completes Step 1 of the matrix process.

For Step 2 we impose the BCs to bring the non-square $\mathbf{Q}_+$ to a square $\mathbf{Q}$. With y_0 and y_N (y_5 in the display example) both equal to zero, we can simply hack off the first and last columns of the matrix and the corresponding elements of $\mathbf{y}_+$ to get a N x N (4 x 4) square. Implementing this step, we have the square second derivative with BCs matrix

$$\mathbf{Q} = \frac{1}{h^2}\begin{bmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \end{bmatrix} \quad .$$

Since $\mathbf{y}'' = \mathbf{Q}\mathbf{y}$, the BVP $y'' = \lambda y$ becomes

$$\mathbf{Q}\mathbf{y} = \lambda\mathbf{y} \quad .$$

For the system to have non-trivial solutions $\mathbf{y}$, λ must be an eigenvalue of $\mathbf{Q}$ with $\mathbf{y}$ being the corresponding eigenvector. This shows that the terminology is appropriate.

(b) This is a checking problem so we consider the action of $\mathbf{Q}$ on the given vector $\mathbf{v}$. We have to separately consider two cases – the first and last row and the 'bulk' or all other rows of the system. Let us consider the latter case first. The i^{th} element of $\mathbf{v}$ is $\sin n\pi x_i$ which, since $x_i = ih$, is $\sin n\pi ih$. The i^{th} row of $\mathbf{Q}$ has the elements 1, -2 and 1 in the positions $i-1$, i and $i+1$ and zeroes everywhere else (all with an outside factor of $1/h^2$). In the i^{th} row of the product $\mathbf{Qv}$, these non-zero elements of $\mathbf{Q}$ will interact with the elements $\#i-1$, $\#i$ and $\#i+1$ of $\mathbf{v}$ respectively. These elements are $\sin n\pi x_{i-1}$, $\sin n\pi x_i$ and $\sin n\pi x_{i+1}$, which evaluate to $\sin n\pi(i-1)h$, $\sin n\pi ih$ and $\sin n\pi(i+1)h$ respectively. Thus, the i^{th} entry of $\mathbf{Qv}$ is

$$\frac{1}{h^2}\left[\sin n\pi(i-1)h - 2\sin n\pi ih + \sin n\pi(i+1)h\right] \quad ,$$

which using the trigs from the review reduces to

$$\frac{2}{h^2}\left(\cos n\pi h - 1\right)\sin n\pi ih \quad .$$

This shows that the i^{th} entry of $\mathbf{Qv}$ is proportional to the corresponding entry of $\mathbf{v}$ through the factor $2(\cos n\pi h - 1)/h^2$, which must be the eigenvalue.

To complete the check, we verify the top and bottom rows. At the top, the first two elements of $\mathbf{v}$ are $\sin n\pi h$ and $\sin 2n\pi h$. Multiplying by $\mathbf{Q}$ gives the first element of $\mathbf{Qv}$ as $(-2\sin n\pi h + \sin 2n\pi h)/h^2$, which using trig identities reduces to $2\sin n\pi h(\cos n\pi h - 1)/h^2$, consistent with what we got earlier. For form's sake we go ahead with the last row, even though we know it's going to work. The last element of $\mathbf{v}$ is $\sin n\pi hN$ and the one before that is $\sin n\pi h(N-1)$. But, since $h = 1/(N+1)$, these are $\sin(n\pi - n\pi h)$ and $\sin(n\pi - 2n\pi h)$ respectively. Expanding the last row followed by basic trig again shows the proportionality relation to be true. Hence $\mathbf{v}$ is indeed an eigenvector of $\mathbf{Q}$ with eigenvalue

* $$\lambda = \frac{2}{h^2}\left(\cos n\pi h - 1\right) \quad ,$$ *

which is the desired answer.

The check holds for all n from 1 to N (evidently, all the steps went through). If we increase n beyond that, then from periodicity of sin function, we start recovering the negatives of the eigenvectors determined here, so there are no new eigenvectors in that region. Of course, a N x N matrix can have only N eigenvectors and not more.

*

COMMENTS : Here we can clearly see how the discrete system transitions to the continuous one. As N becomes larger, the eigenvectors pick up more and more elements of the $\sin n\pi x_i$ functions until they become the functions themselves. As for the eigenvalues, since h is small, we Taylor expand the cosine assuming small argument. Keeping only the first two terms of the series, the eigenvalues reduce to $-n^2\pi^2$. In the finite sized system, the approximation is good for $n = 1$ and becomes worse for high n but as h gets smaller and smaller, the approximation gets better and better for larger and larger n. In the limit, we recover the eigenvalues of the continuous problem.

Where have you seen the matrix $\mathbf{Q}$ before ? Go back through the book and hunt for it. For now I won't elaborate the 'coincidence'. But just this much let me mention that viewed from one angle, this is a complete standalone problem, while viewed from another angle it is one more piece of a grand jigsaw puzzle. This complete puzzle, we'll see when all the pieces are finally put together.

Example 4.1.7

FREDHOLM's alternatives

Consider the ODE

$$y''+k^2y=f(x) \quad , \qquad (*)$$

defined in the interval [0, 1] with the BCs $y(0)=y(1)=0$.

(a) Use the matrix method to write the problem in the form $\mathbf{Ly}=\mathbf{f}$ where $\mathbf{L}$ is a square matrix and $\mathbf{f}$ is the excitation vector.

If $\mathbf{L}$ is invertible, $\mathbf{Ly}=\mathbf{f}$ obviously has a unique solution for all $\mathbf{f}$.

(b) For what value/s of k is $\mathbf{L}$ non-invertible ? Express your answer in terms of the second derivative matrix $\mathbf{Q}$.

(c) In the case that $\mathbf{L}$ is non-invertible, show that $\mathbf{Ly}=\mathbf{f}$ has a solution if and only if $\mathbf{f}$ is orthogonal to the null space of $\mathbf{L}$. Hint : $\mathbf{L}$ is symmetric, so make use of the transposition trick and the spectral theorem from 2.3.

(d) Matrix match type question :

Referring to the BVP (*), the below table gives four choices of k and $f(x)$ in the left hand column and four different solution outcomes in the right hand column. For each choice of $f(x)$, label the solution outcome/s which is/are appropriate. Justify your answers.

A. $k=1, f(x)=1$	P. Trivial solution
B. $k=\pi, f(x)=\sin 2x$	Q. Unique solution
C. $k=\pi, f(x)=(2x+3)\,e^{-x}\sin \pi x$	R. Infinitely many solutions
D. $k=1, f(x)=0$	S. No solution

(e) Consider the ODE

$$y''+k^2y=0 \quad ,$$

defined in the interval [0, 1] with BCs $y(0)=a$ and $y(1)=b$ where a, b, and k are variable. Enumerate all possible outcomes regarding

the existence of one or more solutions, stating clearly the conditions under which each case holds true.

[Level : 3]

(a) From the previous problem, the BVP is $(\mathbf{Q}+k^2\mathbf{I})\mathbf{y}=\mathbf{f}$. Evidently, $\mathbf{L}$ is $\mathbf{Q}+k^2\mathbf{I}$ which is (using $N=4$ for display)

$$\mathbf{L}=\begin{bmatrix} -(2/h^2)+k^2 & 1/h^2 & 0 & 0 \\ 1/h^2 & -(2/h^2)+k^2 & 1/h^2 & 0 \\ 0 & 1/h^2 & -(2/h^2)+k^2 & 1/h^2 \\ 0 & 0 & 1/h^2 & -(2/h^2)+k^2 \end{bmatrix} .$$

Evidently, $\mathbf{L}$ is symmetric, which fact seems to play an important role later.

(b) $\mathbf{L}$ is non-invertible means that $(\mathbf{Q}+k^2\mathbf{I})\mathbf{y}=0$ has a non-trivial solution $\mathbf{y}$. Clearly, this holds true if and only if $-k^2$ is an eigenvalue of $\mathbf{Q}$. Hence the required values of k are the square roots of the negatives of the eigenvalues of $\mathbf{Q}$. In this case, the non-trivial solutions of $\mathbf{L}\mathbf{y}=0$ are the corresponding eigenvectors of $\mathbf{Q}$.

(c) Proving an if-and-only-if means proving two things : (i) if $\mathbf{L}\mathbf{y}=\mathbf{f}$ has a solution then $\mathbf{f}$ is orthogonal to the null space of $\mathbf{L}$, and (ii) if $\mathbf{f}$ is orthogonal to the null space of $\mathbf{L}$ then $\mathbf{L}\mathbf{y}=\mathbf{f}$ has a solution.

We prove (i) first since it's easier. Let $\mathbf{y}_0$ be a solution to $\mathbf{L}\mathbf{y}=\mathbf{f}$, and let $\mathbf{w}$ be a vector in the null space of $\mathbf{L}$. Then, considering the relation

$$\mathbf{L}\mathbf{y}_0=\mathbf{f} \quad ,$$

take the dot product with $\mathbf{w}$ on the both sides. On the RHS we of course have $\mathbf{w}\cdot\mathbf{f}$. On the LHS, $\mathbf{w}\cdot(\mathbf{L}\mathbf{y}_0)=(\mathbf{L}\mathbf{w})\cdot\mathbf{y}_0$ from the transposition trick followed by $\mathbf{L}$ is symmetric (see 2.3). But since $\mathbf{w}$ belongs to the null space of $\mathbf{L}$, $\mathbf{L}\mathbf{w}=0$. Hence $\mathbf{f}\cdot\mathbf{w}=0$ i.e. $\mathbf{f}$ is orthogonal to every vector $\mathbf{w}$ in the null space of $\mathbf{L}$, and therefore to the entire null space of $\mathbf{L}$.

Now we'll prove (ii). We recall that null space of $\mathbf{L}$ is the eigenspace corresponding to the eigenvalue zero. So let $\mathbf{L}$ have the eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots\ldots, \mathbf{v}_m$ corresponding to non-zero eigenvalues $\lambda_1, \lambda_2, \ldots\ldots, \lambda_m$ (which may or may not be repeated) and the eigenvectors $\mathbf{w}_1, \mathbf{w}_2, \ldots\ldots, \mathbf{w}_n$ corresponding to $\lambda=0$. Here, $m+n=N$, the size of $\mathbf{L}$. By the spectral theorem, the $\mathbf{w}$'s form a basis of the n-dimensional null space and can be taken to be orthonormal with no loss of

generality. The $\mathbf{v}$'s also form a basis of the remaining m-dimensional space and can be taken to be orthonormal. Every $\mathbf{w}$ is orthogonal to every $\mathbf{v}$.

An arbitrary vector $\mathbf{f}$ can be written as

$$\mathbf{f} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots.. + c_m\mathbf{v}_m + d_1\mathbf{w}_1 + d_2\mathbf{w}_2 + \ldots.. + d_n\mathbf{w}_n \quad ,$$

where $c_i = \mathbf{f}\cdot\mathbf{v}_i$ and $d_i = \mathbf{f}\cdot\mathbf{w}_i$ from orthonormality of the $\mathbf{v}$'s and $\mathbf{w}$'s. But, for an $\mathbf{f}$ which is orthogonal to the null space of $\mathbf{L}$, $d_i = 0$ for all i and only the c's are non-trivial. Now, consider the vector

$$\mathbf{y}_0 = \frac{c_1}{\lambda_1}\mathbf{v}_1 + \frac{c_2}{\lambda_1}\mathbf{v}_2 + \ldots.. + \frac{c_m}{\lambda_m}\mathbf{v}_m \quad .$$

Since $\mathbf{L}\mathbf{v}_i = \lambda_i\mathbf{v}_i$, $\mathbf{L}$ acting on this vector produces $\mathbf{f}$, and we have obtained a solution $\mathbf{y}_0$ of $\mathbf{L}\mathbf{y} = \mathbf{f}$ by explicit construction. Since the λ_i's are non-zero by definition, our $\mathbf{y}_0$ is well-defined and the two parts of FREDHOLM's theorem have been proved.

(d) Before doing the matrix match, let's assemble what we have on the table. Firstly, the operator $\mathbf{L} = \mathrm{d}^2/\mathrm{d}x^2 + k^2$ with homogeneous BCs at 0 and 1 is 'symmetric' (the mathsy word for this is SELF-ADJOINT). Then, it is invertible if k^2 does not equal an eigenvalue of the BVP and non-invertible if it does. If $\mathbf{L}$ is invertible, $\mathbf{L}y = f(x)$ of course has a unique solution for every f. If $\mathbf{L}$ is non-invertible, the problem has a solution if and only if $f(x)$ is orthogonal to the null space of $\mathbf{L}$. Of course, in this case, whenever there's one solution, there's automatically an infinitude of solutions since we can add any multiple of the homogeneous solution. From 4.1.1, the eigenvalues of the given BVP are $k = n\pi$ with eigenfunctions $\sin n\pi x$, and we can now take the match head on.

A – Q. Since $k = 1$ is not an eigenvalue, $\mathbf{L}$ is invertible and the inhomogeneous problem has a unique solution for every choice of f, including the given one.

B – R. $k = \pi$ is an eigenvalue with eigenfunction $\sin \pi x$. As we argued in Part (b), this eigenfunction belongs to the null space of $\mathbf{L}$, which is non-invertible. It is the only element in this null space, which is thus 1-dimensional. For the given BVP to have a solution, f must be orthogonal to $\sin \pi x$. Recalling the definition of dot product from the Lecture, we must have $\int_0^1 \sin \pi x \sin 2\pi x \; \mathrm{d}x = 0$. Either by

brute force evaluation or by graphing the functions, we can see that the integral is indeed zero. Hence the BVP has a solution. As we just mentioned, in this case one solution means infinite solutions. Please be careful here not to jump at option Q as the answer after applying FREDHOLM correctly.

C – S. Same logic as the last one, it's option R if $\int_0^1 (\sin \pi x)\left[(2x+3)\mathrm{e}^{-x}\sin \pi x\right]\mathrm{d}x$ equals zero and option S otherwise. Now look at that integral. $\sin \pi x$ is positive everywhere between 0 and 1. So is $2x+3$. And so is e^{-x}. Can their product integrate to zero over the interval ? NO. So the BVP has no solution.

D – P, Q. We've already seen that $k=1$ isn't eig so the homogeneous problem can have only the trivial solution. But that also happens to be a unique solution .

(e) The first step is to discretize the problem – since the BCs are new, we'll have to start from the beginning. We have $\mathbf{Q}_+\mathbf{y}_+ + k^2\mathbf{I}\mathbf{y} = 0$ and we'll have to use the BCs to reduce $\mathbf{Q}_+$ to $\mathbf{Q}$. Writing things out in 4-size, we have

$$\frac{1}{h^2}\begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix}\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} + k^2\mathbf{I}\mathbf{y} = 0 \quad . \tag{01}$$

Now use that $y_0 = a$ and $y_5 = b$; this gives

$$\frac{1}{h^2}\begin{bmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} + k^2\mathbf{I}\mathbf{y} = \begin{bmatrix} -a/h^2 \\ 0 \\ 0 \\ -b/h^2 \end{bmatrix} \quad \text{, or}$$

$$\left(\mathbf{Q} + k^2\mathbf{I}\right)\mathbf{y} = \mathbf{f} \quad . \tag{02}$$

If you are sceptical about the upper line here, you can easily verify it by opening it out and checking that you recover the same equations as (01) in each row. The important thing is that the inhomogeneous BCs have turned the problem into

the SAME $\mathbf{Ly}=\mathbf{f}$ structure as we had with a RHS excitation function (such a thing also happened in 4.1.5), so the conditions for existence of solutions shall be very similar to what we have already seen in Parts (a) to (c).

Evidently, if $\mathbf{L}$ is invertible, the homogeneous problem will only the trivial solution and the inhomogeneous problem will have a unique solution. Otherwise the homogeneous problem $\mathbf{Ly}=0$ will have non-trivial solutions $\mathbf{w}_1$, $\mathbf{w}_2$ etc. $\mathbf{L}$ is symmetric, so by FREDHOLM's theorem, the system will have a solution if and only if $\mathbf{f}$ is orthogonal to all the $\mathbf{w}$'s. This implies that

$$\mathbf{w}_i \cdot \left[-\frac{a}{h^2};\quad 0;\quad .....;\quad 0;\quad -\frac{b}{h^2} \right] = 0 \quad ,$$

for every $\mathbf{w}_i$, i.e.

$$-\frac{a}{h^2} w_{i1} - \frac{b}{h^2} w_{iN} = 0 \quad .$$

This expression has h's in the denominator; in the transition from discrete to continuous this tends to zero so we must be careful about how we take the limit.

Firstly, there is the scaling of the dot product; as we've already seen before, the product increases as the number of terms in the vectors so we have to multiply it by h to generate tractable expressions. So our new condition (let's drop the subscript i on $\mathbf{w}$ – we'll remember that this has to hold true for all the homogeneous solutions) is

$$-\frac{a}{h} w_1 - \frac{b}{h} w_N = 0 \quad . \qquad (03)$$

Now what are w_1 and w_N in the continuous limit ? The $\mathbf{w}$'s are non-trivial solutions of the homogeneous problem i.e. solutions which satisfy the ODE with zero BCs. $\#1$ and $\#N$ are the first and last significant points of these solutions. Although it is tempting to say that they approach closer and closer to their BC values (zero) as the discretization gets finer, that clearly turns (03) into a meaningless condition. Rather, it is $\#0$ and $\#N{+}1$ which are the boundary points and $\#1$ and $\#N$ are the BC values plus and minus the derivative times the step size. In maths,

$$w_1 = w_0 + hw'(0) \qquad w_N = w_{N+1} - hw'(L) \quad ,$$

where L denotes the right hand boundary point. In this expression I've let go of the distinction between the discrete vector $\mathbf{w}$ and its limiting function $w(x)$ [although we are doing a mathematical thing here, it's the concepts which matter and not pedantry]. Now, using that $w(0) = w(L) = 0$, and substituting that into (03), we find

$$-aw'(0) + bw'(L) = 0 \quad ,$$

which is a perfectly sensible manifestation of the orthogonal-to-null-space condition in the continuous limit.

From this point onwards, things are easy. As we saw in Part (d), the invertibility condition is $k \neq n\pi$. In the non-invertible case, the null space has only one $\mathbf{w}$, which is $\sin n\pi x$, and the orthogonality condition works out to $-n\pi a \cos 0 + n\pi b \cos n\pi = 0$. Simplifying the trig and putting everything together, we declare our final answer. If $k \neq n\pi$, the BVP has a unique solution for all a and b. If $k = n\pi$, the BVP has infinitely many solutions if $a + (-1)^{n-1} b = 0$ and no solution otherwise.

*

COMMENTS : This is the continuation and elaboration of 4.1.3. FREDHOLM's alternative theorem tells us the conditions under which an inhomogeneous BVP with symmetric operator has solutions. Note however the importance of the symmetry – it was essential for the proof in Part (c) and you cannot apply this theorem to a BVP where the operator doesn't have this property. The matrix method came in useful because it showed that $\mathbf{L}$ of this problem is indeed symmetric. An example of an operator which is not symmetric is $\mathrm{d}^2/\mathrm{d}x^2 + \mathrm{d}/\mathrm{d}x + 1$; the middle term here is represented by the matrix $\mathbf{P}$ (see T061), which isn't symmetric.

A tough examiner can really test your concepts by giving you the third BVP of Part (d) as a standalone problem – many students will plunge headlong into undet coeff (make sure you can pick the trial function!) and, since there's no option but to evaluate the coefficients, end up in an algebraic quagmire. Whenever you are faced with a BVP question, first always think FREDHOLM – go ahead with finding the solution only if you know that there is one !

Part (e) shows how inhomogeneities can enter the ODE as well as the BC on an equal footing. Both forms lead to the same kind of inhomogeneous matrix equation; the orthogonality condition is of course different in the two cases. Now you can go back to 4.1.3 and check that the BVPs with inhomogeneous BC which I'd mentioned in the comments there do indeed satisfy the criteria we obtained here. Of course, if a problem has both an excitation function and inhomogeneous BCs, then the existence criterion has to be that the sum of the integral expression in Part (d) and the derivative expression in Part (e) has to be zero.

Finally, you might wonder why we are going so deep into existence issues now, when I had summarily dismissed the proof of the IVP solution-existence theorem in 2.1 as "beyond our scope". This is because these two theorems are of a fundamentally different nature. The IVP theorem simply tells us that a well-defined IVP always has a unique solution – its proof is completely general and is applicable to linear as well as nonlinear ODEs. The flip side of generality is uselessness – the proof does not help us to actually find the solution to any ODE. It's like saying I guarantee that gold exists somewhere below the earth's surface, now good luck finding it. Here on the other hand we have a situation where the ODEs are simple LCC stuff, and we can find the solution outright; still we are running into a situation where this solution might not fit. Having arrived at the location where gold might exist, we are told from beforehand by FREDHOLM whether digging shall actually produce results. This is a far more down-to-earth phenomenon than the abstract uniqueness theorem, so it falls well within the scope of this book.

4.2

SIMPLE BEAM THEORY

In contrast to the last Unit with its emphasis on theoretical linear algebra and its connections to PDEs, we now do a standalone application of ODE BVPs. This application is beam theory. A beam is a rod shaped object which is subjected to loads perpendicular to its axis. It bends in response to these loads. The simplest theory of beams is called Euler-Bernoulli theory after its inventors, the versatile LEONHARD and the brothers JACOB and DANIEL, both of whom contributed. Here I will bypass most of the derivational stuff – you can see any book on solid mechanics for that. What I will emphasize is the equation satisfied by the deflection of a bended beam. Here, we assume that the beam is in **static equilibrium** – dynamic, out-of-equilibrium situations require a PDE for modeling and must wait for quite some time.

Nevertheless, for the sake of completeness, I will at least sketch some of the derivational features. Under the Euler-Bernoulli (also called simple beam and engineering beam) theory, the following assumptions are made about the bending process :

- The deformation of the beam is entirely perpendicular to the undeformed axis and is small compared to the length of the beam
- Cross-sections of the beam do not deform in a significant manner under application of the load
- During deformation, the cross-section remains planar and normal to the deformed axis of the beam
- The material of the beam is linear elastic

Under these assumptions, one can prove that there is a **neutral axis** inside the beam, which undergoes neither extension nor compression during the bending process. We let x be the direction of the neutral axis of the undeformed beam. The relevant variable during bending is the deflection y of the neutral axis from the x-axis. Deflection is produced by a **transverse line load i.e. a force per unit length applied on the beam, directed perpendicular**

to the neutral axis. This line load is denoted by $q(x)$. Since the deformations are small, x is the same whether measured along the undeformed axis or the deformed one. Similarly, in most cases, it does not matter whether the applied load is perpendicular to the original axis or the deformed one (one case where it does matter is coming up in 4.2.5). The YOUNG 's modulus or elastic modulus of the beam is E and the second moment of cross-sectional area is I . In practice , E and I always occur together ; the product EI is called the **flexural rigidity**. In terms of this, the equation for the shape of the beam, y(x), is

* $$EI\frac{\mathrm{d}^4 y}{\mathrm{d}x^4} = q(x) \quad . \qquad \textbf{(01)*}$$

This equation is what makes beam theory attractive for our course.

Quick note about the dimensions here : E is force per area which is F/l^2, I is area times area which is l^4, so flexural rigidity is Fl^2. Deflection y is of course l, its fourth derivative w.r.t x is l^{-3}, so $EIy^{(4)}$ is F/l. Line load $q(x)$ is defined as force per length which is F/l, so (01) is dimensionally consistent.

Since a beam has a finite length L, the problem is automatically defined in the interval [0, L]. Four conditions are necessary to make the problem complete. They are naturally prescribed at the two ends of the beam, and can be y and/or any of its derivatives upto the third. The physical interpretation of the various conditions is as follows :

- y is of course the **deflection** of the beam. If we specify y at either end, we are constraining that end to not undergo any deflection from a set value. The most common situation is to impose $y = 0$ at either or both ends i.e. to constrain the deflection to its undeformed value.
- y' is the **slope** of the beam. If we specify this at either end, we are constraining that end to not undergo any slope change relative to a set value. Once again, the most common situation is to impose $y' = 0$ at either or both ends i.e. to constrain the slope to its undeformed value.
- EIy'' is the **bending moment** – the torque applied at the end to produce and maintain the deflection.
- EIy''' is the **shear force** – a point load F (as opposed to the line load q) applied at the end to produce and maintain the deflection.

There are some restrictions on what can and cannot be specified. The primary no's are that we cannot prescribe both y and y''' at the same point, and both

y' and y'' at the same point. A few common BCs have physical nomenclature. A **fixed support** or **clamp** at an end means that $y = y' = 0$ at that end. A movable support at an end means $y' = 0$ but y can be anything. A **pin** at an end means $y = y'' = 0$ at that end – it constrains the displacement and does not exert a bending moment. A **free end** means $y''=y'''=0$ at the end – no moment or shear force applied. Finally, a **cantilever beam** is one whose one end is clamped and the other is free.

This basically completes our brief exposition of simple beam theory. We will now see some applications of the theory, and in turn use the applications to develop more BVP theory in terms of concrete examples. Despite the word "simple" in the title, this is actually not too simple a Unit, so get ready for some heavy action.

Example 4.2.1

Cute little cantilever

This is a variation on a problem which appeared in the second mid-semester examination in 2016-17/II.

A cantilever beam of length L and unit flexural rigidity has the end $x = 0$ clamped and the end $x = L$ free. It is subjected to a constant line load of strength $1/L$ all along its length.

(a) Write the BVP satisfied by the cantilever.

(b) Solve the BVP to obtain the shape of the beam.

(c) Determine the deflection of the tip of the cantilever.

[Level : 1]

(a) This is just a test of the concepts and vocabulary introduced in the Lecture. Flexural rigidity means EI and here it has the value 1. The line load is the RHS function $q(x)$, which is the constant function $1/L$. So the equation of the beam is

$$* \qquad y^{(4)} = \frac{1}{L} \quad , \qquad *$$

defined in the interval $[0, L]$. The end $x = 0$ is clamped so $\underline{y(0) = y'(0) = 0}$. The end $x = L$ is free so $\underline{y''(L) = y'''(L) = 0}$. The ODE with these four BCs is the BVP asked for.

(b) The general solution of the ODE is

$$y_G = C_0 + C_1 x + C_2 x^2 + C_3 x^3 + \frac{1}{24L} x^4 \quad .$$

The BC $y(0) = 0$ implies $C_0 = 0$ and $y'(0) = 0$ gives $C_1 = 0$ also. To fit the other BCs we calculate the second and third derivatives of y_G. We have

$$y_G'' = \frac{1}{2L} x^2 + 6C_3 x + 2C_2 \quad ,$$

$$y_G''' = \frac{1}{L} x + 6C_3 \quad .$$

Setting $x = L$ in the second relation above gives $C_3 = -1/6$; then setting that into the first relation gives $C_2 = L/4$. Hence the specific solution is

$$* \qquad y = \frac{1}{24L} x^4 - \frac{1}{6} x^3 + \frac{L}{4} x^2 \quad , \qquad *$$

which is the desired answer.

(c) To calculate the deflection of the tip, we substitute $x = L$ into the shape obtained above, getting $\underline{L^3/8}$.

Example 4.2.2

Not so cute cantilever

For the entire problem, consider a cantilever which satisfies $y^{(4)}=q(x)$ defined in the interval $[0, L]$ with homogeneous BCs $y(0) = y'(0) = y''(L) = y'''(L) = 0$.

(a) A point load (shear force) of strength F_1 is applied at the point $x = x_1$ i.e. $q(x) = F_1\,\delta(x-x_1)$. Use the fact that the deflection on either side of $x = x_1$ satisfies the homogeneous equation, apply the BCs and match the two solutions at $x = x_1$ to determine $y(x)$, the shape of the beam. Express your answer as a function defined in a piecewise manner.

(b) Calculate the deflection of the tip with $F_1 = 1$ and $x_1 = L/2$.

Although the above shape was expressed as $y(x)$, it is actually also a function of x_1. Let $G(x,x_1)$ be the shape if the point load applied at x_1 has unit strength.

(c) Express the answer of Part (a) in terms of $G(x,x_1)$.

(d) Two point loads of strengths F_1 and F_2 are applied at the points x_1 and x_2. Use the facts that the ODE is linear and the BCs are homogeneous to express the shape of the beam in terms of $G(x,x_1)$ and $G(x,x_2)$.

(e) Calculate the deflection of the tip with $F_1 = F_2 = 1/2$, $x_1 = L/3$ and $x_2 = 2L/3$.

[Level : 3]

(a) The BVP we have on hand is

$$y^{(4)} = F_1\delta(x - x_1) \quad ,$$

defined in $[0,\ L]$ with BCs $y(0) = y'(0) = y''(L) = y'''(L) = 0$. The question tells us how we should start off on the solution – we recognize that delta function is zero at all x other than $x = x_1$, so in the regions $x < x_1$ and $x > x_1$, y satisfies the homogeneous equation. Of course, it is not the SAME homogeneous solution in both regions – that would make things trivial and meaningless. The homogeneous solution y_L in the left region $x < x_1$ must satisfy the two BCs at $x = 0$, while the solution y_R in the right region $x > x_1$ must satisfy the two BCs at $x = L$. But there must be more relations connecting y_L and y_R – for one, the beam must be continuous at $x = x_1$ so $y_L(x_1)$ must equal $y_R(x_1)$; higher derivatives might be equal also, we'll have to see. These conditions are the matching conditions; the BCs together with these will determine y_L and y_R completely.

To obtain the matching conditions, we use the familiar trick of integrating across the delta function. We have

$$\int_{(x_1)-\varepsilon}^{(x_1)+\varepsilon} y^{(4)}\,dx = \int_{(x_1)-\varepsilon}^{(x_1)+\varepsilon} F_1\delta(x - x_1)\,dx \quad , \text{ or}$$

$$y'''(x_1 + \varepsilon) - y'''(x_1 - \varepsilon) = F_1 \quad .$$

As $\varepsilon \to 0$, $y'''(x_1+\varepsilon)$ and $y'''(x_1-\varepsilon)$ reduce to $y_L'''(x_1)$ and $y_R'''(x_1)$ respectively, so we have one matching condition between the two functions. Integrating one more time shows that the second derivative y'' jumps by an infinitesimal amount at x_1 i.e. the second derivative is CONTINUOUS at $x = x_1$ (recall here the comments to 1.2.1). The still lower order derivatives are then continuous by default. Hence we have our matching conditions

$$\begin{aligned} y_L(x_1) &= y_R(x_1) \\ y_L'(x_1) &= y_R'(x_1) \\ y_L''(x_1) &= y_R''(x_1) \\ y_L'''(x_1) - y_R'''(x_1) &= -F_1 \end{aligned}$$

The algorithm for finding the beam shape is now on the table.

We let

$$y_L = A_1x^3 + B_1x^2 + C_1x + D_1 \quad \text{, and}$$
$$y_R = A_2x^3 + B_2x^2 + C_2x + D_2 \quad ,$$

where all the constants are unknown. The BCs $y_L(0) = y_L'(0) = 0$ give $C_1 = D_1 = 0$. The BCs $y_R''(L) = y_R'''(L) = 0$ give

$$\begin{aligned} 6A_2 &= 0 \\ 6A_2L + 2B_2 &= 0 \end{aligned} \quad ,$$

implying $A_2 = B_2 = 0$. Hence we are left with A_1, B_1, C_2 and D_2.

The matching conditions, in the order I presented them, yield

$$\begin{aligned} A_1x_1^3 + B_1x_1^2 &= C_2x_1 + D_2 \quad , \\ 3A_1x_1^2 + 2B_1x_1 &= C_2 \quad , \\ 6A_1x_1 + 2B_1 &= 0 \quad , \\ 6A_1 + F_1 &= 0 \quad . \end{aligned}$$

This system has the solution

$$A_1 = -\frac{F_1}{6}, \quad B_1 = \frac{F_1x_1}{2}, \quad C_2 = \frac{F_1x_1^2}{2}, \quad D_2 = -\frac{F_1x_1^3}{6} \quad .$$

Finally, we can use these to write the beam shape as

$$y(x) = \begin{cases} -\dfrac{F_1}{6}x^3 + \dfrac{F_1x_1}{2}x^2 & \text{if } x < x_1 \\ \dfrac{F_1x_1^2}{2}x - \dfrac{F_1x_1^3}{6} & \text{if } x > x_1 \end{cases} \quad , \qquad *$$

which is a piecewise defined function and is the desired answer.

(b) Plugging $F_1 = 1$, $x_1 = L/2$ and $x = L$ into the shape above gives the tip deflection as $\underline{5L^3/48}$.

(c) By definition, $G(x,x_1)$ is the shape with F_1 replaced by unity, hence

$$G(x,x_1) = \begin{cases} -\dfrac{1}{6}x^3 + \dfrac{1}{2}x_1x^2 & \text{if } x < x_1 \\ \dfrac{1}{2}x_1^2x - \dfrac{1}{6}x_1^3 & \text{if } x > x_1 \end{cases} \quad . \qquad (01)$$

Evidently, the shape obtained in Part (a) is $\underline{F_1G(x,x_1)}$.

(d) Although this might appear like an off-topic start, we first verify that the homogeneous BVP has only the trivial solution i.e. the operator $\mathbf{L}=\mathrm{d}^4/\mathrm{d}x^4$ with the given BCs is invertible. The solution to $\mathbf{L}y=0$ is $y=Ax^3+Bx^2+Cx+D$; the BCs at $x=0$ give $C=D=0$ while the BCs at $x=L$ give $A=B=0$. Hence $\mathbf{L}$ is invertible and the inhomogeneous problem has a unique solution.

Our BVP now is

$$y^{(4)} = F_1\delta(x-x_1)+F_2\delta(x-x_2) \quad ,$$

defined in the usual interval $[0,\ L]$ and having the homogeneous BCs $y(0) = y'(0)=y''(L)=y'''(L)=0$. We start by claiming the answer :

* $$y = F_1G(x,x_1)+F_2G(x,x_2) \quad , \qquad (02)^*$$

and then show that it is true. Since the ODE is linear, it is true by definition that $\mathbf{L}(c_1y_1+c_2y_2) = c_1\mathbf{L}y_1 + c_2\mathbf{L}y_2$ for all constants c_1 and c_2. By definition of G, $\mathbf{L}G(x,x_1)=\delta(x-x_1)$ and $\mathbf{L}G(x,x_2)=\delta(x-x_2)$, so

$$\mathbf{L}\left[F_1G(x,x_1)+F_2G(x,x_2)\right]=F_1\delta(x-x_1)+F_2\delta(x-x_2) \quad .$$

Since we already showed that the problem has a unique solution, and since we just found a solution (02) by explicit construction, that must be THE solution of the BVP and is the desired answer.

(e) Plugging in the values gives the tip deflection as $\underline{L^3/9}$.

*

COMMENTS : A point load on a beam is a very realistic physical situation, occurring when say a weight is hung from the beam at some point. All point load problems can be solved this way, whatever the BCs. At the same time, a point load of strength 1 at the centre, and two point loads of strength 1/2 at equal spacings can be thought of as crude discretizations of the uniform line load of total strength 1 considered in the last Tute. For one point load, the deflection of the tip was $5L^3/48$, which is close to and somewhat less than the $L^3/8$ obtained for the continuous load. For two points, the answer of $L^3/9$ is closer. We'll take the limit discrete → continuous in the next problem.

Example 4.2.3

Point and line loads

Again consider the cantilever from the previous problem, $y^{(4)} = q(x)$ defined in the interval [0, L] with homogeneous BCs $y(0) = y'(0) = y''(L) = y'''(L) = 0$.

(a) Suppose N point loads F_i (i goes from 1 to N) are applied at N points x_i. Express the shape of the beam in terms of $G(x,x_i)$ where this function is defined in the previous problem.

(b) For equispaced, equal loads summing to 1, use a computer to calculate the deflection of the tip as N is varied.

N is now allowed go to infinity. The discrete variable x_i becomes a continuous variable x' and a line load $q(x)$ can be treated as an infinitesimal point load at every point x'.

(c) Express the underlined sentiment as a mathematical equation.

(d) Use it to obtain the shape of the beam in terms of $G(x,x')$.

(e) For $q(x) = 1/L$, evaluate the above integral over x'.

(f) Show that for arbitrary applied line load $q(x)$ which may be piecewise zero but does not change sign anywhere along the beam, the maximum deflection occurs at the tip.

[Level : 3]

(a) This is just a continuation of the results of the previous problem; from linearity, the shape is

$$y(x)=\sum_{i=1}^{N}F_i G(x,x_i) \quad , \qquad (01)^*$$

where $G(x,x_i)$ is defined as

$$G(x,x_i)=\begin{cases}-\frac{1}{6}x^3+\frac{1}{2}x_i x^2 & \text{if } x<x_i \\ \frac{1}{2}x_i^2 x-\frac{1}{6}x_i^3 & \text{if } x>x_i\end{cases} \quad , \qquad (02)$$

i.e. G is as in (4.2.2.01) with x_1 replaced by x_i.

(b) A simple Matlab routine achieves this. We report the number obtained after dividing $y(L)$ by L^3; for $N=10$, this number is 0.1212, for $N=30$ it is 0.1237 and for $N=100$ it is 0.1246. The limiting answer is of course 0.125, so the discrete version is approaching that value as the grid gets finer.

(c) This is easy so long as we are careful about which is x and which is x' (the prime notation, same as for derivative, is a bit confusing but universal convention). The line load q is a function of x, as usual. In the discrete form, the point loads F_i are defined at the points x_i so they are functions of x_i, and $q(x)$ has the form

$$q(x)=\sum_{i=1}^{N}F_i\delta(x-x_i) \quad .$$

Now x_i turns into x' in the limit, so the infinitesimal point loads F_i become functions of x' rather than x. These loads are $q(x')\mathrm{d}x'$ (check that this is dimensionally correct : a line load q times a displacement $\mathrm{d}x'$ gives a point load). The delta function remains as it was and the sum from 1 to N becomes an integral from 0 to L :

$$q(x)=\int_0^L q(x')\delta(x-x')\mathrm{d}x' \quad . \qquad *$$

It is a ready check that the integral on the above RHS indeed reduces to the LHS, hence the above is the equation asked for.

(d) This time, we want the continuous equivalent of (01); the process is the same as what we did in the last step. The point loads are $q(x')\,dx'$, while $G(x,x')$ is the beam shape due to a unit point load at x'. It is the same as (02) with x_i replaced by x'. For later reference, I'll again display it explicitly :

$$G(x,x') = \begin{cases} -\frac{1}{6}x^3 + \frac{1}{2}x'x^2 & \text{if } x < x' \\ \frac{1}{2}x'^2 x - \frac{1}{6}x'^3 & \text{if } x > x' \end{cases} \quad . \qquad (03)$$

The sum from 1 to N now becomes an integral from 0 to L, thus

$$* \qquad y(x) = \int_0^L q(x')G(x,x')\,dx' \quad , \qquad *$$

which is the desired answer.

(e) Before starting the evaluation, here's an extremely useful exam strategy. This question could have said "evaluate the expression obtained in Part (d)". Instead it tells you that the result of that part should be an integral over x'. The purpose of this hint is simple – this is a multi-part and hard problem and if you have made a mistake somewhere before, then you will automatically lose all the marks for this and the next part, and hence the bulk of the total credit. So the hint gives you an opportunity for an internal check – if your answer to Part (d) is say an integral over x, you immediately know that you've made a mistake somewhere and you get a chance to go back and correct it before embarking on these latter two parts. In exams, especially when faced with multi-part problems, always be aware of such internal hints which the examiner might give out to keep you on track.

Since $q(x') = 1/L$, the integral is

$$y(x) = \frac{1}{L}\int_0^L G(x,x')\,dx' \quad .$$

$G(x,x')$ is defined piecewise, so the integral will be defined piecewise also. This time, since the integral is over x', we 'reverse the roles' of x and x' in the definition of $G(x,x')$, getting

$$G(x,x') = \begin{cases} -\frac{1}{6}x'^3 + \frac{1}{2}xx'^2 & \text{if } x' < x \\ \frac{1}{2}x^2x' - \frac{1}{6}x^3 & \text{if } x' > x \end{cases} \quad . \qquad (04)$$

Note that the above is the SAME $G(x,x')$ as in (03) just written in a different manner ! In particular, it is NOT " $G(x',x)$ " – that is a garbage quantity as there is no such thing as the shape of the beam as a function of x' with a point load applied at x !

Equation (04) suggests that the integral over x' must be split at the value $x' = x$. The integral then is

$$y(x) = \frac{1}{L}\left[\int_0^x \left(-\frac{1}{6}x'^3 + \frac{1}{2}xx'^2\right)dx' + \int_x^L \left(\frac{1}{2}x^2x' - \frac{1}{6}x^3\right)dx'\right] \quad ,$$

which evaluates to

$$y(x) = \frac{1}{L}\left(\left[-\frac{1}{24}x'^4 + \frac{1}{6}xx'^3\right]_0^x + \left[\frac{1}{4}x^2x'^2 - \frac{1}{6}x^3x'\right]_x^L\right) \quad .$$

This last expression has the value

* $$y(x) = \frac{1}{24L}x^4 - \frac{1}{6}x^3 + \frac{L}{4}x^2 \quad ,$$ *

which is the desired answer. We note that this is identical to the result of 4.2.1, which had better be the case – that problem and this part are the same !

(f) We assume without loss of generality that $q(x)$ is zero or positive everywhere, which is called positive semi-definite. Start from a point load of unit strength being applied at arbitrary point x' so that the response is $G(x,x')$. Using the definition (03) of $G(x,x')$ we compute the derivative dG/dx. Since G is a function of two variables, it is better to express this as a partial derivative rather than an ordinary one (note however that the problem does NOT become a PDE!). We have

$$\frac{\partial}{\partial x}G(x,x') = \begin{cases} -\frac{1}{2}x^2 + x'x & \text{if } x < x' \\ \frac{1}{2}x'^2 & \text{if } x > x' \end{cases} \quad .$$

In the first region, the derivative is $x(x'-x/2)$ which is positive since x is positive and smaller than x'. In the second region, the derivative is trivially positive. These relations hold for all x', so we have

$$\frac{\partial G}{\partial x} > 0 \quad ,$$

for all x and x'.

This now implies that if a unit point load is applied at any x', the maximum deflection occurs at the tip $x = L$. Evidently, the same holds true if the unit point load is replaced by any positive point load. Since the positive semi-definite line load $q(x)$ can be split up into a (infinite) number of positive (and zero – a trivial case) point loads and the maximum deflection for each contributing point load will occur at the tip, and since the resulting shape of the beam is the sum of the shapes in response to each individual point load, the maximum deflection of this sum will also occur at the tip.

We note that the argument also holds true if $q(x)$ is negative or zero throughout – the tip will be the point with minimum value of y (again the maximum deflection). If $q(x)$ changes sign over the beam however, then for some x' the tip will be a maximum, for some x' it will be a minimum and no conclusion can be drawn about their sum.

*

COMMENTS : If I were in your place, I would be livid at the author of the book after being shown Part (e) – five or six pages of fancy footwork to arrive at a result which was unquestionably simple before. So why did I do it ?

What I was doing was introducing the technique called GREEN's function (you guessed it, that's G), which is a powerful method of solving inhomogeneous problems with homogeneous boundary conditions. The method is useful for both ODEs and PDEs, so it makes sense to get acquainted with it in the easier situation. In Part (f) here, we got a reward for our efforts. There we proved a conclusion which is non-trivial because it is valid for ARBITRARY load functions – I don't know of any other technique which can demonstrate this. GREEN's function is typically considered a difficult topic and rarely if ever included in courses below the exclusively graduate level. It is true that the evaluation of G

for a PDE given the BCs is often a hair-raising task. But the basic concepts underlying this method are nothing fancy.

Note that although we used a specific example of cantilever to develop the method, it works for ANY linear problem with homogeneous BCs. One of the steps involved is the expression of the continuous function (the inhomogeneity or 'excitation') as the limit of a series of delta functions – this is very general and has no relation to the specific problem under consideration. The other key step is being able to write the combined response as the sum of (or integral over) the responses to the various delta functions. This step was initiated in Part (d) of the previous problem and extended in Parts (a) to (d) of this one. All these steps required only the facts that the ODE was linear and that the BCs were homogeneous; nowhere did we invoke the specific forms of the ODE or of the solutions. The assumption which we DID make was that the operator $\mathbf{L}$ is invertible [in 4.2.2 Part (d)]; if that doesn't hold true also, then we're looking at trouble. Fortunately, practical situations do feature invertible operators so we will neglect this case from further consideration.

We can look at GREEN's function from another perspective if we think in terms of the discretized representation of the problem. The delta function $\delta(x–x')$ is equivalent to the vector $\mathbf{d}_k = (1/h)\,\mathbf{e}_k$ where $\mathbf{e}_k$ is the vector with 1 in position k and zero everywhere else (x_k corresponds to the location x' and k can be anything between 1 and N, the system size). The scaling of $1/h$ ensures that h times the dot product (i.e. the discretized integral) of any arbitrary vector $\mathbf{f}$ and $\mathbf{d}_k$ gives f_k as it should. Then, the discretized GREEN's function $\mathbf{g}_k$ is a vector such that $\mathbf{L}\mathbf{g}_k = \mathbf{d}_k$. Since $\mathbf{L}$ is invertible, $\mathbf{g}_k$ exists for all k.

Now suppose we want to solve $\mathbf{L}\mathbf{y} = \mathbf{f}$, where $\mathbf{f} = [f_1; f_2; \ldots..; f_N]$ is an arbitrary vector. Evidently, we can write $\mathbf{f} = f_1\,\mathbf{e}_1 + f_2\,\mathbf{e}_2 + \ldots.. + f_N\,\mathbf{e}_N$. But by definition of $\mathbf{g}_k$, $\mathbf{L}(h\mathbf{g}_k) = \mathbf{e}_k$. Then by linearity of $\mathbf{L}$, $\mathbf{L}\mathbf{y} = \mathbf{f}$ must be solved by $\mathbf{y} = f_1\,h\mathbf{g}_1 + f_2\,h\mathbf{g}_2 + \ldots.. + f_N\,h\mathbf{g}_N = h\sum f_k\mathbf{g}_k$. Since the subscript k corresponds to x', the continuous limit of this solution is the integral $y = \int_0^L f(x')G(x,x')\,\mathrm{d}x'$.

Example 4.2.4

Cantilever by matrix method

A cantilever satisfies the ODE

$$\frac{\mathrm{d}^4 y}{\mathrm{d}x^4} = c \quad ,$$

in the interval $[0, L]$ with BCs $y(0) = y'(0) = y''(L) = y'''(L) = 0$. This ODE will be solved using matrix method. The numerical values to be taken are $L = 7$ and $c = -0.0012544$ (which will be useful in the next problem).

(a) Analytically determine the deflection of the tip.

(b) Obtain an expression for the centred discrete fourth derivative and hence for the raw fourth derivative matrix $\mathbf{S}_+$.

The vector $\mathbf{y}_+$ is defined with two additional boundary points at each end of the significant region : #–1 and #0 at the left and #N+1 and #N+2 at the right. To implement the BCs, y_{-1} is assumed to be zero, a right side first derivative is taken at site #–1, a centred second derivative taken at site #N and a right side third derivative again at site #N.

(c) Impose the BCs as prescribed above to convert $\mathbf{S}_+$ to a square $\mathbf{S}$.

(d) Complete the solution of the problem using matrix method and determine the deflection of the tip.

[Level : 3]

(a) In 4.2.1 we saw that with a constant line load of strength $1/L$, the tip deflection came out to be $L^3/8$. Since the deflection is a linear function of the load, we can easily see that if the load strength is c, then the deflection will be $cL^4/8$. Plugging in the numbers, this evaluates to 0.3765 (if we ignore the sign which merely indicates that the beam hangs "downwards" as per some arbitrarily defined convention). The purpose of this simple calculation is of course to establish a comparison benchmark for the numerical computation.

Parts (b) and (c) of the question are essentially the two steps of the algorithm outlined in 4.1.5.

(b) In the preliminary step, we discretize the interval $[0, L]$ is into N significant points and two boundary points on each side as the question says (of course, four extra points are needed for a problem with fourth derivative). Since $x = 0$ corresponds to site $\#{-1}$ and $x = L$ to site $\#N{+}2$, the step size h thus becomes $L/(N{+}3)$. We then start the calculation proper by deriving the expressions for higher order discrete derivatives. For third derivative, we need the derivative of the second derivative. Immediately we have a choice of left or right side, and for now we do both. In the left side form we have

$$y'''_i = \frac{1}{h}\left(y''_i - y''_{i-1}\right) \quad \text{, or}$$

$$y'''_i = \frac{1}{h}\left(\frac{y_{i+1} - 2y_i + y_{i-1}}{h^2} - \frac{y_i - 2y_{i-1} + y_{i-2}}{h^2}\right) = \frac{y_{i+1} - 3y_i + 3y_{i-1} - y_{i-2}}{h^3} \quad . \qquad \text{(01L)}$$

In the right side form we have

$$y'''_i = \frac{1}{h}\left(y''_{i+1} - y''_i\right) = \frac{y_{i+2} - 3y_{i+1} + 3y_i - y_{i-1}}{h^3} \quad . \qquad \text{(01R)}$$

Just as for the second derivative, for the centred fourth derivative we either take a right side on (01L) or a left side on (01R). If you are good at number pattern recognition, you will see a trend here. In the denominator, n^{th} derivative has h^n. In the numerator, starting from the appropriate leading term, first derivative has

the coefficients $1\ -1$, second derivative has $1\ -2\ 1$, third derivative has $1\ -3\ 3\ -1$ and so on. Either by guesswork or by calculation, we arrive at the result

$$y^{(4)}{}_i = \frac{y_{i+2} - 4y_{i+1} + 6y_i - 4y_{i-1} + y_{i-2}}{h^4} \quad , \qquad (02C)$$

the 'C' standing for 'centred', and the solution is underway.

Just as in 4.1.5-6, the raw fourth derivative matrix $\mathbf{S}_+$ contains the formula (02C) in every row, arranged in a stepwise manner. For $N = 4$ it looks like

$$\mathbf{S}_+ = \frac{1}{h^4}\begin{bmatrix} 1 & -4 & 6 & -4 & 1 & 0 & 0 & 0 \\ 0 & 1 & -4 & 6 & -4 & 1 & 0 & 0 \\ 0 & 0 & 1 & -4 & 6 & -4 & 1 & 0 \\ 0 & 0 & 0 & 1 & -4 & 6 & -4 & 1 \end{bmatrix} \quad .$$

and is the desired answer.

(c) As in 4.1.5-6, we have $\mathbf{y}_{(4)} = \mathbf{S}_+\mathbf{y}_+$ and we now have to impose the BCs to convert $\mathbf{S}_+$ to $\mathbf{S}$. For conceptual clarity, I will display the product $\mathbf{S}_+\mathbf{y}_+$ below :

$$\mathbf{y}^{(4)} = \mathbf{S}_+\mathbf{y}_+ = \frac{1}{h^4}\begin{bmatrix} 1 & -4 & 6 & -4 & 1 & 0 & 0 & 0 \\ 0 & 1 & -4 & 6 & -4 & 1 & 0 & 0 \\ 0 & 0 & 1 & -4 & 6 & -4 & 1 & 0 \\ 0 & 0 & 0 & 1 & -4 & 6 & -4 & 1 \end{bmatrix} \begin{bmatrix} y_{-1} \\ y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} \quad . \qquad (03)$$

The first BC is of course $y_{-1} = 0$. The next is $y'(0) = 0$. The question tells us to take a right side first derivative at site $\#-1$; this yields $y'_{-1} = (y_0 - y_{-1})/h$. Setting it equal to zero gives $y_0 = 0$ also. So y_{-1} and y_0 are both zero and when we impose these conditions, the first two columns of $\mathbf{S}_+$ go away with no modification.

The next BC is $y''(L) = 0$. As per instructions we take the double derivative at the last significant lattice site $\#N$, getting

$$\frac{y_{N+1} - 2y_N + y_{N-1}}{h^2} = 0 \quad ,$$

which is a relation connecting the boundary point y_{N+1} to two internal points y_N and y_{N-1}. Finally we have the BC $y'''(L)=0$. We take the right side third derivative (01R) at site $\#N$, getting

$$\frac{y_{N+2}-3y_{N+1}+3y_N-y_{N-1}}{h^3}=0 \quad ,$$

a second relation between the internal and boundary points. These two equations now yield

$$y_{N+1}=-y_{N-1}+2y_N \quad ,$$
$$y_{N+2}=-2y_{N-1}+3y_N \quad .$$

As an aside we note that if we'd taken the BC derivatives differently, say second derivative at site $\#N+1$ instead of N, then the calculations would have been more complicated but the end result would have been the same in the limit $h\to 0$.

We plug these results into (03) with the first two columns of $\mathbf{S}_+$ and the first two elements of $\mathbf{y}_+$ lopped off :

$$\mathbf{y}^{(4)}=\frac{1}{h^4}\begin{bmatrix} 6 & -4 & 1 & 0 & 0 & 0 \\ -4 & 6 & -4 & 1 & 0 & 0 \\ 1 & -4 & 6 & -4 & 1 & 0 \\ 0 & 1 & -4 & 6 & -4 & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ -y_3+2y_4 \\ -2y_3+3y_4 \end{bmatrix} \quad .$$

To generate the square $\mathbf{S}$, we now try to absorb the last two columns of the above matrix into the remaining ones. The first two rows are trivial since in these rows the offending fifth and sixth columns there have zeroes anyway. So we expand the last two rows of the matrix. Opening out the third row, we get the equation $y^{(4)}{}_3 = y_1-4y_2+6y_3-4y_4+(-y_3+2y_4)$ or $y^{(4)}{}_3 = y_1-4y_2+5y_3-2y_4$. Similarly, the fourth row opens out to $y^{(4)}{}_4=y_2-2y_3+y_4$. We can write these in a matrix form as

$$\mathbf{y}^{(4)} = \begin{bmatrix} 6 & -4 & 1 & 0 \\ -4 & 6 & 4 & 1 \\ 1 & -4 & 5 & -2 \\ 0 & 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \quad .$$

Although this was for a 4x4 matrix, the procedure will remain the same for arbitrary N since only the four bottom right corner elements will get modified on account of the BC. Hence we can write our target matrix $\mathbf{S}$ as follows (with $N = 6$ instead of 4 for increased clarity) :

$$\mathbf{y}^{(4)} = \mathbf{S}\mathbf{y} \quad \text{, where}$$

$$\mathbf{S} = \frac{1}{h^4} \begin{bmatrix} 6 & -4 & 1 & 0 & 0 & 0 \\ -4 & 6 & -4 & 1 & 0 & 0 \\ 1 & -4 & 6 & -4 & 1 & 0 \\ 0 & 1 & -4 & 6 & -4 & 1 \\ 0 & 0 & 1 & -4 & 5 & -2 \\ 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix} \quad .$$

This completes the big step of the problem. The discretized ODE will have the form $\mathbf{S}\mathbf{y} = \mathbf{b}$ where $\mathbf{b}$ is the excitation vector.

(d) To wrap up the problem, we determine $\mathbf{b}$. That is almost trivial :

$$\mathbf{b} = -0.0012544\left[1; \quad 1; \quad .....; \quad 1\right] \quad ,$$

where the '1's repeat N times. The code which implements $\mathbf{S}\mathbf{y} = \mathbf{b}$ is pretty simple to write. I varied N based on hit and trial comparison with the answer of Part (a). At $N = 3000$, I am getting a static deflection of 0.3754, which shows an accuracy level of about 3 in 1000. This is close to the maximum accuracy which I found while varying N.

*

COMMENTS : Initially the accuracy of the numerical solution increases with increasing N, which is perfectly plausible. At very high N however (greater than 10000), the solution again tends to become less accurate, probably on account of errors accrued while Matlab carries out the step $\mathbf{S}\backslash\mathbf{b}$. A more detailed analysis of this process and the resulting error will take us into hardcore numerical linear algebra, about which you know as much as I do. So, for the

present purposes, we'll just be satisfied with the accuracy check and won't go into the details of wherefrom the errors arise and how they can be minimized.

This problem was a demo of the matrix method – the technique wasn't really indispensable because we had the analytical solution of the ODE to begin with. But in the next Tute we'll use the matrix $\mathbf{S}$ to numerically solve a meaningful problem which has no analytical solution.

Example 4.2.5

Helicopter fan blade

Each blade of the main fan (or rotor) of a helicopter can be modelled as a cantilever of uniform cross-section which is clamped at the centre of the fan ($x = 0$) and free at the far end ($x = L$). For a typical helicopter, the blade length is 7 m and the cross-section approximately rectangular with breadth 0.8 m and thickness 0.004 m (4 mm). The second moment of cross-sectional area is $7.5\text{x}10^{-6}$ m^4, the elastic modulus of the material is 10^{11} N/m^2 (100 GPa) and the density is 3000 kg/m^3. The undeformed shape of the blade in the absence of all forces is perfectly horizontal.

When the helicopter is dead, the blade hangs under its own weight.

(a) Determine the deflection of the tip.

When the helicopter is live, the blade rotates with angular velocity ω about a vertical axis, giving rise to a purely horizontal centrifugal force $m\omega^2 x$ on each mass element m of the blade. This force acts in addition to the weight. Take the rotation speed to be 35 rad/s (340 rpm).

(b) Write a BVP for the shape of the rotating blade.

(c) Try a solution to the above BVP in the form of a power series with unknown coefficients. Take the first four non-trivial terms of this series and use a computer to evaluate the coefficients. Make sure that your trial function satisfies all the BCs.

(d) Plot your solution and report the deflection of the tip.

(e) Since the four-term series is grossly approximate, discuss whether the solution is plausible on physical grounds. This analysis must be standalone; no reference may be drawn to the numerical solution of the subsequent part.

(f) Drawing reference to the previous problem, solve the BVP for the rotating blade using the matrix method. Plot your solution and report the deflection of the tip.

[Level : 4]

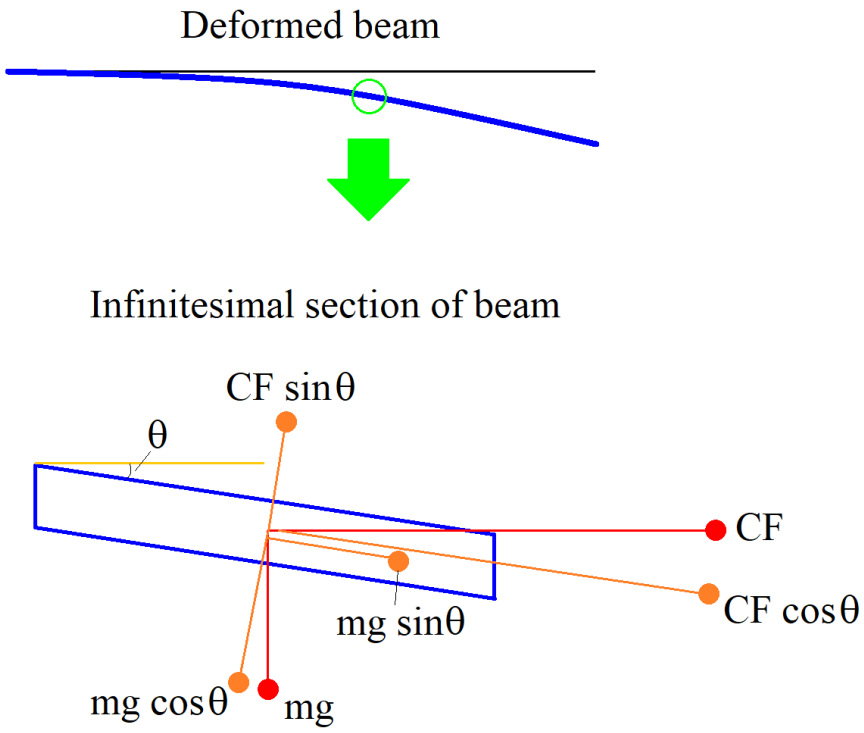
Our first four has taken quite a while in coming – we are more than halfway through the book so far as Tutorial numbers are concerned. But only a rash or desperate batsman goes for big hits too quickly. The mature hitter sizes up the bowlers, the pitch and the outfield, settles into his groove and only then aims for the boundary. Once set however, he can shoot for glory whenever an opportunity presents itself. The same holds true for us – we have to be solid on the basics before getting ambitious. But after the foundations are built, there are no holds on our aspirations – indeed, in the second Part of the book we're going to smash that ball all around the park.

(a) The previous example featured that $y^{(4)} = -0.0012544$ with cantilever BCs at $x = 0$ and $x = 7$. It clearly mentioned that this equation appears in the next (i.e. this) problem. Sure enough, the length of the blade here is 7 m and gravity acts downwards so the load should have a negative sign. The tarot cards tell us that q/EI is 0.0012544 SI Units, so the deflection of the tip is 0.3765 m as per T067 Part (a).

In case you aren't happy with fortune-telling, let's get that figure 0.0012544 rigorously. E and I are given directly, so out comes the flexural rigidity, equal to 750000 Nm^2. The total mass of the blade is volume times density which is 7 x 0.8 x 0.004 x 3000 giving 672 kg. The total weight is g times the mass, which is 9.8 x 672 i.e. 6586 N, and dividing by the length gives the line load 940.9 N/m. Now dividing by EI gives the magic number 0.0012544 on the RHS of the ODE and the tip deflection of 37.65 cm follows. Since all parameters are hardly specified to four-place accuracy, it makes more sense to report the deflection as 38 cm.

(b) Let y denote the vertically upward direction so that the beam weight and deflection are in the negative y-direction. Now, first a tempting but garbage answer : since the centrifugal force is purely horizontal, that is directed along the axis of the beam. Bending of beams is produced by transverse loads and not axial ones, so the centrifugal force goes out of the bending equation, leaving us with $EIy^{(4)} = -mg/L$.

If nothing else, the brouhaha being made in Parts (c-e) about the BVP which you should obtain here should immediately alert you that this equation is wrong. Now, let's see what's right. First we consider the free body diagram of an infinitesimal section of the beam, shown in the Figure.



This section is inclined at angle θ to the horizontal, where we take θ to be finite rather than small – we'll do the limit only at the end. The length of the section is $\mathrm{d}x/\cos\theta$ and its mass is ρ_1 times that where ρ_1, the "one-dimensional density", denotes mass per unit length. The forces acting on it are gravity, which is vertical and centrifugal force (CF) which is horizontal. Gravity splits into a $\cos\theta$ component transverse to the element and a $\sin\theta$ component parallel to the element. The former produces bending while the latter contributes to axial stress and is rejected. Centrifugal force splits into a $\sin\theta$ bender and a $\cos\theta$ axial puller which latter is again thrown away. We are given that (and hopefully know from mechanics that) centrifugal force on the element is $(\mathrm{d}m)\,\omega^2 x$. Thus, the bending forces are $-(\mathrm{d}m)g\cos\theta$ and $(\mathrm{d}m)\,\omega^2 x\sin\theta$. Only now, take the limit θ is very small. The infinitesimal mass becomes $\rho_1\,\mathrm{d}x$, the bending component of gravity becomes $-(\rho_1\mathrm{d}x)g$ and the bending component of centrifugal force becomes $(\rho_1\,\mathrm{d}x)\,\omega^2 x\theta$. But, for small angles, $\mathrm{d}y/\mathrm{d}x = -\tan\theta$ (the negative because we have taken θ to be clockwise positive), and $\tan\theta$ also tends to θ, so θ is actually $-y'$ and the centrifugal bending term at location x is $-(\rho_1\mathrm{d}x)\,\omega^2 xy'$.

Finally, to get the force per unit length, we throw out the factor of dx from the force on the element and write the bending equation as

$$EIy^{(4)} = -\rho_1 g - \rho_1 \omega^2 xy' \quad \text{, or}$$
$$EIy^{(4)} + \rho_1 \omega^2 xy' = -\rho_1 g \quad .$$

This is unlike any other equation we have seen so far.

Now we have to plug in numerical values. From Part (a), the mass of 7 m of blade is 672 kg so the mass of 1 m is 96 kg and $\rho_1 = 96$. With $\omega = 35$, we get

$$750000\,y^{(4)} + 117600\,xy' = -940.8 \quad \text{, or} \qquad (01a)$$

* $$y^{(4)} + 0.1568\,xy' = -0.0012544 \quad . \qquad (01b)^*$$

The BCs are usual cantilever and $L = 7$, so $\underline{y(0) = y'(0) = y''(7) = y'''(7) = 0}$. The ODE together with these BCs make up the BVP asked for.

(c) Since the ODE (01b) cannot (at least I don't know how to) be solved in closed form, we try the power series solution. This means to take a trial solution

$$y = A_0 + A_1 x + A_2 x^2 + \ldots\ldots \quad ,$$

where the coefficients are unknown and can be determined by plugging the trial into the ODE and BCs. Evidently, since $y(0)$ and $y'(0)$ are zero, the terms A_0 and A_1 must both be zero, giving a trial solution

$$y = Ax^2 + Bx^3 + Cx^4 + Dx^5 \quad . \qquad (02)$$

(For serieses with small number of terms, I find the A, B, C….. notation easier than A_1, A_2, A_3…..). The ODE then becomes

$$24C + 120Dx + \alpha\left(2Ax^2 + 3Bx^3 + 4Cx^4\right) = -\beta \quad ,$$

where α and β denote 0.1568 and 0.0012544 respectively. We must now equate the LHS and RHS order by order starting from the lowest. Each order will yield one equation connecting A, B, C and D. However, we have to fit the BCs at the free end so we must keep some of the unknowns in reserve for that. Two BCs will imply two constraints among A, B, C and D, so the equating terms process must produce only two additional equations. Hence, in this step we should equate only the x^0 and x^1 terms on each side. The first one says $24C = -\beta$ while the second gives $D = 0$. This makes the series (02) have three non-trivial terms rather than four, as required in the question.

So we add to the trial (02) another term Ex^6 and again plug into the ODE to get

$$24C + 360Ex^2 + \alpha\left(2Ax^2 + 3Bx^3 + 4Cx^4 + 6Ex^6\right) = -\beta \quad .$$

Now we get $24C = -\beta$ at the constant level and $2\alpha A + 360E = 0$ at the x^2 level. The two additional constraints come from the BCs. We have

$$y'' = 2A + 6Bx + 12Cx^2 + 30Ex^4 \quad ,$$

$$y''' = 6B + 36Cx + 120Ex^3 \quad .$$

Substituting $x = L$ into this relation and then plugging in numbers for everything yields the following linear system for A, B, C and E :

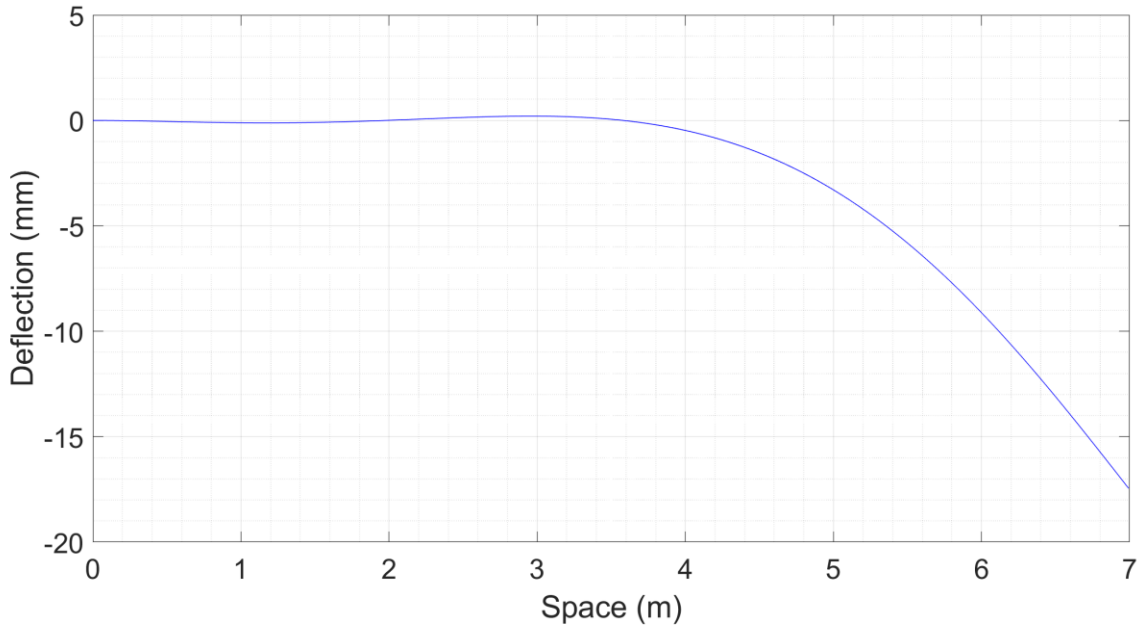
$$\begin{bmatrix} 0 & 0 & 24 & 0 \\ 0.3136 & 0 & 0 & 360 \\ 2 & 42 & 588 & 72030 \\ 0 & 6 & 252 & 41160 \end{bmatrix} \begin{bmatrix} A \\ B \\ C \\ E \end{bmatrix} = \begin{bmatrix} -0.001254 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad .$$

A computer is one of the wonders of the modern world; it takes a second to solve the above system and spit out the answer $[A;\ B;\ C;\ E] = 10^{-3}[-0.3231;\ 0.2644;\ -0.0523;\ 0.0003]$. So we claim the four-term power series solution

$$* \qquad y = \frac{1}{1000}\left(-0.3231x^2 + 0.2644x^3 - 0.0523x^4 + 0.0003x^6\right) \quad , \qquad *$$

which, accurate or otherwise, is the desired answer.

(d) The plot is shown below. The axes are NOT in 1:1 scale !



The deflection of the tip is 1.75 cm.

(e) The most striking feature of the solution is that the deflection of the tip has reduced by a factor of 20 (recall it was 38 cm before). This is consistent with what we know from observation – that the blades of a helicopter droop when static and appear straight as ramrods when running. It is also logical – centrifugal force provides a load opposite to the direction of gravity so it should raise the beam rather than lowering it. The force is also many times stronger than gravity – at the midpoint of the blade, $\omega^2 x = 4500$ or so, while g is merely 10 ! Very crudely, the average slope of the beam is about 1/400 (1.75 cm drop in 700 cm run) and the average centrifugal force times the average slope is certainly of the same order of magnitude as gravity. Hence a balance between these two is eminently plausible.

That said, the series solution is also not cause for euphoria. The section of beam between $x = 2$ and $x = 3$ has a positive slope. This is impossible in reality – if the beam slopes upwards, then the centrifugal force acquires a negative or DOWNWARD transverse component. Gravity is already downwards, so who keeps the beam aloft ? This feature dents the credibility of the solution.

The bottom line is that four is a really small number of terms to take in the power series (we equated only TWO terms and called it solving the ODE!). So we cannot expect it to produce a perfect solution. The order of magnitude balancing

convinces us that the output is not garbage (which it might well have been) but the positive slope region shows clear scope for improvement. The actual deflection of the tip will probably be off by a factor of 2 or 3 and we should certainly try to generate a more accurate solution. The more so as the problem features an aircraft where safety is critical, and we should be able to use our solution to calculate stresses etc; in such a situation, errors of 100 percent are just not acceptable. That's why it's matrix method to the rescue.

(f) The most important component of the matrix formulation of (01b) we already have from the last Tute – it's the fourth derivative plus BC matrix $\mathbf{S}$. Here, we need to take care of the extra $\rho\omega^2 xy$' or $0.1568\,xy$' term. In T061 we covered the matrix representation of the first derivative in detail; now we adapt the formula to suit the present purposes. We express $\mathbf{y}$' as $\mathbf{P}_+\mathbf{y}_+$ where, in accordance with the "Point to note" in example 4.1.5, $\mathbf{y}_+$ this time has only one boundary point over and above the significant points. Conversion of $\mathbf{P}_+$ to $\mathbf{P}$ is most easily done by considering the extra point y_0, whose value is already known to be zero. Then, the first derivative is

$$\mathbf{y}' = \mathbf{P}\mathbf{y} \quad \text{, where}$$

$$\mathbf{P} = \frac{1}{h}\begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix} .$$

The factor of x before y' is represented as a matrix whose $(i,i)^{\text{th}}$ element is x_i and all other elements are zero i.e. $\mathbf{X} = \text{diag}\,(x_1\ x_2\ \ldots..\ x_N)$. Thus this term is

$$0.1568xy' \to 0.1568\mathbf{X}\mathbf{P}\mathbf{y} \quad \text{, where}$$

$$\mathbf{X} = \begin{bmatrix} x_1 & 0 & 0 & 0 \\ 0 & x_2 & 0 & 0 \\ 0 & 0 & x_3 & 0 \\ 0 & 0 & 0 & x_4 \end{bmatrix} .$$

The LHS of the ODE is now complete.

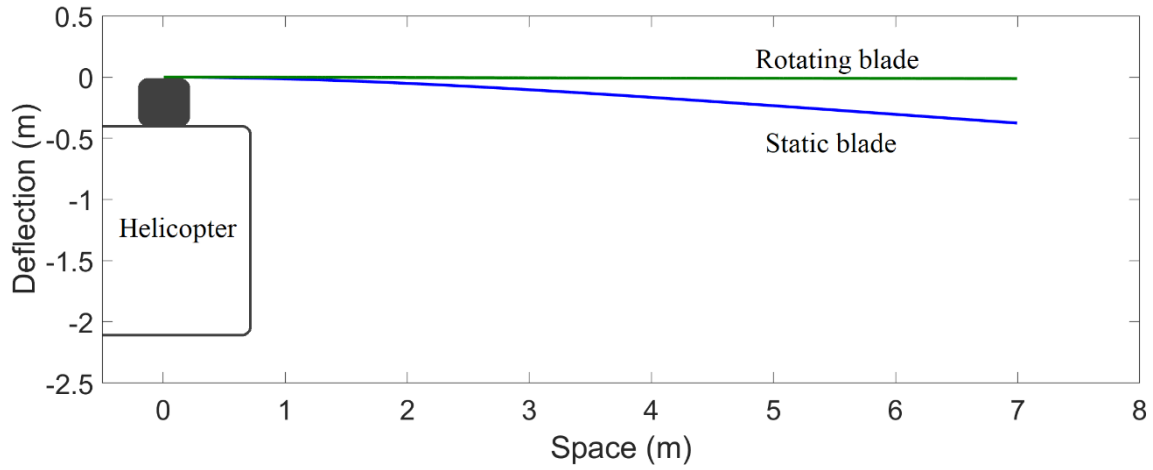
Carrying over the simple RHS from the last example, we have

$$\mathbf{b} = -0.0012544\begin{bmatrix} 1; & 1; & .....; & 1 \end{bmatrix} \ ,$$

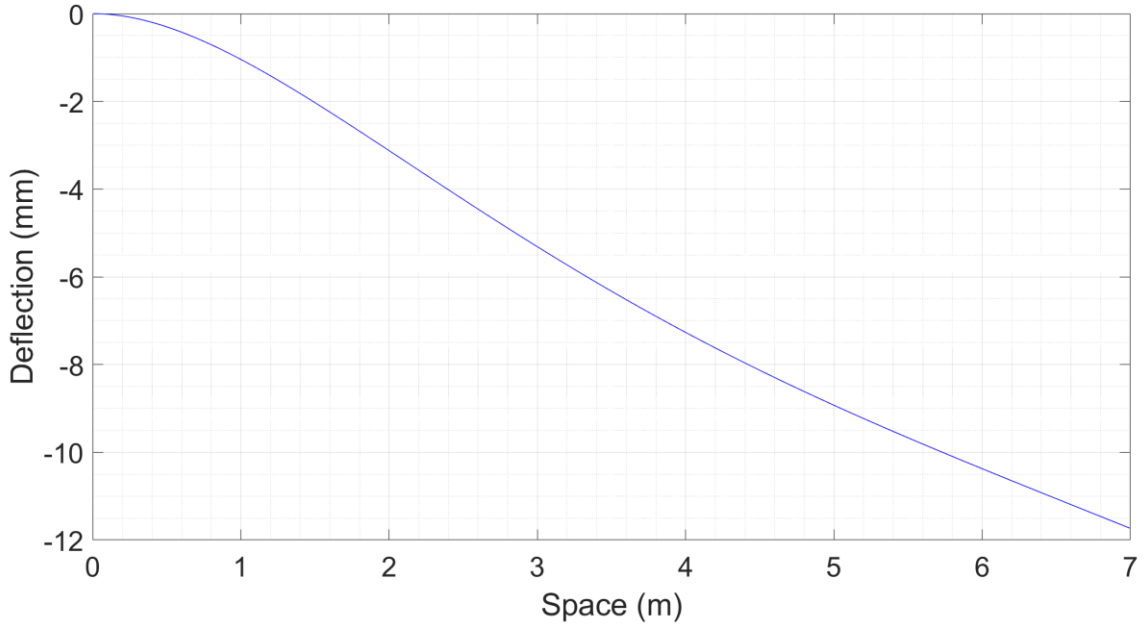
and the matrix method representation of the BVP :

$$\left[\mathbf{S}+0.1568\mathbf{XP}\right]\mathbf{y}=\mathbf{b} \quad . \tag{04}$$

This is what enters into the computer. We've already seen that $N=3000$ is good for accuracy, so we stick to that value here. The profiles of the static and rotating beams are shown in the plot below, together with a schematic of the helicopter fuselage. This time the axes ARE to 1:1 scale !



Now we present an enlarged plot of the shape of the beam as obtained numerically.



The deflection of the tip is 1.17 cm. This time, the plot shows no unphysical features like positively sloped region, which happened with the power series solution. The deflection obtained in Part (d) turns out to be about 50 percent off from true, which is very much consistent with our error analysis in Part (e).

*

COMMENTS : Part (c) here is an elementary example of the method of series solutions to ODEs – the basic power series method and the more sophisticated method of FROBENIUS. These methods are required to obtain special functions like Legendre polynomials, Bessel functions etc. Intentionally I didn't cover that topic here – an introduction to those functions doesn't really serve much purpose unless you are shown the full scope and power of those techniques. But such an exposition is outside the scope of this our little book, and if we can't push it, there's not much point in just starting it.

Between this and 4.1.7, you can see both the 'knife' and 'hammer' aspects of the matrix method . Also note how we ran multiple checks on the answers which were returned by the computer – the static cantilever is the most important one while comparison with the power series solution also convinces us that the answers are in the right ballpark . Such checks are extremely important whenever we solve a problem numerically – since we're using the computer as a black box, they are the only thing which ensures that we don't fall for garbage returned by it.

The straightening of helicopter blades as they are spun up is apparent on observation. For an excellent demonstration, watch a video of startup of Aerospatiale SA315B helicopter uploaded on Youtube by user "Ed Whisenant". The section of the video relevant for our purposes is reproduced at the following link : www.shayak.in/Shayakpapers/DELTA/T068_V1.mp4. When the helicopter is on the ground with its fan on, the relevant forces are more or less only the weight and the centrifugal force since the thrust is kept close to zero. When the helicopter is in the air, there's also the thrust or lift (they are the same for a helicopter) so the net force on the blade is vertically upwards. Even so, the centrifugal force is many times stronger than the lift and tends to straighten the blade out, to an extent that the deflections are virtually invisible to the eye.

4.3

ENGINEERING FOURIER SERIES

A periodic function $f(t)$ with period T means that $f(t+T) = f(t)$ for all t. The frequency of such a function is $\omega = 2\pi / T$. At the heart of Fourier analysis lies the theorem discovered by JOSEPH FOURIER, the originator of the topic : any function $f(t)$ which is periodic with frequency ω can be expanded in a series

$$f(t) = A_0 + \sum_{n=1,2,3.....}^{\infty} \left(A_n \cos n\omega t + B_n \sin n\omega t \right) \quad ,$$

where the A_n's and B_n's are constants. Herein let us define the convention, valid permanently, that **a summation index *n* with no other information specified means a sum over the natural numbers 1,2,3….. all the way to infinity**. This is by far the most common summation we'll need, and it will save me a lot of typing time if we can leave the limits etc implicit. Exceptions of course will have to be stated very clearly. Thus we can write the above as

$$f(t) = A_0 + \sum_{n} \left(A_n \cos n\omega t + B_n \sin n\omega t \right) \quad . \qquad \textbf{(01)*}$$

In words, **a periodic function of a given frequency can be expanded as a constant (or dc term) plus a sum of cosines and sines at the fundamental frequency and at higher harmonics which are integer multiples of this frequency**.

Fourier analysis is one of the core areas of pure mathematics and often features topics which are on first sight completely unrelated to this theorem. For example, a course on this topic which I had taken in my final year of undergraduation featured one third of the lectures on proving that if p and q are relatively prime natural numbers, then the sequence $p+q$, $p+2q$, $p+3q$ etc out to infinity contains infinitely many primes. (This course had been French to me then, and still is now.) The word "engineering" in the Unit title indicates that unlike in most Units, our treatment here will only be surface deep – we shall stay kilometres clear of such mathematical developments,

fascinating though they may be. Rather we'll be content with only those parts of Fourier series which are directly useful for solving practically motivated problems in ordinary (and later partial) differential equations.

Given the theorem, the natural question to ask is, given f, how to find the coefficients A_n and B_n. We do this using the following integral relations satisfied by the sines and cosines, where τ is any arbitrary time (recall that T is the period), and m and n are arbitrary natural numbers :

$$\int_{\tau}^{\tau+T} \cos m\omega t \cos n\omega t \; \mathrm{d}t = \begin{cases} T/2 & \text{if } m = n \\ 0 & \text{otherwise} \end{cases} ,$$

$$\int_{\tau}^{\tau+T} \sin m\omega t \sin n\omega t \; \mathrm{d}t = \begin{cases} T/2 & \text{if } m = n \\ 0 & \text{otherwise} \end{cases} ,$$

$$\int_{\tau}^{\tau+T} \cos m\omega t \sin n\omega t \; \mathrm{d}t = 0 \quad \text{with no condition} \quad .$$

These integral formulae can be obtained relatively easily by using the trig sum and difference formulas; we saw an example of this while evaluating the power of the oscillatory external force in 3.3.12. The most important thing is that they hold true.

To find A_n and B_n we take (01), multiply with $\cos m\omega t$ on the both sides for some m, and integrate from τ to $\tau+T$. All terms on the RHS drop out except the one with $n = m$; equating the survivors readily yields

* $$A_0 = \frac{1}{T} \int_{\tau}^{\tau+T} f(t)\mathrm{d}t \quad , \qquad \textbf{(02a)*}$$

* $$A_n = \frac{2}{T} \int_{\tau}^{\tau+T} f(t) \cos n\omega t \; \mathrm{d}t \quad , \qquad \textbf{(02b)*}$$

* $$B_n = \frac{2}{T} \int_{\tau}^{\tau+T} f(t) \sin n\omega t \; \mathrm{d}t \quad . \qquad \textbf{(02c)*}$$

These relations are valid for arbitrary τ and are the most general formulae for the Fourier coefficients of an arbitrary function $f(t)$.

Practically, the value of τ which we use most often for the integrations is $\tau = -\pi/\omega$. The integral then runs from $-\pi/\omega$ to $+\pi/\omega$ and enables us to exploit

the **symmetry properties** (symmetry or antisymmetry) of the function f whenever they are present. A function is called symmetric (or "even" in mathsy books) if $f(-t) = f(t)$ for all t. Examples are constant function, t^2 and $\cos t$. The two halves of the function for $t > 0$ and $t < 0$ look like mirror images reflected about the 'f-axis'. A function is called antisymmetric (or "odd") if $f(-t) = -f(t)$ for all t. Examples are t, t^3 and $\sin t$. The two halves of the function for $t > 0$ and $t < 0$ look like mirror images reflected about the origin.

The symmetry properties relevant for us are :

- $\cos n\omega t$ is symmetric for all n
- $\sin n\omega t$ is antisymmetric for all n
- The product of two symmetric functions is symmetric
- The product of a symmetric and an antisymmetric function is antisymmetric
- The product of two antisymmetric functions is symmetric
- The integral of a symmetric function over an interval $[-\tau, \tau]$ for arbitrary τ is twice the integral of the function over the interval $[0, \tau]$
- The integral of an antisymmetric function over an interval $[-\tau, \tau]$ for arbitrary τ is zero

These properties often enable us to get the Fourier coefficients with far less effort than would be involved in a mindless application of (02). The main conclusions are that the **Fourier series of a symmetric function consists of A_n's only (including A_0)** and the **Fourier series of an antisymmetric function consists of B_n's only**.

Finally, sometimes it happens that we are given functions which are not periodic but are defined in a finite interval. This situation generally occurs with the independent variable being space x rather than time t. Depending on the situation, it might be profitable to perform an **extension** of the given function to a periodic function. Suppose we have the function f defined in $[0, L]$. Then we can extend it to a periodic function over the entire real line by several ways. Three of the most common are :

- Create replicas of f in $[L, 2L]$, $[2L, 3L]$, $[-L, 0]$ etc.
- First extend f symmetrically into the interval $[-L, 0]$ and then create replicas of the new function in $[L, 3L]$, $[-3L, -L]$ etc.

- First extend f antisymmetrically into the interval $[-L, 0]$ and then create replicas as above

Extension happens to be a useful concept while handling PDEs, so we'll become familiar with the technique right now.

A natural question to ask is whether the equation (01) is valid everywhere, for all f. In other words, does the Fourier series of any arbitrary function converge everywhere to the function value ? The answer (a very famous theorem by LENNART CARLESON whose proof belongs to the 'French' parts of Fourier analysis) is as follows : the Fourier series for f converges to f on all intervals where f is continuous. So if f is itself continuous, its Fourier series converges to it everywhere. If f has a finite-sized discontinuity at a point t_0, say $f(t_0^-) = a$ and $f(t_0^+) = b$ then at the point t_0, the Fourier series converges to $(a+b)/2$. Moreover, the Fourier series for f shows GIBBS' phenomenon around the point t_0. This phenomenon is best illustrated with an example, so we'll do it in the Tute coming up.

Finally, another thing we should know is that the **continuity properties of a function depend on the asymptotic size of its Fourier coefficients**. Any function with one or more finite-sized discontinuities has its Fourier coefficients going as $1/n$ for large n. Any function which is continuous but has one or more finite-sized discontinuities in its derivative has its large n Fourier coefficients going as $1/n^2$. In general, the faster the asymptotic rate of decay of the coefficients, the more tractable is the function. Although the proof of this is not difficult (the primary step is integration by parts), the fact is more important than the proof so we'll just take it for granted.

Example 4.3.1

Square wave

The square wave $f(t)$ is defined as

$$f(t)=\begin{cases}-1 \text{ if } -\pi < t \le 0 \\ 1 \text{ if } 0 < t \le \pi\end{cases},$$

and $f(t)$ is periodic with period $T=2\pi$.

Expand $f(t)$ in a Fourier series.

[Level : 1]

We can see that $f(t)$ is antisymmetric hence $A_n = 0$ for all n. For the B_n's we use the integral formula (4.3.02c) with $T = 2\pi$ so $\omega = 1$. We have

$$B_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin nt \ \mathrm{d}t \quad .$$

Since f and sin are both antisymmetric, their product is symmetric so the integral is twice that from 0 to π. Thus we have

$$B_n = \frac{2}{\pi}\int_{0}^{\pi} 1\sin nt \ \mathrm{d}t \quad ,$$

which evaluates to $4/n\pi$ if n is odd and zero if n is even. Note that this is consistent with the continuity properties we mentioned in the Lecture – the square wave has a finite-sized discontinuity and its coefficients are $1/n$. Thus the Fourier expansion of f is

$$* \qquad f(t) = \sum_{n=1,3,5.....} \frac{4}{n\pi}\sin nt \quad , \qquad *$$

which we can also write as

$$* \qquad f(t) = \sum_{n} \frac{4}{(2n-1)\pi}\sin(2n-1)t \quad , \qquad *$$

which adheres to our convention about an unlabelled n. Either of the above is the desired answer.

*

COMMENTS : Fourier decomposition of a periodic function has great importance in electrical engineering, where it is used to calculate, among other things, the quantity of fundamental harmonic vis a vis other harmonics present in a signal. (The word "wave" for a time-periodic function is a bit of a misnomer since the hallmark of a true wave is periodicity in both space and time, as we'll see in Chapter 7. Nevertheless, the term is very popular and doesn't really cause confusion.) The decomposition process invariably tends to become mechanical and tedious, as in this example and the ones that follow. From an exam viewpoint, the primary thing to remember is to exploit symmetry and reduce the calculation as much as possible. An error on the symmetry score is senseless and is generally penalized heavily; thus, if your answer features zero A_n and

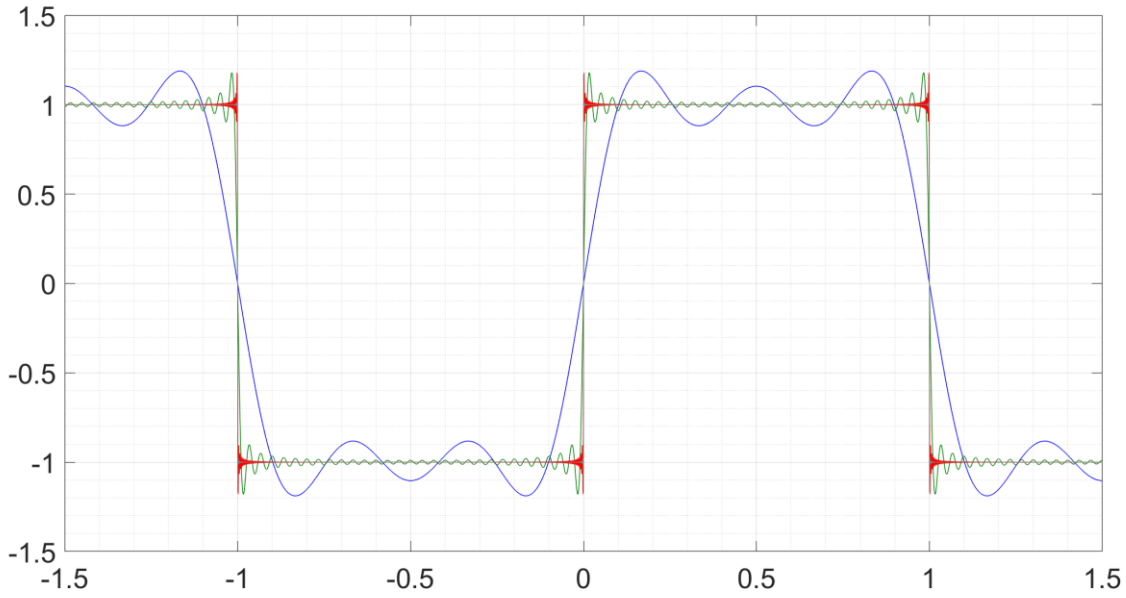
wrong B_n, you will probably get 90 percent of the credit for this problem, while if your answer features non-zero A_n arising from a botched integral, you're likely staring at 50 percent and not more.

An important thing to note is that the Fourier coefficients of a function are the same whatever its frequency. The ARGUMENT of the cos and sin terms changes but the COEFFICIENTS don't. Thus, the square wave of period 2 instead of 2π will look like

$$g(t) = \sum_n \frac{4}{(2n-1)\pi} \sin(2n-1)\pi t \quad .$$

Because of this, most books on Fourier analysis work with functions of period 2π alone.

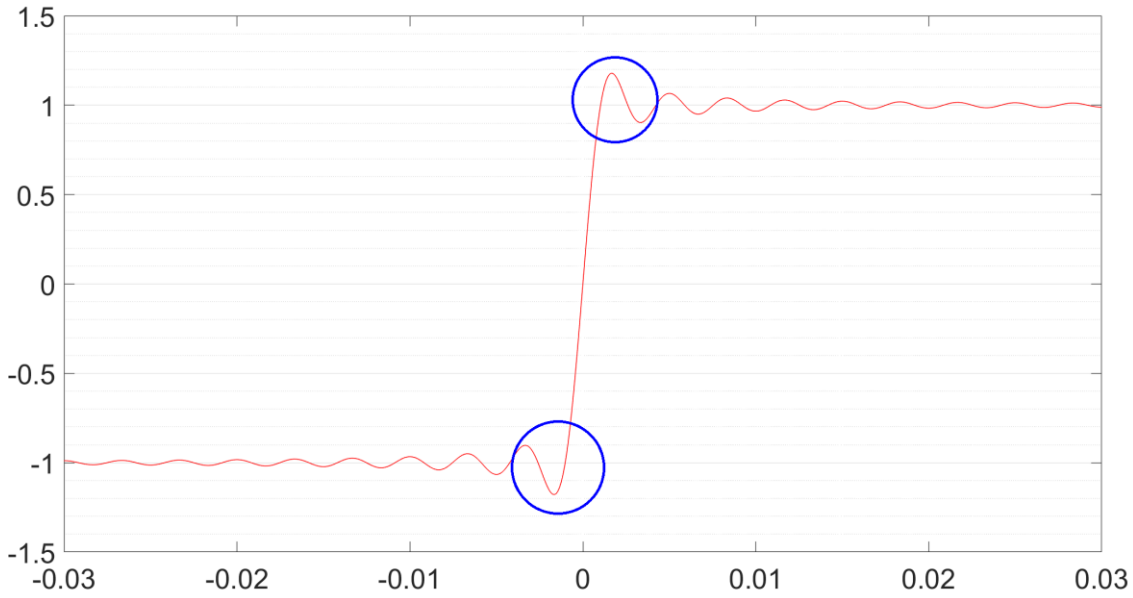
We'll now use this example to look at GIBBS' phenomenon. Let's try plotting the first N terms of the Fourier series of $g(t)$ with N taking the values 3, 30, 300 and 3000. The four plots are shown below in our standard livery of blue, green, red and grey.



The blue plot is obviously a very crude approximation of the square wave. The green one is much better but there seems to be a bit of overshoot at $t = -1, 0, 1$ i.e. the points where g has a discontinuity. There's some oscillation everywhere else also but that's much smaller. The red one is smooth everywhere except at the jumps of g – again there's a bit of overshoot and oscillation before things

settle down. If you look hard enough, you can just see a fine grey line on top of the red one – that too has an overshoot at the jumps.

Here's a zoomed in plot of the red curve showing the overshoot at $t = 0$ (encircled, note the high level of zooming in the x-axis!).



On the negative side, the curve quite definitely hits a value of -1.15 or so before jumping up to positive. Then again it hits +1.15 before subsiding back to 1. This overshoot is GIBBS' phenomenon.

The formal description of this phenomenon is that, if $f(t)$ has a jump at $t = t_0$ with $f(t_0^-) = a$ and $f(t_0^+) = b$, and if $F_N(t)$ be the sum of the first N terms of its Fourier series, then $F_N(t)$ exhibits an overshoot on either side of the jump. As $N \to \infty$, the size of the maximal overshoot tends to the asymptotic value $0.089(b–a)$ while the location of this max goes as $t_0 \pm O(1/N)$. (Note that the 0.089 here is the numerical value of an analytical expression and not an empirical factor.) Further details of this phenomenon are not of interest to us. The $O(1/N)$ term in the location of the overshoot means that the practical implications of GIBBS' phenomenon are limited. But while plotting a Fourier series we should be aware of its existence, unlike some students and instructors whom I have seen.

Example 4.3.2

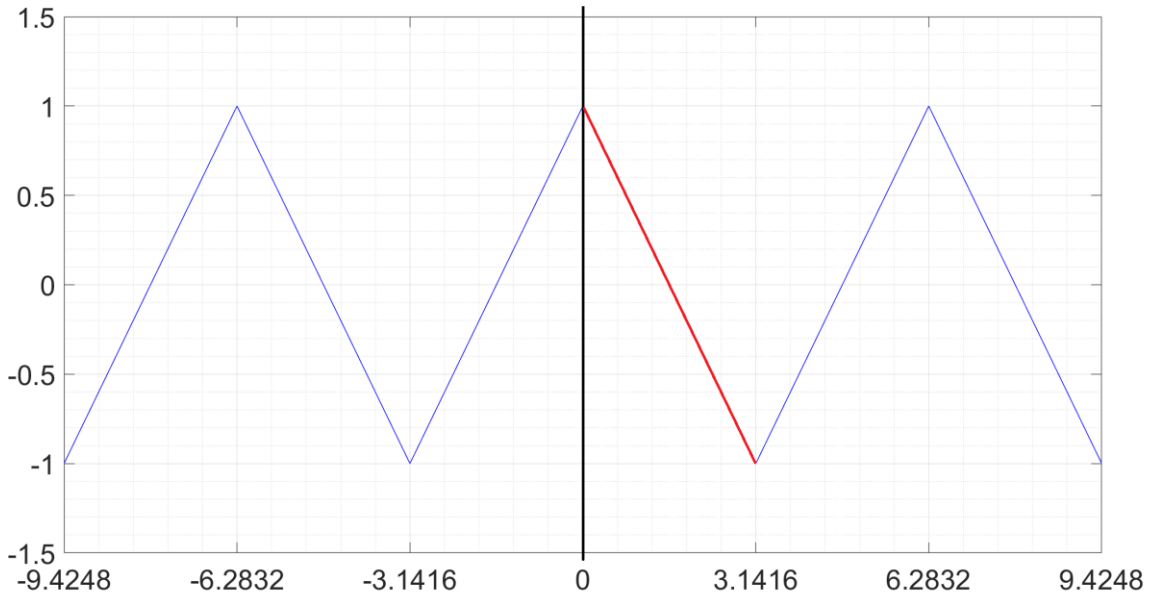
Waves with straight lines

(a) The triangle wave $f(x)$ goes in a straight line from $(0,1)$ to $(\pi,-1)$. Extend it symmetrically to create a function of period 2π and expand the resulting function in a Fourier series.

(b) The sawtooth wave $g(x)$ goes in a straight line from $(-\pi,-1)$ to $(\pi-\varepsilon,1)$ where ε is small tending to zero, and then jumps back abruptly to $(\pi,-1)$. It is periodic with period 2π. Expand $g(x)$ in a Fourier series.

[Level : 1]

(a) The given $f(x)$ is defined in the interval $[0, \pi]$. We extend it symmetrically into $[-\pi, 0]$ by reflecting it in the vertical axis. Then we replicate it periodically. The schematic representation of this process is the below Figure. The red line is the given function while the blue line is the extended function, which we call $F(x)$.



We can see that $F(x)$ is symmetric hence $B_n = 0$ for all n. The term A_0 is just the average value of the function over the period (note this fact), which is clearly zero in this case. For the other terms A_n, we use the fact that due to symmetry (both F and cos are symmetric hence so is their product), the integral over the whole period will be double of the integral from 0 to π. In this region the function is a straight line with slope $-2/\pi$ and intercept 1 i.e.

$$F(x) = -\frac{2}{\pi}x + 1 \quad .$$

The integral formula then says

$$A_n = \frac{2}{\pi}\int_0^{\pi}\left(-\frac{2}{\pi}x + 1\right)\cos nx \; \mathrm{d}x \quad .$$

We evaluate the integral by parts, and substitute the trig for the limits; the answer is $8/n^2\pi^2$ if n is odd and zero if n is even. We thus have

* $$f(x) = \sum_{\text{odd } n} \frac{8}{n^2\pi^2} \cos nx \quad , \qquad *$$

which is the desired answer.

(b) We can see that $g(x)$ is antisymmetric hence $A_n = 0$ for all n. For B_n, we use the symmetry conditions and get

$$B_n = \frac{2}{\pi} \int_0^{\pi} \frac{x}{\pi} \sin nx \ \mathrm{d}x \quad ,$$

which evaluates to $2/n\pi$ if n is odd and $-2/n\pi$ if n is even. We thus have

* $$g(x) = \sum_n (-1)^{n+1} \frac{2}{n\pi} \sin nx \quad , \qquad *$$

which is the desired answer and another boring Level 1 problem is over.

*

COMMENTS : Serieses with odd numbers only, ones with alternating signs etc. can be represented either in a mathsy notation or in an intuitive one. For example, over the last two problems we saw three ways of writing the sum over odd numbers – the $2n-1$ notation is the most pedantic while the other two are more informal. Likewise, you can write the result of Part (b) here in the form we gave above or in the alternative form

$$g(x) = \frac{2}{\pi} \sin x - \frac{2}{2\pi} \sin 2x + \frac{2}{3\pi} \sin 3x + \ldots \quad ,$$

which is less rigorous but just as convincing. Personally I have no preference, and both representations will fetch you full marks in exams. However, if you are going for the mathsy notation, please make sure that what you write is correct. For instance, in the expansion for $g(x)$ if you write $(-1)^n$ instead of $(-1)^{n+1}$, that will become a mistake and attract a penalty.

Again we note that both the above answers are consistent with the relations between the continuity properties of a function and the size of its Fourier coefficients.

Example 4.3.3

Chopper wave

An electrical chopper converts a dc voltage into an ac voltage by turning it ON and OFF periodically. The ratio of ON time to the total period is called the duty cycle d, which thus lies between zero and unity. A particular chopper is operating at frequency ω. Its output is

$$g(t)=\begin{cases}1 \text{ if } 0<t\le\dfrac{2\pi d}{\omega}\\ 0 \text{ if } \dfrac{2\pi d}{\omega}<t\le\dfrac{2\pi}{\omega}\end{cases},$$

and $g(t)$ is periodic with period $T=2\pi/\omega$.

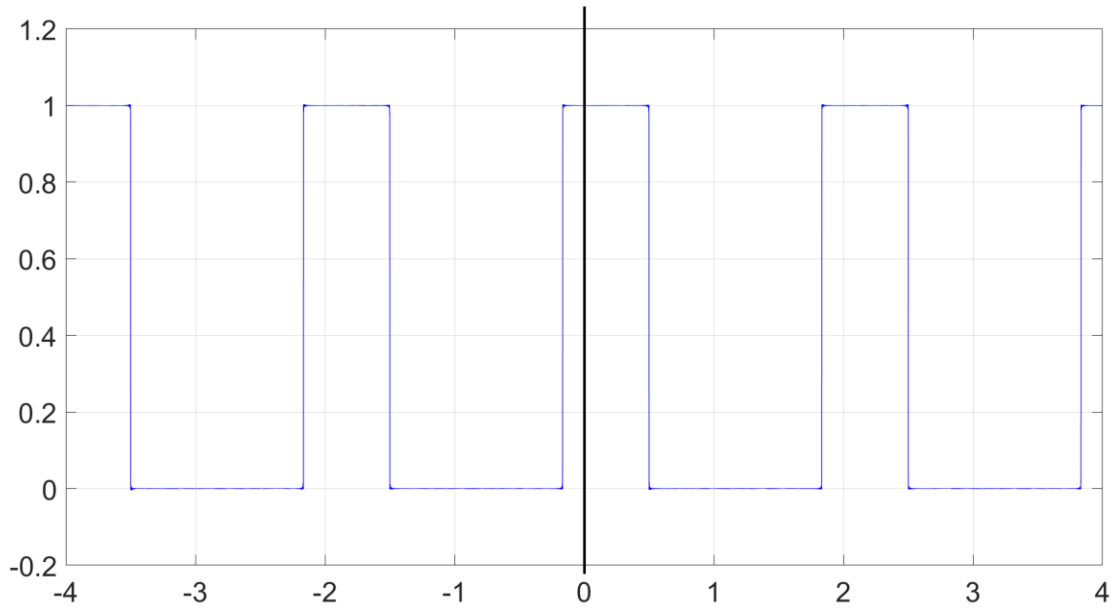
(a) Define a new function $f(t)$ which is $g(t)$ with a suitable time shift so that its Fourier series consists of either cosines only or sines only.

(b) Expand the function $f(t)$ in a Fourier series.

(c) The current $i(t)$ in the winding of a dc motor connected to a chopper satisfies $i'+i=f(t)$. Obtain a formal solution for $i(t)$ in the form of a series.

[Level : 2]

(a) The Fourier series of f will consist of cosines only if the function is symmetric. It will consist of sines only if f is antisymmetric. The graph of the original function g is shown in the Figure below.



We can see that if we shift the graph leftward so that the dependent variable axis passes through the centre of a pulse, then we will get a symmetric function. No shifting of g will convert it to an antisymmetric function. Hence we perform a shift to make f symmetric, defining it in the fundamental interval $[-\pi/\omega, \pi/\omega]$ where we have

$$f(t) = \begin{cases} 1 & \text{if } -\dfrac{\pi d}{\omega} \le t \le \dfrac{\pi d}{\omega} \\ 0 & \text{otherwise} \end{cases} \quad .$$

Of course, f is periodic with period $2\pi/\omega$.

(b) I'll present the answer straightaway – typing up the trivial details is just TOO boring. The required expansion is

$$f(t) = d + \sum_n \frac{2}{n\pi} \sin n\pi d \cos n\omega t \quad .$$

For some night-before-exam feel-good practice, verify that I did the above integral correctly. At the very least, it is certainly consistent with the continuity properties.

(c) For this problem we note that the homogeneous equation is $i_h' + i_h = 0$ whose solution $i_h = e^{-t}$ dies out in time, so only the particular solution is of interest. To find that we use the method of undetermined coefficients (and I hope it's not become unremembered coefficients). The equation is of the form

$$i' + i = d + \sum_n A_n \cos n\omega t \quad ,$$

and the problem asks for a solution in the form of a series. So we take our trial function to be the sum of trials corresponding to each term in the excitation. The trial for the constant is of course a constant (and inspection reveals the constant to be d). The trial for the cos terms is $D_n \cos n\omega t + E_n \sin n\omega t$, with a D_n and E_n defined for each n. Using C_n and S_n to denote $\cos n\omega t$ and $\sin n\omega t$ respectively and plugging into the equation, we have

$$n\omega D_n S_n + n\omega E_n C_n + D_n C_n + E_n S_n = A_n C_n \quad .$$

Equating cos and sin terms and solving yields

$$D_n = \frac{A_n}{1 + n^2\omega^2}$$
$$E_n = \frac{n\omega A_n}{1 + n^2\omega^2} \quad .$$

Substituting these as well as the values of A_n into the series gives

$$* \qquad i = d + \sum_n \left(\frac{2}{n\pi} \frac{\sin n\pi d}{n^2\omega^2 + 1} \cos n\omega t + \frac{2\omega}{\pi} \frac{\sin n\pi d}{n^2\omega^2 + 1} \sin n\omega t \right) \quad , \qquad (01)*$$

which is the desired answer.

*

COMMENTS : Although this problem had a reasonable starting point with choppers and motors, it degenerated into a homework assignment as it progressed. The model of the dc motor here is not trivialized. DC motors really are controlled by choppers in high-end applications so nothing fishy there also. Rather, the primary shortcoming of the problem is that the series (01) is impossible to sum explicitly to get a more meaningful answer. This is actually a very typical problem with the Fourier approach. It happens because Fourier expansion is a quintessentially one-way phenomenon – given a function $f(t)$ we

can find the series trivially but given a series we can hardly ever find the function. A secondary deficiency of this problem is that choppers don't operate in a blindly periodic manner OFF-ON-OFF-ON like an automaton but change their state depending on the motor parameters.

In the next Unit we'll overcome the problem of series summation by developing the method of matched boundary conditions. Then we'll see what the chopper-controlled dc motor really looks like.

4.4

PERIODIC EXCITATION AND MATCHED BOUNDARY CONDITIONS

Here, we return to the dynamical systems which have been our focus in most of this course. In particular, we consider systems excited by some sort of periodic functions. In Chapter 3 we saw a lot of these with the excitation being external and having the form of pure cosines and sines. Sometimes it happens that the external excitation is periodic but not pure cos or sin and an exact (or at least accurate even if approximate) solution is desirable on physical grounds. In such a case, term by term Fourier expansion is more often than not impractical, as in 4.3.3. Some problems may feature the excitation parametrically instead of as an inhomogeneous term.

The method of boundary matching works if we have an excitation function which is piecewise tractable. For example, the square wave "sq t" of 4.3.1 is piecewise constant – solving $\dot{x} + x = \mathrm{sq}\, t$ explicitly may be impossible, but solving $\dot{x} + x = \pm 1$ is trivial. Starting from a particular time, we can solve the simpler equations in successive intervals, and apply suitable matching conditions at each transition to generate the overall solution. If we are interested in a periodic "steady state" solution, we can apply periodic boundary conditions.

Qualitative descriptions of the method sound wishy-washy and I'm not going to try elucidating further. Instead we'll go to the Tutes where the examples will automatically make the procedure clear. But let's just note that long back, in Chapter 1, we have already solved a problem using this method. That was the pharmacokinetics problem, 1.2.8 – because the equation was first order, I did not need to make a song and dance about it.

Example 4.4.1

Trampoline gymnast

A gymnastic trampoline can be modelled as a mass-spring-damper system. The point-mass gymnast stands on the trampoline for part of the time and is in the air for the rest of the time. Her action can be modelled as follows. When she is on the trampoline, she waits for the first time that it reaches its static equilibrium position and has upward velocity; she then exerts an impulsive force P which causes her to detach from the trampoline and fly into the air. When she comes back down, the collision with the trampoline is perfectly inelastic and the cycle begins all over again. A particular setup has the following parameter values : mass of gymnast $M = 50$ kg, mass of trampoline $m = 5$ kg, damping constant $C = 110$ Ns/m, spring constant $k = 5000$ N/m, impulse strength $P = 150$ kgm/s. Take $g = 10$ m/s^2.

(a) Start from the instant just before the gymnast jumps. Letting the unknown v_0 be her speed at this time, obtain IVPs for her and for the trampoline immediately after the jump.

(b) Determine the time of collision between the trampoline and the gymnast on her way down, and hence the speed of the two bodies at this instant, immediately prior to the collision.

(c) Obtain and solve an IVP for the combined system valid from the collision onwards.

(d) Use a computer to plot the above solution for various values of v_0 and hence obtain the gymnast's speed just before the jump. Impose the condition that this speed equal v_0 to determine v_0 and therefrom the height and period of the jump in the steady state.

[Level : 3]

(a) Let z be the vertically upward direction and the datum $z=0$ be the position of the trampoline in the absence of all forces. We are given that the gymnast jumps when the trampoline is at its equilibrium position. This equilibrium is $-(M+m)g/k$ (recall that m is trampoline mass and M is gymnast mass) which evaluates to -0.11 m. Her velocity at this point is the unknown v_0, which gets determined only at the last step. The impulse exerted by the gymnast for the jump is 150 kgm/s; since her mass is 50 kg, this means that she imparts to herself a speed of 3 m/s over and above what it was previously. Hence her speed just after the jump is v_0+3. In the air her equation of motion is trivial since her acceleration is constant. We thus have for the gymnast

$$\ddot{z}=-10$$
$$z(0)=-0.11,\quad \dot{z}(0)=v_0+3$$

which is one of the two IVPs asked for.

For the trampoline on the other hand, we have a forced damped oscillator equation, $m\ddot{z}+C\dot{z}+kz=-mg$. Before the gymnast jumped, the tramp's position and velocity were those of the gymnast; after the jump the position does not change but the velocity does on account of the impulse. The change in velocity is $-P/m$ which -30 m/s; the trampoline's velocity right after the jump is thus v_0-30 m/s. Plugging in the numbers, we get

$$\ddot{z}+22\dot{z}+1000z=-10$$
$$z(0)=-0.11,\quad \dot{z}(0)=v_0-30$$

which is the second IVP asked for.

(b) The time of collision will be the first point that the trajectories of the gymnast and the trampoline intersect, finding which is no easy task. However, we see that the trampoline equation has extremely heavy but subcritical damping, so both homogeneous solutions decay at the same rate e^{-11t}. Unless the time of the jump is really short (which we know from practical experience it isn't), the tramp is going to be back to its equilibrium (particular solution, $z_p=-0.01$ and zero velocity) by the time the gymnast arrives. So we make that assumption here, that the collision occurs at $z=-0.01$ with the trampoline being static.

A middle school calculation then gives the trajectory of the gymnast : time to max height is $(v_0+3)/g$, height attained is $(v_0+3)^2/2g-0.22$, time to come back

satisfies $t^2=(v_0+3)^2/g^2-0.1/g$ and speed at the point of impact is $v_1=[(v_0+3)^2-2]^{1/2}$. We'll define v_1 this way and add a negative sign explicitly when relevant.

(c) The equation is standard mass spring damper with gravity, the mass now being that of the tramp plus the athlete. Since the equations don't feature time explicitly as a parameter (see 2.1.1) we can reset time to zero at the instant of the collision with no loss of generality . That gives $z(0)=-0.01$. For initial velocity we note that the problem says that the collision is perfectly inelastic i. e. the gymnast sticks to the trampoline upon impact . From elementary mechanics we know that the total momentum is conserved , so the post-collision velocity v_2 satisfies $(M+m)v_2=Mv_1+m(0)$, or $v_2=(10/11)v_1$. Plugging in the numbers, we have

$$\ddot{z}+2\dot{z}+90.1z=-10$$

$$z(0)=-0.01,\quad \dot{z}(0)=-v_2=-\frac{10}{11}\sqrt{(v_0+3)^2-2}\quad , \qquad *$$

which is the IVP asked for.

The damping constant is 2 rad/s, damped frequency is 9.48 rad/s, the particular solution is $z_p=-0.11$ (constant) and the homogeneous problem thus has ICs $z_h(0)=0.10,\ \dot{z}_h(0)=-v_2$. The solution is

$$z=\mathrm{e}^{-t}\left[0.1\cos\omega_d t+\left(\frac{-v_2+0.1}{\omega_d}\right)\sin\omega_d t\right]-0.11\quad , \qquad (01)*$$

which is the desired answer.

(d) We write a computer program to plot the above $z(t)$ as well as its time derivative. The speed immediately prior to the jump will be the velocity at the instant when the position first becomes equal to –0.11 (the static equilibrium) and the gymnast is moving upwards. This velocity is of course a function of v_2, which is contained in the ICs; it's some messy transcendental function which we are bypassing with the help of the computer, and let's just call it $f(v_2)$. Now from Part (c), v_2 is already known in terms of v_0; inverting the relation we have $v_0=\sqrt{1.2v_2^2+2}-3$. On the other hand, from periodicity of the jump in the steady state, the gymnast's velocity just before the jump must equal v_0; thus $v_0=f(v_2)$. This yields

$$f(v_2) = \sqrt{1.2v_2^2 + 2} - 3 \quad ,$$

which is the boundary matching condition in the steady state of jumping.

We try various values of v_2 on the computer until we get a good satisfaction of the matching condition; the value which works best is $v_2 = 7.7$, giving $v_0 = 5.55$ m/s, speed immediately after jump 8.55 m/s and maximum height 3.55 m. The time spent in air is 1.7 s while the time spent on the tramp is 0.3 s, giving a total period 2.0 s for each jump.

Before claiming these as final answers, we go back to the approximation we made on the way, in Part (b), and check that these numbers are compliant with the approximation (recall the similar step we took in 3.3.1). The simplification was to assume that the motions of the tramp die down completely while the gymnast is in air. Since the time of flight is 1.7 s, the initial motions get to decay by a factor of $e^{-11(1.7)}$ during this time. This is an enormous reduction, so the approximation is for all practical purposes exact and we can declare $v_0 = 5.55$ m/s, height of jump 3.55 m and period of jump 2.0 s to be our final answers.

*

COMMENTS : Although this model is extremely simple, it can account for a lot of quantitative features of the actual trampoline jump (the parameter values are taken from standard literature). A good video reference can be found on Youtube, uploaded by user “Olympic”, which shows the women’s trampoline jump event in the 2016 Olympics in Rio de Janeiro. A section of this video relevant for our purposes is reproduced at the following link : www.shayak.in/Shayakpapers/DELTA/T072_V1.mp4. Here we can see the British athlete BRYONY PAGE on the tramp. The jump height is indeed close to a couple of human heights and the period is very close to 2 seconds.

Example 4.4.2

Mathieu equation

The Mathieu equation is $\ddot{x}+(1+a\cos t)x=0$ and can be used to model parametric excitation phenomena. It cannot be solved in closed form, and no inroads can be made using elementary techniques. Consider the related but simpler system

$$\ddot{x}+(1+a\,\mathrm{sq}\,\omega t)x=0 \quad ,$$

where a is a constant between zero and unity and $\mathrm{sq}\,\omega t$ denotes the square wave of height unity and period $2\pi/\omega$, encountered in example 4.3.1. Let x_n and v_n be the position and velocity at the start of the nth period of excitation.

(a) Given x_i and v_i for some i, solve the ODE over one period to determine x_{i+1} and v_{i+1}. Express your answer as a matrix equation $[x_{i+1};\, v_{i+1}]=\mathbf{T}[x_i;\, v_i]$.

(b) Using the above results, determine x_{i+k} and v_{i+k} in terms of x_i, v_i and $\mathbf{T}$.

(c) What conditions must the eigenvalues of $\mathbf{T}$ satisfy for the system to be stable or unstable ? Justify your answer. Hint : Write $\mathbf{T}=\mathbf{P}^{-1}\mathbf{D}\mathbf{P}$ which is valid for complex eigenvalues also.

(d) Assuming $a<<1$, obtain the characteristic polynomial for the eigenvalues of $\mathbf{T}$ and use it to plot the stability of the system at various points in the ω,a parameter space with ω belongs to [0, 3] and a belongs to [0, 0.1].

(e) Using a computer, calculate the eigenvalues of $\mathbf{T}$ exactly and replot the stability of the system in the parameter space ω belongs to [0, 3] and a belongs to [0, 0.5].

[Level : 3]

(a) We are given the position and velocity at the start of a period and want to find them at the start of the next period. Because of the discontinuous nature of the excitation, the ODE is different in the first and second halves of the period. In the first half, i.e. from $t=0$ to $t=\pi/\omega$, it is $\ddot{x}+(1+a)x=0$ while in the second half, i.e. from $t=\pi/\omega$ to $t=2\pi/\omega$ it is $\ddot{x}+(1-a)x=0$. Since there are no delta function excitations anywhere, both position and velocity remain continuous at the changeover point.

Now we solve the first ODE. Letting $\omega_+=\sqrt{1+a}$, the ODE is $\ddot{x}+\omega_+^2 x=0$ with ICs x_i and v_i. The solution is $x=x_i\cos\omega_+ t+(v_i/\omega_+)\sin\omega_+ t$. Setting $t=\pi/\omega$ gives $x(\pi/\omega)$ as

$$x_{i+1/2}=x_i\cos\frac{\pi\omega_+}{\omega}+\frac{v_i}{\omega_+}\sin\frac{\pi\omega_+}{\omega} \quad .$$

The subscript $i+1/2$ shows that this x is halfway between the i^{th} and $i+1^{\text{th}}$ periods. Likewise, differentiating x and setting $t=\pi/\omega$ into the resulting expression gives

$$v_{i+1/2}=-\omega_+ x_i\sin\frac{\pi\omega_+}{\omega}+v_i\cos\frac{\pi\omega_+}{\omega} \quad .$$

These values act as ICs for the second ODE.

Letting $\omega_-=\sqrt{1-a}$, this ODE is $\ddot{x}-\omega_-^2 x=0$. Since the new ODE is independent of time, we can again set the time to zero with no loss, and the new ICs are $x_{i+1/2}$ and $v_{i+1/2}$. The solution is $x=x_{i+1/2}\cos\omega_- t+(v_{i+1/2}/\omega_-)\sin\omega_- t$ and setting $t=\pi/\omega$ in this as well as its derivative gives

$$x_{i+1}=\left(x_i\cos\frac{\pi\omega_+}{\omega}+\frac{v_i}{\omega_+}\sin\frac{\pi\omega_+}{\omega}\right)\cos\frac{\pi\omega_-}{\omega}+\left(-\frac{\omega_-}{\omega_+}x_i\sin\frac{\pi\omega_+}{\omega}+\frac{v_i}{\omega_-}\cos\frac{\pi\omega_+}{\omega}\right)\sin\frac{\pi\omega_-}{\omega}$$

$$v_{i+1}=\left(-\omega_- x_i\cos\frac{\pi\omega_+}{\omega}+\frac{\omega_-}{\omega_+}v_i\sin\frac{\pi\omega_+}{\omega}\right)\sin\frac{\pi\omega_-}{\omega}+\left(\omega_+ x_i\sin\frac{\pi\omega_+}{\omega}+v_i\cos\frac{\pi\omega_+}{\omega}\right)\cos\frac{\pi\omega_-}{\omega}$$

We are almost ready except that the question asks for a matrix. We can easily rearrange terms to get the thing into a matrix form. We also use the shorthand $C_+ = \cos(\pi\omega_+/\omega)$, $C_- = \cos(\pi\omega_-/\omega)$, $S_+ = \sin(\pi\omega_+/\omega)$ and $S_- = \sin(\pi\omega_-/\omega)$ to simplify the appearance of the expressions. We then have

$$\begin{bmatrix} x_{i+1} \\ v_{i+1} \end{bmatrix} = \mathbf{T}\begin{bmatrix} x_i \\ v_i \end{bmatrix} \quad \text{, where} \qquad \text{(01a)*}$$

$$\mathbf{T} = \begin{bmatrix} C_+C_- - \dfrac{\omega_-}{\omega_+}S_+S_- & \dfrac{1}{\omega_+}S_+C_- + \dfrac{1}{\omega_-}C_+S_- \\ -\omega_-C_+S_- - \omega_+S_+C_- & -\dfrac{\omega_-}{\omega_+}S_+S_- + C_+C_- \end{bmatrix} \quad , \qquad \text{(01b)*}$$

which is the desired answer.

(b) Evidently, $[x_{i+2};\ v_{i+2}] = \mathbf{T}[x_{i+1};\ v_{i+1}] = \mathbf{T}^2[x_i;\ v_i]$, similarly $[x_{i+3};\ v_{i+3}] = \mathbf{T}^3[x_i;\ v_i]$ and so on. In the general case, $\underline{[x_{i+k};\ v_{i+k}] = \mathbf{T}^k[x_i;\ v_i]}$.

(c) The system will be stable if x and v decay to zero or remain bounded as time goes on. Since the solution remains of the same size within each period of excitation, this means that x and v at the start of each period must also be decaying or bounded i.e. x_{i+k} and v_{i+k} must remain bounded as k becomes large. Conversely, the system will be unstable if x and v are divergent, or x_{i+k} and v_{i+k} grow unboundedly as k becomes large.

Now, we have obtained that $[x_{i+k};\ v_{i+k}] = \mathbf{T}^k[x_i;\ v_i]$. The hint tells us to use the formula $\mathbf{T}^k = \mathbf{P}^{-1}\mathbf{D}^k\mathbf{P}$ where $\mathbf{D}$ is the diagonal matrix consisting of the eigenvalues of $\mathbf{T}$. Evidently, the entries of $\mathbf{T}^k$ will grow unboundedly with k if and only if at least one entry of $\mathbf{D}^k$ is very large. Writing the eigenvalue λ in the most general form of $R\,\mathrm{e}^{\mathrm{j}\theta}$ (where θ equals 0 or π if λ is real), we can see that its successive powers decrease in size if $R < 1$, increase in size if $R > 1$ and remain of same size if $R = 1$.

Substituting this into the result of Part (b) we see that $[x_{i+k};\ v_{i+k}]$ goes to zero for large k if $|\lambda_1|$ and $|\lambda_2|$ are both less than unity, it diverges if either of $|\lambda_1|$ or $|\lambda_2|$ is greater than unity, and it remains bounded if either of $|\lambda_1|$ or $|\lambda_2|$ is exactly unity and the other one is equal or smaller. Thus the condition for stability depends solely on the eigenvalue $\lambda_{\max}$ with the larger absolute value; the system is $\underline{\text{stable if}}$ $|\lambda_{\max}| \le 1$ $\underline{\text{and unstable if}}$ $|\lambda_{\max}| > 1$. Note the

difference from the stability conditions in 2.4, where the real part of the largest eigenvalue must be less than or equal to ZERO. Recall that we've already seen such a thing in example 4.1.4 Part (b), unless you skipped that.

(d) If $a << 1$, ω_+ becomes $1+a/2$ and ω_- becomes $1-a/2$. The trig products in (01b) become trigs of sums and differences; the sum frequency is $\omega_+ + \omega_- = 2$ while the difference frequency is $\omega_+ - \omega_- = a$. We further note that ω_+ and ω_- are reciprocals of each other, which simplifies the quotient terms in $\mathbf{T}$. Then, hacking at the trig and checking/repeating the calculation over and over again until the signs converge, we get

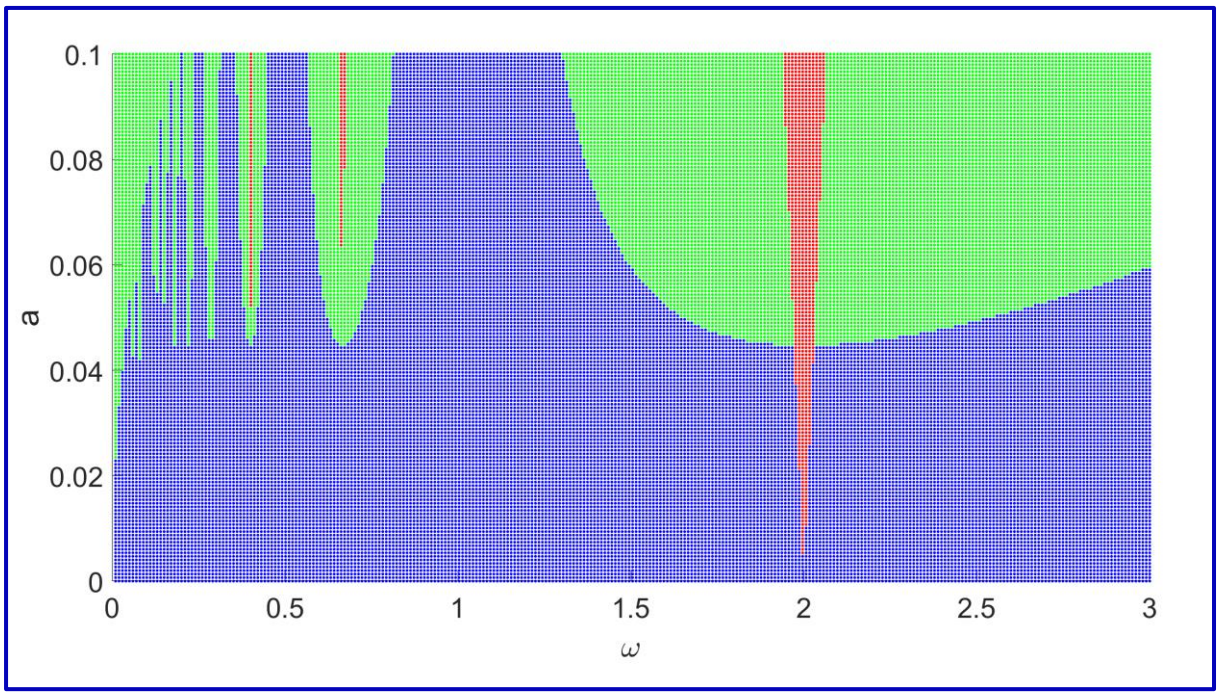
$$\mathbf{T} = \begin{bmatrix} \cos\frac{2\pi}{\omega} - \frac{a}{2}\left(1-\cos\frac{2\pi}{\omega}\right) & \sin\frac{2\pi}{\omega} - \frac{\pi a^2}{2\omega} \\ -\sin\frac{2\pi}{\omega} - \frac{\pi a^2}{2\omega} & \cos\frac{2\pi}{\omega} + \frac{a}{2}\left(1-\cos\frac{2\pi}{\omega}\right) \end{bmatrix} .$$

The characteristic polynomial is $\det(\mathbf{T}-\lambda\mathbf{I})=0$, which after some manipulation yields

$$\lambda^2 + \lambda\left(-2\cos\frac{2\pi}{\omega}\right) + 1 - \frac{a^2}{4}\left(1-\cos\frac{2\pi}{\omega}\right)^2 - \frac{\pi^2 a^4}{16\omega^2} = 0 \quad . \qquad (02)$$

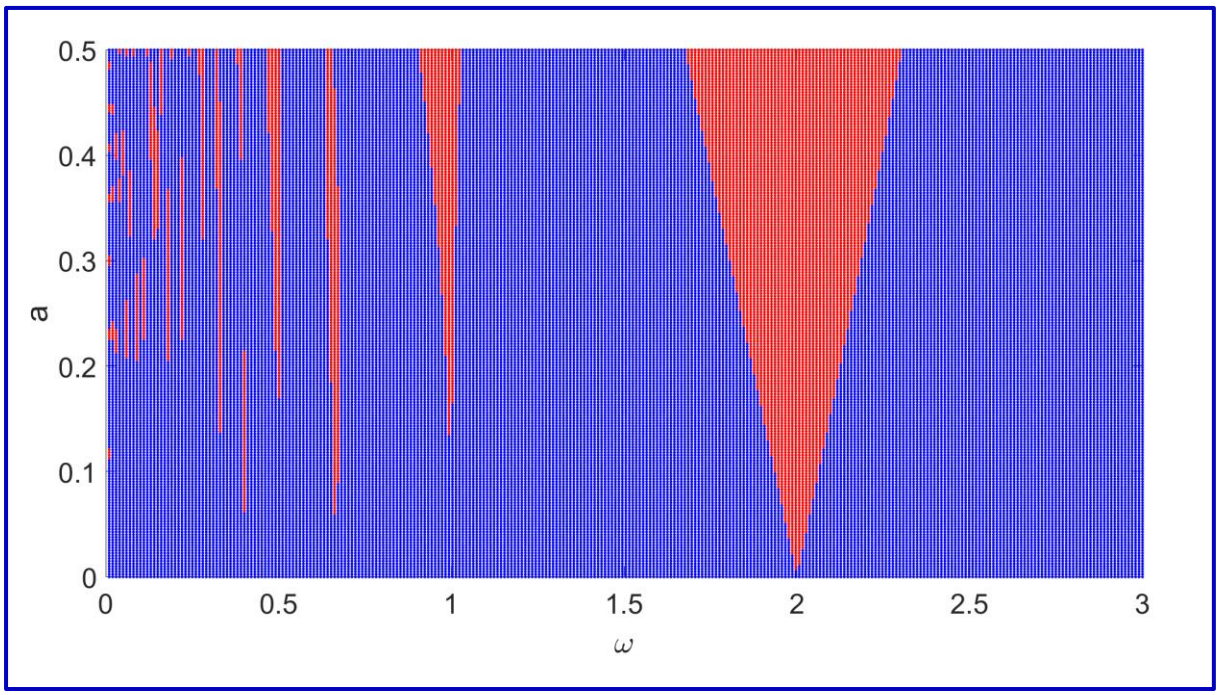
We see that in the limit $a = 0$ (where the Mathieu equation becomes harmonic oscillator equation with natural frequency 1), the eigenvalues reduce to $\exp \mathrm{j}(\pm 2\pi/\omega)$. This makes perfect sense because the system evolves periodically with period 2π, so if we sample at frequency ω, we should expect to see a phase shift or 'rotation' of the ICs through angle $2\pi/\omega$. This limiting case acts as a necessary check on our work so far.

Although (02) can be solved analytically, it is just a tedious exercise in algebra so we feed it to the computer and have it plot the eigenvalues in the region of parameter space asked for. We break the ω-axis into 300 points and the a-axis into 150 points and colour each point green if the largest eigenvalue is less than unity, red if greater than unity and blue if exactly unity (to within machine precision). The plot is shown below.



We can see a very pronounced instability at $\omega = 2$; the other two red lines might be inauthentic because they are surrounded by green regions. The more accurate plot of the next part of this problem will be required to resolve this issue.

(e) We write a simple Matlab code which does the required job. The ω and a axes are now split into 300 and 200 points respectively and the colour code remains the same as in Part (d). The results are shown below.



The big instability at $\omega = 2$ remains, as do the smaller instabilities to the left of $\omega = 1$. A new instability emerges at $\omega = 1$, which is in fact the second-largest instability zone in the diagram. The green regions i.e. decaying solutions disappear entirely, giving way to blue regions i.e. bounded solutions only.

*

COMMENTS : This simplified analysis captures most features of the actual Mathieu equation. The most prominent feature of the plot is the huge red tongue of instability coming out from $\omega = 2$. Since the natural frequency of the oscillator (in the absence of the excitation) is $\omega_0 = 1$, this instability occurs at excitation frequency twice of the natural frequency. This is a faithful reproduction of the parent system – that too is characterized by a massive instability at excitation twice of natural, called the 2:1 subharmonic resonance. Note the contrasts between this resonance and the ordinary linear resonance – that occurs at exactly $\omega = \omega_0$ and nowhere else (a LINE in the ω-a plane) while this occurs in a whole region of parameter space centred round $\omega = 2\omega_0$. That resonance is linear growth while this is exponential growth (eigenvalues larger than unity so x_{i+k} and v_{i+k} go as something to the power of k). There are smaller instability bands at other frequencies; these phenomena too are quite accurate. The second-strongest instability in Mathieu equation occurs at excitation

frequency equals natural frequency; this is called the 1:1 resonance and is clearly visible here. In reality, the instability tongues reach all the way down to $a=0$; here they appear to stop short because the eigenvalue is greater than unity by an amount below the precision of the computer.

Within the stable regions, the solutions are in general quasiperiodic. The eigenvalues here are $\exp \mathrm{j}(\pm\theta)$ for some θ, which implies phase shift of x_{i+k} and v_{i+k} through angle θ over every cycle. This is like a modulation term; this modulation has an effective characteristic frequency of $\theta\omega/2\pi$. On the other hand, within each cycle, there are motions at frequency ω_+ and ω_-. In general the internal frequencies and the modulating frequency are incommensurable (have an irrational ratio) so the overall motion is quasiperiodic.

At this point let me mention another "approximate Mathieu" equation which is popular in Literature. This replaces the sinusoidal excitation with a series of delta functions – instead of $\mathrm{sq}\,\omega t$ like we had here, that version has something like $\sum_n \delta(t - 2n\pi/\omega)$. Although that is easier to work with, the stability zones have very little resemblance to those of the actual Mathieu equation, and so we did not discuss it here.

For the real Mathieu equation, the transition curves between stable and unstable can be obtained using perturbation theory, a topic which lies in the field of nonlinear dynamics and is outside the scope here – for a good exposition see RICHARD RAND's lecture notes. The behaviour in the stable zones (which IS quasiperiodic) can be obtained from Floquet theory – again outside our scope but introduced in RAND and covered in detail elsewhere. Our primary purpose here was to create a simple equation which can nevertheless explain the observations I stated as facts in the comments to 3.3.10 Part (c). Moreover , this method – replacing a complicated oscillatory function with square wave and then constructing the matrix $\mathbf{T}$ – can give you good approximate results in any kind of system with parametric excitation.

A variation on the Mathieu equation is the inverted pendulum or Kapitza pendulum, first discovered by PYOTR LEONIDOVICH KAPITZA. This consists of a pendulum whose base is oscillated rapidly at some frequency ω in the vertical direction. We saw in T054 that its equation of motion is

$$\ddot{\theta} + \frac{1}{l}\left(g + a_0 \cos \omega t\right) \sin \theta = 0 \quad ,$$

where a_0 is the amplitude of the acceleration of the suspension point. For $a_0 = 0$, the inverted position $\theta = \pi$ is evidently unstable. For sufficiently high a_0 however, it turns out that the inverted position becomes stable. Although we can get this by the square wave method, there's a much easier nonlinear dynamical method which leads to the result – see any book on the subject for that.

Example 4.4.3

Chopper control of dc motor

The current $i(t)$ in the winding of a dc motor connected to a chopper satisfies the equation

$$i' + \gamma i = V(t) \quad ,$$

where γ is a constant and V is the externally applied voltage. With a chopper, $V(t)$ has two possible values : $V = V_0$ (a constant) if the chopper is ON and $V = 0$ if the chopper is OFF. In hysteresis control, also called bang-bang or two-limit control, two current values are specified. One is the output desired from the motor, called the set point i_{set}, while the other is the tolerable fluctuation level, i_{tol}. Typically $i_{tol} << i_{set}$. The chopper switches from ON to OFF if $i = i_{set} + i_{tol}$ and it switches from OFF to ON if $i = i_{set} - i_{tol}$. The chopper is ON if $i < i_{set} - i_{tol}$ and OFF if $i > i_{set} + i_{tol}$.

For a particular motor, γ, V_0, i_{set} and i_{tol} ($<< i_{set}$) are given. The circuit is quiescent when the apparatus is switched on.

(a) Determine the condition under which the steady state operation features the chopper in an alternating ON-OFF-ON-OFF state.

(b) If the above condition is satisfied, determine the switching period (time interval between two successive ON→OFF transitions) and duty cycle of the chopper and the trajectory of the current in the steady state.

[Level : 2]

(a) Since the system is initially quiescent, the chopper comes ON when it is started. The motor equation then is $i'+\gamma i = V_0$ and IC is $i(0)=0$. Such a simple ODE at this late stage in the book might be a surprise but don't get confused and start searching for hidden complications – there aren't any ! We immediately know what its solution looks like – the current increases monotonically towards its fixed point, which is V_0/γ. The hysteresis control is essentially an oven thermostat – the coils (chopper) turn ON when the temperature (current) is too low and turn OFF when the temperature becomes too high. Since the ON→OFF transition takes place at $i = i_{\text{set}}+i_{\text{tol}}$, the chopper will turn OFF if i becomes equal to this value on its way to the fixed point. In other words, the upper cutoff must be LESS than the fixed point. Once the chopper turns off, the current equation becomes $i'+i=0$ which features decay of current. Sooner or later, the lower cutoff will be hit and the chopper will turn ON again, thus establishing an alternating state. Hence the condition for alternating operation is $\underline{i_{\text{set}}+i_{\text{tol}} < V_0/\gamma}$. (By an analogous argument, the chopper will never turn ON if $i_{\text{set}}-i_{\text{tol}}<0$; this is a mathematically bonafide but practically senseless condition and is neglected from consideration.)

(b) We start from the instant that the chopper turns OFF. The ODE is $i'+\gamma i=0$, the current at this time (which is our IC) is $i_{\text{set}}+i_{\text{tol}}$ so the solution is $i=(i_{\text{set}}+i_{\text{tol}})\mathrm{e}^{-\gamma t}$. This mode of operation persists until i becomes equal to $i_{\text{set}}-i_{\text{tol}}$. The time taken for this to happen satisfies a transcendental equation but we simplify using the information that $i_{\text{tol}} << i_{\text{set}}$. Basis this inequality we expect that the current will hit the lower cutoff a very short time after the chopper turns OFF (short compared to $1/\gamma$). Accordingly, we can Taylor expand the exponential and retain the first non-trivial term, getting $i=(i_{\text{set}}+i_{\text{tol}})(1-\gamma t)$. Now the equation for the time to hit the lower cutoff is almost trivial to solve; we have

$$t_{\text{OFF}} = \frac{2i_{\text{tol}}}{\gamma\left(i_{\text{set}}+i_{\text{tol}}\right)} \quad .$$

Likewise, we can handle the ON phase of the chopper. Here, the ODE is $i'+\gamma i=V_0$ and IC is $i_{\text{set}}-i_{\text{tol}}$ giving the solution

$$i = \frac{V_0}{\gamma} + \left(i_{\text{set}} - i_{\text{tol}} - \frac{V_0}{\gamma}\right)\mathrm{e}^{-\gamma t} \quad .$$

This solution remains valid until i becomes $i_{\text{set}}+i_{\text{tol}}$, when the chopper turns OFF. Again assuming small time and Taylor expanding the exponential, we have

$$t_{\text{ON}} = \frac{2i_{\text{tol}}}{V_0 - \gamma\left(i_{\text{set}} - i_{\text{tol}}\right)} \quad .$$

We are now ready to report answers.

The switching period is the sum of t_{OFF} and t_{ON}; dropping i_{tol} relative to i_{set} in the two denominators we have

$$T = 2i_{\text{tol}}\left(\frac{1}{\gamma i_{\text{set}}} + \frac{1}{V_0 - \gamma i_{\text{set}}}\right) \quad . \qquad *$$

The duty cycle is the ratio of ON time to total time, which is t_{ON}/T; simplifying fractions gives $\underline{d = \gamma i_{\text{set}}/V_0}$. Finally, the trajectory of $i(t)$ is just two pieces of exponentials (which look almost like straight lines) from the upper cutoff to the lower cutoff and back again.

*

COMMENTS : The full equation of the series wound dc motor is

$$Li' + Ri = V - k\omega i \quad ,$$
$$I\dot{\omega} = Ci^2 \quad ,$$

where L is the winding inductance, R the winding resistance, k the "motor constant", ω the rotation speed, I the moment of inertia of the load and C the "torque constant". Basically, what it means is that the winding is like an LR circuit which sees the applied voltage as well as a back-emf arising from the rotation of the rotor coils through the stator magnetic field (the speed is ω and the field proportional to i hence this term is proportional to their product). The rate of change of speed is proportional to the torque, which is the product of the stator and rotor currents (both equal to i) and hence goes as i^2.

This is a second order nonlinear equation – how did we reduce it to the kiddo ODE with which we started the whole book ? We did so by the constant ω approximation. If the rate of variation of ω is much smaller than that of i then ω can be treated as a constant in the i-equation, and we recover the starting point of this problem. This constant ω assumption is almost universal in Literature; here it is bonafide because we are interested in the dynamics of i over only very

small intervals (t_{ON} and t_{OFF}). During these intervals we can reasonably expect the speed to stay almost constant. The assumption breaks down completely when ω is very close to zero i.e. when the motor is just starting – for this reason, ultra-low speed control of motors is a hair-raising task.

The extremely simple equation of the motor (except at very low speeds) means that currents increase or decrease monotonically even in the absence of control; this feature enables uncontrolled dc motors to be used even in demanding applications like electric locomotives. An example of such loco is WAP4; the motors in the loco are connected directly via a rectifier (ac to dc) to a tap changer (variable turns ratio transformer) which links to the throttle. Changing the throttle settings causes smooth changes in currents and torque and correspondingly jerk-free acceleration. Compare with what happened when the induction motor was line started in 3.3.1 – the currents shot up rapidly to twice their set values and went up and down before settling into their steady states.

With even a simple control scheme, the regulation of speed and torque of dc motor is extremely precise. The ostensibly crude algorithm we saw here actually forms the core of most high-end dc motor control systems. Power electronic devices are such that it is easy to practically implement this algorithm with a tolerance level of a few parts in thousand of the set point. These features have resulted in the enduring popularity of dc motor in industry and technology, despite its drawbacks compared to ac motor (presence of commutator and brushes which have high wear and tear rate).

Nevertheless, the plus points of ac motor are undeniable – it gives you much greater power to weight ratio, higher speed of operation, and if it is manufactured well, it lasts a lifetime. This means that for many applications, ac motors are more desirable than their dc counterparts, even though their uncontrolled performance is lousy and a good control scheme is far more complicated. The industry standard algorithm which achieves this control is the subject of the next Tute. That by the way uses hysteresis control and is the hardest ODE problem in the book, so make sure you understand the concepts here really well before you take on the challenge.

Example 4.4.4

Direct torque control of induction motor

An induction motor can be thought of as two sets of windings, called direct and quadrature (*d* and *q*) on each of the rotor and stator (*r* and *s*). When the motor is controlled by a voltage source, the dynamical variables are the currents in all four windings – i_{rd}, i_{rq}, i_{sd} and i_{sq} – and external voltage V_d and V_q is applied on the stator (the subscript *s* on *V* is typically omitted). In the space phasor model, the direct and quadrature components of current and voltage are combined into complex quantities called phasors or vectors. Thus, rotor current vector $\mathbf{i}_r = i_{rd} + \mathrm{j} i_{rq}$, stator current vector $\mathbf{i}_s = i_{sd} + \mathrm{j} i_{sq}$, voltage vector $\mathbf{V} = V_d + \mathrm{j} V_q$ etc.

The equation of the motor in space phasor variables is

$$\mathbf{i}_r' + (1 - \mathrm{j}\omega)\mathbf{i}_r + \delta(\mathbf{i}_s' - \mathrm{j}\omega\mathbf{i}_s) = 0 \quad ,$$
$$\mathbf{i}_s' + \delta\mathbf{i}_r' = \mathbf{V} \quad ,$$

where δ is a (real – all complex quantities are non-italicized and in bold face) constant between zero and unity, and ω is a positive constant. The torque output of the motor is

$$\Gamma = i_{rd} i_{sq} - i_{sd} i_{rq} \quad ,$$

which can be thought of as the "cross product" of the rotor and stator current phasors if one is willing to treat them like vectors in the *d*-*q* plane. The similarities between phasors and "real vectors" often affords insight into induction motor dynamics and will likely be useful for this problem.

(a) Define the combined flux phasor $\boldsymbol{\psi}_+ = \delta\mathbf{i}_r + \mathbf{i}_s$ and obtain the equations of motion as well as the torque in terms of $\boldsymbol{\psi}_+$ and $\mathbf{i}_r$.

(b) A three-phase sinusoidal voltage at frequency Ω, as obtained from a typical supply line, can be written as $\mathbf{V} = \mathbf{U}_0 \mathrm{e}^{\mathrm{j}\Omega t}$ where $\mathbf{U}_0$ is a constant vector. With this $\mathbf{V}$, obtain the particular solutions for $\boldsymbol{\psi}_+$ and $\mathbf{i}_r$. Defining the slip frequency to be $\varepsilon = \Omega - \omega$, obtain the torque as a function of ε.

(c) For a line-supplied motor with $\omega = 50$, $\delta = 0.9$, and $|\boldsymbol{\psi}_+| = 1$, determine the lower of the two values of ε which produce a torque output of 0.25. The upper value is an unstable operating point of the motor so it is not of practical interest.

In direct torque control (DTC), the applied voltage vector is not continuous in time but changes in discrete steps. It is held constant at some vector value $\mathbf{V}_0$ for a small duration, called the switching interval T. T is considerably smaller than all other times involved in the problem, and its minimum possible value is a property of the control system. Let $\boldsymbol{\psi}_0$ and $\mathbf{i}_0$ denote the values of $\boldsymbol{\psi}_+$ and $\mathbf{i}_r$ at the start of a particular switching interval. The value of $\mathbf{V}_0$ to be held during the interval is determined basis the values of $\boldsymbol{\psi}_0$ and $\mathbf{i}_0$, and the behaviour desired from the motor during the interval.

When it is desired to build flux at zero torque, the voltage vectors are chosen such that the size of the flux vector increases while the rotor current vector remains diametrically opposite to the flux vector.

(d) Given $\boldsymbol{\psi}_0$ and $\mathbf{i}_0$ diametrically opposite, show that while certain conditions are satisfied (which you need not determine explicitly) it is possible to choose $\mathbf{V}_0$ such that at the end of the interval, the size of the flux vector increases while the rotor current remains diametrically opposite.

Once the rated flux level is built up, the torque must be generated. This is done by causing the vector $\boldsymbol{\psi}_+$ to rotate while keeping its magnitude constant. For this purpose, the voltage vector selected is $\mathbf{V}_0 = j\alpha\boldsymbol{\psi}_0$ where α is a positive real number.

(e) Show that while $|\mathbf{i}_r|$ is sufficiently small compared to $|\boldsymbol{\psi}_+|$ (a condition which you again need not examine in greater detail), a high value of α causes an increase in torque during the switching interval while the value α=0 (zero applied voltage) causes a reduction in torque during the interval.

In DTC, the result of Part (e) is employed for hysteresis control of the torque while keeping the flux constant at the rated value. The

set point and tolerance values Γ_{set} and Γ_{tol} are prescribed and two values of α are employed, $\alpha = \alpha_{max}$, a very high value limited only by the system parameters, and $\alpha = 0$. The former corresponds to the ON state and the latter to the OFF state. The ON→OFF transition is made at $\Gamma = \Gamma_{set} + \Gamma_{tol}$, the OFF→ON transition at $\Gamma = \Gamma_{set} - \Gamma_{tol}$ and the system remains ON if $\Gamma < \Gamma_{set} - \Gamma_{tol}$ and OFF if $\Gamma > \Gamma_{set} + \Gamma_{tol}$.

(f) The motor of Part (c) having $\omega = 50$, $\delta = 0.9$ and $|\boldsymbol{\psi}_+| = 1$ is now operated using direct torque control with $\Gamma_{set} = 0.25$, $\Gamma_{tol} = 0.01$ and $\alpha_{max}=\infty$. Using a computer where appropriate, determine the average angular velocities of $\boldsymbol{\psi}_+$ and $\mathbf{i}_r$ in the steady state.

[Level : 6]

Six off the last ball ! A problem like this is excellent for examinations like the erstwhile SMITH's Prize exam in Cambridge where the prof gives you unlimited time, paper, food etc (in one instance, Sir ARTHUR CAYLEY had plied the candidates with booze!) and stretches you to your limits in a relaxed environment. When faced with such a monster, the best strategy is of course to make full use of the unlimited available time. First read the problem and get a hang of the overall picture. Then narrow down your focus and take each part systematically while not bothering about what comes later. In particular, Parts (a) to (c) are simple variable transformation questions which do not in any way require the control concepts – the phrase "direct torque control" appears only after Part (c). So get them out of the way before focussing on the control.

(a) The second equation becomes almost trivial in the new variables : $\dot{\boldsymbol{\psi}}_+ = \mathbf{V}$. To transform the first one, do some simple addition and subtraction,

$$\begin{aligned}
&\mathbf{i}_r' + (1 - \mathrm{j}\omega)\mathbf{i}_r + \delta \mathbf{i}_s' - \mathrm{j}\delta\omega\mathbf{i}_s \\
&= \delta^2 \mathbf{i}_r' + (1-\delta^2)\mathbf{i}_r' + \delta \mathbf{i}_s' - \mathrm{j}\delta\omega\mathbf{i}_s - \mathrm{j}\delta^2\omega\mathbf{i}_r + (1 - \mathrm{j}\omega + \mathrm{j}\delta^2\omega)\mathbf{i}_r \quad . \\
&= \delta\dot{\boldsymbol{\psi}}_+ + (1-\delta^2)\mathbf{i}_r' - \mathrm{j}\delta\omega\boldsymbol{\psi}_+ + (1 - \mathrm{j}\omega(1-\delta^2))\mathbf{i}_r
\end{aligned}$$

Then, collecting terms and letting $\lambda = 1/(1-\delta^2)$, we have the two equations

* $$\dot{\boldsymbol{\psi}}_+ = \mathbf{V} \quad , \qquad (01a)^*$$

* $$\mathbf{i}_r' + (\lambda - \mathrm{j}\omega)\mathbf{i}_r = \delta(-\dot{\boldsymbol{\psi}}_+ + \mathrm{j}\omega\boldsymbol{\psi}_+) \quad , \qquad (01b)^*$$

which is the ODE system asked for.

To find the torque, the cross product interpretation is the easiest handle. We are given that $\Gamma = "\mathbf{i}_r \times \mathbf{i}_s"$ (the cross product in quotes as the definition works but it's just a mathematical trick) and we must now introduce $\boldsymbol{\psi}_+$ in place of $\mathbf{i}_s$. We guess that torque is the cross product of rotor current and combined flux; since $\boldsymbol{\psi}_+ = \delta\mathbf{i}_r + \mathbf{i}_s$ and the cross product of a vector with itself is zero, this reduces to the original answer and is correct. Hence

* $$\Gamma = "\mathbf{i}_r \times \boldsymbol{\psi}_+" = i_{rd}\psi_{+q} - \psi_{+d}i_{rq} \quad , \qquad *$$

and the change of variables is complete.

(b) For this we try particular solutions of the form $\mathrm{e}^{\mathrm{j}\Omega t}$ for both $\boldsymbol{\psi}_+$ and $\mathbf{i}_r$. An easy undetermined coefficient calculation (but with COMPLEX constants!) gives

* $$\boldsymbol{\psi}_+ = \frac{1}{\mathrm{j}\Omega}\mathbf{U}_0\,\mathrm{e}^{\mathrm{j}\Omega t} \quad , \qquad (02a)^*$$

* $$\mathbf{i}_r = -\frac{\varepsilon\delta\left(\lambda - \mathrm{j}\varepsilon\right)}{\Omega\left(\lambda^2 + \varepsilon^2\right)}\mathbf{U}_0\,\mathrm{e}^{\mathrm{j}\Omega t} \quad , \qquad (02b)^*$$

where $\varepsilon = \Omega - \omega$ as specified in the problem.

An $\mathrm{e}^{\mathrm{j}\Omega t}$ term rotates in the *d*-*q* plane with frequency Ω. Equation (02) shows that both $\boldsymbol{\psi}_+$ and $\mathbf{i}_r$ have constant magnitude and make the same angle relative to each other; thus their cross product the torque is a constant. To evaluate that constant, let's assume without loss of generality that $\mathbf{U}_0$ is purely real; then $\boldsymbol{\psi}_+$ is purely imaginary and the torque comes from the real part of the rotor current. Doing the multiplication, the torque is

* $$\Gamma = \frac{\lambda\delta\,|\,\mathbf{U}_0\,|^2}{\Omega^2}\frac{\varepsilon}{\lambda^2 + \varepsilon^2} \quad , \qquad *$$

which is the desired answer. If you are familiar with electric machines then of course you know this result, likely arising from a different derivation which bypasses the space phasor dynamic model. (If you know the answer from beforehand and see that you did NOT recover it now, go back and check your work. The answer has to be the same whatever the process involved in getting there.)

(c) A simple calculation (by either hand or computer) gives $\underline{\varepsilon = 1.596}$. The preliminary calculations are now done with – we have gained some familiarity with the induction motor system and have earned some partial credits to boot.

(d) Exam tip : the text before the question gives a hint about the answers to Parts (a) and (b). It tells you that if the (combined) flux and (rotor) current vectors are diametrically opposite, then the torque should be zero. Check that your previous answers satisfy this property before moving on (they do). Then, we are given $\boldsymbol{\psi}_0$ and $\mathbf{i}_0$; since $\mathbf{i}_0$ is diametrically opposite to $\boldsymbol{\psi}_0$, we can write it as $\mathbf{i}_0 = -c_1\boldsymbol{\psi}_0$ where c_1 is positive and real. Given these, we now try to find the response of $\boldsymbol{\psi}_+$ and $\mathbf{i}_r$ to an arbitrary but constant applied voltage vector $\mathbf{V}_0$. The ODE (01a) with the given IC has the obvious solution

$$\boldsymbol{\psi}_+ = \boldsymbol{\psi}_0 + \mathbf{V}_0 t \quad , \qquad (03)$$

and substituting that into (01b) gives

$$\mathbf{i}_r' + (\lambda - \mathrm{j}\omega)\mathbf{i}_r = -\delta\mathbf{V}_0 + \mathrm{j}\delta\omega\boldsymbol{\psi}_0 + \mathrm{j}\delta\omega\mathbf{V}_0 t \quad ,$$

with the IC $\mathbf{i}_r(0) = \mathbf{i}_0$.

The solution of this ODE (homogeneous plus particular using undet coeff – one cannot help but marvel at the versatility of this method, it has given us solutions for heaven knows how many real life problems) is

$$\mathbf{i}_r = \mathbf{c}\mathrm{e}^{(-\lambda+\mathrm{j}\omega)t} + \mathbf{a} + \mathbf{b}t \quad \text{, where} \tag{04a}$$

$$\mathbf{a} = \frac{\delta}{\lambda - \mathrm{j}\omega}\left[\left(-1 - \frac{\mathrm{j}\omega}{\lambda - \mathrm{j}\omega}\right)\mathbf{V}_0 + \mathrm{j}\omega\boldsymbol{\psi}_0\right] \quad , \tag{04b}$$

$$\mathbf{b} = \frac{\mathrm{j}\delta\omega\mathbf{V}_0}{\lambda - \mathrm{j}\omega} \quad , \tag{04c}$$

$$\mathbf{c} = \mathbf{i}_0 - \frac{\delta}{\lambda - \mathrm{j}\omega}\left[\left(-1 - \frac{\mathrm{j}\omega}{\lambda - \mathrm{j}\omega}\right)\mathbf{V}_0 + \mathrm{j}\omega\boldsymbol{\psi}_0\right] \quad . \tag{04d}$$

Note that $\mathbf{a}+\mathbf{c} = \mathbf{i}_0$, as required to satisfy the IC.

What we are interested in however is the solution only at a SMALL time T after the vector is switched on; the general solution above is not of any significance. To obtain the small time solution, we Taylor expand (04a) and retain the first non-trivial terms; that gives $\mathbf{i}_r(T) = \mathbf{a}+\mathbf{c}+\mathbf{b}T+(-\lambda+\mathrm{j}\omega)\mathbf{c}T$. Substituting the values and simplifying the algebra we get

$$\mathbf{i}_r(T) = \mathbf{i}_0 + T\left[(-\lambda + \mathrm{j}\omega)\mathbf{i}_0 - \delta\mathbf{V}_0 + \mathrm{j}\delta\omega\boldsymbol{\psi}_0\right] \quad . \tag{05}$$

With the clairvoyance typical of the good problem solver, we have refrained from substituting in the condition $\mathbf{i}_0 = -c_1\boldsymbol{\psi}_0$ so far. The above is a completely general solution for arbitrary $\mathbf{V}_0$, $\boldsymbol{\psi}_0$ and $\mathbf{i}_0$ and it will be of use in solving the next part of this question. Only now plug in the oppositeness condition and get

$$\mathbf{i}_r(T) = \mathbf{i}_0 + T\left[-\delta\mathbf{V}_0 - c_1(-\lambda + \mathrm{j}\omega)\boldsymbol{\psi}_0 + \mathrm{j}\delta\omega\boldsymbol{\psi}_0\right] \quad .$$

Coming over to the flux (03), we see that during the switching interval, $\boldsymbol{\psi}_+$ has changed by the amount $\mathbf{V}_0 T$; since this causes an increase in $|\boldsymbol{\psi}_+|$, the vector $\mathbf{V}_0$ must be having a POSITIVE component parallel to $\boldsymbol{\psi}_0$. On the other hand, for the current to remain diametrically opposite to the flux, we want the change in $\mathbf{i}_r$ to be the change in $\boldsymbol{\psi}_+$ times some NEGATIVE real number. (To really split hairs, a positive real would also work over one switching interval because the ICs are

aligned oppositely; however it would fall through after a large number of intervals. A negative real would work always, so we go with that.) This means that (factoring out the T from the changes in $\mathbf{i}_r$ and $\boldsymbol{\psi}_+$ during the interval)

$$-\delta\mathbf{V}_0 + \left(-c_1(-\lambda + \mathrm{j}\omega) + \mathrm{j}\delta\omega\right)\boldsymbol{\psi}_0 = -c_2\mathbf{V}_0 \quad ,$$

where c_2 is real and positive. In other words,

$$\mathbf{V}_0 = \frac{-c_1\left(-\lambda + \mathrm{j}\omega\right) + \mathrm{j}\delta\omega}{\delta - c_2}\boldsymbol{\psi}_\mathbf{0} \quad .$$

So long as $c_2 < \delta$, this indeed has a positive component parallel to $\boldsymbol{\psi}_0$ hence it is feasible to achieve the increase in flux while keeping the current diametrically opposite, which was to be shown. Fortunately the problem does not ask us to investigate the circumstances under which $c_2 < \delta$ holds true, and we are done.

(e) We are given that for torque generation $\mathbf{V}_0 = \mathrm{j}\alpha\boldsymbol{\psi}_0$. Without loss of generality let us assume that $\boldsymbol{\psi}_0$ lies along the d-axis (real axis). Then, before the torque generation starts, $\mathbf{i}_r$ lies along the negative d-axis; the torque generation steps presumably push it into the second or third quadrant. Let us see which one.

At the end of the switching interval, the new flux vector is obviously $\boldsymbol{\psi}_0 + \mathrm{j}\alpha\boldsymbol{\psi}_0 T$, so it moves up along the q-axis. As for the current vector, we use (05) to get the change in $\mathbf{i}_r$ during the interval as

$$\mathbf{i}_r(T) - \mathbf{i}_0 = T\left[\left(-\lambda + \mathrm{j}\omega\right)\mathbf{i}_0 - \mathrm{j}\delta\alpha\boldsymbol{\psi}_0 + \mathrm{j}\delta\omega\boldsymbol{\psi}_0\right] \quad . \qquad (06)$$

When α is very high, the largest term in the above RHS is the second one, and this causes the current vector to acquire a negative quadrature component, pushing it into the third quadrant. In other words, the current rotates counter-clockwise (CCW). Because of the α, the flux vector has also rotated CCW, but since by the problem assumptions it is sufficiently larger than the current vector, the angle it covers is less. Hence at the end of the interval, the angle between the current and flux changes from 180° to something less than that; the cross product torque which depends on the sine of the angle then becomes positive. The higher the value of α, the greater the extent to which the current vector gets pushed into the third quadrant, and the greater the increase in torque during the interval. The same reasoning holds true if $\mathbf{i}_0$ is not exactly along negative d-axis but somewhere in the bulk of the third quadrant – high α still means a CCW rotation relative to the flux vector and an increase in torque.

When $\alpha = 0$ on the other hand, only the first and third terms in (06) are non-trivial. The first one causes the third-quadrant rotor current to decrease along its length and turn CCW. The third term however makes pushes it towards the negative d-axis, making it turn clockwise. Once again since the flux is sufficiently greater than the current, the second effect dominates and the torque decreases. Thus, both the statements given in the question have been shown. Fortunately, the problem again spares us the tedium of analysing exactly how small the current must be compared to the flux for this to work.

(f) There is hysteresis control of the torque with $\alpha = \alpha_{\max}$ and $\alpha = 0$ playing the role of the ON and OFF states of the chopper. The steady state solution is of course periodic and looks something like the following. Without loss of generality we consider an instant of time when $\boldsymbol{\psi}_+$ is real. When the chopper comes ON, the torque must be at its lower cutoff value, which is 0.24. The current vector is somewhere in the third quadrant – we don't know where exactly. During the ON phase, both flux and current turn CCW, current going faster than flux so that the torque increases. When torque becomes 0.26, chopper turns OFF. Flux vector becomes stationary and rotor current goes clockwise so the torque decreases. When the chopper turns ON again, the torque is of course 0.24. But we have more. By definition of the steady state, not only must the torque be equal to what it was last time but the current must be in the SAME CONFIGURATION AS IT WAS LAST TIME, RELATIVE TO THE NEW POSITION of the flux vector. This is the boundary matching condition; expressing it quantitatively will give us the value of $\mathbf{i}_r$ at the instant the chopper turned ON (and hence immediately at any other point in its trajectory).

Since the size of $\boldsymbol{\psi}_+$ is 1 and since we are assuming it to be along the d-axis when the chopper turns ON, its value at this time is $\boldsymbol{\psi}_0 = 1+\mathrm{j}0$. Since the torque at this instant is 0.24, the quadrature component of $\mathbf{i}_r$ must be -0.24; the direct component is some unknown quantity which we call A. Hence $\mathbf{i}_0 = A-\mathrm{j}0.24$ when the chopper comes ON.

Let T_1 be the duration that the chopper remains ON. In this interval, (06) will be the appropriate equation for $\mathbf{i}_r$, and we now use the information that $\alpha_{\max} = \infty$. This means that all terms in the RHS of (06) will be negligible compared to the second term and can be dropped. Further, since the rate of change is infinite, the ON phase will last only for an infinitesimal interval, such that the total

change in $\mathbf{i}_r$ is finite. Thus, T_1 is a meaningless variable and it is the product $\alpha_{\max} T_1$ which makes sense. Let $\beta = \alpha_{\max} T_1$, and let $\boldsymbol{\psi}_1$ and $\mathbf{i}_1$ denote the flux and current at the end of the interval. Then, from (06),

$$\mathbf{i}_1 = A - \mathrm{j}0.24 - \mathrm{j}0.9\beta \quad ,$$

where we have used that $\delta = 0.9$. At this time, integrating the flux equation gives

$$\boldsymbol{\psi}_1 = 1 + \mathrm{j}\beta \quad .$$

Evaluating the torque and setting it equal to the upper cutoff for the chopper gives $0.24 + A\beta + 0.9\beta = 0.26$ or

$$A\beta + 0.9\beta = 0.02 \quad . \tag{07}$$

This is one equation for the unknowns and β. It is nonlinear, which is probably why the question allows a computer at some stage. But it's too early for that yet, now we'll have to handle the OFF phase. A quick observation is that A is likely to be of order unity (same size as i_{rq} which is 0.24) and hence β is quite small. This is consistent with the fact that a rotation through angle of order β produces a small change in torque (Γ_{tol}) relative to the basal torque (Γ_{set}) so we would expect β to be of the same size as $\Gamma_{\mathrm{tol}}/\Gamma_{\mathrm{set}}$.

Let an unknown time T_2 be the duration of the OFF phase. During this phase $\boldsymbol{\psi}_+$ remains constant at its value of $1+\mathrm{j}\beta$. The current evolves as per (06) and at time T_2 it becomes equal to

$$\begin{aligned} \mathbf{i}_1 &= \mathbf{i}_r(T_2) \\ &= A - \mathrm{j}\big(0.24 + 0.9\beta\big) + T_2\big[-5.26A + 12 + \mathrm{j}\big(50A + 46.26 + 4.5\beta\big)\big] \end{aligned} \quad . \tag{08}$$

Since β is already a small quantity, we can drop the 4.5β term from the above RHS.

Now to match the boundary conditions. By definition of steady state, the current at time T_2 must look identical to what it was at $t = 0$ but relative to the new position of the flux vector. This has rotated CCW from its original position by an angle $\arctan(\beta/1)$ which for small β equals β. Now treating the phasors fully like vectors, we define a new basis d',q' where the d'-axis is aligned with the new flux vector and the q'-axis is orthogonal. Evidently, the new basis is obtained from the old one by a rotation through β about "z-axis" (out of the plane). The boundary matching condition states that the coordinates of $\mathbf{i}_1$ in the new basis must equal A and -0.24. The change of basis formula is

$$\begin{bmatrix} d \\ q \end{bmatrix} = \begin{bmatrix} 1 & -\beta \\ \beta & 1 \end{bmatrix} \begin{bmatrix} d' \\ q' \end{bmatrix} \quad ,$$

and it gives the current components in the old basis as $A+0.24\beta$ direct and $-0.24+A\beta$ quadrature.

Finally, we equate these to the direct and quadrature components of (08) to get two more equations for A, β and T_2. Plugging in the numbers, these equations are

$$-5.26AT_2 + 12T_2 = 0.24\beta \quad , \qquad \text{(09a)}$$

$$-0.9\beta + 50AT_2 + 46.26T_2 = A\beta \quad . \qquad \text{(09b)}$$

These two equations together with (07) can be solved for A, β and T_2.

This is the system which we need to solve on a computer. We can use various methods for this, like Newton-Rhaphson method, PICARD's iteration etc; here I have just used a hit and trial method, varying A, finding the corresponding β using (07) and then solving for T_2 using each of (09a,b); the desired solution is where the two values of T_2 are the same. That point is $A = -0.0692$, $\beta = 0.0241$ and $T_2 = 0.0004678$.

If we've managed to get to this point, full credit is assured. But nevertheless, like good students, let's make the examiner happy by supplying the answer s/he wants. The desired quantity is the average angular velocity of $\boldsymbol{\psi}_+$ and $\mathbf{i}_r$ in the steady state. Let's first focus on $\boldsymbol{\psi}_+$. We saw that it rotates through angle β instantaneously and then holds stationary for time T_2. This hop-step continues ad infinitum so the average angular velocity is β / T_2 which works out to 51.52 units. By definition of the steady state, the flux and current maintain constant angle relative to each other on the long term so their average angular velocities are the same. Hence the average angular velocity of both $\boldsymbol{\psi}_+$ and $\mathbf{i}_r$ is 51.52 units in the steady state.

*

COMMENTS : A barrage of them. First, the motor model here is considerably different from that of example 3.3.1. What accounts for the difference ? The difference is that the motor of 3.3.1 was controlled by a current source and this by a voltage source. The current source equation is

$$L_r \mathbf{i}_r' + \left(R_r - \mathrm{j}\omega L_r\right)\mathbf{i}_r = M\mathbf{i}_s' - \mathrm{j}\omega M\mathbf{i}_s \quad , \tag{10}$$

where L_r is rotor inductance, R_r is rotor resistance, M is mutual inductance and the stator current $\mathbf{i}_s$ is a known function of time because it is controlled externally. Expressing this in d,q form, non-dimensionalizing, and using the fact that the stator currents in that problem are constant leads to the equation in 3.3.1.

The voltage source equation is

$$L_r \mathbf{i}_r' + \left(R_r - \mathrm{j}\omega L_r\right)\mathbf{i}_r + M\left(\mathbf{i}_s' - \mathrm{j}\omega \mathbf{i}_s\right) = 0 \quad , \tag{11a}$$

$$L_s \mathbf{i}_s' + R_s \mathbf{i}_s + M\mathbf{i}_r' = \mathbf{V} \quad , \tag{11b}$$

where L_s and R_s are stator inductance and resistance. Dropping R_s and assuming $L_r = L_s$ (both harmless approximations) followed by non-dimensionalizing gave the equations which acted as the starting point of this problem. Both (10) and (11) are burdened by the additional equation

$$I\dot{\omega} = \Gamma \quad ,$$

which makes the systems nonlinear; considerations similar to the one for the dc motor in the previous Tute enable us to sweep this nuisance under the carpet in many situations. The vector I've called $\boldsymbol{\psi}_+$ is the magnetic field (or flux – the engineering jargon) at the surface of the stator, arising from both stator and rotor currents. In the Literature, "stator flux" and $\boldsymbol{\psi}_s$ are more common than "combined flux" and $\boldsymbol{\psi}_+$, but I have avoided that terminology as it creates the mistaken impression that the field is created by the stator alone.

The results for line-supplied motor in Parts (b) to (c) are very standard. You will likely have noticed that the homogeneous system does not have all decaying solutions : $\mathrm{d}\boldsymbol{\psi}_+/\mathrm{d}t = 0$ means a zero eigenvalue. So are the initial conditions relevant for line supplied motors ? No because the second equation uses the resistanceless stator approximation ($R_s = 0$). In the full model (11) you will see that the second equation (11b) also has a damping term, which we ignored in the problem. The small but finite stator resistance causes negligible change to the particular solution (even more so in DTC where the times involved are small) but it does kill off the homogeneous solutions and cause the motor to settle into the steady state.

Direct torque control was invented by ISAO TAKAHASHI and TOSHIHIKO NOGUCHI. Their first presentation was in an obscure 1984 conference in Japan;

the standard paper dates from 1986. By some corporate actions, the patent went instead to ABB of Switzerland rather than to the original inventors. This was one of the single most significant inventions in the field of power electronics; in one fell swoop it enabled induction motors to occupy the central place in industry and technology which they were previously denied on account of the difficulties associated with the existing field oriented control (FELIX BLASCHKE, 1972). DTC also made possible the use of induction motors in railway locomotives; one such locomotive is WAP7 with whom we are already very familiar. (In the first Tute you took the loco characteristics as a given and calculated the train acceleration, now you saw what makes the loco run; this is surely progress.) In Parts (d) and (e) we saw two modes of operation in DTC – creating flux at zero torque and creating torque at constant flux. Using these two sequences (and their negatives) it is possible to take the motor from any operating point (given flux and torque) to any other operating point in a very short time along a controlled path.

There are a couple of minor simplifications in the presentation of DTC I gave above. Here, the applied voltage vector $\mathbf{V}_0$ could be anything; in the actual version it can only be one of the six vectors $U\exp \mathrm{j}\,(2k\pi/6)$ for U real constant and $k=0,1,2,3,4,5$. To create the effect of an arbitrary $\mathbf{V}_0$, the switching interval is further partioned into a number of sub-intervals. Since the change in $\boldsymbol{\psi}_+$ and $\mathbf{i}_r$ over the interval is linear in both t and $\mathbf{V}$, the effect of applying $\mathbf{V}_1$ for time T_1 and $\mathbf{V}_2$ for time T_2 is equivalent to applying $T_1\mathbf{V}_1+T_2\mathbf{V}_2$ for time T_1+T_2. Using this averaging principle, the arbitrary vector $\mathbf{V}_0$ of our model is further broken into a linear combination of the six available voltage vectors applied over appropriate sub-intervals of T. Because of the discrete nature of the voltage vectors, the control of the flux is also hysteretic (here we had it held to a set magnitude ψ_0). Both these points are minor and do not change the core physics and maths behind DTC, which is all covered here.

We note however that our presentation refers to the regime of operation where the average voltage applied on the motor is less than the inverter's maximum U. In the FLUX WEAKENING regime [so called because the magnitude of the flux decreases with increase in motor speed as per (02a)], the inverter operates continuously with no zero voltage vectors – the six non-trivial inverter voltages

are applied sequentially to produce a simulation of line-supplied ac voltage and the torque of the motor acquires a significant oscillatory component.

The original paper doesn't do the derivation this way. The authors have used Laplace transform approach and then performed a hellish contour integration to invert it. Their calculation is so difficult (and the steps shown by them so scant) that it is not reproduced in any power electronics book; the results are treated as given even in graduate level texts. The simplification here arises not from any of the approximations but because the pioneering authors obtained the equivalents of (06) etc. by a different procedure. They started from continuous voltage vectors and considered the response of the flux and current to a step change in slip frequency. Using the Laplace transform, they were able to get these things as well as the torque as functions of time. Only then did they take the limit that the time of interest, T, is actually very small. Here I have imposed the limit first, and derived approximate solutions which are good in the short interval but lousy at longer times. However we don't really need the long time solutions to understand how DTC works. I hope that this new, easier derivation of TAKAHASHI and NOGUCHI's momentous result makes the work more accessible, both within the electrical engineering community and outside.

Our approach has also cast some light on the asymptotic behaviour of a DTC controlled motor, in the calculation of Part (f). With a set point torque of 0.25 (and tolerance 0.01), the average angular velocity of $\boldsymbol{\psi}_+$ and $\mathbf{i}_r$ worked out to 51.52 units. On the other hand, in Part (c) we saw that with line supply, a continuous torque of 0.25 results if the slip frequency is 1.60 units i.e. the rate of rotation Ω of $\boldsymbol{\psi}_+$ and $\mathbf{i}_r$ is 51.60 units. The answer from DTC is extremely close to this. Just as close are the currents; in the continuous mode, when $\boldsymbol{\psi}_+ = 1+\mathrm{j}0$, the rotor current is $-0.0759-\mathrm{j}0.25$ while in DTC, the rotor current is $-0.0692-\mathrm{j}0.24$ at the lower cutoff point. The finer we make the tolerance, the closer do the values of the steady state DTC currents approach to their continuous counterparts, until they become equal in the limit.

* * * * *

This page intentionally left blank

This page intentionally left blank

PART TWO

****

Partial

Differential

Equations

5

THE 1+1 HEAT EQUATION

This Chapter introduces partial differential equations through the example of the heat equation in one space and one time dimension. A theoretical discussion of the solution of homogeneous and inhomogeneous problems is followed by applications of the equation.

5.1

INTRODUCTION; SEPARATION OF VARIABLES

In this Unit (and in fact the entire Chapter) we concentrate on the simple partial differential equation (PDE)

$$\alpha^2 \frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t} \quad , \tag{01}$$

where $u(x,t)$ is the dependent variable (u is a popular choice of dependent variable when discussing a PDE in an abstract context, while α^2 is just the notational convention used in the version of the course which I teach). Using dot for time and prime for space, we can write the equation as $\alpha^2 u'' = \dot{u}$. It is also quite common, especially in mathematical literature, to use subscripts to denote partial derivatives, so a third way of writing (01) is $\alpha^2 u_{xx} = u_t$.

Equation (01) is called the 1+1 heat equation, since it features one space and one time dimension (sometimes it is also called "1-dimensional" heat equation, with the time dimension being implicit), and since it describes the evolution of temperature inside a body. It can also describe the evolution of chemical concentrations inside a volume filled with a suitable solvent, in which context it is called diffusion equation. However, we will leave the physical interpretation of the equation until later. For now we will simply treat (01) as a mathematical equation and acquire the technical skills necessary to crack a PDE problem. As you will see, PDEs require a special machinery which is quite different from that of ODEs, although there are of course conceptual overlaps.

A whole Chapter dedicated to 1+1 heat equation might seem excessive. But the thing is, with two independent variables and a first derivative on one of them, this is the simplest PDE in the curriculum. Thus, over the course of four sections with 18 examples you will see almost all the typical PDE manipulations and calculations in the easiest possible setting. Then, having learnt the basics, you'll have no difficulty in applying the concepts to other, more complicated equations.

SUFFICIENCY OF CONDITIONS

A standalone PDE equation is meaningless. Just as we need ICs or BCs to completely define an ODE problem, similarly we need additional conditions over and above the equation. The nature of these conditions is quite elaborate, and is the same for linear, nonlinear, homogeneous and inhomogeneous PDEs. As an approximate but useful starting guideline, **for EVERY ORDER of derivative with respect to any particular independent variable, we need ONE CONDITION specified at ONE value of the concerned variable and ALL possible values of ALL other independent variables.** Here, condition can mean the value of either the function or of any partial derivative of the function w.r.t the concerned variable and of an order lower than the highest present in the equation. If a PDE problem is well defined, its solution is **unique**. Conversely, ill-defined PDE problems cannot be solved at all.

Like many technical definitions, the last one is quite dense so let me elaborate a bit. Considering 1+1 heat, there is a first partial derivative w.r.t time, so we need one condition pertaining to t. This must take the form of u being specified at one value of t and all values of x. Typically, we take the time to be zero so $u(x,0)$ becomes one additional piece of info which comes with (01). For obvious reasons, we call it the **initial condition**. The heat equation also has a second partial derivative w.r.t space so we need two conditions pertaining to x. Since the partial on x is the second, these conditions might feature u as well as u'. They can take the form of u or u' or some combination like $u'+\gamma u$ being specified at two values of x and all values of t. Specifying u and u' at the same spatial point might also appear like an OK prescription, but it turns out, that is actually a malafide condition – see section 5.4.

So if the problem (01) is defined in x belongs to [0, 1] we can have something like $u(0,t) = 0$ and $u(1,t) = 0$. The spatial conditions are typically called the **boundary conditions**. Thus, 1+1 heat equation requires one IC and two BCs for the problem to be defined completely – any lower amount of information and you are never going to be able to find a solution. Exam PDE problems typically have all required ICs and BCs handed to you on a platter, and, unless the question is a trick, you don't need to see whether they are proper. But when you are modeling phenomena using PDEs and have to write the equation and conditions yourself, then it is vital to check for a match between

the two – if there's a mismatch then any effort to solve the problem will be doomed from the start.

Why did I call the BC/IC prescription an approximate guideline ? Because reality is more complicated. For example, if a PDE in Cartesian coordinates has BCs specified on the surface of a cylinder, then you cannot write the BC as a prescription at one value of x and all possible y etc. but the problem may still be complete. Sometimes, a transformation of coordinates (like say polar coordinates for this example) makes things transparent. In some other cases, the problem is complete even if it is not obvious at first sight. A formally rigorous definition of completeness of conditions will become extremely pedantic and will exceed the limits of my knowledge, so I will not go into the issue here. But, the above statement is definitely a good starting point and as we look at PDE problems throughout this and the next two Chapters, you will automatically get a feel for what constitute doable and undoable BCs.

SEPARATION OF VARIABLES

Suppose that the PDE problem which you've been given is well-defined, and you want to solve it. First thing to see is whether the PDE is linear. The definition is the same as for linear ODEs (recall 1.2). **In a linear PDE, the DEPENDENT VARIABLE and ALL its partial derivatives appear to FIRST OR ZERO POWER alone , with the coefficients being ARBITRARY functions of the independent variables.** We can also define linearity in terms of differential operator – for that, we view the combination of partial derivatives etc. as an operator which acts on the function u to produce some function f of the independent variables. The equation is linear if this operator satisfies the three properties $\mathbf{L}0=0$, $\mathbf{L}(cu)=c\mathbf{L}u$ and $\mathbf{L}(u_1+u_2)=\mathbf{L}u_1+\mathbf{L}u_2$. Linearity is vital : linear PDEs are difficult, nonlinear ones are impossible. Almost all PDEs we work with in this book will be linear.

After linear comes **homogeneous** – a linear PDE problem is homogeneous if and only if (a) the equation contains no functions of independent variables alone and (b) all **boundary conditions** have zero on the RHS. Thus, (01) together with $u(0,t)=0$, $u(1,t)=0$ and $u(0,x)=x-x^2$ is homogeneous (note that the IC need not be $u(x,0)=0$; that would produce the trivial solution!) while (01) with $u(0,t)=\cos 2t$ and whatever else is not.

If a PDE problem is well-defined, linear and homogeneous, then a method which sometimes works is called **separation of variables**. From a statistical viewpoint, if we take at random a linear homogeneous PDE and ask whether sep var will work, the answer will probably be 'no' rather than 'yes'. But it so happens that almost all of the PDEs which actually occur in the laws of Nature are amenable to this method (a thank you to Uparwala for that). So we consider that method in full detail. Before embarking on the exercise, note that this has no relation to separable first order ODEs (1.1) – those had to be first order but could be nonlinear, while this has to be linear and homogeneous but there's no restriction on the order. There's also very little similarity between the two methods. Although sep var is very general, a presentation of the method will become too abstract unless we work in terms of a specific example. So for now I will base my discussion on (01) and we'll generalize it later, as and when the situation requires.

We start off the method by claiming a trial solution u which is a product of a function of x and a function of t. Following standard notation, we denote our trial by $u(x,t) = X(x)T(t)$. This very line might make you want to stop reading on : surely we can recover only a handful of the solutions to (01) by starting from this special form ! But the method redeems itself at a later stage – just for your information, we can use it to find a solution of (01) which is

$$u(x,t) = \frac{1}{2\alpha\sqrt{\pi}\sqrt{t}} \exp\left(-\frac{x^2}{4\alpha^2 t}\right) \quad .$$

This is as far from a product solution as New Delhi is from New York, and it should convince you completely that the method is not restrictive and induce you to read on.

Next, we plug the trial into the PDE. We have

$$\alpha^2 X''T = X\dot{T} \quad .$$

Since X and T are functions of x and t alone, the primes and dots in the above can now be replaced by **ordinary** derivatives rather than partials. (A popular notation, quite common in literature, uses primes to indicate the derivatives w.r.t both x and t, with the independent variable being understood implicitly. This notation I find rather confusing, and I have seen it as a source of error, especially among beginning students.) Then, divide both LHS and RHS by $\alpha^2 XT$ to get

$$\frac{1}{X}\frac{\mathrm{d}^2X}{\mathrm{d}x^2} = \frac{1}{\alpha^2}\frac{1}{T}\frac{\mathrm{d}T}{\mathrm{d}t} \quad . \tag{02}$$

If you are fastidious then you might object that what happens if X or T is zero at any point. The most appropriate response to that is "we are engineers" – mathematicians can worry about the case where X or T is zero at a point but so long as we get results, we won't bother. (For the record, the method does generate solutions where X and T have one or more zeroes.)

Now is the important step. Note that the LHS of (02) is a function of x alone while the RHS is a function of t alone. Let's call them $f(x)$ and $g(t)$. Now, for two functions to be equal, they must be equal to each other at every possible value of the argument of either function. Considering one point (x_0,t_0) we have $f(x_0) = g(t_0)$. But considering another point (x_0,t_1), we get $f(x_0) = g(t_1)$, and combining these two yields $g(t_0) = g(t_1)$. This equality must hold true for any pairs of values t_0, t_1, hence $g(t)$ is in fact a constant function of time. Evidently, $f(x)$ is also a constant function of space. **Hence, in (02), not only are the both sides equal but they are equal to a constant.** This is the crux of the method of separation of variables. The argument is a little subtle, so read it and ponder over it until you are completely sure of it.

Letting the constant be λ, (02) now gives

$$\frac{1}{X}\frac{\mathrm{d}^2X}{\mathrm{d}x^2} = \lambda \quad , \tag{03a}$$

$$\frac{1}{T}\frac{\mathrm{d}T}{\mathrm{d}t} = \alpha^2\lambda \quad . \tag{03b}$$

These are just ODEs for X and T; in fact, clearing the fraction from the denominator we see that they are our old friends from 1.2.1 and 2.1.1 !

So is the PDE just two coupled ODEs in fancy dress ? NO. Because we don't know what λ is. For **every** value of λ, (03a) will have a solution $X(x)$ and (03b) will have a solution $T(t)$. Any and all of the products XT are possible solutions of (01). Now invoke linear and homogeneous. If the PDE problem satisfies these two properties, then any linear combination of solutions is itself a solution. So the most general solution of the problem is a linear combination of product solutions $X(x)T(t)$ corresponding to every value of λ from minus to plus infinity ! This is a complicated situation.

The way we narrow down the choices of λ is by invoking plausibility and boundary conditions. Suppose λ is positive i.e. $\lambda=k^2$. Then, (03b) has the solution

$$T(t)=\exp\left(\alpha^2k^2t\right) \quad ,$$

which is the exponentially growing solution we have been rejecting repeatedly in the first Part of this book. There's no need to act differently now – a physically motivated equation can never have an exponentially growing solution. And the argument that the $X(x)$ part will somehow counteract the growth is fallacious – X is a function of x alone and can't do anything about a divergent time-dependence. Hence $\lambda>0$ corresponds to an implausible case. The other cases $\lambda<0$ and $\lambda=0$ clearly admit decaying and constant in time solutions so they are plausible. In the first case, the solutions of (03a) for X are $\cos kx$ and $\sin kx$ where $k^2=-\lambda$; in the second case, the solutions are 1 and x. Thus the separable solutions XT look like

$$\{\cos kx,\sin kx\}\exp\left(-\alpha^2k^2t\right) \quad \text{, and}$$

$$\{1,x\} \quad ,$$

corresponding to the two cases. To form an arbitrary linear combination of these, we multiply each possible solution by an arbitrary constant and add them all; this combination looks like

$$u_G(x,t)=A+Bx+\sum_k\left(C_k\cos kx+D_k\sin kx\right)\exp\left(-\alpha^2k^2t\right) \qquad \textbf{(04)*}$$

*

.

The summation runs over all k greater than zero. The case $k=0$ is of course covered explicitly in the $A+Bx$ outside; for non-zero k, the T's are the same for positive and negative k and the X's for negative k are linear combinations of those for positive k. Equation (04) is thus the **general solution** of the 1+1 heat equation. If k is a discrete variable, then the summation of course makes perfect sense. If k is continuous, then the sum is actually an integral – C and D become functions of k and the general solution looks like

$$u_G(x,t)=A+Bx+\int_0^\infty \mathrm{d}k\left(C(k)\cos kx+D(k)\sin kx\right)\exp\left(-\alpha^2k^2t\right) \quad .$$

This is of course the natural extension of (04) to the case of continuous k. Whether the sum or the integral is the more accurate representation depends

on the BCs; nevertheless you can always write (04) as the representation of the general linear combination of physically plausible separable solutions of the 1+1 heat equation. It is in fact this summation stage where separation of variables is capable of generating more than simple product solutions – the $\left(1/\sqrt{t}\right)\exp\left(-x^2/t\right)$ solution I quoted above comes about as an integral of separable solutions over k (see 5.3).

BC FIT

Now come to the BCs. Here the key step lies the fact that they are homogeneous. Suppose that the problem is defined in x belongs to [0, L] with u given zero at both ends for all time t. Then, $X(0)T(t)$ and $X(L)T(t)$ both equal zero. Since T is not the zero function, this implies that $X(0)$ and $X(L)$ must both be zero. Combining this info with (03a) gives us an ODE BVP for $X(x)$, just like the ones in the last Chapter. For most values of λ this BVP will have the trivial solution $X = 0$; these λ's are useless. The useful λ's are the eigenvalues of this BVP and the corresponding X's are the eigenfunctions. If the BCs are different, then the eigenvalues and functions will be different. But either way, (04) will reduce to a sum over eigenfunctions corresponding to some special values of k. This sum is called the **eigenfunction expansion**. For problems defined on finite domains, the eigenvalues usually form a discrete set, called the **spectrum**. For problems defined on infinite domains, the eigenvalue spectrum may be continuous, a case which entails greater technical complexity but no additional conceptual difficulty.

Does the method break down completely if the BCs aren't homogeneous ? Fortunately, no. Like in ODEs, we can again break up an inhomogeneous problem into a homogeneous and a particular part and then work on the two parts separately. Inhomogeneous PDEs are complicated however so we defer consideration of the issue to a later section (5.3 again).

IC FIT

After obtaining the eigenfunction expansion, how do we evaluate the unknown coefficients in it ? For example, suppose our eig exp looks like

$$u(x,t)=\sum_n C_n X_n(x)T_n(t) \quad , \tag{05}$$

then how do we find the coefficients C_n ?

We do it by using the IC. If we are given $u(x,0)=f(x)$, we set $t=0$ in the expansion (05). This gives $f(x)$ as a sum over the spatial eigenfunctions. Immediate question : can you at all write $f(x)$ as a sum over eigenfunctions for arbitrary f? The answer is yes, in many cases (details later – for now we just want to get the basics running). In such case, eigenfunctions are said to be **complete**. In most cases of completeness, what also happens is that the eigenfunctions are **orthogonal** over the domain. Recall from 4.1 that the dot product "$f(x)\cdot g(x)$" means $\int_0^L fg\ \mathrm{d}x$ and two functions are orthogonal if this dot product is zero. Thus orthogonality of eigenfunctions means that, if $X_m(x)$ and $X_n(x)$ be two eigenfunctions in the expansion (05), then

$$\int_0^L X_m(x)X_m(x)\mathrm{d}x=\begin{cases}\text{finite if } m=n\\ 0 \text{ otherwise}\end{cases} \quad .$$

If the spectrum is continuous, then the “finite” actually becomes “infinite” but the dot product relation for the coefficients still holds.

Given orthogonality of the eigenfunctions, finding the coefficients in the expansion is an easy task. Consider a general relation

$$f(x)=\sum_k C_k X_k(x) \quad ,$$

where f is an arbitrary function and the X_k's are orthogonal eigenfunctions. Then, to find the C's we use the $c_k=\mathbf{x}_0\cdot\mathbf{v}_k$ trick we saw back in 2.3. We first normalize the eigenfunctions to unit vectors by dividing by their lengths. Thus, a unit eigenfunction is $v_k X_k$ where

$$v_k=\frac{1}{\left(\int_0^L \left[X_k(x)\right]^2\mathrm{d}x\right)^{1/2}} \quad .$$

Then, taking the dot product shows that

$$C_k=v_k^2\int_0^L f(x)X_k(x)\ \mathrm{d}x \quad .$$

This completes the determination of the unknown coefficients in the eigenfunction expansion, and hence the solution of the PDE problem.

To recap, there are three basic steps in the solution of the 1+1 heat equation problem – finding the **general solution**, fitting the BCs and writing the **eigenfunction expansion**, and fitting the IC and finding the **coefficients** in this expansion.

This is the general algorithm for solving 1+1 heat equation problems. You will have noticed that the treatment was very detailed upto (04) but became more brief after that. This happened because (04) is completely general while the subsequent steps are problem-specific. The moment you have 1+1 heat equation, you can start your solution process from (04). But then the eigenvalues and eigenfunctions will depend on the exact BCs given in the problem at hand. We'll elaborate this part in the Tutes.

A couple more things before going on to those. One is that $u = 0$ does not necessarily refer to zero degrees on the Centigrade, Fahrenheit, Kelvin or any other fixed temperature scale. You can always **define your own temperature scale** so that $u = 0$ becomes the most convenient value, relative to the problem at hand. For example, suppose the ends of a rod are maintained at 20 °C while the initial internal temperature is $50 \sin 2x$ °C. Then you can define $u = 0$ to be 20 °C and let the IC internal temperature be $30 \sin 2x$ on your new scale. The second thing is that if an end of a rod is **insulated**, then that means $\partial u / \partial x = 0$ at that end. In the same way, if the rod represents a chemical tube, the $u_x = 0$ BC is referred to as a **sealed** end. Why these hold true, we'll see later, but exam problems unrelated to modeling sometimes use this jargon just to test whether you know it or not. Finally, although we aren't going into modeling yet, it is often helpful if we are aware of the basic physical picture of heat conduction in a rod, which the 1+1 heat equation aims to describe. Elementary physical considerations often act as a test of the correctness of a mathematical solution.

Example 5.1.1

Good and bad PDE problems

Each of the five questions below gives you a PDE equation and a set of possible initial and boundary conditions. The dependent variable is u and the independent variables are listed explicitly. Assume that an independent variable x or y belongs to the interval [0, 1] while an independent variable t can take all positive values. Determine whether the BCs and/or ICs together with the equation completely define the PDE problem. Independent of the outcome, also write whether the equation can be solved by separation of variables. Do NOT attempt to solve the problems even in the most favourable case.

(a) $u(x,t)$; $\dfrac{\partial^2 u}{\partial x^2} - \dfrac{\partial u}{\partial t} - u = 0$; $u(0,0) = u(1,0) = 0$, $u(x,0) = x - x^3$

(b) $u(x,y,t)$; $\dfrac{\partial^2 u}{\partial x^2} + \dfrac{\partial^2 u}{\partial y^2} - \dfrac{\partial u}{\partial t} = 0$; $u(0,y,t) = u(1,y,t) = 0$, $u(x,y,0) = \sin 2\pi x \cos 3\pi y$

(c) $u(x,y)$; $y^2 \dfrac{\partial^2 u}{\partial x^2} + x^2 \dfrac{\partial^2 u}{\partial y^2} = 0$; $u(0,y) = u(1,y) = 0$, $u_y(x,0) = u_y(x,1) = 0$

(d) $u(x,y)$; $\dfrac{\partial^4 u}{\partial x^4} + 2\dfrac{\partial^4 u}{\partial x^2 \partial y^2} + \dfrac{\partial^4 u}{\partial y^4} = 0$; $u_x(0,y) = u_x(1,y) = 0$, $u_y(x,0) = u_y(x,1) = 0$

(e) $u(x,t)$; $\dfrac{\partial u}{\partial t} + u\dfrac{\partial u}{\partial x} - \dfrac{\partial^2 u}{\partial x^2} = 0$; $u(0,t) = 0$, $u'(1,t) = 0$, $u(x,0) = \sin 4x$

[Level : 2]

(a) The equation looks like a variation on the 1+1 heat equation. Correct BCs should provide the value of u at $x = 0$ and $x = 1$ at ALL possible values of t. However the ones given in the problem only specify u at the two ends at $t = 0$. Hence the BCs are insufficient to generate a solution. The IC is OK. To check separability, we first see whether the equation is linear and the BCs are homogeneous. They are, so we now try $u=X(x)T(t)$. We get $X"T - \dot{T}X - XT = 0$, and dividing through by XT,

$$\frac{X"}{X} - \frac{\dot{T}}{T} - 1 = 0 \quad .$$

Taking $\dot{T} / T$ over to the RHS, LHS becomes a function of x alone (since the constant "1" is as legit a function of x as it is of t), RHS a function of t alone so both sides equal a constant and hence the equation has separated. Bottom line, well-defined : NO, separable : YES.

(b) Corresponding to the second partial derivative on x, there are two BCs specifying u at two values of x and every possible y and t. Corresponding to the second partial on y however there is nothing. Corresponding to the first partial on t, we have a proper IC but the absence of the y BC has killed it already. To check separability, we first see whether the equation is linear and the BCs are homogeneous. They are, so we now try $u=X(x)Y(y)T(t)$. We get $X_{xx}YT + XY_{yy}T - XY\dot{T} = 0$, and dividing through by XYT,

$$\frac{1}{X}\frac{\mathrm{d}^2X}{\mathrm{d}x^2} + \frac{1}{Y}\frac{\mathrm{d}^2Y}{\mathrm{d}y^2} - \frac{1}{T}\frac{\mathrm{d}T}{\mathrm{d}t} = 0 \quad .$$

Transferring the T term to the RHS, a function of t alone must equal a function of x and y alone, so the T thing is a constant function. Letting the constant be λ, a function of x alone plus a function of y alone equals a constant. Same situation as in Part (a) above, so each of them must again equal a constant, and the equation has separated. Bottom line, well-defined : NO, separable : YES.

(c) Corresponding to the second partial derivative on x, there are two BCs specifying u at two values of x and every possible y. Corresponding to the second partial derivative on y, there are two BCs specifying $\partial u/\partial y$ at two values of y and every possible x. Since the equation is second order in y, a first partial is an acceptable condition and the BCs are complete. To check separability, we first

see whether the equation is linear and the BCs are homogeneous. They are, so we now try $u = X(x)Y(y)$. We get $y^2X_{xx}Y + x^2XY_{yy} = 0$. Although dividing by XY doesn't do the job this time, dividing by x^2y^2XY does. Doing that,

$$\frac{1}{x^2X}\frac{\mathrm{d}^2X}{\mathrm{d}x^2} + \frac{1}{y^2Y}\frac{\mathrm{d}^2Y}{\mathrm{d}y^2} = 0 \quad .$$

Once again, function of x alone equals function of y alone so the equation has separated. Bottom line, well-defined : YES, separable : YES.

(d) Corresponding to the fourth partial derivative on x, there should have been four conditions specified w.r.t x. However, only two are given. Likewise, corresponding to the fourth partial on y, there should have been four more conditions specified w.r.t y, but again there are only two. Hence the BCs are hopelessly inadequate. To check separability, we first see whether the equation is linear and the BCs are homogeneous. They are, so we now try $u = X(x)Y(y)$. We get

$$Y\frac{\mathrm{d}^4X}{\mathrm{d}x^4} + 2\frac{\mathrm{d}^2X}{\mathrm{d}x^2}\frac{\mathrm{d}^2Y}{\mathrm{d}y^2} + X\frac{\mathrm{d}^4Y}{\mathrm{d}y^4} = 0 \quad .$$

To attempt separation, we must divide by something or other to generate functions of x and y alone in each term. The first and third terms become tractable if we divide by XY. However, this does nothing to the second term which remains a mixture of x and y. Conversely, if we divide by X_{xx} or Y_{yy}, we can clear the second term but can't make progress with the first and third terms. So this equation is not separable. Bottom line, well-defined : NO, separable : NO.

(e) Corresponding to the second partial derivative on x, there are two BCs specifying u or u' at two values of x and every possible t. Hence these BCs are complete. Corresponding to the first partial on t, there is an IC specifying u at $t = 0$ and every possible x. Hence the conditions are complete. To check separability, we first see whether the equation is linear and the BCs are homogeneous. Red flag is raised immediately – the equation here is nonlinear on account of the $u(\partial u/\partial x)$ term. Hence separation of variables is not going to work in this case. Bottom line, well-defined : YES, separable : NO.

*

COMMENTS : Practically, it is meaningless to check the equation for separability if the associated conditions are incomplete. Here, I made you do it only so that I could get this preliminary stuff over with in a single Tute, and start cracking PDEs from the next one. But in reality, until you have a full set of conditions, you don't get off the ground.

Some of the PDEs in this problem are standard equations, and are going to get elaborated later in the book. Part (a) is a 1+1 heat equation with a dissipation term. Part (b) is the 2+1 heat equation, which describes heat conduction in two space and one time dimension. Part (d) is the biharmonic equation, which arises in elasticity theory and in linearized fluid flow theory. Although it couldn't be separated, the situation is not altogether hopeless because we can write it as a Laplace equation on top of a Laplace equation, and that is tractable. Part (e) is a pot-pourri of terms taken from the Navier-Stokes equation of fluid mechanics. The variable u in this case is the flow velocity. The nonlinearity here makes the equation impossible to solve, and contributes to the extreme technical difficulty of fluid mechanics as a subject.

Example 5.1.2

Simple sep var

Consider the PDE problem for $u(x,t)$

$$4\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t} \quad ,$$

with x belongs to [0, 1], $t > 0$, $u(0,t) = u(1,t) = 0$ and $u(x,0) = 2\sin 3\pi x$.

Determine $u(x,t)$.

[Read Only]

Because this is our first time, I will quickly run through the sep var process again. We seek solution of the form $u(x,t)=X(x)T(t)$. Plugging into the PDE gives

$$4X''T = X\dot{T} \quad ,$$

where prime and dot now denote ordinary derivatives because X and T are functions of x and t alone. Then dividing through by $4XT$,

$$\frac{1}{X}\frac{\mathrm{d}^2X}{\mathrm{d}x^2} = \frac{1}{4T}\frac{\mathrm{d}T}{\mathrm{d}t} \quad .$$

Since the LHS is a function of x alone and the RHS a function of t alone, they must both be equal to a constant, which we call λ. Imposing this condition leads to ODEs for X and T

$$\frac{\mathrm{d}^2X}{\mathrm{d}x^2} = \lambda X \quad ,$$

$$\frac{\mathrm{d}T}{\mathrm{d}t} = 4\lambda T \quad .$$

We now consider the cases $\lambda > 0$, $\lambda < 0$ and $\lambda = 0$.

In the first case, the T equation has the solution $T = \mathrm{e}^{4\lambda t}$, with positive λ. This represents exponential growth and a physically meaningless solution, so this case is rejected from consideration. In the second case, $\lambda = -k^2$. The X equation has the solutions $X = \cos kx$ and $X = \sin kx$ and the T equation has the corresponding solution $T = \exp(-4k^2t)$. In the third case, the X equation has the solutions $X = 1$ and $X = x$ and the T equation has the corresponding solution $T = 1$. The general solution is an arbitrary linear combination of all possible products $X(x)T(t)$ which is

$$u_G(x,t) = A + Bx + \sum_k \left(C_k \cos kx + D_k \sin kx\right)\exp\left(-4k^2t\right) \quad , \qquad (01)$$

where k may be discrete or continuous.

Whenever you start from 1+1 heat equation, you will always end up at this step. The paths to solution of different problems (corresponding to different BCs and ICs) start diverging from here. So it is completely acceptable if you use (01) as a formula and start off your solution with something along the lines of "Separation of variables on the 1+1 heat equation followed by rejection of unphysical solutions leads to the general solution

$$u_G(x,t) = A + Bx + \sum_k \left(C_k \cos kx + D_k \sin kx\right) \exp\left(-4k^2 t\right) \quad ,$$

where k may be discrete or continuous." Two reasons why I showed the intermediate steps : firstly, it gives you a chance to see the separation process again, and secondly, if you do not remember the form (01) then you should be able to carry out these steps quickly and arrive here. Corollary : you may think that if you learn the general solution of heat equation (and a few other equations like 2D Laplace and 1+1 wave) then you can get by without having to understand how sep var works. That notion is fallacious. You can always encounter a PDE, either in an exam or in your research, which you've never seen before, and then if you don't understand the logic behind sep var, you won't be able to get off the ground.

There are two ways we can argue the BC fit. The first is directly derived from the general solution and, given that we already have the general solution, is logically pleasing although it is more conceptual. The second goes back to the ODEs which came out of separating the PDE and is easier to implement. Logically, it might appear like a step backward from the general solution but it enables us to bypass the general solution altogether.

For the first approach, we argue that the general solution has the form of a sum $u_G(x,t) = \sum_k X_k(x) T_k(t)$ over infinitely many k's (we can bring the $A+Bx$ terms into the case $k = 0$). This has to satisfy the BCs $u(0,t) = 0$ and $u(1,t) = 0$ for all time. Now as time evolves, the T_k's evolve in their own sweet way, each k independent of the other, and we get infinitely many different linear combinations of the functions $X_k(x)$. The only way all these combinations can satisfy the BCs is if every X_k satisfies the BC on its own. This takes us to 4.1.1 – there we saw that $A+Bx$ can't satisfy the double zero BCs (non-trivially) while $C_k \cos kx + D_k \sin kx$ can satisfy them if and only if $k = n\pi$ in which case C_k must be zero but D_k can be anything. The sum over all k now becomes a sum over n.

For the second approach to BC fit, we say that $u(0,t) = u(1,t) = 0$ implies $X(0)T(t) = X(1)T(t) = 0$. Since $T(t)$ is not the zero function, $X(0)$ and $X(1)$ must both be equal to zero. X then satisfies the ODE BVP $X'' - \lambda X = 0$ with these BCs. From T057 we already know what happens now. The case $k = 0$, which produces the $A+Bx$ term in (01) yields $A = B = 0$. The case $k \neq 0$ produces the eigenvalues

$k = n\pi$ ($n = 1,2,3.....$) with the corresponding eigenfunctions $X = \sin n\pi x$. Hence we can whittle down the sum (01) over all k to a sum over the eigenvalues. The terms in the sum consist of the spatial eigenfunctions multiplied by their corresponding time evolutions.

After either of the two approaches above we have

$$u(x,t) = \sum_n D_n \sin n\pi x \exp\left(-4\pi^2 n^2 t\right) \quad , \qquad (02)$$

which is the eigenfunction expansion of $u(x,t)$. We stick with the convention, defined already in 4.3, that an unlabelled summation index n refers to the natural numbers 1, 2, 3..... – in case the sum index is different, we'll always be careful to specify the same explicitly.

To find the coefficients D_n, we now use the IC. We are given that $u(x,0) = 2\sin 3\pi x$. Setting $t = 0$ in (02) we get

$$u(x,0) = \sum_n D_n \sin n\pi x = 2\sin 3\pi x \quad .$$

Evidently, $D_3 = 2$ and $D_n = 0$ for all n different from 3. Substituting this into (02) we get the specific solution

* $$u(x,t) = 2\sin 3\pi x \exp\left(-36\pi^2 t\right) \quad ,$$ *

which is the desired answer.

*

COMMENTS : A problem like this is sometimes called an initial boundary value problem (IBVP). This terminology by itself is fine, but what do you call something like $u_{xx}+u_{yy} = 0$ with BCs ? BBVP ? Because of this, I will stay away from this jargon. I will let IVP and BVP denote ODE problems only and refer to all combinations of PDE and compatible BCs and/or ICs as "PDE problems".

There are a couple of minor differences between many conventional treatments and what I did above. The first is that I wrote the general solution explicitly – most people bypass this step altogether and go straight to the ODE eigenvalue problem. The only reason I wrote the general solution is that it highlights the similarity between the processes of solving an ODE and solving a PDE – both have a general solution which is an arbitrary linear combination of fundamental solutions and which you then narrow down by imposing the given conditions.

Moreover, on one occasion after I had shown the ODE eigenvalue approach in class, one student had asked that why must the $X(x)$ function satisfy the BCs by itself – can't it happen that every X doesn't satisfy the BC individually but their sum does ? This question can be resolved only by arguing basis the general solution as I did above. The second point of difference is that in the sep var step, I have called the separation constant λ. Many call it $-\lambda$. This choice of sign is motivated by the prior knowledge that the eigenvalues will come out positive if defined this way. At the sep var step however, there is absolutely no information about the sign or size of the constant – all we know is that it IS a constant. Because of this, I prefer to call it λ. Let the BCs tell us what its sign will be.

In the ODE-based approach to the BC fit, we saw one instance of a very general principle – a homogeneous BC carries over from the PDE to the separated ODE. So whenever you have a problem with a BC featuring zero on the RHS, that BC will come over to the ODE for the concerned variable, and you can get the eigenfunctions by solving the BVP for that variable. We'll see more of this when we do other PDEs in the next Chapter onwards.

Example 5.1.3

Complicated IC

The temperature $u(x,t)$ of the rod x belongs to $[0, \pi]$ satisfies the 1+1 heat equation

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t} \quad ,$$

with BCs $u(0,t) = u(\pi,t) = 0$ and IC $u(x,0) = f(x)$ such that $f(x)$ rises in a straight line from $(0,0)$ to $(\pi-\varepsilon,1)$ where ε is extremely small and then jumps back abruptly to $(\pi,0)$.

(a) Determine $u(x,t)$.

(b) Describe the shape of $u(x,t)$ at moderately large times.

[Read Only]

(a) We know that separation of variables on the 1+1 heat equation followed by rejection of unphysical solutions leads to the general solution

$$u_G(x,t) = A + Bx + \sum_k \left(C_k \cos kx + D_k \sin kx\right) \exp\left(-k^2 t\right) \quad ,$$

where k may be discrete or continuous (note that $\alpha = 1$ in this problem). BCs are almost identical to the previous example , and we readily have $A = B = 0$, eigenvalues are $k = n$ and corresponding eigenfunctions are $\sin nx$. Hence the eigenfunction expansion of u is

$$u(x,t) = \sum_n D_n \sin nx \exp\left(-n^2 t\right) \quad . \tag{01}$$

Setting $t = 0$ in this expression gives

$$\sum_n D_n \sin nx = f(x) \quad , \tag{02}$$

where $f(x)$ is the function given in the question. Of course, the calculation upto this point was nearly a carbon copy of the last Tute; it's now that things start to diverge (which is why this is a separate problem). There are two ways we can proceed from here.

Method 1 : Linear algebraic method

We think of the eigenfunctions $\sin x,\ \sin 2x,\ \sin 3x \ldots..$ as vectors (if you want, eigenvectors of some matrix) $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_3$..... and want to express the IC function $f(x)$, which we now think of as a vector $\mathbf{f}$, as a linear combination of these vectors. In other words, we have an equation like

$$\mathbf{f} = c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + c_3\mathbf{u}_3 + \ldots.. \quad , \tag{03}$$

and want to determine the constants c (we've called them D in the specific example above but never mind). We take for granted that these vectors are orthogonal i.e. $\mathbf{u}_i \cdot \mathbf{u}_j$ is finite if $i = j$ and zero otherwise (the conditions for orthogonality, we'll see later but the bottom line is that they'll hold for almost every problem in this book). Recall from 4.1 that the dot product of two function-vectors is the integral of their product over the interval of definition.

The vectors $\mathbf{u}_n$ are not ortho-NORMAL though, so we normalize $\mathbf{u}_n$ to a unit vector $\mathbf{v}_n$ by multiplying the normalization constant v_n, i.e. we define $\mathbf{v}_n = v_n\mathbf{u}_n$.

Here v_n is the reciprocal of the length of $\mathbf{u}_n$, i.e. $v_n = 1/\left(\mathbf{u}_n \cdot \mathbf{u}_n\right)^{1/2}$. Then, (03) becomes

$$\mathbf{f} = \frac{c_1}{v_1}\mathbf{v}_1 + \frac{c_2}{v_2}\mathbf{v}_2 + \frac{c_3}{v_3}\mathbf{v}_3 + \ldots\ldots \quad .$$

Now where have we seen this type of equation before ? In 2.3. Just as in there, we take the dot product with $\mathbf{v}_n$ on the both sides; since $\mathbf{v}_n \cdot \mathbf{v}_n = 1$ and $\mathbf{v}_n$ dotted with everything else is zero, we have $c_n / v_n = \mathbf{f} \cdot \mathbf{v}_n$. Finally, since we want things in terms of $\mathbf{u}$ rather than $\mathbf{v}$, we write $\mathbf{v}_n$ as $v_n\mathbf{u}_n$, which yields

$$c_n = v_n^2 \mathbf{f} \cdot \mathbf{u}_n \quad .$$

The complete generality of the argument leading to this result means that it is applicable not just to the $\sin nx$'s of this example but to any eigenfunction expansion in terms of orthogonal eigenfunctions $X_n(x)$.

Let's now use this to calculate the D_n's in (02). We have

$$v_n = \frac{1}{\left(\int_0^\pi \sin^2 nx \ \mathrm{d}x\right)^{1/2}} = \frac{1}{\sqrt{\pi/2}} = \sqrt{\frac{2}{\pi}} \quad .$$

Note that v_n is the same for ALL n. Then, the coefficient formula yields

$$D_n = v_n^2 \mathbf{f} \cdot \mathbf{u}_n = \frac{2}{\pi}\int_0^\pi f(x) \sin nx \ \mathrm{d}x \quad . \qquad (04)$$

Evaluating the integral for the given f, we get

$$D_n = (-1)^{n+1}\frac{2}{n\pi} \quad ,$$

which is the desired expression. Now you can see why back in 2.3 I was making such a brouhaha about using dot products as opposed to matrix inversion wherever possible. Here there's no matrix to invert but you can exploit the eigenfunction orthogonality relations to obtain the coefficients.

This is the most general way of going about the evaluation of the coefficients. It's also the most abstract since it requires us to think of the functions as vectors. I presented it first because it turns out to be the only method available when things get hard, for example if the BCs are non-standard or if the dimensionality

of the PDE happens to be more than 1+1. For most 1+1 problems however, there is a simpler method which uses Fourier series of extended functions and thus relates to the material of 4.3 rather than the more conceptual 2.3. I'll elaborate on that method now.

Method 2 : Extended function method

For this, we recognize that $\cos nx$ and $\sin nx$ are nothing but the components of a Fourier series and (02) looks like some kind of Fourier expansion. In fact we have already Fourier expanded something remarkably similar to the function $f(x)$ in 4.3.2Part (b) and we should be able to use that result with no calculation at all. The catch is that $f(x)$ is not periodic – it is defined only in the interval $[0, \pi]$ and nothing is known about it outside, while Fourier expansions can only be done on periodic functions.

So what we now do is perform an extension. We redefine the function f so that its domain becomes the entire real line and it becomes periodic. But there are several ways of doing the extension (see 4.3) – we can brute force replicate it in $[\pi, 2\pi]$, $[-\pi, 0]$ etc, we can extend antisymmetrically to $[-\pi, 0]$ and then replicate in 2π sized intervals and we can also extend symmetrically to $[-\pi, 0]$ before replicating in 2π sized intervals. Presumably there are more elaborate extension schemes also, listing which will confuse rather than clarify. But among all these possibilities, which one do we actually do ?

The criterion for choosing the extension algorithm is that the solution of the extended problem in the interval $[0, \pi]$ must be identical to the solution of the actual problem. The extension cannot change the solution ! Now, the eigenfunctions of the actual problem are $\sin nx$'s, which, if we extend them beyond the problem boundaries and to infinity, are all antisymmetric with fundamental period 2π. Any linear combination of those must also be antisymmetric with period 2π. The IC $f(x)$ is one such combination, so its extension must also be antisymmetric and periodic with period 2π.

Looking at it the other way, suppose we start from the extended problem i.e. we take the heat equation valid on the entire real line and have as our IC the antisymmetrically 2π-periodically extended $f(x)$. Because the problem has an infinite domain of definition, it makes no sense to impose BCs at the ends – the periodicity condition is effectively a substitute for end BCs. Then, in the sum over

separable solutions, the only terms which will survive are $k=n$ with the corresponding space eigenfunctions being $\sin nx$. Hence our solution will look like

$$u(x,t)=\sum_n D_n \sin nx \exp\left(-n^2 t\right) \quad ,$$

this time valid over the entire real line. Now consider the portion of the solution in the interval $[0, \pi]$. Evidently, since $\sin 0$ and $\sin n\pi$ are zero, $u(0,t)$ and $u(\pi,t)$ will both be zero for all time. Thus the extended solution satisfies both the BCs and the IC of the original problem and hence the portion of the extended solution in the interval $[0, \pi]$ is the solution to our original problem.

Given the extension, the Fourier formula is

$$D_n=\frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin nx \ \mathrm{d}x \quad ,$$

which, since f and sin are both antisymmetric, is actually twice of the integral from 0 to π and we recover the formula (04). In this case, thanks to 4.3.2 we don't need to actually evaluate the integral. The antisymmetric extension of f is exactly the sawthooth wave of 4.3.2 Part (b), for which we already know that $D_n=(-1)_{n+1}(2/n\pi)$. This brings us to the same point as the linear algebraic approach did.

Now, we go back to the eigenfunction expansion (01) and plug in the D_n's we got by either procedure. This generates the specific solution

* $$u(x,t)=\sum_n (-1)^{n+1}\frac{2}{n\pi}\sin nx \exp\left(-n^2 t\right) \quad ,$$ *

which is the desired answer. Note that it is almost never possible to sum these serieses explicitly, so the final answer stays in the form of a summation. In the exam hall, never ever waste time trying to sum a series like this, unless you are explicitly asked to do so !

(b) At very large time of course, $u(x,t)$ looks like the zero function because everything decays exponentially. However, at times smaller than that, the largest surviving component of u will be the eigenfunction which has the slowest decay rate. This is the first one – the decay rate increases quadratically in n. Although u starts off with a mixture of all harmonics, the higher the harmonic,

the faster it gets killed off. The last to disappear is the first harmonic, so at moderately large times u looks like

$$u(x,t) = \frac{4}{\pi}(\sin x)\mathrm{e}^{-t} \quad . \qquad *$$

Its shape is that of a pure sine function between $x = 0$ and $x = \pi$.

*

COMMENTS : Although we aren't really going into the physics of heat transfer yet, the result of this problem is eminently plausible if we accept that the heat equation indeed models heat flow in a body. If a rod with some complicated initial temperature is connected to a temperature bath $u = 0$ at both ends, we do expect on physical grounds that at large times, the temperature inside the rod will become zero everywhere. Note that the zero is on an arbitrary scale – it might well be 100 degrees on the Centigrade scale !

In this problem, I defined the IC in a pedantically correct manner, with the ε and all. A more lax statement is $u(x,0) = x/\pi$. If posed like this, it might appear that the IC and the right side BC are in contradiction. However, what it actually means is this – the initial temperature is x/π until almost the end of the rod, and then jumps to zero at the very end of the rod. Since we are not hardcore mathematicians, the informal statement is perfectly acceptable.

If the right end of the domain is made $x = L$ rather than π, the formula for the coefficients D_n is

$$D_n = \frac{2}{L}\int_0^L f(x)\sin\frac{n\pi x}{L}\mathrm{d}x \quad .$$

To see how this arises, perform the extension and translate the Fourier coefficient formulae (4.3.02) from time to space domain. For an arbitrary function which is periodic with period $2L$, the Fourier coefficient formulae become

$$A_0 = \frac{1}{2L}\int_{-L}^{L} f(x)\mathrm{d}x \quad , \qquad (05a)$$

$$A_n = \frac{1}{L}\int_{-L}^{L} f(x)\cos\frac{n\pi x}{L}\mathrm{d}x \quad , \qquad (05b)$$

$$B_n = \frac{1}{L}\int_{-L}^{L} f(x)\sin\frac{n\pi x}{L}\,\mathrm{d}x \quad . \tag{05c}$$

In the above, I could have used any interval of size $2L$ as the integration limits, but $-L$ to L is more or less universal choice since that enables us to exploit symmetry properties. Then, we use that the product of f and sin is symmetric so the integral from $-L$ to L is double that from 0 to L. Most students remember the D_n formula for the exams; that's fine so long as you remember it correctly. But you should be familiar with the derivation so that you don't land in trouble if your memory blanks out in the exam hall.

An important thing to note here is that although the IC was discontinuous, u at all subsequent times is nice and smooth. This is in fact a very general property of the heat equation – solutions at any finite time are always continuous, free of kinks etc. We can understand this phenomenon by looking at the eigenfunction expansions of the solutions. Recall from 4.3 that a discontinuous function has its Fourier coefficients going as $1/n$, a kinky function as $1/n^2$ etc. In heat equation, the eigenfunction expansions feature $\exp(-\alpha^2 n^2 t)$ so, whatever the Fourier coefficients of the IC, at any finite time the size of the coefficients drops precipitously to zero for large n. Such serieses cannot generate anything other than nice and smooth functions with no discontinuities, kinks or any other kind of "bad behaviour".

The limiting shape of $u(x,t)$ is a general feature of the 1+1 heat equation. The higher the value of k, the greater the decay rate. So at long times, an arbitrary IC changes shape to the lowest harmonic present in it. An easy mnemonic for this is "smallest k, slowest decay". The case $k = 0$ produces a solution which doesn't decay at all; so far we haven't seen this solution in action.

Example 5.1.4

Insulation BC

The temperature $u(x,t)$ of the rod x belongs to $[0, \pi]$ satisfies the 1+1 heat equation

$$9\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t} \quad .$$

The two ends of the rod are insulated. At $t = 0$, the system is set off with the initial temperature profile $u(x,0) = f(x)$ where $f(x) = 1$ in the region $[0, \pi/3]$ and $f(x) = 0$ in the region $[\pi/3, \pi]$.

(a) Determine $u(x,t)$, optionally using the result of 4.3.3.

(b) Determine the temperature inside the rod at very large times.

[Level : 1]

(a) We first translate the BCs into mathematical language – insulated ends means that $\partial u/\partial x = 0$ at $x = 0$ and $x = \pi$. We know that separation of variables on the 1+1 heat equation followed by rejection of unphysical solutions leads to the general solution

$$u_G(x,t) = A + Bx + \sum_k \left(C_k \cos kx + D_k \sin kx\right) \exp\left(-9k^2 t\right) \quad , \qquad (01)$$

where k may be discrete or continuous (note that $\alpha = 3$ in this problem). To fit the BCs we go back to the separable solutions – we have $\partial u/\partial x = X'(x)T(t)$; since T is not the zero function, $\mathrm{d}X/\mathrm{d}x$ must be zero at $x = 0$ and $x = \pi$. The function X now satisfies the BVP of 4.1.2 Part (a) [which explains my motivation behind selecting that problem at that juncture]; the eigenvalues are $k = 0$ and $k = n$ and the corresponding eigenfunctions are 1 and $\cos nx$. Hence the eigen-function expansion of u is

$$u(x,t) = A + \sum_n C_n \cos nx \exp\left(-9n^2 t\right) \quad .$$

In an exam, eigenfunction expansion is the most important waypoint on the path to the solution of a PDE. So always be sure to write it explicitly and label it as what it is. In any marking scheme with partial credits, a correct eig exp is bound to fetch a very good fraction of the total marks, even if the subsequent coefficient determination is wrong. However, many students write up solutions in such a way that it becomes very difficult for the examiner to identify the eig exp and check whether it's correct. Then, the part marks stand in jeopardy.

Setting $t = 0$ in this expression gives

$$A + \sum_n C_n \cos nx = f(x) \quad ,$$

where $f(x)$ is the function given in the question.

We solve for the coefficients using the extended function method. Since the constant function A and all the $\cos nx$'s are symmetric and 2π-periodic, we extend f SYMMETRICALLY and 2π-periodically . We are told to refer to 4.3.3; indeed we see that the extended f is identical to the function in Part (a) of that problem if we make the correspondence $t \rightarrow x$, $\omega = 1$ and $d = 1/3$. The function is Fourier expanded in Part (b) there; the result is $A = 1/3$ and $C_n = (2/n\pi)\sin(n\pi/3)$. Hence the specific solution to our problem is

$$* \qquad u(x,t) = \frac{1}{3} + \sum_n \frac{2}{n\pi} \sin\frac{n\pi}{3} \cos nx \exp\left(-9n^2 t\right) \quad , \qquad *$$

which is the desired answer.

(b) At large time, all the exponentials go to zero and $u(x,t)$ evidently tends to the constant function 1/3.

*

COMMENTS : The primary difference between the solutions of the previous problem and this one is that u went to zero the last time but to a constant this time. This difference occurred because we changed the BC. Last time we had both ends at temperature zero; this time, we have insulated the ends of the rod. On physical grounds we expect that because of the insulation, the initial heat cannot escape from the rod; it just spreads out all over the rod. This is just the behaviour we get by solving the equation – the limiting temperature profile is $u = \text{const.}$ and the value of the constant is the average of the initial temperature distribution. Simple physical reasoning like this is an excellent fault diagnostic tool – if for instance you have found a solution which goes to zero in time, you immediately know that you've made a mistake.

With insulated BC at both ends of a rod of length L, the general formula for the Fourier coefficient works out to

$$C_n = \frac{2}{L} \int_0^L f(x) \cos\frac{n\pi x}{L} \, dx \quad ,$$

exactly analogous to that for the rod with $u = 0$ BC at both ends. Once again, it is a memorization favourite. No objection so long as you can derive it, and you don't misapply it (for example, carrying it over to a problem whose BCs are given at $-L$ and L instead of 0 and L).

The solutions $A+Bx$ of the heat equation, which do not depend on time, are called STEADY STATE solutions. At very large time, the steady state is the only solution which survives because all other terms in the general solution are attached to decaying exponentials. Since this fact is based on the general solution rather than the eigenfunction expansion, it is extremely robust and holds true whenever a 1+1 heat equation is involved. The standard way of seeing the A term is through insulation, like in this example. A non-trivial Bx term arises

only if we have an inhomogeneous problem so we aren't going to encounter it for a while yet.

Now for another important feature of the solution. Since all eigenfunctions of the 1+1 heat equation are decaying, many times what happens is that the time trace of temperature at any point is monotonic. If the rod is being heated, the temperature at each point keeps going up and up until it hits the steady state. If it's being cooled, it keeps going down and down until it reaches zero. There is no oscillation of temperature anywhere before settling into the steady state. Now, this phenomenon is not ALWAYS true – for example if you take a rod which is initially hot near the ends and zero temperature in the middle, and you set it off with $u = 0$ at the two ends, then points in the centre will first become warmer as the initial heat spreads out and then again become cool as the heat exits the rod. But, more often than not it does happen to hold true, and provides insight into the temperature dynamics of a given problem. More on this when we actually see examples.

Example 5.1.5

Mixed BC

The temperature $u(x,t)$ of the rod x belongs to $[0, L]$ satisfies the 1+1 heat equation

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t} \quad ,$$

with the BCs $u'(0,t) = 0$ and $u(L,t) = 0$ (note the prime on u present in the first BC but absent in the second).

(a) Obtain the eigenfunction expansion of $u(x,t)$.

(b) If the IC is $u(x,0) = f(x)$, obtain the Fourier coefficient formula.

(c) If $f(x) = 12\cos(3\pi x/2L)$, determine $u(x,t)$.

(d) For some choice of $f(x)$, would you expect to see a non-zero steady state temperature inside the rod at large times ? If yes, provide an example $f(x)$. If no, justify your answer.

[Level : 2]

(a) We know that separation of variables on the 1+1 heat equation followed by rejection of unphysical solutions leads to the general solution

$$u_G(x,t) = A + Bx + \sum_k \left(C_k \cos kx + D_k \sin kx\right) \exp\left(-k^2 t\right) \quad ,$$

where k may be discrete or continuous (note that α=1 in this problem). Now we fit the BCs to the general solution. The $x = 0$ BC forces $A = 0$ while the $x = L$ BC forces $B = 0$. $C_k \cos kx$ satisfies the left side BC automatically. With non-zero C_k it satisfies the right side BC if $\cos kL = 0$. Now the zeroes of cos are $\pi/2$, $3\pi/2$ etc. so we get $kL = (n-1/2)\pi$ where $n = 1,2,3,\ldots..$ as the eigenvalue condition. $D_k \sin kx$ again infringes the left side BC unless $D_k = 0$. Hence the eigenvalues are $k = (n-1/2)\pi/L$, corresponding eigenfunctions are $\cos kx$ and the eigenfunction expansion of u is

$$u(x,t) = \sum_n C_n \cos\left(\frac{2n-1}{2}\frac{\pi}{L}x\right) \exp\left(-\left[\frac{2n-1}{2}\right]^2 \frac{\pi^2}{L^2} t\right) \quad , \qquad (01)^*$$

which is the desired answer. Since the eigenvalues involve half-integers, check the limiting case to ensure that the series starts from 1/2 rather than −1/2 or 3/2 or something. Recall our convention that an unqualified n is 1,2,3 etc.

(b) Let's do this using the extended function method. To perform the extension we note that the fundamental period of the eigenfunctions is $4L$ (the first one describes a quarter wave from $x = 0$ to $x = L$). So we need to do TWO extensions before replicating $f(x)$ periodically. Since cos is symmetric, the first extension is symmetric into the interval $[-L, 0]$. Now we observe that the eigenfunctions (quarter wave, three-quarters wave, five quarters wave etc.) are antisymmetric w.r.t the point $x = L$, so we extend $[-L, L]$ antisymmetrically about L into $[L, 3L]$. Only then we can replicate the IC function $f(x)$ periodically over the entire real line.

Given this, now we employ the Fourier formulae (5.1.3.05). We have,

$$C_n = \frac{1}{2L} \int_{-2L}^{2L} f(x) \cos\left(\frac{2n-1}{2}\frac{\pi x}{L}\right) \mathrm{d}x \quad .$$

But from symmetry about origin, the above integral is double of that from 0 to $2L$. Then, both f and the eigenfunctions are antisymmetric about the centre of

the interval $x = L$, so the resultant is again double of the contribution from 0 to L alone. Doing this we have

* $$C_n = \frac{2}{L}\int_0^L f(x)\cos\left(\frac{2n-1}{2}\frac{\pi x}{L}\right)\mathrm{d}x \quad , \qquad (02)^*$$

which is the desired answer.

In this example, the extension process became somewhat tedious. Just for comparison, you can try doing obtaining (02) yourself the linear algebraic way, and see which of the two options you find easier.

(c) We resist the temptation to blindly plug things into the expansion formula we got above . Instead we see that , just as in 5.1.2, the given IC is a SINGLE eigenfunction corresponding to $n = 2$. Evidently, C_2 is 12 and $C_n = 0$ for all other n. Hence, only the term with $n = 2$ survives in the eigenfunction expansion (01) and we have $u(x,t) = 12\cos([3\pi/2L]x)\exp(-[9\pi^2/4L^2]t)$.

(d) We can do this two ways : mathematically and physically. For the mathematical solution, we note that we can see a non-zero steady state temperature only if the $A+Bx$ term is present in the eigenfunction expansion. Then, we can choose an IC which has a non-zero coefficient attached to this term. However, the steady state solution is clearly absent in the expansion (01) of the present problem. So NO $f(x)$ exists which produces a non-zero steady state temperature.

Physically, we see that the right end of the rod is maintained at temperature zero. Although the left end is insulated, that doesn't matter – all the initial heat will flow out through the right end until the temperature becomes zero everywhere. This will of course hold true whatever the initial temperature profile. So NO $f(x)$ exists which produces a non-zero steady state temperature.

*

COMMENTS : If the IC is given as a single eigenfunction as in Part (c) or as an obvious sum of two or three eigenfunctions [say something like $12\cos(3\pi x/2L) - 9\cos(5\pi x/2L)$] then it is by far the best option to solve straight by inspection instead of attempting Fourier expansion. Of course, Fourier expansion isn't wrong, and if you can do the integral correctly, your answer will be right (it had

better be if the theory is to make sense). But doing the integral is completely unnecessary and entails a risk of error. Moreover, since the needless integration shows reliance on memory rather than concepts, examiners often come down hard on you if you try it and get it wrong.

Example 5.1.6

Periodic IC

The temperature $u(x,t)$ in an infinite rod satisfies the 1+1 heat equation

$$2\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t} \quad .$$

The initial condition is $u(x,0) = f(x)$ where $f(x)$ is the triangle wave which goes in a straight line from $(-\pi,-1)$ to $(0,1)$ and then in another straight line from $(0,1)$ to $(\pi,-1)$ and is periodic with period 2π.

Using the results of 4.3.2, determine $u(x,t)$.

[Level : 1]

Although no BCs are given explicitly, the problem is not incomplete in this particular case ! With an infinite rod, it is impossible to prescribe BCs like constant-temperature or insulated ends (if the rod is infinite, where are its ends?). Rather, the BC either takes the form of a plausibility condition (that $u \to 0$ as $|x| \to \infty$) or is somehow rendered redundant by some other information in the problem. Here, this information is that u is periodic with period 2π. We've already seen while discussing extensions in 5.1.3 how periodicity can act as an effective BC; now we'll use that insight to solve this problem.

We know that separation of variables on the 1+1 heat equation followed by rejection of unphysical solutions leads to the general solution

$$u_G(x,t) = A + Bx + \sum_k \left(C_k \cos kx + D_k \sin kx\right) \exp\left(-2k^2 t\right) \quad , \qquad (01)$$

where k may be discrete or continuous (note that $\alpha^2 = 2$ in this problem). Since there are no obvious BCs on $X(x)$, we cannot whittle down the choices of k right away, and we proceed to the IC stage. Setting $t = 0$, the exponential becomes unity, and we are told that the resulting function $f(x)$ is periodic with period 2π. This immediately kills off the Bx term. Moreover, since f is 2π-periodic, it can be expanded in a Fourier series $A_0 + A_1 \cos x + B_1 \sin x + A_2 \cos 2x + B_2 \sin 2x + \ldots..$ and at least at $t = 0$, these are the ONLY terms which are actually present in the sum (01). But if the other k's are absent at $t = 0$, then they can't make an appearance at a later stage. Hence the eigenfunction expansion of u is

$$u(x,t) = A + \sum_n \left(C_n \cos nx + D_n \sin nx\right) \exp\left(-2n^2 t\right) \quad .$$

This is how the periodicity condition acted as a BC substitute, reducing the general solution to an eigenfunction expansion.

Now, we use the result of 4.3.2 which is the Fourier expansion of the particular $f(x)$ we have here. We see that $A = 0$, $C_n = 8/n^2\pi^2$ if n is odd and zero if n is even , and $D_n = 0$. Hence we have

* $$u(x,t) = \sum_{\text{odd } n} \frac{8}{n^2\pi^2} \cos nx \exp\left(-2n^2 t\right) \quad ,$$ *

which is the desired answer.

*

COMMENTS : Another way of posing this problem is to say that a wire of length 2π is bent into a circle and then give its initial temperature. The wraparound nature of the wire automatically enforces on the problem the periodicity BC we used here. These last five Tutes are very typical Final exam problems on heat equation. Of course there are a lot of variations – a different IC is the most common, while a modified equation (say heat equation with dissipation) is another possibility. Once you master these five, your base is ready and you are well prepared to handle any simple problem of this type. We will now embark on a gradual increase of the sophistication so that we can reach a level where we use 1+1 heat equation to model and describe real phenomena rather than merely calculate temperatures in toy rods.

5.2

DISCRETIZATION AND MATRIX METHOD

In N°. 12 we have seen how we can discretize a BVP using the matrix method. The same concept carries over to PDEs as well. To discretize the 1+1 heat equation defined in a finite spatial interval, we partion the space into N lattice sites and define a temperature u_i at each site. Following standard matrix method procedure, we let $\mathbf{u}$ be the vector of all the u_i's. Recall that the second spatial derivative together with boundary conditions is represented by the matrix $\mathbf{Q}$, which consists of rows containing $(1/h^2)$ [….. 1 –2 1 …..] together with some modification at the corners to accommodate the BCs. Thus, $\partial^2 u/\partial x^2 \rightarrow \mathbf{Qu}$. This discretization reduces an ODE problem to an algebraic equation; here there's an extra complication because of the $\partial/\partial t$ term.

In the most general case, we would need to implement a numerical method like EULER's method to handle the time part. However, because heat equation is linear and constant-coefficient, we don't tamper with it in this case. Rather, we treat each u_i as a continuous dynamical variable and express the heat equation as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{u} = \mathbf{Qu} \quad ,$$

which is a coupled LCCH ODE system like the ones in N°. 07. We can easily solve this system (with the aid of a computer where appropriate) and thus generate an approximate numerical solution to the PDE.

Other than the numerical utility, the discretization also has some theoretical value. In the previous Lecture, I mentioned that the spatial eigenfunctions are "often" complete and orthogonal – the matrix representation tells us whether they are or not in a given situation. The spatial eigenfunctions are of course the limit of the eigenvectors of $\mathbf{Q}$. Now if $\mathbf{Q}$ is symmetric, its eigenvectors are orthonormal and they span the space. Spanning space means that every arbitrary vector can be written as a linear combination of the eigenvectors, or, in the continuous limit, the arbitrary initial condition can be written as a

linear combination of the spatial eigenfunctions. Hence a symmetric **Q** gives us both completeness and orthonormality at one go. In the Literature, the condition which is given for eigenfunction orthonormality is that the differential operator must be **self-adjoint**. Understanding this condition in terms of the PDE and the BCs requires considerably sophisticated manipulations, which make it suitable for graduate level courses only. But what the fancy maths boils down to is essentially this. Recall that we've already worked with symmetric (self-adjoint) operators once, in T063.

Now you can understand why, back in L07, I had said that the very thought of the practically useful matrices not being symmetric is hair-raising. If they weren't symmetric, then there would be no one-shot integral formula for the coefficients in the eigenfunction expansion. Every time, we would have had to find the coefficients by considering some certain number of eigenfunctions, Gram Schmidting them to orthonormality and then using the dot product. Now let's forget about this dystopia and come back to our more or less well-behaved world.

Example 5.2.1

Baby and big heat equations

Consider the problem posed in T078, which is repeated below.

"The temperature $u(x,t)$ of the rod x belongs to $[0, \pi]$ satisfies the 1+1 heat equation

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t} \quad ,$$

with BCs $u(0,t) = u(\pi,t) = 0$ and IC $u(x,0) = x/\pi$."

Reference may also be drawn to T033 and T062 for solving the present problem.

(a) Use the matrix method to write the PDE as $\mathrm{d}\mathbf{u}/\mathrm{d}t = \mathbf{Q}\mathbf{u}$.

(b) Suitably discretizing the IC, solve the resulting system on a computer and with the help of appropriate plots compare between the numerical and analytical solutions as the number of lattice sites increases.

[Level : 2]

(a) We use the results of T062. We know from there that $h = \pi/(N+1)$ [here the BCs are at 0 and π rather than 0 and 1] and that the matrix $\mathbf{Q}$ which implements zero BC at each end has the form (using $N = 4$ for display purposes)

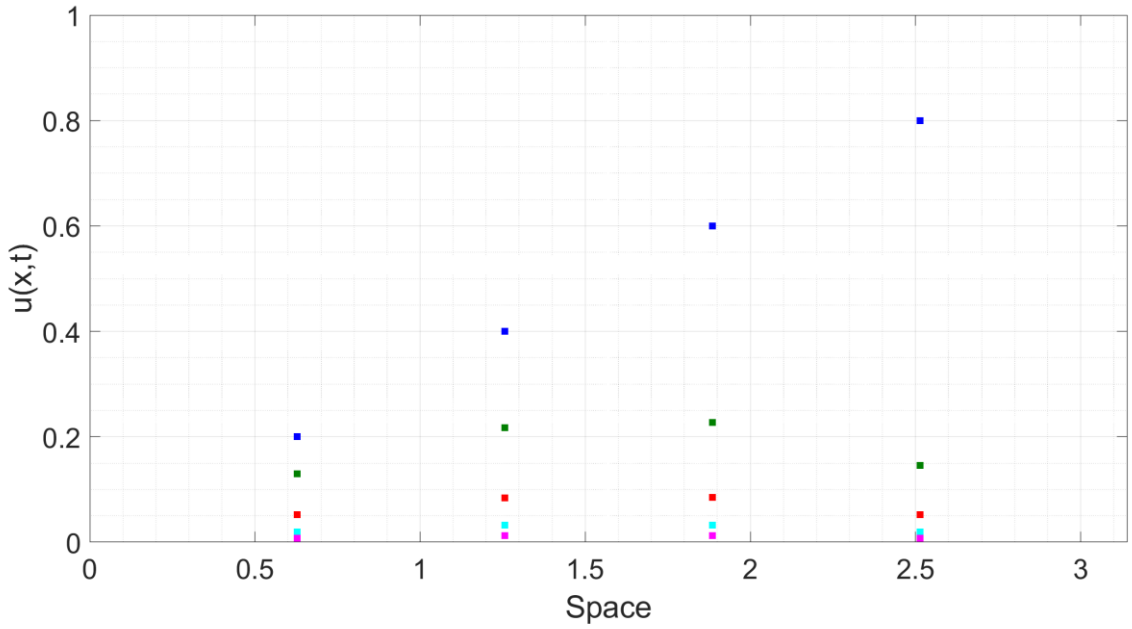
$$\mathbf{Q} = \frac{1}{h^2}\begin{bmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \end{bmatrix} .$$

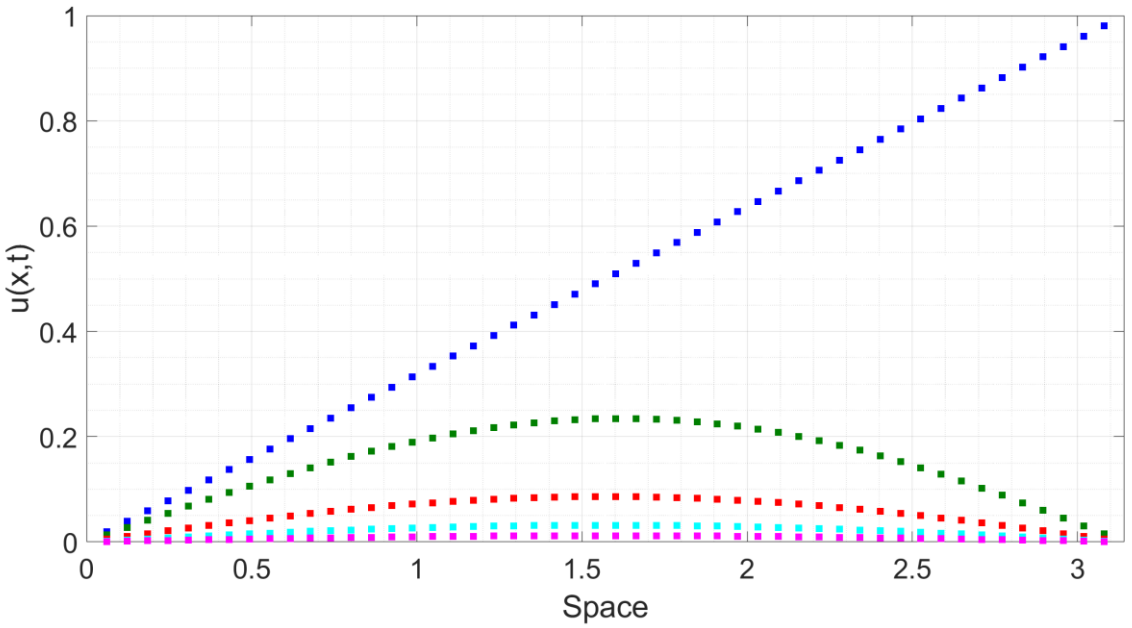
Hence our discretized PDE looks like

$$* \qquad \frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \frac{1}{h^2}\begin{bmatrix} -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} , \qquad *$$

which is the same as the baby heat equation from T033 and is the desired answer.

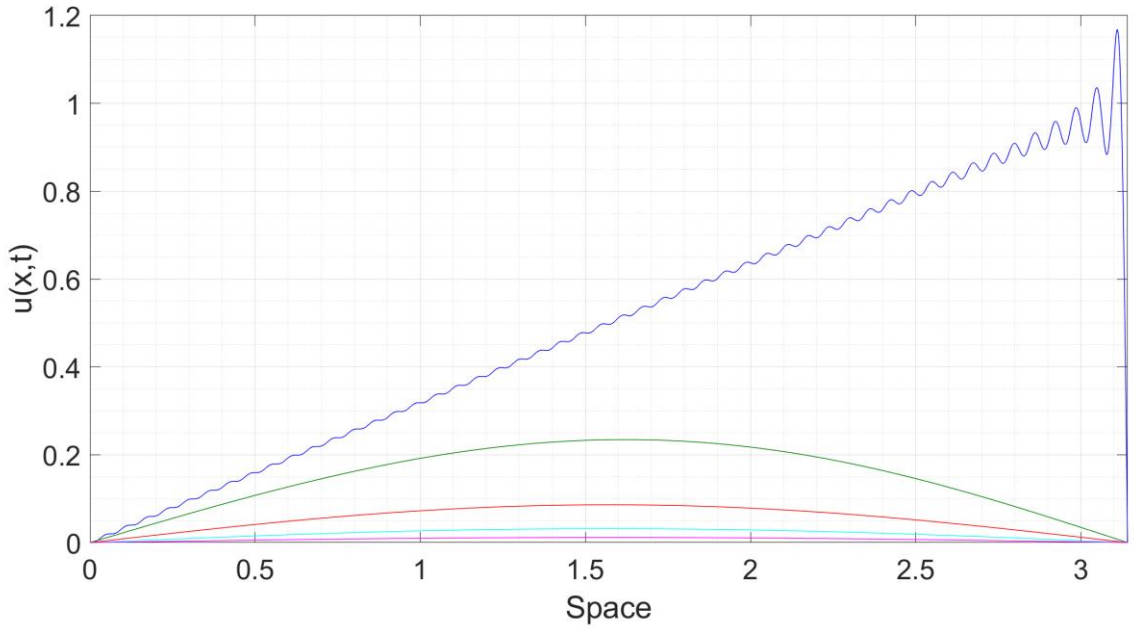
(b) The actual IC corresponds to the straight line $y = x/\pi$ in the interval $[0, \pi]$. In the discretized form, we let $u_i = (1/\pi)x_i$ at every significant lattice site, noting that the BCs are taken care of by the additional sites. This done, we use the machinery of N°. 07 on the system. Since I have already elaborated the process in full while solving T033, I will skip all the details now and go straight to the results. Below we show plots of the discrete solution as a function of space, and at five instants of time, $t = 0,1,2,3,4$. The successive time traces are in blue, green, red, cyan and magenta respectively. We have split the space into four lattice sites for the first plot, and 50 for the second.





In both these plots we can see that the solution decays to zero as time goes on and that at moderately large times $u(x,t)$ looks like a pure sine function of ever-decreasing amplitude, as we predicted in T078 Part (b).

Finally, we present a plot of the analytical solution of T078. Since the solution is the in the form of a sum, we have taken the first 100 terms of the series to generate the plot.



As we would expect, this is nearly a carbon copy of the plot with $N = 50$. There is a GIBBS' phenomenon wiggle in the IC near $x = \pi$ because the sawthooth wave is discontinuous there. Fortunately, this phenomenon is extremely localized so it doesn't affect the accuracy of solutions based on Fourier expansions. The time traces of u at the subsequent times do not suffer from this problem as $u(x,t)$ doesn't have discontinuities at any t greater than zero. In these plots, we can clearly see that the temperature at all points in the rod is decreasing monotonically in time.

*

COMMENTS : If your course has emphasized the LCCH ODE systems of N°. 07, now should be an eureka moment for you. This problem shows you that a PDE is in fact the limit of an ODE system. When you see for the first time the concepts of separation of variables and eigenfunction expansion, they might appear like rabbits pulled out of a hat, topics which are having no relation to anything covered previously in the course. But that's not true. The LCCH ODE systems automatically produce 'separable solutions' because the eigenvectors are getting multiplied by functions of time. The general solution is an arbitrary linear combination of these solutions, and the combination gets determined from the IC. The basic machinery is the same here. Since the PDE is continuous in space,

there are of course infinitely many eigenvalues and functions involved here. If the matrix is symmetric (or the operator self-adjoint), the space eigenfunctions are orthonormal and we can use dot product to calculate the unknown coefficients. This too is a carryover from the LCCH ODE system case. Specifically considering the matrix $\mathbf{Q}$, we evaluated its eigenvalues and vectors in T062, showing that they indeed turn into their continuous counterparts in the limit and making the analogy between the two systems complete.

Finally, coming back to the matrix method, note that it is also a very easy but robust method for numerically obtaining solutions to PDE problems, and we'll keep it up our sleeve to deploy whenever analytical solution might prove to be intractable.

Example 5.2.2

Leakage BC

The temperature $u(x,t)$ of the rod x belongs to $[0, L]$ satisfies the 1+1 heat equation

$$\alpha^2 \frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t} \quad ,$$

with BCs $u = 0$ at $x = 0$ and $u' + \gamma u = 0$ at $x = L$ for some positive constant γ. Such a boundary condition represents heat leaking out of the end of the rod at a rate proportional to the temperature there.

(a) Obtain the eigenfunction expansion for $u(x,t)$. Your answer should feature the roots of a certain transcendental equation, which do not have simple algebraic expressions.

(b) Are the eigenfunctions orthogonal ? Justify your answer. If they are, also determine the normalization constants.

[Level : 2]

(a) We know that separation of variables on the 1+1 heat equation followed by rejection of unphysical solutions leads to the general solution

$$u_G(x,t) = A + Bx + \sum_k \left(C_k \cos kx + D_k \sin kx\right) \exp\left(-\alpha^2 k^2 t\right)$$

,

where k may be discrete or continuous. The BC at $x = 0$ immediately kills off the A and $C_k \cos kx$ terms and we go over to the BC at $x = L$. Considering the Bx term, $u' + \gamma u = 0$ implies $B + \gamma BL = 0$ or $B(1+\gamma L) = 0$. Since γ is positive, the only way for this to be satisfied is $B = 0$ and we have the trivial case. Finally, look at $D_k \sin kx$. Imposing the BC yields

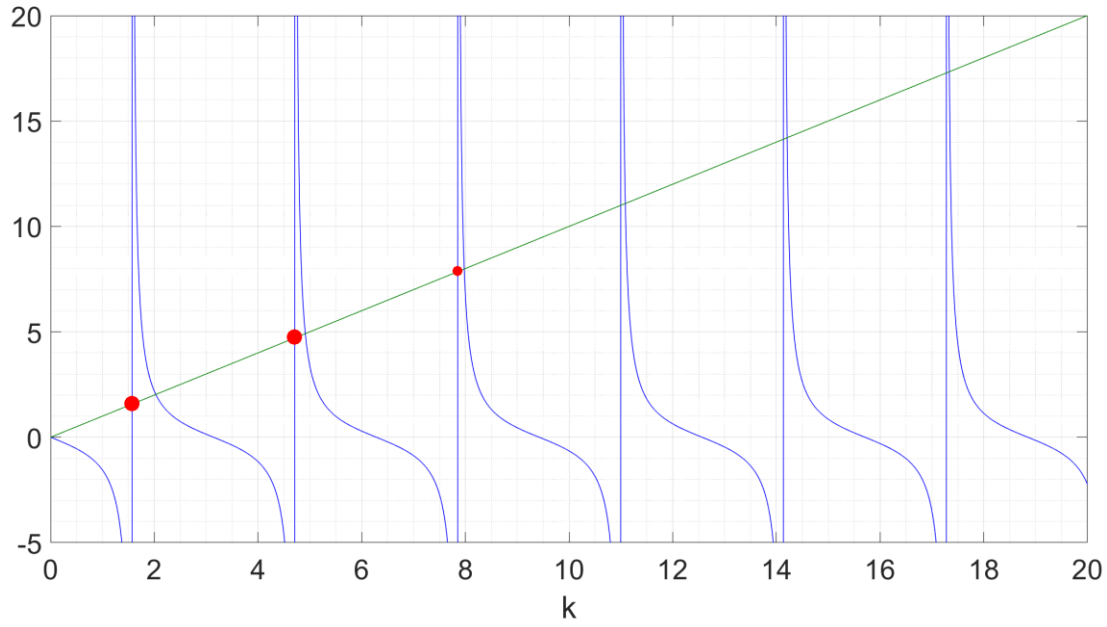
$$kD_k \cos kL + \gamma D_k \sin kL = 0 \quad .$$

Since $D_k = 0$ is the trivial case, we exclude that possibility and get the eigenvalue equation

$$k = -\gamma \tan kL \quad . \qquad (01)$$

This is the unsolvable transcendental equation mentioned in the question.

To visualize the nature of the roots, we plot the functions k and $-\gamma \tan kL$ (with $\gamma = L = 1$) simultaneously; the roots are the intersections of the two curves.



Note however that there are some spurious intersections ! The computer plots tan with a straight line joining $(\pi/2, -\infty)$ to $(\pi/2, \infty)$ [whatever its definition of ∞ may be] but the actual function has a jump there and not a line. The plot has a

spurious line for every jump of tan. So the intersections of k with these fictitious lines are clearly false. I have highlighted the first three of these fakes with red dots. Clearly, there is a sequence of roots with the first one at around $k = 2$ and the successive ones at higher and higher values. Changing γ and L will not affect the qualitative characteristics of the plot, and we'll still have an increasing sequence. Since the functions k and tan are both antisymmetric, the roots on the negative side will be the exact negatives of those on the positive side and the corresponding eigenfunctions will also be the negatives of those picked up from positive k. Hence the negative k region is redundant.

We let β_n denote the n^{th} root of (01) with β_1 being the one near 2 in the plot. Then the eigenfunction expansion is

$$u(x,t) = \sum_n D_n \sin \beta_n x \exp\left(-\alpha^2 \beta_n^2 t\right) \quad , \tag{*}$$

which is the desired answer. As an aside, note that the ODE eigenvalue BVP derived from this problem, $y'' + \lambda y = 0$ with BCs $y(0) = 0$ and $y'(L) + \gamma y(L) = 0$, is a popular exam question.

(b) To determine whether eigenfunctions are orthogonal, we discretize the equation using matrix method and see if the matrix is symmetric. Since we haven't worked with these BCs before, we start from the raw second derivative matrix $\mathbf{Q}_+$. Defining the N significant sites and adding site #0 and #N+1 at the two ends we have (using $N = 4$ for display purposes)

$$\mathbf{u}'' = \left[u''_1;\quad u''_2;\quad u''_3;\quad u''_4\right] = \mathbf{Q}_+ \mathbf{u}_+$$

$$= \frac{1}{h^2}\begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} \quad .$$

Since the left side BC is just $u = 0$, u_0 becomes zero and the first column of $\mathbf{Q}_+$ goes away with no modification. At the right end we have $u' + \gamma u = 0$. Taking a right side first derivative at site #N, we can write that as

$$\frac{u_{N+1}-u_N}{h}+\gamma u_N=0 \quad \text{, or}$$
$$u_{N+1}=(1-\gamma h)u_N \quad .$$

Now we use this information to eliminate the last column of $\mathbf{Q}_+$ (recall the similar step at the right side BC in T067). Expanding along the last row, we see that we can get rid of the offending column if we replace the -2 by $-1-\gamma h$. Hence our matrix $\mathbf{Q}$ is

$$\mathbf{Q}=\frac{1}{h^2}\begin{bmatrix}-2 & 1 & 0 & 0\\ 1 & -2 & 1 & 0\\ 0 & 1 & -2 & 1\\ 0 & 0 & 1 & -1-\gamma h\end{bmatrix} \quad .$$

This matrix is symmetric so its eigenvectors are indeed orthogonal. Since the eigenfunctions are the limiting cases of the eigenvectors, we conclude that they are orthogonal also.

To obtain the normalization constant, we calculate the "length" of each eigenfunction i.e. the square root of each eigenfunction dotted with itself. We have

$$\int_0^L \sin^2\beta_n x\,dx=\int_0^L\frac{1}{2}(1-\cos 2\beta_n x)\,dx$$
$$=\frac{L}{2}-\frac{1}{4\beta_n}\sin 2\beta_n L \quad .$$

(To get the double angle trig formula from our trig review, set $D=C$ in formula (12) there.) Denoting the last expression above by μ_n, the length of the n^{th} eigenfunction must be $\sqrt{\mu_n}$. Hence the normalization constant for the n^{th} eigenfunction is $v_n=1/\sqrt{\mu_n}$, which is

$$v_n=\frac{1}{\left(\frac{L}{2}-\frac{1}{4\beta_n}\sin 2\beta_n L\right)^{1/2}} \quad ,$$

and is the desired answer.

*

COMMENTS : The most surprising feature of this problem is that the eigenfunctions are orthogonal. On the face of it, you would never guess that – $\sin \beta_n x$'s where β_n's are some weird transcendental numbers don't certainly look orthogonal. Even proving orthogonality by direct calculation is not at all easy because the β_n's have no explicit expressions. However, by looking at the operator rather than its eigenfunctions, we were able to show orthogonality with very little effort.

A second thing is that the normalization constant is different for each eigenfunction. In the earlier problems, if you try doing them without using extension, you will find that $v_n = \sqrt{(2/L)}$ for all the eigenfunctions. This time however, that cute relation no longer holds. These intractable eigenfunctions also mean that if we have to fit an IC, we can no longer do it by performing extensions – although the sines are all antisymmetric, each is periodic with a different period. To practically solve such a problem, a computer will be essential (it enters the picture the moment the eigenvalue equation becomes transcendental). Nevertheless the orthogonality condition will be a significant help while working the machine.

Finally, why does this BC represent leakage ? Just hold your patience a little longer, while I finish building up the mathematical arsenal. When that is fully stocked, we are going to get the physical pictures behind the heat equation and the different boundary conditions.

5.3

INHOMOGENEOUS PDE PROBLEMS

When we add excitation to a homogeneous ODE, it becomes quite exciting – the drab LCCH ODEs of N°. 05 turned into something completely different from N°. 11 onwards. When we add excitation to a homogeneous PDE, it can become really really exciting. Here we see some methods of cracking (or at least making inroads before handing over to a computer) some of these problems.

There are two places where excitation can enter a PDE problem – **either in the equation or in the BC** (and of course it can enter both). A problem of the former type looks something like $\alpha^2 u'' = \dot{u}$ with BCs $u(0,t) = f(t)$, $u(L,t) = g(t)$ and some IC. A problem of the latter type looks like

$$\frac{\partial u}{\partial t} - \alpha^2 \frac{\partial^2 u}{\partial x^2} = f(x,t) \quad , \tag{01}$$

with homogeneous BCs and some IC.

As with ODEs, the trick is to **split the solution $u(x,t)$ into a homogeneous part u_h and a particular part u_p.** Absorb the inhomogeneity into the particular solution and use the homogeneous solution to fit the IC. In most cases, what happens is that all eigenfunctions of the homogeneous problem feature decaying exponentials so they go away at large time, leaving behind the particular solution alone.

With inhomogeneous BC, the particular solution is generally found by some kind of guesswork. Looking at the BC we guess the time-dependence of the particular solution and then solve for the spatial dependence. This is an extension of the method of undetermined coefficients – however, instead of unknown constants, this time we have unknown functions. We'll see examples in the Tutes.

GREEN'S FUNCTION

Suppose we have an inhomogeneous PDE of the form (01) in a certain spatial interval with homogeneous boundary conditions at the ends. Then, we can define the **GREEN's function** $G(x,x',t,t')$ to be a function which satisfies

$$\left(\frac{\partial}{\partial t}-\alpha^2\frac{\partial^2}{\partial x^2}\right)G(x,x',t,t')=\delta(x-x')\delta(t-t') \quad ,$$

together with the homogeneous BCs. Here $\delta(\ldots)$ is of course the Dirac delta function. In terms of this function, a solution of (01) is

$$* \qquad u(x,t)=\int_{\text{interval}} dx' \int_0^t dt'\, f(x',t')G(x,x',t,t') \quad . \qquad \textbf{(02)*}$$

Here the x' integral runs over the entire region where the problem is defined, while the t' integral runs from the initial time (taken as zero here) to the current time t. This fact stems from causality of the GREEN's function. This solution also assumes $u(x,0) = 0$; if that is not the case, we can add on a suitable homogeneous solution. We'll derive all this in detail in a Tute.

One particular GREEN's function for 1+1 heat equation is very famous. It corresponds to the case where the rod is infinitely long i.e. the interval of definition is x belongs to $[-\infty, \infty]$. The BCs are that as $|x| \to \infty$, $u \to 0$. In this situation,

$$* \qquad G(x,x',t,t')=\frac{1}{2\alpha\sqrt{\pi}\sqrt{t-t'}}\exp\left(-\frac{(x-x')^2}{4\alpha^2(t-t')}\right) \quad . \qquad \textbf{(03)*}$$

This is in fact the solution I had quoted in L16 as an example of something which looks unrelated to separable solutions (there I had chosen $x' = t' = 0$). The complicated derivation of this formula is again the subject of one of the Tutes.

Example 5.3.1

Inhomogeneous BC

The temperature $u(x,t)$ of the rod x belongs to [0, 1] satisfies the 1+1 heat equation

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t} \quad .$$

The left end is maintained at room temperature 30 ^{o}C while the right end is immersed in boiling water. The initial temperature distribution is $u(x,0) = 50\ ^{o}$C.

(a) Write the PDE problems satisfied by the homogeneous and particular solutions.

(b) Make a reasonable guess for the temporal behaviour of the particular solution, and hence determine $u_p(x,t)$.

(c) Determine the temperature inside the rod at large times.

[Level : 2]

(a) Immersion in boiling water is just a different way of saying that the right end is maintained at 100 °C. We can redefine temperature scale to set u at either end to zero, but not both at once. So we let 30 °C be zero on the new scale. Then the new BCs are $u(0,t)=0$ and $u(1,t)=70$, and the IC is $u(x,0)=20$. Here there is an apparent incompatibility between IC and BCs, so we reason as in T078 and agree that the IC is actually something which is zero exactly at the left end, jumps instantaneously to 20 there, and again jumps instantaneously to 70 at exactly the right end.

Evidently, we have an inhomogeneous problem here, so we decompose $u = u_h + u_p$. We note however that the inhomogeneity lies only in the BC – since the equation itself is homogeneous, both u_h and u_p must be satisfying it. By definition, u_h must be satisfying homogeneous BCs, so we have $u_h(0,t) = u_h(1,t) = 0$. Now we fit the BCs to the particular solution so that their sum can satisfy the overall BCs given on u. At the left end we have $u(0,t) = 0$, so $u_h(0,t) + u_p(0,t) = 0$. Since $u_h(0,t) = 0$, $u_p(0,t)$ must be zero also. At the right end, $u(1,t) = 70$, so $u_h(1,t) + u_p(1,t) = 70$, or since $u_h(1,t) = 0$, $u_p(1,t)$ must equal 70. Thus, u_p satisfies both the BCs of u as a whole.

Of course, the part which u_p should NOT satisfy is the IC (if we force both BCs and IC upon u_p then the decomposition becomes meaningless). At $t = 0$, $u_p = u_p(x,0)$, which is some function $f(x)$. We can calculate $f(x)$ if we can obtain u_p exactly, otherwise it will just remain as "some function" – but for now it doesn't matter. Now, fit the IC to the homogeneous solution so that their sum can satisfy the overall IC given on u. We have, $u(x,0) = 20$, or, $u_h(x,0) + u_p(x,0) = 20$ or, $u_h(x,0) = 20 - f(x)$.

Combining all of this, we have the following. Both u_h and u_p satisfy the 1+1 heat equation $u'' = \dot{u}$. The particular solution satisfies the BCs $\underline{u_p(0,t) = 0}$ and $\underline{u_p(1,t) = 70}$. The homogeneous solution satisfies the BCs $\underline{u_h(0,t) = u_h(1,t) = 0}$ and the IC $\underline{u_h(x,0) = 20 - u_p(x,0)}$.

(b) The behaviour of the particular solution will of course be related to the excitation which produces it. Now, this excitation is a constant function of time – the right end is just maintained at $u = 70$. Hence we guess that u_p will be constant in time also. So we try something like $u_p = g(x)$ for some unknown function g. Plugging into the heat equation, $\partial u_p/\partial t = 0$ so $\partial^2 u_p/\partial x^2 = 0$ also and

$g'' = 0$. The solution of this is $g(x) = A + Bx$; fitting the two BCs we get $A = 0$ and $B = 70$. Hence the particular solution is $u_p = 70x$. Since the problem is posed in the Centigrade scale, we report the answer in those units : $\underline{u_p(x,t) = 70x + 30\ ^\circ C}$.

(c) We know from T078 that all eigenfunctions of the homogeneous problem have negative temporal exponentials, so they die out in time. Since the IC on u_h gets expanded in these eigenfunctions, $u_h(x,t)$ dies out also, and at large times we see only the particular solution which is $\underline{70x + 30\ ^\circ C}$.

*

COMMENTS : We see several similarities (and differences) between inhomogeneous PDEs and the more familiar inhomogeneous ODEs. The biggest similarity is undoubtedly that the specific solution is the sum of homogeneous and particular. At a more microscopic level, the IC goes into the homogeneous solution in both situations. The primary difference is of course the extra dimension here. In ODE, the particular solution doesn't satisfy the IC and that is all – there is no extra BC which it should or shouldn't satisfy. The calculation of Part (b) here is the PDE equivalent of undet coeff. For an ODE with constant excitation, you would guess a constant response and plug into the equation to find that constant; for the PDE the constant gets replaced by a function of space. But again, the concept of plugging a guess solution into the equation and then hacking it is common to both situations.

Part (c) here is a oneliner. Many students however embark immediately on a Fourier expansion of the IC on u_h, determine $u_h(x,t)$ and hence $u(x,t)$ completely and only then take the limit $t \to \infty$. If the process goes correctly, they arrive where we did after countless minutes lost; else they end up at a garbled point and throw their marks into the water. Whenever the question says "at large time" or something equivalent, the probability is very high that there's some trick involved – go for a brute force calculation only after making sure that you can't find any short cut.

Example 5.3.2

Free space GREEN's function

Derive the formula (L1803).

[Read Only]

We start with a statement of what exactly we want to show. This is that, if the function $G(x,x',t,t')$ satisfies the equation

$$\left(\frac{\partial}{\partial t}-\alpha^2\frac{\partial^2}{\partial x^2}\right)G(x,x',t,t')=\delta(x-x')\delta(t-t') \quad ,$$

defined in the interval x belongs to $[-\infty, \infty]$ with BCs that $G \to 0$ as $|x| \to \infty$, then

$$G(x,x',t,t')=\frac{1}{2\alpha\sqrt{\pi}\sqrt{t-t'}}\exp\left(-\frac{(x-x')^2}{4\alpha^2(t-t')}\right) \quad .$$

Because the interval is the entire real line and the BCs are plausibility conditions, this G is typically called the free space GREEN's function for the 1+1 heat equation.

As a first step, let's take x' and t' to be zero and try to get G in this case. This simplified G will be a function of x and t alone. Since the heat equation does not feature x and t explicitly, it is invariant under the transformation $x = x + \xi$ and $t = t + \tau$ for all ξ and τ. This is the extension of what we saw in T020 for the harmonic oscillator. Then, after getting the simplified G, we can simply replace x and t there with $x-x'$ and $t-t'$ respectively. So we try to solve

$$\left(\frac{\partial}{\partial t}-\alpha^2\frac{\partial^2}{\partial x^2}\right)G(x,t)=\delta(x)\delta(t) \quad , \qquad (01)$$

with BCs as previously.

To better understand what happens now, let us replace for a moment the function $\delta(x)$ by some "more tractable" $f(x)$. Note that at all $t > 0$, the RHS of (01) is identically zero, so at these times, G satisfies the homogeneous equation. We claim that the effect of the $\delta(t)$ is to generate the IC for this equation as $G(x,0) = f(x)$. Check the claim manually – if $G = 0$ upto $t = 0$ and $G = f(x)$ at $t = 0$ then $\partial G/\partial t$ is $f(x)\,\delta(t)$. As for $\partial^2 f/\partial x^2$, since we took f to be a tractable function, this quantity is finite at $t = 0$ (it is zero before). But at $t = 0$, the delta function is infinite, so the f'' term is negligible by comparison and can be dropped. Hence, $G = 0$ before $t = 0$ and $G = f(x)$ at $t = 0$ does satisfy that $\dot{G} - \alpha^2 G'' = f(x)\delta(t)$. Now, although the delta on x is not a typical tractable function, it is well-behaved enough for the present purposes, and we still have $G(x,0) = \delta(x)$ as the IC for the homogeneous equation.

We know that separation of variables on the 1+1 heat equation followed by rejection of unphysical solutions leads to the general solution

$$G(x,t) = A + Bx + \sum_k \left(C_k \cos kx + D_k \sin kx\right) \exp\left(-\alpha^2 k^2 t\right) \quad , \qquad (02)$$

where k may be discrete or continuous. The A and B terms evidently flout the BCs unless they are both zero, so we can drop them both from the general solution. As regards the cos and sin terms, they too don't appear to go to zero at large x. But it so happens that it is possible to represent decaying functions as continuous linear combinations of cosines and sines. This is called the FOURIER TRANSFORM.

In Fourier transform, what we do is we write a non-periodic function $f(x)$ as

$$f(x) = \int_{-\infty}^{\infty} \tilde{f}(k) \mathrm{e}^{\mathrm{j}kx} \, \mathrm{d}k \quad .$$

(For periodic functions of course, the Fourier series replaces the transform.) Here, $\tilde{f}(k)$ can be complex and is called the Fourier transform of $f(x)$. The fundamental theorem in this area states that

$$\tilde{f}(k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(x) \mathrm{e}^{-\mathrm{j}kx} \, \mathrm{d}x \quad .$$

As regards the proof of this theorem, I had seen it once (in that lovely course on mathematical Fourier analysis mentioned in L14) and remember nothing of it. But for us, it suffices to know that this works.

Given this, every possible k in the general solution (02) becomes an eigenfunction, so the sum turns into an integral and the eigenfunction expansion becomes

$$G(x,t) = \int \mathrm{d}k \left(C(k) \cos kx + D(k) \sin kx\right) \exp\left(-\alpha^2 k^2 t\right) \quad .$$

Like the sep var solutions, this integral is over $k>0$. We now write $C(k)\cos kx + D(k)\sin kx$ as $g(k)\,\mathrm{e}^{\mathrm{j}kx}$ where $g(k)$ is complex. This immediately raises the question of how to cancel the imaginary part. To make that happen, we include negative k also in the integral. Then, we will pick up a term $g(-k)\,\mathrm{e}^{-\mathrm{j}kx}$ in addition to $g(k)\,\mathrm{e}^{\mathrm{j}kx}$; if $g(-k)$ is the complex conjugate of $g(k)$, then their resultant will become a real valued linear combination of $\cos kx$ and $\sin kx$. Finally, since

$k = 0$ is just a point on the k-line, there's no harm if we take it on board our integral. With these modifications, we have

$$G(x,t) = \int_{-\infty}^{\infty} \mathrm{d}k \; g(k) \exp(\mathrm{j}kx) \exp\left(-\alpha^2 k^2 t\right) \quad , \qquad (03)$$

as our new eigenfunction expansion.

Now for the IC fit. Setting $t = 0$ in (03),

$$G(x,0) = \int_{-\infty}^{\infty} g(k) \mathrm{e}^{\mathrm{j}kx} \, \mathrm{d}k = \delta(x) \quad .$$

Although I can't prove FOURIER's transformation theorem even if paid to, I know when to use it. Here is one such instance – $g(k)$ is evidently the Fourier transform of $\delta(x)$. So,

$$g(k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \delta(x) \mathrm{e}^{-\mathrm{j}kx} \, \mathrm{d}x \quad .$$

From the properties of delta function, the integral evaluates to $\mathrm{e}^{-\mathrm{j}k.0}$ which is 1, so $g(k) = 1/2\pi$. Plugging this into (03) we get

$$\begin{aligned} G(x,t) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}k \;\; \exp(\mathrm{j}kx) \exp\left(-\alpha^2 k^2 t\right) \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}k \;\; \exp\left(-\alpha^2 k^2 t + \mathrm{j}kx\right) \end{aligned} \quad .$$

We have now reduced $G(x,t)$ to an integral; it remains only to evaluate this integral. Nowadays, every integral can be found on a website, and we will hardly ever need to do an integral like this again. So I will just outline the steps involved. First define variable $K = k/(\alpha\sqrt{t})$, then complete the square inside the exponential, and then use that $\int_{-\infty}^{\infty} \exp\left(-x^2\right) \mathrm{d}x = \sqrt{\pi}$. After hacking all the steps we get

$$G(x,t) = \frac{1}{2\alpha\sqrt{\pi}\sqrt{t}} \exp\left(-\frac{x^2}{4\alpha^2 t}\right) \quad .$$

Now check that G satisfies the BCs at infinity. We can see that G is a Gaussian function having time-varying height and width. Since Gaussians always go to zero at infinity, G satisfies the BCs of the problem.

Finally, we make the substitution $x \rightarrow x - x'$ and $t \rightarrow t - t'$ to get

$$G(x,x',t,t') = \frac{1}{2\alpha\sqrt{\pi}\sqrt{t-t'}} \exp\left(-\frac{(x-x')^2}{4\alpha^2 (t-t')} \right) \quad ,$$

which is the formula (L1803). This completes the derivation.

*

COMMENTS : This derivation is technically quite difficult, but it serves some important purposes. Firstly, it shows you how we can derive an apparently unseparable solution using separation of variables – merely quoting the result doesn't really satisfy. The same holds for the GREEN's function itself – we don't need the derivation to use the function but still, some of you certainly will want to know wherefrom it came about. Secondly, this problem introduces you to Fourier transform methods. Although we won't be using it too much in this course, more conventionally advanced math methods courses, especially in physics and electrical engineering, do Fourier transforms and inverse transforms like anything. So this gives you a heads-up for that kind of course.

Example 5.3.3

GREEN's function approach

Drawing reference to T065-6, derive the formula (L1802).

[Level : 1]

What we want to show this time is that : if $u(x,t)$ satisfies the inhomogeneous PDE

$$\frac{\partial u}{\partial t} - \alpha^2 \frac{\partial^2 u}{\partial x^2} = f(x,t) \quad ,$$

in some spatial interval with homogeneous BCs at the two ends, and if $G(x,x',t,t')$ be a solution to

$$\frac{\partial G}{\partial t} - \alpha^2 \frac{\partial^2 G}{\partial x^2} = \delta(x-x')\delta(t-t') \quad ,$$

in the same interval with the same BCs, then

$$u(x,t) = \int_{\text{interval}} dx' \int_0^t dt'\, f(x',t') G(x,x',t,t') \quad .$$

This formula is evidently very similar to the result of T066 Part (d), so we more or less copy the steps which led to that conclusion.

In the ODE situation, we started by treating the continuous line load as a limit of a sequence of point loads, i.e. we wrote

$$q(x) = \text{Lt} \sum_i F_i \delta(x - x_i) = \int_0^L dx'\; q(x')\delta(x-x') \quad .$$

We do the same here, treating the continuous $f(x,t)$ as the limit of a sequence of impulsive functions. Here, the impulses are in both space and time. The strength of the impulse at point (x',t') is $f(x',t')\,dx'\,dt'$ and we have,

$$f(x,t) = \int_{\text{interval}} dx' \int_0^\infty dt'\, f(x',t')\delta(x-x')\delta(t-t') \quad .$$

Evidently, this is a mathematical identity. For now I have let the upper limit on t' be infinity because there's nothing to suggest a cutoff – I'll bring that in later.

Next, since the equation is linear and the BCs homogeneous, we write the combined response as the sum of the responses to each individual impulse. By definition of G, the response to unit impulse at point (x',t') is $G(x,x',t,t')$ so the response to an impulse of strength $f(x',t')dx'dt'$ at point (x',t') is $f(x',t')\,dx'\,dt'\,G(x,x',t,t')$. The total response is the sum of all these responses which is

$$u(x,t) = \int_{\text{interval}} dx' \int_0^{\infty} dt' \, f(x',t') G(x,x',t,t') \quad .$$

But now think about the time integration. Temporal causality is a basic tenet of Nature – an action taken at a certain time can only generate a response at all future times. Hence the impulse applied at time t' can generate response of the system only at times $t > t'$. This means that t' must be less than or equal to t, so the integration on t' should run only from 0 to t. (Since there is no principle of spatial causality, there's no such restriction imposed on the x' integral.) Hence we finally get

$$u(x,t) = \int_{\text{interval}} dx' \int_0^{t} dt' \, f(x',t') G(x,x',t,t') \quad ,$$

which is the formula (L1802).

Note that the free space GREEN's function (L1803) forces causality automatically – for $t' > t$, it is imaginary. But in a different interval and with a different BC even if the GREEN's function does not enforce causality by default, then also you must impose this condition on the integral formula to get meaningful results.

Finally, at $t = 0^+$, you can see that the t' integral runs over an infinitesimal interval. If f is a well-behaved function, this implies that $u(x,0) = 0$. Hence the formula (L1802) gives a solution satisfying this IC. If the given IC is different, we can always add to this expression a solution of the homogeneous equation so that the IC gets satisfied.

What happens if both the PDE and the BCs are inhomogeneous i.e. you have something like $\dot{u} - \alpha^2 u'' = f(x,t)$ with BCs $u(0,t) = g(t)$ and $u(L,t) = 0$? Then, you split u into $u_1 + u_2$ where u_1 satisfies the homogeneous PDE with inhomogeneous BC and u_2 satisfies the inhomogeneous PDE with homogeneous BC, and obtain u_1 and u_2 separately.

*

COMMENTS : We have reached the point where the mathematical discussion of how to solve 1+1 heat equation (and at least in theory any other separable PDE) is more or less complete. This part was a little dry but very necessary, just as almost every pianist needs to practice his/her scales and arpeggios before going

on to actual musical pieces (the few pianists who can rattle off Mozarts and Beethovens with no prior practice fall way outside the normal system). Now we come to the interesting part where we will use this knowledge to model and solve realistic problems of heat flow.

5.4

MODELING WITH THE 1+1 HEAT EQUATION

Here we finally look at how the 1+1 heat equation is derived. Consider a rod of constant cross-sectional area A in the y-z plane, through which the flow of heat is one-dimensional only, being directed along the x-axis. Although the derivation can easily be generalized to accommodate two or three space dimensions, we ignore that case for now. We denote by q the **heat current** (by analogy with electromagnetism, we should probably call it the heat current density), which is the amount of energy flowing through unit cross-sectional area of the rod per unit time. In general it's a vector but in one space dimension it becomes a scalar. Its unit is J/m^2s which is W/m^2. The three relevant material properties of the rod are **density** ρ (kg/m^3), **specific heat capacity** C (J/kgK) and **thermal conductivity** κ (W/mK). The SI Unit of temperature is Kelvin, where 1 ^oK equals 1^oC but 0 ^oK equals -273 ^oC (and is the lowest possible temperature attainable in Nature). Hence, SI Units for C, κ etc might just as well be stated in terms of Centigrade degrees, and you can use this scale in a calculation with no worries about units and dimensions.

Now we consider an infinitesimal section of the rod between x and $x+\Delta x$. In a time Δt, suppose the change in temperature of this section is Δu. (Normally I would have used T for temperature as in 1.2.6 but in this case that causes confusion with the $T(t)$ separable function, so we'll continue with u. For other equations however, we'll use u for the abstract version of the equation and go over to the physically relevant symbol when modeling .) Then, during this time, the change in energy of this section is $(\rho A\Delta x)C\Delta u$. This change must equal the total energy which is brought in and carried away by the heat currents (recall here how the change in quantity of salt in a chemical tank was the difference between the amount entering via the inlet and exiting via the outlet, 1.2.3). Following usual sign convention, the current at the left end x brings in energy if it is positive and the current at the right end $x+\Delta x$ takes out energy if it is positive. Hence, in time Δt, the change in energy is $A[-q(x+\Delta x)+q(x)]\Delta t$. So we have

$$\rho A\Delta x C\Delta u = A\left[-q\left(x+\Delta x\right)+q(x)\right]\Delta t \quad \text{, or}$$

$$\rho C\frac{\Delta u}{\Delta t} = \frac{-q\left(x+\Delta x\right)+q(x)}{\Delta x} \quad .$$

Taking the obvious limit,

$$\rho C\frac{\partial u}{\partial t} = -\frac{\partial q}{\partial x} \quad . \tag{01}$$

The LHS looks like the eventual target but the RHS is still some way off.

Now, we invoke **FOURIER's law** of heat conduction (this very problem was in fact what set old JOSEPH off on the path to his series and transform), which states that **the heat current is proportional to the negative of the spatial gradient of the temperature**. The constant of proportionality is the thermal conductivity [as a quick check on the units, $q = -\kappa u'$ so $\mathrm{W/m^2} = \kappa.\mathrm{K/m}$ or κ must have the units $(\mathrm{W/m^2})/(\mathrm{m/K})$ which is W/mK]. Like OHM's law, FOURIER's law is a typical material property – most common materials show this proportionality relation for reasonably small temperature gradients etc. and it might be possible to synthesize materials which will violate the law completely. Violations of FOURIER's law however are much less common than OHM's, and, assuming the material to be well-behaved (typically the least restrictive assumption in any modeling problem involving the heat equation), we have $q = -\kappa(\partial u/\partial x)$ whereby (01) becomes

$$\rho C\frac{\partial u}{\partial t} = \kappa\frac{\partial^2 u}{\partial x^2} \quad \text{, or}$$

$$\frac{\kappa}{\rho C}\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t} \quad , \tag{02}$$

which last expression is of course the 1+1 heat equation if we make the identification $\alpha^2 = \kappa/\rho C$.

Basis this derivation we can now understand the various boundary conditions. A constant temperature BC is obvious and needs no further explanation. An **insulation BC** means that heat is not allowed to flow out of that end of the rod i.e. the heat current at the end is zero. Since $q = -\kappa u'$, this shows that u' must be zero. An **imperfect insulator** leaks heat to the atmosphere. Here, the current at the end is not zero but is proportional to the temperature difference between the end and the atmosphere, so we have something like $q = \gamma(u-u_0)$ or $\kappa u' = -\gamma(u-u_0)$ where u_0 is the atmospheric

temperature (check the signs : if the right end is leaking and rod is hotter than atmosphere then current goes to the right as it should). Taking $u_0 = 0$, we recover the structure of the BC of example 5.2.2.

It is impossible to determine γ in the above expression in terms of material properties. This is because heat flow from the rod to the atmosphere is not a fundamental physical process but just another thermal conduction problem which we can model rigorously if we work in three space dimensions instead of one. However, calculation and experiments show that we can often substitute this complex end problem with the simpler ad hoc (and physically plausible) BC and still get very good results for the temperature inside the rod. The effective value of γ for each situation will have to be found from experiments involving the apparatus.

In terms of this physical interpretation, we can also understand why a prescription of u and u' at the same spatial point constitutes a malafide BC, as I mentioned in 5.1. Evidently, if the homogeneous "BCs" are $u(0) = 0$ and $u'(0) = 0$ then the only solution $X(x)$ which satisfies them is the trivial one. This happens because you cannot prescribe both the temperature and the current at a given point. Maintaining one end at $u = 0$ means that you are pumping in or sucking out whatever heat is required to ensure that the end stays at that temperature – the current through the end becomes a self-adjusting variable, dependent on the temperature inside the rod. On top of the temperature constraint if you impose that the current also be fixed, then that becomes over-specification and you reach a contradiction.

Now suppose we have a **source** of heat inside the rod, as in a heating coil. In the most general case, this source can have a different strength in every infinitesimal volume element of the body, hence its dimension is power per unit volume i.e. W/m^3. Here of course, the source strength is constant in all directions perpendicular to the length of the rod. Let this strength be $S(x,t)$. Then, plugging this term into our starting equation for the 1-dimensional rod, we have

$$\rho A \Delta x C \Delta u = A\left[-q\left(x+\Delta x\right)+q(x)\right]\Delta t + A\Delta x S\left(x,t\right)\Delta t \quad .$$

Taking the continuous limit,

$$\rho C \frac{\partial u}{\partial t} + \frac{\partial q}{\partial x} = S\left(x,t\right) \quad ,$$

and then from FOURIER's law,

$$\rho C \frac{\partial u}{\partial t} - \kappa \frac{\partial^2 u}{\partial x^2} = S(x,t) \quad .$$

Rearranging terms,

$$\frac{\partial u}{\partial t} - \frac{\kappa}{\rho C} \frac{\partial^2 u}{\partial x^2} = \frac{1}{\rho C} S(x,t) \quad ,$$

which is the inhomogeneous heat equation.

Although the final form of the equation is intuitively almost obvious, my purpose in doing the derivation in detail was to account for all the constants involved. The $f(x,t)$ RHS term is actually the **source power density divided by the product of material density and specific heat**. This factor is impossible to obtain unless we derive it, and in actual modeling problems, the correctness of your answer depends as much on your conceptual understanding and technical skill as on your precision in accounting for every numerical factor attaching to every single expression. A negative source is called a heat **sink** – it is a drain of heat.

Now suppose we have a rod which loses heat to the atmosphere ($u = u_0$) through the lateral surface. Then, we can model it reasonably well as a sink inside the rod, which takes heat away at a rate proportional to the difference between u and u_0. In such a case, the heat equation looks like

$$\frac{\partial u}{\partial t} - \frac{\kappa}{\rho C} \frac{\partial^2 u}{\partial x^2} = -\beta (u - u_0) \quad .$$

The minus sign on the RHS accounts for the fact that if the rod is hotter than the atmosphere, it loses heat. Like γ in the leakage BC, β is a proportionality constant whose value needs to be determined from the properties of the setup at hand – we cannot find it from theory as a rod with leaking sides is not a 1+1 problem but a 3+1 problem ! This lateral loss term is thus a stand-in for a more accurate (and hundred times more difficult) model and calculation.

Finally, note that sometimes the density and specific heat capacity of materials are specified relative to that of water. The values for water are 1000 kg/m^3 and 4180 J/kgK respectively. Relative density (or "specific gravity" – a term I'll avoid here) of a material means the ratio of the density of the material to that of water.

The derivation of the **diffusion equation** (spread of chemicals through a solution) is almost identical, with chemical concentration taking the place of temperature, and chemical quantity substituting for energy. FOURIER's law is replaced by **FICK's**, which says the same thing – diffusion current is proportional to negative of concentration gradient. The constant this time is called the diffusivity. A rod whose ends are maintained at constant concentration is like the one with BCs on u. A rod whose ends are sealed prevents chemical current from flowing across, as the insulated end does for heat. Likewise, ends can leak chemical, there can be sources of chemical (as in a reaction chamber) and lateral loss of chemical from the tube. Here we'll do all applications based on heat transfer rather than chemical diffusion because those are more readily apparent in everyday life. But since the maths is the same, you can easily extend what you learn here to diffusion problems also.

Problems on heat flow are often categorized into two kinds – **steady state** problems and **transient** problems. The first class of problems deals with solutions of the heat equation which are independent of time while the second deals with the complete spatio-temporal behaviour. The former class is significantly easier than the latter – steady state means $\partial u/\partial t = 0$ so in 1+1 case it reduces the PDE to an ODE, and a trivial one at that. In higher dimensions, steady state problems are also PDEs but again they are considerably simpler than with the dynamics added on. Here we are going to see both kinds of problems.

This brief discussion gives you the theoretical foundation necessary to set up and solve problems on heat conduction – now is the task of actually building the house on top of the foundation.

Example 5.4.1

Cooking a steak

A steak (slab) of pork/ beef/ fish/ paneer, initially at room temperature 30 ^{o}C, is being cooked on a griddle (tawa). The steak has thickness 0.05 m (5 cm), constant cross-sectional area 0.01 m^2 (100 cm^2), thermal conductivity 1 W/mK, specific heat capacity 1000 J/kgK and density 1000 kg/m^3. The griddle is maintained at constant temperature 200 ^{o}C. It is observed that if the steak is cooked indefinitely without being flipped, the eventual temperature attained by the top surface is 95 ^{o}C.

(a) Write a PDE problem for the temperature of the steak as a function of space and time. Motivate carefully the boundary conditions.

(b) Decompose the solution into homogeneous and particular components and write the BCs and IC satisfied by each of them.

A food is certified safe to eat if its temperature becomes at least 75 ^{o}C throughout its bulk.

(c) If the steak is not flipped on the griddle, determine the time taken for it to become certifiably safe to eat.

(d) If the steak is flipped and cooked for a long time on both sides, determine the maximum temperature u^* such that the entirety of the steak has attained a temperature greater than or equal to u^* at some stage of the cooking process. Mathematically, the quantity asked for is $\min_x \left\{ \max_t u(x,t) \right\}$.

[Level : 4]

I did say in the last Chapter that in the PDE part of this book, we'd be smashing the ball all around the park. Having had enough of ones and twos, it's time to open up the heavy hitting. This is a transient problem since we have to calculate the time taken for the food to cook – a steady state analysis will give only the asymptotic temperature distribution inside the steak and nothing else.

(a) No different from ODE problems, systematic approach is the key. First let's write down what is obvious. We have to model the temperature dynamics using 1+1 heat equation with the relevant space dimension being the vertical (i.e. normal to the griddle surface). Let $x = 0$ be the griddle, so that $x = L$ $(=0.05)$ becomes the top of the steak. Evidently, the steak cannot be having heat sources or sinks inside it, so the governing equation for $u(x,t)$ is

$$\frac{\kappa}{\rho C}\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t} \quad \text{, or}$$

$$10^{-6}\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t} \quad . \qquad (01)^*$$

(This being my first time working a heat equation with realistic numbers rather than toys, I was very surprised by the extremely low value of α^2 and went back to check my derivation – nothing wrong here.) Since atmospheric temperature is 30 °C, we let $u = 0$ denote 30 °C – so long as the size of the degree remains that of the Centigrade degree, there's no change in material parameter values.

Now for the BCs and IC. The bottom of the steak is in direct contact with the griddle, which is 200 °C. So it is very reasonable to impose that the bottom be maintained at this temperature throughout i.e. $u(0,t) = 170$. The BC at the top is more complicated. Although contact with atmosphere might suggest $u(0.05,t) = 0$, that is clearly wrong. The top surface of a real steak is hardly at room temperature and we are also told in the problem that the asymptotic temperature there is 95 °C. (The bottom acquires the griddle's temperature because the griddle metal is an excellent conductor but the top doesn't acquire the air's temperature because air is a lousy conductor.) Insulated end is a better candidate, but again it contradicts the problem information – if we insulated one end of a rod and maintained the other at 200 °C, the whole rod would eventually become 200 °C. So what's left is imperfect insulation i.e. leaky BC. This is also physically the most plausible – if we keep our hand a few centimetres above the steak, we can clearly feel the heat that it radiates i.e. leaks to the atmosphere

(if the steak had been an insulator, this heat would have been zero). Hence we go for a BC of the form $u'+\gamma u = 0$ at the top end $x = 0.05$.

To find the value of γ, we use the asymptotic information. We know that a steady state solution of 1+1 heat equation always has the form $A+Bx$ – in the present case, $u(0,t) = 170$ and $u(0.05,t) = 65$, so $u_{SS} = -2100x + 170$. Hence u' at the end (and everywhere else) is -2100; equating this to $-\gamma u$ gives $\gamma = 32.3$. Finally, the initial temperature of the steak is given to be room temperature, so $u(x,0) = 0$. Hence the complete PDE problem is the 1+1 heat equation (01) together with the BCs $\underline{u(0,t) = 170}$, $\underline{u'(0.05,t) + 32.3\,u(0.05,t) = 0}$ and the IC $\underline{u(x,0) = 0}$. Note that the cross-sectional area of the steak didn't enter the picture. With modeling problems, it can happen that the question gives extra information – in such a case don't wrack your brains trying to fit it in !

(b) Although the equation itself is homogeneous, the problem is inhomogeneous due to the presence of the 170 in the BC at $x = 0$. So we partion the solution $u(x,t)$ into a homogeneous solution u_h and a particular solution u_p. Since the BC at the top is already homogeneous ($u'+\gamma u$ equals ZERO), both parts satisfy this BC. The inhomogeneity is at the griddle, so we let u_p satisfy this BC while u_h satisfies the griddle BC $u_h(0,t) = 0$. The particular solution is of course the steady state solution $-2100x+170$ we already found above; it satisfies both the BCs by definition, and doesn't care for the IC. Finally, we fit the IC to the homogeneous solution, using $u_h(x,0) = u(x,0) - u_p(x,0)$. This gives $u_h(x,0) = 2100x - 170$, and PDE problems for both parts are complete.

(c) We have already solved homogeneous problem in 5.2.2, so we now carry over the relevant steps. The eigenvalues are the roots of $k = -\gamma \tan kL$ which is $k = -32.3 \tan 0.05k$ in this case. Since this equation is transcendental, the computer enters the proceedings here. By plotting the graphs of the both sides, we find the eigenvalues – here I have gone with the first ten of them, thus planning to include the first ten terms of the eventual eigenfunction expansion. The first ten eigenvalues are $\beta_1 = 44.6$, $\beta_2 = 100.6$, $\beta_3 = 161$, $\beta_4 = 223$, $\beta_5 = 285$, $\beta_6 = 348$, $\beta_7 = 411$, $\beta_8 = 473$, $\beta_9 = 536$ and $\beta_{10} = 598$. The corresponding spatial eigenfunctions are $\sin \beta_n x$; by the result of Part (b) of 5.2.2, they are orthogonal, and the normalization constant for the n^{th} eigenfunction is

$$v_n = \frac{1}{\left(\frac{L}{2} - \frac{1}{4\beta_n}\sin 2\beta_n L\right)^{1/2}} \quad .$$

The eigenfunction expansion for $u_h(x,t)$ is

$$u_h(x,t) = \sum_1^{10} D_n \sin \beta_n x \exp\left(-10^{-6}\beta_n^2 t\right) \quad .$$

To find the D_n's we set $t = 0$ and exploit the orthonormality of the eigenfunctions – each coefficient must be the dot product of the IC and the corresponding eigenfunction. Since the IC on the homogeneous problem is $2100x - 170$, we have

$$D_n = v_n^2 \int_0^{0.05} (2100x - 170)\sin \beta_n x \, \mathrm{d}x \quad .$$

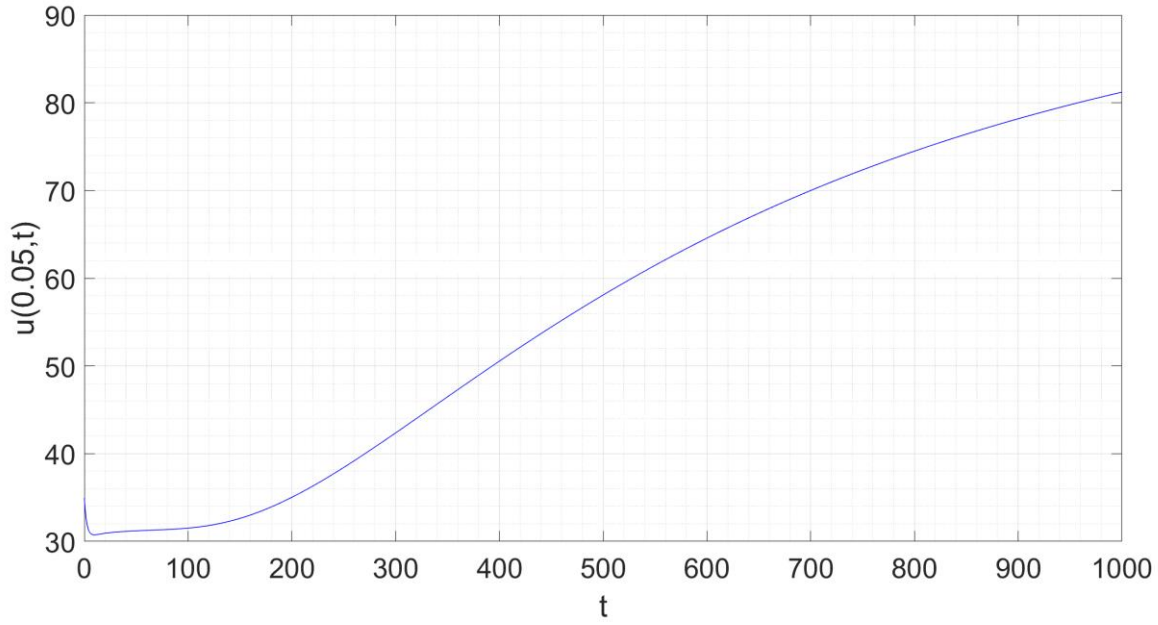
The computer being already on board, we use it to do the integration for us. The integration procedure is the simplest of all – the rectangle method from 1.1.4. (Just as a check, I used the computer to also verify that the eigenfunctions are indeed orthonormal – although 5.2.2 showed it clearly, the result is counter-intuitive enough to warrant an extra check. They are.)

The last three formulae together give us $u_h(x,t)$, and adding the particular solution turns it into $u(x,t)$. The question however wants the time taken for the temperature in the entire steak to become 75 °C or higher, and that's a different ball game. We could of course get the computer to plot $u(x,t)$ at a lot of t's, find the minimum u in each plot and see when that minimum crosses the threshold. There is however a slick and accurate way of doing this exactly, which in fact is what bumps this problem up to a Level 4.

Suppose that at any time t, $u(x,t)$ has a local minima at some point inside the steak. Then, by definition of minima, $\partial u/\partial x = 0$ at that point, negative before that point and positive after. But κu' is the negative of the heat current. A local minima inside the steak would mean that the current is zero there and flows TOWARDS the point from BOTH sides. For an initially cool steak being heated from below, this is an inadmissible current configuration. In reality, the current flows in from the griddle, passes through the steak and leaks out to the atmosphere. Throughout the bulk of the steak, the current always goes from

bottom to top, so the temperature of the steak decreases monotonically from bottom to top. Hence, it is sufficient to find the time at which the TOP of the steak i.e. the point $x = 0.05$ attains the temperature of 75 °C.

This done, now things are easy. Using the formula for $u(x,t)$ we generate a plot of $u(0.05,t)$. We also add the constant quantity 30 so that the plot really shows temperature in Centigrade degrees. This plot is shown below.



There is a little wiggle right at $t = 0$ because of the finite number of terms in the eigenfunction expansion and GIBBS' phenomenon. Ignoring this, we can see that the time required for the top of the steak to attain 75 °C is about 14 minutes.

(d) If the steak is flipped and cooked for long time on both sides, then it first attains the steady state distribution from the bottom to the top and then after flipping, again attains the steady state distribution from the new bottom to the new top. Letting end 1 be the original bottom and end 2 the new bottom, the first configuration is a straight line from 200 °C at end 1 to 95 °C at end 2 and the second configuration is a straight line from 200 °C at end 2 to 95 °C at end 1. The maximum temperature attained by any internal point is the higher of the two values corresponding to these two steady states. The minimum of the max occurs at the midpoint of the steak and is (95+200)/2 which equals 148 °C.

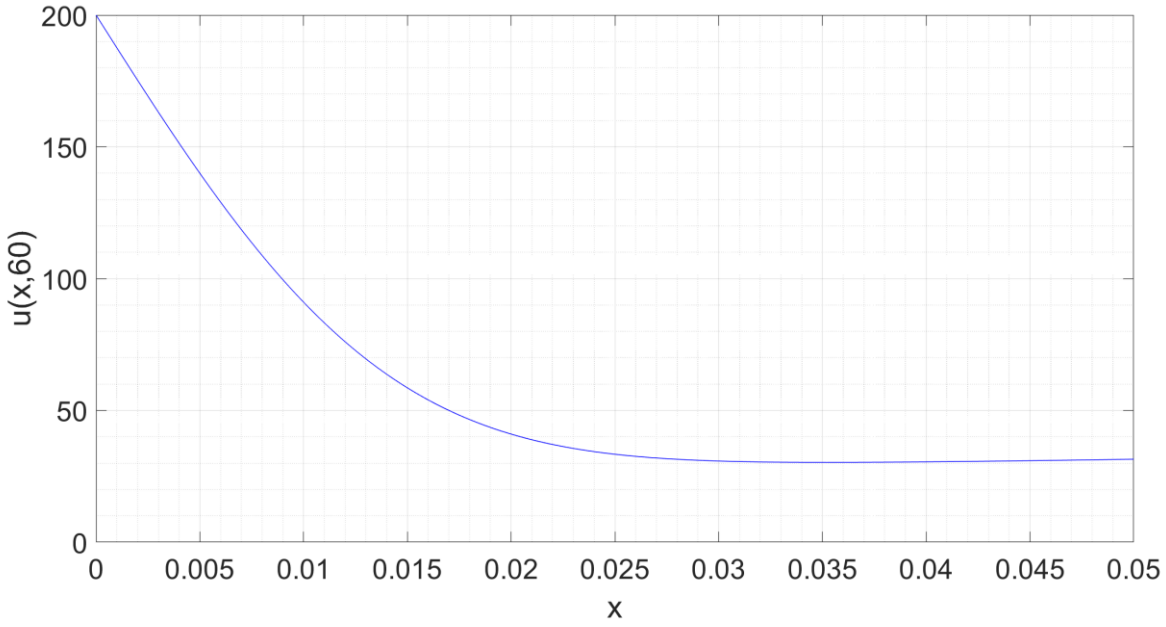
*

COMMENTS : The answers of Part (c) and (d) are implausible – 14 minutes is too long and 150 °C everywhere is too hot. Just to clarify, there is nothing wrong with the material parameter values given in the question, which are taken from relevant tables and averaged over the various choices of food mentioned (and simplified a bit to produce a neat value of α^2). So we seem to have reached a worrisome situation where a decent model yields implausible answers. Now what to do ?

The problem which has occurred is that, although we have a good model, we haven't asked the right questions so far. In Part (c), instead of thinking about the top of the steak, let's focus on the CENTRE. Getting Matlab to plot $u(0.025,t)$ we can see that it takes just four minutes to attain the safety temperature. At this point if you FLIP the steak, it will take less than four more minutes to attain the safety temperature from the other side because of the initial heat inside. The whole thing will get done (at least certified safe) in seven or so minutes. It took fourteen minutes for the whole thing to become safe without flipping, but half of that with flipping – this shows how important flipping is. As an aside, in the time trace, we can again see that the temperature of the top is increasing monotonically – no ups and downs on the way to the steady state value of 95 °C.

Part (d) shows that the result of cooking for long time on both sides is to cook all parts to a more or less uniform temperature (there's only 50 degrees difference between the hottest and coldest parts). With the griddle at 200 °C (raging hot), the steak becomes a brickbat but with the griddle at 100 °C for example, the minimum would be 80 °C and would produce a really uniform cooking.

To generate the characteristic crust on the other hand, we need a high temperature for a short time. As an example, here is the plot of $u(x,t)$ of Part (c) as a function of x at a time value of one minute.



We see that although the bulk of the steak is stone cold, the first one-tenth of the thickness is at or above 150 °C. This is a BOUNDARY LAYER, like the one we saw in 2.1.6 Part (d). At a very small time, the bulk of the steak is cool because the heat has not reached it, but the end must be hot because it's in contact with the hot griddle. So the temperature has no choice but to jump up sharply from cool in the bulk to very hot at the bottom. The practical manifestation of the boundary layer is the crust on the steak (next time you go to a steakhouse, refer to the crust as "boundary layer" and impress your friends). At this point if we flip the steak, we will generate a crust on the other side while keeping the centre cool.

Because of the complementary properties of fast-and-hot and slow-and-warm cooking, many steakhouses use a combination of the two to achieve the best results. They either grill it first on medium and then sear off the ends full throttle or do it in the reverse order. They take the extra trouble especially if the steak is thick, which it is in our case (5 cm is close to the thickest available). At least the 1+1 heat equation can explain their action quantitatively, although cooking a steak is far easier than calculating its temperature !

Example 5.4.2

Oven coil

An oven coil having length 0.5 m and circular cross-section of radius 0.01 m (1 cm) develops a constant power density inside it when the current flows through it. At full throttle, the total power consumed by it is 1 kW. The material properties are $\kappa = 11.3$, $\rho = 8400$ and $C = 450$ SI Units. The coil is laterally insulated, its two ends are maintained at room temperature which is 30 ^{o}C, and the coil itself is also at room temperature before the current is turned ON.

(a) Write a PDE problem for the temperature of the coil as a function of space and time.

(b) Discuss the effect of the initial condition on the eventual temperature profile.

(c) Determine the maximum temperature inside the coil at large times.

(d) Is the above answer physically plausible ? Discuss with reference to the assumptions made in the modeling process.

[Level : 2]

(a) Since room temperature is 30 °C, we let it correspond to $u = 0$. The oven coil looks like ideal candidate for 1+1 heat equation with source term. Indeed, the currents flowing through the coil generate heat, which acts as source in addition to the usual heat diffusion mechanism. The lateral insulation prevents leakage from the sides so there is no sink or leakage term. Hence we write the equation

$$\rho C \frac{\partial u}{\partial t} - \kappa \frac{\partial^2 u}{\partial x^2} = S(x,t) \quad . \tag{01}$$

S is evidently a constant function; its value is the total power output divided by the volume of the coil. Plugging in the numbers we get

$$* \qquad 3780000 \frac{\partial u}{\partial t} - 11.3 \frac{\partial^2 u}{\partial x^2} = 25500000 \quad , \qquad *$$

which is the PDE asked for. The BCs are $u(0,t) = u(0.5,t) = 0$ and IC is $u(x,0)=0$.

(b) We split the solution into homogeneous and particular components. The particular solution will be a constant function of time because the source is constant in time. The homogeneous solution will absorb the IC. Since this solution satisfies $\alpha^2 u_h'' = \dot{u}_h$ with BCs $u(0,t) = u(L,t) = 0$, we know that all terms in its eigenfunction expansion have decaying exponentials. Hence the homogeneous solution dies out in time and the IC has no effect on the eventual temperature distribution inside the coil. This reduces the problem to a steady state calculation, which is the why the level is considerably lower than the last Tute.

(c) To find the maximum temperature, we must first find the temperature profile. Since the numbers in the problem are extremely cumbersome, we work in symbolic form and plug in values only at the end (this is a very useful strategy which you should keep in mind). Trying the steady state solution $u_{ss}(x,t) = g(x)$ and plugging into (01) we have $-\kappa g'' = S$. The general solution is $(-S/2\kappa)x^2 + Ax + B$; $g(0) = 0$ implies $B = 0$ and $g(L) = 0$ implies $A = SL/2\kappa$. Hence the specific solution is

$$u_{SS}(x,t) = -\frac{S}{2\kappa} x^2 + \frac{SL}{2\kappa} x \quad ,$$

which is the steady state temperature profile inside the coil.

Finding the maxima of this function is a trivial task, but just to give you some insight on heat flow, I will do it another way. A local maxima of temperature means a point where $\partial u/\partial x = 0$. To the left of this point, u' is positive while to its right, u' is negative. Since u' is the negative of heat current, this means that current flows AWAY from the maximal point in both directions. Now imagine the heat flow pattern inside the coil. Heat is being generated everywhere, and is being sucked away through the two ends which are at temperature zero. So near the left end, the heat current is flowing towards the left, and near the right end it is flowing towards the right. From the symmetry of the setup, the only point which can be the separator between the two is the midpoint of the coil. Hence the zero current, and the maxima of temperature, must be occurring at $x = L/2$.

Substituting this into the expression for u, we can see that the maximal u is $SL^2/4\kappa$. Finally, plugging the numbers back in, the maximal value works out to 141000 °C.

(d) The above answer is physically implausible. In fact, that's a euphemism – the answer is absolute shit. It is off by TWO ORDERS of magnitude. A typical oven coil glows red hot and has a temperature of 500-1000 °C. One lakh Centigrade would probably cause the coil metal to dissociate into electrons and protons.

Evidently, the coil model is garbage – now our task is to find the fallacy. The problem is the assumption of lateral insulation. In reality, the coils radiate copious heat to the air inside the oven. In fact, that is the primary purpose of the coil – to provide heat to the air which in turn can heat the food. The air leaks heat constantly to the room through the oven walls, and it takes this heat back from the coils. This draining of heat to the air keeps the coils at practical temperatures. By insulating the coils, not only did we take away the heat removal mechanism but we also ensured that, even if the coil could somehow withstand the six-figure temperature, the oven itself would remain at room temperature. This is a suitably grievous modeling error, and can account for the gigantic disparity between model and reality. To correct the model, the first thing we need to do is insert a lateral leakage term.

*

COMMENTS : This is a toy version of something which can very well happen when you are modeling a hitherto unknown phenomenon. The most important thing is to be able to recognize implausible answers when you get them, and use them as starting steps in rectification of the model. As for the leakage term, I could have added a few more parts to this question which take that effect into account. But I haven't bothered because we are going to see that term in action in the very next Tute. Finally, the material properties of this problem correspond to nichrome, which is the standard material for making heating elements.

Example 5.4.3

Film boiling of glycerine

Easier form : Solve the standalone equation $\dot{u} - \alpha^2 u'' = -ku$ *in a finite interval with BCs and IC. Such an equation arises in Part (b) here.*

In a laboratory experiment, a slender metallic rod is immersed in a beaker of hot glycerine and a current is passed through it. The bulk of the glycerine is at a temperature of 300 °C which is taken as the datum $u = 0$. The rod has length 10 cm, diameter 2.5 mm, and material properties $\kappa = 11.3$, $\rho = 8400$ and $C = 450$ SI Units. The equation for temperature inside the rod is

$$\rho CA\frac{\partial u}{\partial t} - \kappa A\frac{\partial^2 u}{\partial x^2} = P - f(u) \quad , \qquad (*)$$

where A is the cross-sectional area, P is the power generated (input power) and $f(u)$ is the power dissipated per unit length of the rod. It is found that $f(u)$ has the form $f(u) = \alpha u - \beta u^2 + \gamma u^4$ where $\alpha = 4.63$, $\beta = 0.00734$ and $\gamma = 4.88\text{x}10^{-9}$ SI Units.

Assume for now that the rod has infinite length.

(a) Describe the nature of steady state solutions $u^*(x,t)$ of the governing equation (*).

Since (*) is nonlinear, the system goes into a steady state solution if and only if it is stable. To obtain the stability, one considers a small perturbation $\Delta u(x,t)$ added onto the steady state, substitutes it into (*) and linearizes. One then performs eigenfunction expansion on the linear equation and draws one's conclusions basis the behaviour of the temporal parts.

(b) Determine the condition for stability of a steady state solution of (*).

(c) In a particular run, the input power P is raised from 750 to 850 W/m in steps of 10, and the system allowed to settle into steady

state after each increment. Describe the response qualitatively and quantitatively, and highlight its most important feature.

Now factor in the finite length of the rod. When an input power of $P=1420$ is applied [with $f(900)=1420$], it is found that the two ends of the rod are actually at temperature 850.

(d) Linearizing (*) about $u(x,t) = 900$, determine the steady state temperature profile and plot its graph. You might wish to choose your origin at the centre of the rod for this calculation.

(e) Basis the above result, discuss the validity of the infinite length assumption made in Parts (a-c).

[Level : 4]

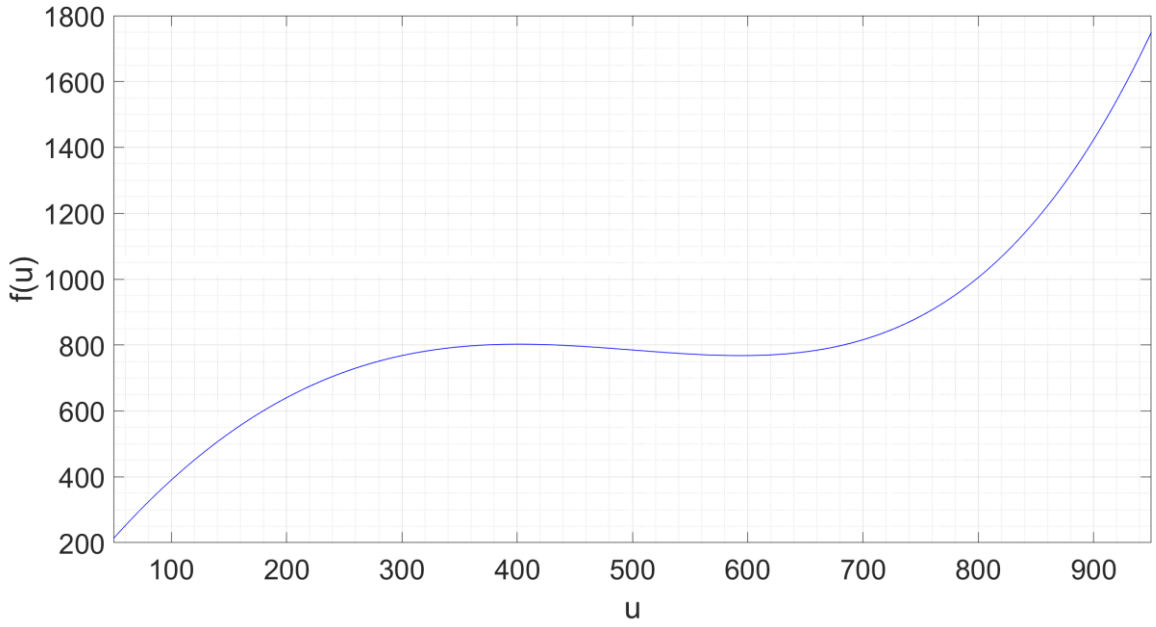
Since this is a problem on stability analysis, Sections 1.4 and 2.4 should be a useful review.

(a) We start by writing the governing equation. We have

$$\rho CA\frac{\partial u}{\partial t}-\kappa A\frac{\partial^2 u}{\partial x^2}=P-f(u) \qquad (01)$$
$$=P-\alpha u+\beta u^2-\gamma u^4$$

For steady state solutions, we must of course have u^* independent of time. What the infinite length tells us is that u^* will be independent of x also. For if the rod is endless then every part of it is identical to every other, so there is no physical reason to distinguish between any two points on the rod. Hence the steady state solution must be having the form $\underline{u^*(x,t)=\text{const}}$. Plugging this into (01), the LHS vanishes identically hence so must the RHS, giving $P=f(u^*)$. Thus, the steady state temperature depends on the input power P.

We present below a plot of $f(u)$ vs. u.



We can see that for $u<250$ and $u>700$ (approximately) corresponding to $P<750$ and $P>850$, the function is single valued while in the intermediate range it is multivalued. Hence for sufficiently low and sufficiently high power we have a unique steady state while in the intermediate range we can have more than one steady state. This is very likely why Part (c) of this question highlights this power range – let's see how the calculation pans out.

(b) We plug $u = u^* + \Delta u(x,t)$ into (01), and linearize. We have

$$\rho CA \frac{\partial}{\partial t}(u^* + \Delta u) - \kappa A \frac{\partial^2}{\partial x^2}(u^* + \Delta u) = P - f(u^* + \Delta u) \quad . \tag{02}$$

As in L04 etc, on the LHS, u^* is a constant so its derivatives vanish. On the RHS, we Taylor expand $f(u^* + \Delta u)$ getting

$$f(u^* + \Delta u) = f(u^*) + \left.\frac{\partial f}{\partial u}\right|_{u^*} \Delta u + \ldots\ldots \quad .$$

Since $P = f(u^*)$, the first term above cancels P on the RHS of (02) and we are left with

$$\rho CA \frac{\partial}{\partial t} \Delta u - \kappa A \frac{\partial^2}{\partial x^2} = -f_u(u^*) \Delta u \quad ,$$

where f_u denotes $\partial f / \partial u$. Denoting $f_u(u^*)$ by F, Δu by U, and setting the leading coefficient to unity, we can write this as

$$\frac{\partial U}{\partial t} - Q \frac{\partial^2 U}{\partial x^2} = -RFU \quad , \tag{03}$$

where Q and R are positive numbers. This is the linearized stability equation.

This is a new equation so we solve by separating variables. Trying $U = X(x)T(t)$ we have

$$X\dot{T} - QX''T = -RFXT \quad .$$

Dividing by XT and invoking the sep var argument,

$$\frac{\dot{T}}{T} + RF = Q\frac{X''}{X} = \text{const.} \quad \text{, or}$$

$$\frac{\dot{T}}{QT} + \frac{RF}{Q} = \frac{X''}{X} = \lambda \quad . \tag{04}$$

This is where things start to differ from the ordinary 1+1 heat equation case. There, we rejected solutions which grew exponentially in time. We could do that because 1+1 heat is a model of a real physical process, and that can't show an exponentially growing solution. Equation (03) however is the result of linearization of a nonlinear equation about a steady state. Its solution CAN grow

in time if the steady state is unstable – when the perturbation becomes too big, the linearization is no longer valid. Hence this plausibility condition falls through.

The new condition is that the solution must be plausible in SPACE i.e. it cannot be going to infinity at large x. Such a solution would be huge even at small time, would represent something completely unphysical and would also contradict the definition of a perturbation (we assumed that Δu was SMALL compared to u^*, not infinitely large compared to u^*). Hence, we must start working the separable solutions using this as our constraint.

As usual, partion into cases $\lambda>0$, $\lambda<0$ and $\lambda=0$. In the first case the solutions are $e^{\pm kx}$ where $k=\sqrt{\lambda}$; at $x\rightarrow+\infty$ one of them is trouble while at $x\rightarrow-\infty$ the other one is trouble. Hence this case shows unphysical solutions only. In the second case, the solutions are $\cos kx$ and $\sin kx$, which are themselves tractable and can be used to make up even more tractable functions via the Fourier transform (T085). In the third case, the solutions are $A+Bx$; the latter is intractable but a constant is just about acceptable (we WOULD probably like a perturbation to go to zero at large x, but let it be – it won't matter anyway). Hence λ can be negative or zero.

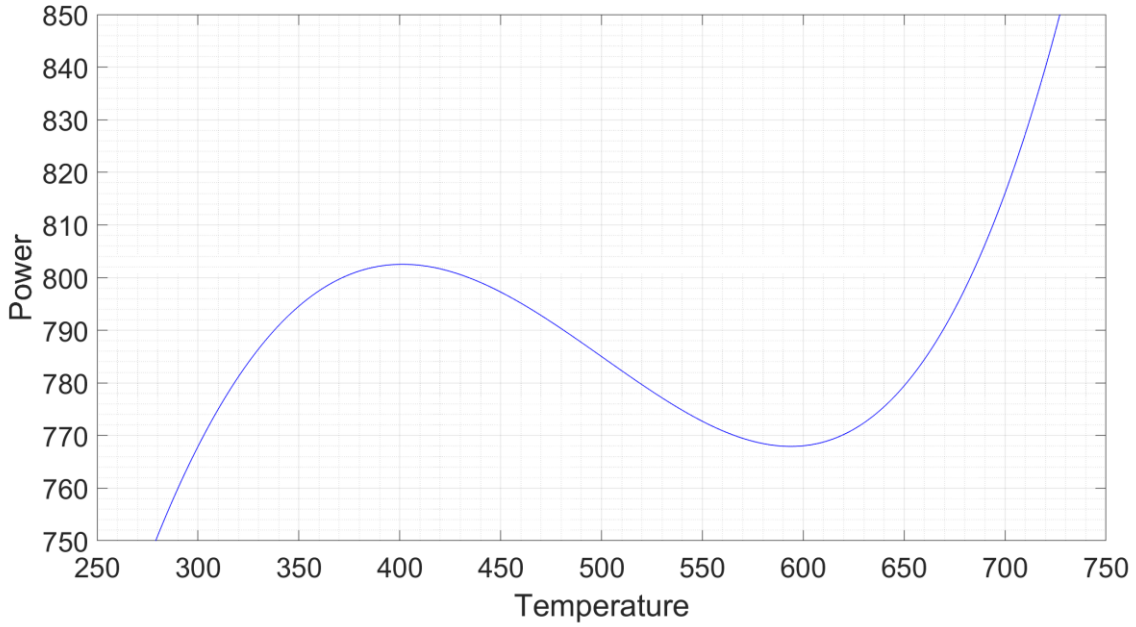
Now considering (04), we have

$$\frac{\dot{T}}{T}=Q\lambda-RF \quad ,$$

$$\dot{T}=(Q\lambda-RF)T \quad .$$

Evidently we get decaying solutions if $Q\lambda-RF$ is negative. For stability of the steady state we want decaying T's for EVERY PERMISSIBLE value of λ. This would imply that, if a random perturbation were to be tacked on, its every single constituent eigenfunction would die down in time. Conversely, even if we have growing solutions for ANY ONE λ, then the corresponding eigenfunction present in the perturbation would grow unboundedly and destabilize the system. (This condition is of course similar to that of stability of the multivariate systems in N°. 08 – there too, we cannot have a single eigenvector attached to a positive eigenvalue. Since a PDE is a limit of such a system, the similarity is natural.) Since Q and R are positive and λ is zero or negative, we are assured of all decaying solutions if F is positive. Hence, for the steady state to be stable, $f(u)$ must have a positive slope at the steady state value u^*.

(c) We must now use the above conclusion to predict the result of the experiment. First we plot u vs P [which equals $f(u)$] in the relevant region.



At $P = 750$, $u^* = 275$ while at $P = 760$, $u^* = 290$. No doubt about these two since $f(u)$ is single valued there; since the graph slopes upwards, these steady states are obviously stable. The next increment enters the multivalued region – $P = 770$ can mean $u^* = 300$, $u^* = 570$ and $u^* = 630$. Among these, the points $u^* = 300$ and $u^* = 630$ are stable while 570 is unstable. But, u was at 290 before so we expect that it will amble over to the stable steady state at 300 instead of crossing the unstable 570 (it will get repelled away the moment it comes close to 570) and going all the way to 630. Repeating the argument for $P = 780$ to $P = 800$, we can see that the system slowly crawls up the left hump as the throttle is advanced, reaching $u^* = 370$ at $P = 800$.

Drama occurs now. $P = 810$ flies straight over the hump and hits the curve only at $u^* = 690$. Since this steady state is stable, the system indeed goes over to it. Thereafter, as P is raised further, the temperature continues to rise along the right hill, reaching $u^* = 730$ at $P = 850$. The most important feature is of course the sudden jump in temperature by more than 300 °C for a mere 10 W/m increase in input power.

(d) We now have to linearize the equation about the constant value $u = 900$, even though this time it is not a steady state solution (evidently, since that's our

TARGET). The motivation behind this linearization is that the temperature in the entire rod is sufficiently close to 900 for the linearization to be valid. The OUTCOME of the linearization however is the same (03); this time, since we are going to need the numbers, I'll write it with all parameters in place. Letting U now denote the difference from 900, we have

$$\rho CA\frac{\partial U}{\partial t}-\kappa A\frac{\partial^2 U}{\partial x^2}=-f_u(900)U \quad .$$

Since we want steady state solution, the $\partial/\partial t$ automatically evaluates to zero and U becomes a function of x alone. Substituting numbers for the other things, we have

$$-0.000552\frac{\mathrm{d}^2U}{\mathrm{d}x^2}=-5.65U \quad .$$

The question suggests that we "might wish" to take the origin at the centre of the rod – what it means is we'd better take that origin at the centre else we'll never ever get a result. Since we're using SI Units, we convert the length of rod to metre, hence x belongs to $[-0.05,\ 0.05]$. The boundary conditions are given in the question – we have $U(-0.05)=U(0.05)=-50$ (remember that U denotes deviation from 900). Hence we finally have the BVP

$$U''-102400U=0 \quad ,$$

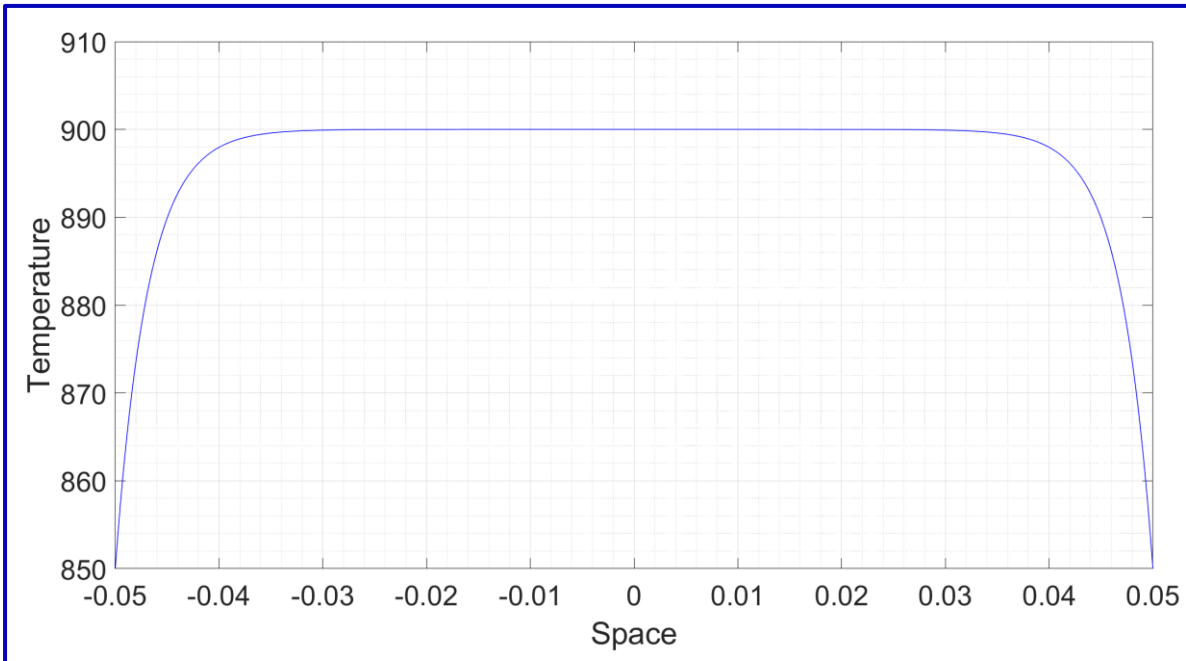
together with the BCs given above.

What a relief to encounter something from the first Part of the book (2.1.1 to be precise) – the general solution is $C_1\cosh 320x+C_2\sinh 320x$. Now look at the BCs – U at -0.05 equals U at $+0.05$ so they are mirror image BCs. The function cosh being symmetric satisfies them automatically, while sinh being antisymmetric can never fit them. Hence $C_2=0$. Fitting the BC gives $C_1=-50\ /\ \cosh 16$ which is $-1.1254\text{x}10^{-5}$, and we have the specific solution $U=-1.1254\text{x}10^{-5}\cosh 320x$. Finally, since $u=900+U$, the actual temperature profile is

$$* \qquad u_{SS}(x,t)=900-1.1254\times10^{-5}\cosh 320x \quad , \qquad *$$

which is the desired answer.

The plot of u_{SS} vs. x is shown below.



This is the required plot.

(e) The assumption of infinite length is in effect an assumption that the steady state temperature is CONSTANT along the rod. Now, we just plotted the steady state temperature given the information that it is actually lower at the ends by 50 degrees. What we can see is that, nevertheless, the temperature is very nearly constant in the bulk of the rod – the deviation is less than 2 degrees over the middle 80 percent of the length and is practically zero over the middle 60 percent. Indeed, all the variation occurs right at the ends themselves, and the bulk is almost completely unaffected by the end effects. Hence the assumption of infinite length and constant temperature is an excellent approximation in this case.

*

COMMENTS : Here we will understand where the governing equation came from. As mentioned in the question, the apparatus consists of a current carrying rod immersed in boiling glycerine. As the power is increased and the rod made hotter than the glycerine, initially the liquid boils in a regular manner (“nuclear boiling”) around the rod. The boiling glycerine takes away heat from the rod via conduction and convection, modelled by the $-\alpha u$ lateral leakage term (the term which we intentionally let go of in the oven coil of the last Tute). As the temperature continues to rise however, a film of glycerine vapour starts to

develop around the rod. This film reduces the ability of the glycerine to carry heat away from the rod, modelled by the $+\beta u^2$ term. (These terms are necessarily somewhat of the nature of "stand-in" terms, since 1+1 heat equation with leakage is itself a stand-in for a more complicated system. But they have clear physical motivation and produce reasonable results.) The rod also loses heat to the surroundings by radiation – as per the Stefan-Boltzmann law, this loss is proportional to the fourth power of temperature and is accounted for by the $-\gamma u^4$ term.

In the experiment, as the power is raised, the vapour film becomes more and more pronounced as the temperature rises. Then, at the jumping point, the film suddenly becomes completely developed, the rod transitions from conductive and convective heat loss to radiative heat loss and becomes white-hot. This is the film-boiling regime. Its onset is indeed extremely sudden and dramatic.

Suppose you now run Part (c) backwards, coming down from 850 to 750 – what would happen ? Evidently, the system would start from the top right corner and slide down the curve, entering the valley. At $P = 810$ it would be faced with three choices, but, as during the power increase, it would go to the nearest stable one, which is $u = 690$. This sliding down would continue until $P = 770$. Then it would have no other option but to jump back to $u = 290$. Thus, during power escalation, the forward transition from nuclear to film boiling occurs at $P = 810$ but during de-escalation the backward transition occurs at $P = 770$. This dependence of the transition temperature on the previous state of the system is called HYSTERESIS. It is indeed observed in the film boiling experiment. (Hysteresis can be seen in less exotic setups also – see STROGATZ for an example from first order ODEs.)

Finally, Part (d) shows a boundary layer, which we saw for the steak and is also there in 2.1.6. Indeed, the structure of the equation is the same as in 2.1.6 part (c) – in the equation for U, an extremely small U'' term balances an $O(1)\,U$ term. If we drop the derivative term, we get the trivial solution $U = 0$. This is indeed what prevails in the bulk of the rod – only at the two ends do we see a sharp change in U so as to account for the boundary conditions.

Example 5.4.4

Japanese paper pot

A Japanese paper pot consists of a paper bowl containing boiling soup which is heated from below using a flame. The pot is cylindrical, having a base diameter of 0.15 m, while the paper has properties thickness 0.0002 m (0.2 mm), $\kappa = 0.05$, $\rho = 1000$ and $C = 1500$ SI Units. The auto-ignition point of the paper (temperature at which it catches fire spontaneously) is 230 ^{o}C.

Assuming the heating to be uniform all across the base, determine the maximum power which can be input to the pot without the paper catching fire.

[Level : 2]

This is an easy problem once we get the model right (meaning : it comes without warning on an exam and there will be trouble). Since the problem is one of heat conduction (more to the point since the Chapter is on heat equation) we write the 1+1 heat equation for heat transfer from the flame through the base of the pot to the soup inside. The x direction is vertically upwards, normal to the base of the pot. The 1+1 model is valid since the heating is uniform all across the base – then there is no temperature variation across the base. Since there are no heat sources inside the paper, and we don't expect significant lateral leakage from the base, there will be no additional terms on the basic equation. So we have

$$\rho C \frac{\partial u}{\partial t} = \kappa \frac{\partial^2 u}{\partial x^2} \quad ,$$

as the governing equation.

Since temperatures generally increase or decrease monotonically, we expect that the after the heating starts, the initially cool pot will get hotter and hotter as time goes on, becoming hottest in the steady state. Hence we can throw away the $\partial/\partial t$ from the above equation, and focus on the steady state solutions (what a relief). These will of course get determined by the boundary conditions.

The top side of the base is in contact with boiling water, which is 100 °C. Since water is excellent at carrying heat away (although it's only an average conductor, it takes heat away by conVECtion – see 5.4.7), it is very reasonable to have the top of the base maintained at 100 °C, which we now define to be $u = 0$. The bottom of the base is exposed to the flame. It acquires some temperature greater than 100 °C so that the heat current flows from the bottom to the top. The higher the temperature of the bottom, the stronger the current, and hence the greater the power input across the base (remember, power equals heat current times surface area). The maximum temperature we can have at the bottom is the auto-ignition point of the paper, which is $u = 130$, so the maximum permissible power will correspond to this BC at the bottom.

Now we have a well-defined PDE problem (in fact ODE problem) with BCs. The steady state solutions are of course $u_{SS} = A + Bx$; $u = 130$ at $x = 0$ (the bottom surface) and $u = 0$ at $x = 0.0002$ (the top surface) means a thermal gradient of -650000 and a heat current $q = -\kappa u' = 32500$. Finally, the total input power is qA, which evaluates to 574 W. *

COMMENTS : Since the paper pot (called kami nabe in Japanese) is a safety critical equipment (a malfunction can turn pot and eaters into kami kaze), we would probably like to have a safety factor of at least 2, and not pump more than 200-250 W into the pot. This safety factor accounts for the paper being thicker than quoted, the auto-ignition point being lower than quoted, the heating being stronger at some points on the base than others etc etc etc. Even so, this is not an inconsiderable amount of power. Induction cookers are generally calibrated by power – a setting of 500 W can boil a cup of water in about four minutes or so. Double that time or a little more in a paper pot is perfectly reasonable.

The auto-ignition point of paper can apparently show a 50 ^{o}C variation in a range centred at the value I've used here. Indeed, this is the most commonly quoted value, perhaps on account of the novel Fahrenheit 451 by RAY BRADBURY, in which he describes brainwashing of society by burning books and other literature. The temperature in the title, which translates to 233 ^{o}C, is BRADBURY's version of the auto-ignition point of paper, but seems to have become a de facto standard. (Nowadays, mass brainwashing of society occurs automatically through computers and smartphones which, when connected to certain websites, are capable of supplying their users with tidbits and trivialities – and worse – at the rate of a dozen a minute. So 33 instead of 233 ^{o}C is sufficient to achieve BRADBURY's objective, a sure sign of technological progress.)

At first sight, a paper pot being held to flame and not catching fire appears utterly paradoxical. It is generally explained by saying that the boiling water keeps the paper at 100 ^{o}C. This explanation is almost perfect; the only catch is that paper is an INSULATOR (its thermal conductivity is three orders of magnitude lower than that of metals). Hence, 100 ^{o}C inside does not mean 100 ^{o}C outside, and it is not an obvious or foregone conclusion that the paper won't catch fire (if we put the pot on a modern day gas stove, it very likely will).

The thinner the paper, the lower the external temperature corresponding to a given power influx. The flip side of course is that a pot made of tissue won't be able to carry soup inside it without breaking. If you want an equation for the stresses on the base, that becomes a fourth order PDE. So to design a pot which will not only not catch fire but also withstand the weight of the soup, you must solve that on top of this. As often happens, the people who first invented these

paper pots (in the 16th century, before PDEs existed), got the design correct by hit and trial. Today also, chefs all over Japan make pots basis tradition and instinct – even though it's Japan we're dealing with, we still don't expect that chefs will be juggling partial differential equations. Which just goes to show….. something.

Example 5.4.5

Hot and cold

On a typical winter evening in the city of Ithaca, the temperature outside is below freezing and there is a blowing wind. At a busy Chinese restaurant, customers are entering and exiting continuously. Every time the door opens, the wind comes in from outside, while heaters located inside the restaurant try to maintain its temperature to a tolerable level. An attempt will be made in this problem to model and solve for the temperature inside the restaurant.

For modeling, the assumption is made that the cross-sectional area of the restaurant is constant and the length is considerably greater than the width and height, so that the 1+1 heat equation applies. From $t=0$ to $t=t_1$, the door is assumed to be open so that the wind flows in. Since this removes heat from the restaurant, it can be modelled as a heat sink of spatiotemporally constant strength S_1 (W/m^3). From $t=t_1$ to $t=T$, the door is closed. During this time, the heaters add heat to the restaurant so that their effect can be modelled as a heat source of constant strength S_2. The motion of the door is now assumed to be periodic with period T. Since the restaurant neither gains nor loses heat on the average, one must have $S_1(t_1)+S_2(T-t_1) = 0$. The boundary conditions are assumed that the interior end of the restaurant $x=L$ is maintained at a constant temperature, the reference value $u=0$, while the door end $x=0$ is insulated. This means that all heat exchange takes place between the restaurant and the source and/or sink – since the wind is represented as a sink, the assumption is reasonable.

The method of matched boundary conditions will be employed to obtain a periodic solution for $u(x,t)$, using the following steps.

(a) Considering the equation $\dot{u}-\alpha^2 u''=P$ where P is a constant, obtain the particular solution $u_p(x,t)$ with the BCs as given in the question.

(b) Let the temperature at $t = 0$ be expressible as $u(x,0) = \sum_n C_n X_n(x)$ where $X_n(x)$ are the system eigenfunctions and the constants C_n are unknown. Determine $u(x,t)$ for $0 \le t \le t_1$ in terms of the constants C_n. The Fourier expansion of the answer to Part (a) will be relevant here.

(c) Imposing the condition that $u(x,t)$ must be continuous at $t = t_1$, determine $u(x,t)$ for $t_1 \le t \le T$, again in terms of the constants C_n.

(d) Imposing periodicity of $u(x,t)$ with period T, determine the constants C_n.

(e) Taking the reference temperature $u = 0$ to be 15 °C and using the parameter values $L = 8$, $\kappa = 0.026$, $\rho = 1.3$, $C = 1000$, $S_1 = -225$, $S_2 = 75$, $t_1 = 15$ and $T = 60$ SI Units, plot the temperature of the midpoint of the restaurant as a function of time.

[Level : 4]

Although this is not a six, it's no pushover either – almost a page of reading followed by rapidfire questions. So we start with a quick summary of the problem. Basically, the restaurant alternates between two states – door open and door closed. In the first state, the wind absorbs heat from it, which is modelled as a sink. In the second state, the heater supplies heat to it, which is modelled as a source. The alternation of source and sink gives rise to a temporally periodic particular solution $u(x,t)$, which is our target. Since the sink and source are like the OFF and ON states of a chopper, it is no surprise that boundary matching is the relevant technique . 4.4, examples 1-3 should provide a helpful review (no harm in reading 4.4.4 either although that's harder than this so can't be a prereq).

(a) The equation is $\dot{u} - \alpha^2 u'' = P$ with the BCs $u'(0,t) = 0$ and $u(L,t) = 0$. Since the excitation function is constant in time, the particular solution should be independent of time also. Hence, we try $u_p(x,t) = g(x)$. Evidently, g satisfies the ODE $g'' = -P/\alpha^2$. Its solution is $g = -(P/2\alpha^2)x^2 + Ax + B$; the $x = 0$ BC implies $A = 0$ while the $x = L$ BC implies $B = PL^2/2\alpha^2$. Hence we have

$$u_p(x,t) = \frac{P}{2\alpha^2}\left(L^2 - x^2\right) \quad , \qquad (01)^*$$

which is the desired answer.

(b) The equation for the temperature in the time interval $[0,\ t_1]$ is

$$\rho C \frac{\partial u^{(1)}}{\partial t} - \kappa \frac{\partial^2 u^{(1)}}{\partial x^2} = S_1 \quad ,$$

where the superscript 1 highlights that this is the first phase of the dynamics. We write this in the equivalent form

$$\frac{\partial u^{(1)}}{\partial t} - \alpha^2 \frac{\partial^2 u^{(1)}}{\partial x^2} = P_1 \quad .$$

This has a particular or steady state solution of the form (01) with P_1 in place of P, plus some homogeneous solution, thus

$$u^{(1)}(x,t) = u_{SS}^{(1)}(x,t) + u_h^{(1)}(x,t) \quad \text{, where}$$

$$u_{SS}^{(1)}(x,t) = \frac{P_1}{2\alpha^2}\left(L^2 - x^2\right) \quad .$$

To find $u_h^{(1)}$, we need the IC on $u^{(1)}$ at $t = 0$.

What the question gives us are the coefficients C_n of the IC when expanded in the system eigenfunctions. 5.1.5 tells us that these are $X_n(x) = \cos([2n-1]\pi x/2L)$. Since they are complete, any arbitrary function which satisfies the BCs at $x = 0$ and $x = L$ can be expanded in their terms, one such function being $u^{(1)}(x,0)$. Thus we are given that

$$u^{(1)}(x,0) = \sum_n C_n \cos\frac{(2n-1)\pi x}{2L} \quad . \tag{02}$$

The C's are unknown at this step because that's just how matched boundary conditions works – you start off with unknown constants, work with them, and they become known only at the end when you apply the matching condition.

So, $u_h^{(1)}(x,0)$ becomes the RHS of (02) minus $u_{SS}^{(1)}$. To make further progress, we must have $u_h^{(1)}$ entirely in terms of the eigenfunctions. For this, we expand $u_{SS}^{(1)}$ in a series of these functions (which is why the question gives the hint about it) – since they are orthogonal with $v_n = \sqrt{(2/L)}$ for each one (again see 5.1.5), we have

$$u_{SS}^{(1)} = \sum_n E_n^{(1)} \cos\frac{(2n-1)\pi x}{2L} \quad \text{, where}$$

$$E_n^{(1)} = \frac{2}{L}\int_0^L \frac{P_1}{2\alpha^2}\left(-x^2 + L^2\right)\cos\frac{(2n-1)\pi x}{2L}\,\mathrm{d}x \quad .$$

Looking up the messy integral and plugging in the limits gives

$$E_n^{(1)} = (-1)^{n+1}\frac{16P_1L^2}{\alpha^2(2n-1)^3\pi^3} \quad .$$

At once we have

$$u_h^{(1)}(x,0) = \sum_n \left(C_n - E_n^{(1)}\right)\cos\frac{(2n-1)\pi x}{2L} \quad .$$

After getting the expansion of $u_h^{(1)}(x,0)$, it is a second's work to make it $u_h^{(1)}(x,t)$ – each spatial eigenfunction just adds on the exponential temporal evolution which we've already calculated in 5.1.5. So we have

$$u_h^{(1)}(x,t) = \sum_n \left(C_n - E_n^{(1)}\right)\cos\frac{(2n-1)\pi x}{2L}\exp\left(-\frac{\alpha^2(2n-1)^2\pi^2}{4L^2}t\right) \quad .$$

In the subsequent parts of the calculation, as the expressions get longer, the exponential with its elaborate argument is going to look horribly cumbersome. So we'll define the shorthand notation

$$f_n(t) = \exp\left(-\frac{\alpha^2 (2n-1)^2 \pi^2}{4L^2} t \right) \quad .$$

(I would have called it T_n except that T is the period of the excitation.) Finally, $u^{(1)}(x,t)$ is $u_h{}^{(1)}+u_{SS}{}^{(1)}$. We express this latter term in its expanded form rather than the algebraic form and get

$$u^{(1)}(x,t) = \sum_n \cos\frac{(2n-1)\pi x}{2L}\left[E_n^{(1)} + \left(C_n - E_n^{(1)} \right) f_n(t) \right] \quad , \qquad (03)^*$$

which is the desired answer.

(c) We now enter the second phase of the dynamics, so the superscript on u becomes 2. The governing equation is

$$\frac{\partial u^{(2)}}{\partial t} - \alpha^2 \frac{\partial^2 u^{(2)}}{\partial x^2} = P_2 \quad .$$

As usual, the complete solution is the sum of the steady state solution

$$u_{SS}^{(2)} = \frac{P_2}{2\alpha^2}\left(L^2 - x^2 \right) \quad ,$$

and a homogeneous solution $u_h{}^{(2)}$. Following the steps of Part (b), we can expand $u_{SS}{}^{(2)}$ as

$$u_{SS}^{(2)} = \sum_n E_n^{(2)} \cos\frac{(2n-1)\pi x}{2L} \quad \text{, where}$$

$$E_n^{(2)} = (-1)^{n+1} \frac{16 P_2 L^2}{\alpha^2 (2n-1)^3 \pi^3} \quad .$$

The IC for $u^{(2)}$ is defined at time t_1; indeed, its value is (03) evaluated at $t = t_1$. Thus,

$$u^{(2)}(x,t_1) = \sum_n \cos\frac{(2n-1)\pi x}{2L}\left[E_n^{(1)} + \left(C_n - E_n^{(1)} \right) f_n(t_1) \right] \quad .$$

Subtracting $u_{SS}{}^{(2)}$ from this yields $u_h{}^{(2)}(x,t_1)$ as

$$u_h^{(2)}(x,t_1)=\sum_n \cos\frac{(2n-1)\pi x}{2L}\left[E_n^{(1)}+\left(C_n-E_n^{(1)}\right)f_n(t_1)-E_n^{(2)}\right] \quad .$$

Then for $u_h^{(2)}(x,t)$, we tack onto each eigenfunction the corresponding time evolution; since the IC is at $t = t_1$ rather than $t = 0$, the arguments of the exponentials must be $t–t_1$ rather than just t. This gives

$$u_h^{(2)}(x,t)=\sum_n \cos\frac{(2n-1)\pi x}{2L}\left[E_n^{(1)}+\left(C_n-E_n^{(1)}\right)f_n(t_1)-E_n^{(2)}\right]f_n(t-t_1) \text{ , or}$$

$$u_h^{(2)}(x,t)=\sum_n \cos\frac{(2n-1)\pi x}{2L}\left[\left(E_n^{(1)}-E_n^{(2)}\right)f_n(t-t_1)+\left(C_n-E_n^{(1)}\right)f_n(t)\right] \quad .$$

Finally, for $u^{(2)}(x,t)$, we add the steady state solution (in the expanded form) to the homogeneous solution, getting

$$u^{(2)}(x,t)=\sum_n \cos\frac{(2n-1)\pi x}{2L}\times \left[E_n^{(2)}+\left(E_n^{(1)}-E_n^{(2)}\right)f_n(t-t_1)+\left(C_n-E_n^{(1)}\right)f_n(t)\right] \quad , \qquad (04)$$

which is the desired answer.

(d) The periodicity condition means that the temperature at $t = T$ equals that at $t = 0$. Hence, (04) evaluated at $t = T$ must be equal to the starting temperature (02). This means that

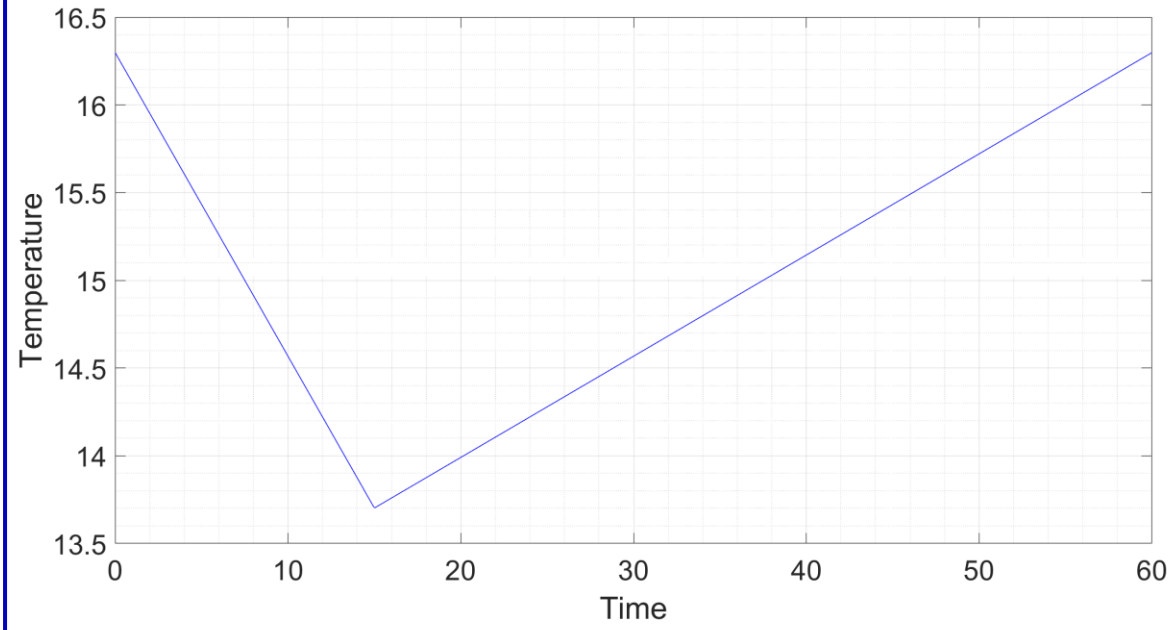
$$\sum_n \cos\frac{(2n-1)\pi x}{2L}\left[\left(E_n^{(1)}-E_n^{(2)}\right)f_n(T-t_1)+\left(C_n-E_n^{(1)}\right)f_n(T)+E_n^{(2)}\right]$$
$$=\sum_n C_n \cos\frac{(2n-1)\pi x}{2L} \quad .$$

Since eigenfunction expansion is unique, the coefficients of $\cos([2n–1]\pi x/2L)$ in this series equation must be equal each to each; this reduces the series equation to a simple algebraic equation for C_n. Letting $t_2 = T–t_1$, this equation has the solution

$$C_n=\frac{E_n^{(1)}\left[f_n(t_2)-f_n(T)\right]+E_n^{(2)}\left[1-f_n(t_2)\right]}{1-f_n(T)} \quad ,$$

which is the desired answer.

(e) The long calculation is finally over; now we have to plot the results and see if they make sense. Since everything is analytical, the plotting is simple. The midpoint of the restaurant corresponds to $x = 4$. We generate $u^{(1)}(4,t)$ and $u^{(2)}(4,t)$ using the formulae (03) and (04) together with the coefficients (05). Since it hardly makes any difference in run time, we take 1000 terms in the eigenfunction series. Finally, since $u = 0$ corresponds to 15 °C, we add 15 to the result for u and show the temperature in Centigrade.



We run a few plausibility checks on this plot before declaring it final. Firstly, the temperature decreases monotonically during the interval $[0, 15]$ which is when the door is open. It increases monotonically during $[15, 60]$ which is when the door is closed. Both of these are very much what we would expect. Then, the temperature is continuous at 15 and is equal at the start and end of the interval, consistent with the continuity and periodicity conditions we imposed. Basis these encouraging results, we declare this plot to be our answer.

*

COMMENTS : The most surprising feature of this calculation is that the heat equation with a piecewise constant, periodic, zero-average source function and some physically plausible BCs can be solved analytically. Remember, (03-5) are an EXACT SOLUTION of the mathematical model of the restaurant – we didn't

throw out any term anywhere. Here the source was constant in two pieces 1 and 2. If it were constant in k pieces of durations t_1, t_2, ….., t_k, then the solutions $u^{(1)}$, $u^{(2)}$, ….., $u^{(k)}$ would become the natural generalizations of what we got here. The C_n formula would also generalize; this generalization is worth quoting in full. Let $E_n^{(k)}$ be the Fourier coefficients of the steady state solution in the k^{th} piece of excitation; then we would have

$$C_n = \frac{1}{1-f_n(T)}\begin{pmatrix} E_n^{(k)}\left[1-f_n(t_k)\right]+E_n^{(k-1)}\left[f_n(t_k)-f_n\left(t_k+t_{k-1}\right)\right] \\ +.....+E_n^{1}\left[f_n\left(t_k+t_{k-1}+.....+t_2\right)-f_n(T)\right] \end{pmatrix} .$$

Of course, this is for one particular BC choice – with other BCs the details would be different but the big picture would remain the same. The assumption of zero average of the source is necessary to ensure a tractable periodic solution u; a non-zero average of the source would mean heat being pumped in or sucked out during each period and would preclude a physically meaningful periodic particular solution for u.

Now let's discuss the validity of the restaurant model. The source as a substitute for the wind is of course a stand-in term for an extremely complicated phenomenon. While setting the problem, I'd chosen the cross-sectional area of the restaurant to be 8 m^2 and had taken the wind to be at –15 °C (yes, cold even for Ithaca) and flowing in at about 50 m^3 per minute (a very reasonable flow rate). Assuming instant and perfect mixing of internal air and wind gives the equivalent sink strength of this problem. The right side BC assumes that the effect of the wind at the furthest end of the restaurant is negligible. The left side BC assumes that there is no heat flow through the door. Although this might seem ridiculous if the door is open, the wind has already been accounted for in the sink term, so any benign BC on the door will be good enough.

We can see that the fluctuations in internal temperature are quite large. After the door is opened, the temperature drops by almost 3 °C in 15 seconds. This is by no means an insignificant amount. What this model doesn't capture is that the amplitude of the fluctuations becomes larger as one moves closer to the door. This has happened because we assumed the sink to be spatially constant. In reality, the wind will suck more heat from the places near the door than from the places in the interior. A more complicated sink term which is stronger near the door and weaker away from it would have shown the spatially varying

fluctuations. In this case, I didn't find that extra feature worth the extra mess – it would have added no new conceptual subtlety but only some hassles in finding and expanding $u_{SS}^{(1)}$.

So how do fancier restaurants protect their generous patrons from large-amplitude high-frequency temperature fluctuations ? They use either or both of two possible solutions. One solution is to have the interior separated from the outside through two doors – the first leads from the road into a small porch and then the second opens into the dining area. This porch acts as a buffer between the customers and the elements. The second solution is to mount an air curtain on top of the door. This is a fan which blows a vertical jet of air from ceiling to ground, all along the width of the door. The cold wind finds it very difficult to cross the curtain and harass the diners inside.

Example 5.4.6

NEWTON's rule again

The temperature $u(x,t)$ of the rod x belongs to $[0, \pi]$ satisfies the 1+1 heat equation

$$\alpha^2 \frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t} \quad ,$$

with BCs $u(0,t) = u(\pi,t) = 0$ and IC $u(x,0) = u_0$.

(a) Determine $u(x,t)$.

(b) Determine the spatially averaged temperature of the rod as a function of time.

(c) If the rod's average temperature were to be modelled using NEWTON's rule of cooling as $\dot{T} + \gamma T = 0$, determine the value of γ which would be most appropriate.

(d) Would the above value of γ be optimal for all initial temperature distributions of the rod ? Justify your answer.

[Level : 2]

(a) If you see a problem like this right after the last one, you might be tempted to search for hidden complexities which do not exist. But this is really just 5.1.3 again, with a different IC. That too is nothing new – it's the square wave from 4 .3.1. Putting all this together, the specific solution is a one-liner :

$$u(x,t) = \sum_{\text{odd } n} \frac{4}{n\pi} \sin nx \exp\left(-\alpha^2 n^2 t\right) \quad , \qquad *$$

which is the desired answer.

(b) To obtain the spatially averaged temperature, we integrate $u(x,t)$ over the rod and divide by its length i.e.

$$u_{\text{avg}}(t) = \frac{1}{\pi} \int_0^\pi u(x,t)\,\mathrm{d}x \quad .$$

Plugging in the $u(x,t)$ we found with such little effort, and doing the easy integration,

$$u_{\text{avg}}(t) = \sum_{\text{odd } n} \frac{8}{n^2\pi^2} \exp\left(-n^2 t\right) \quad , \qquad (01)*$$

which is the desired answer.

(c) The ODE based on NEWTON's rule has the solution $T = T_0\,\mathrm{e}^{-\gamma t}$ (and I'm really not going to refer to 1.2.1 at this advanced stage of our course). Now we have to compare this with the series (01). We can see that the largest term in the series is the first term – at $t = 0$ the second term is smaller by 1/9th, the third by 1/25th and so on. Moreover, since those terms decay far more quickly, at times larger than zero they will become smaller still and the only surviving component of $u_{\text{avg}}(t)$ will be the first term. This behaves as e^{-t}, so to achieve agreement with the ODE we must have $\underline{\gamma = 1}$.

(d) The value $\gamma = 1$ would <u>NOT</u> be optimal for all ICs. Considering just one example, suppose the IC were $u(x,0) = \sin 3x$ then the temperature as a whole would decay as e^{-9t}, so would the average, and the ideal γ would have been $\gamma = 9$.

*

COMMENTS : This problem shows the application and limitation of NEWTON's rule in describing heat conduction. For the IC of uniform initial distribution, the

rule works quite well to describe the average temperature of the body (the only temperature which an ODE model can be expected to calculate). For the different IC considered in Part (d), the rule would work well with a different γ. But suppose the IC were something like $\sin x + 100 \sin 3x$? Then, at small times, the e^{-9t} behaviour would dominate, making the ODE have an effective γ close to 9. At large times however, due to the larger decay rate, the $\sin 3x$ component would become smaller than the $\sin x$ component and the ODE would have an effective γ close to 1. Thus, an ODE with a single value of γ would do a bad job of predicting the dynamics of the average temperature in such a case.

Here we chose BCs to be at the atmospheric temperature itself. We could also have chosen something like leakage BC or one end insulation BC without qualitative differences. In all cases, the IC would get expanded in a series of eigenfunctions, and the average temperature would become a series of more and more negative exponentials. We would get a good agreement with the ODE if the first term happened to be dominant from $t = 0$, and a poor agreement if that was not the case.

This problem further justifies the case for not calling NEWTON's statement a "law", an issue we already elaborated when we first saw the rule in 1.2.6. Three hundred years back when Sir ISAAC made this discovery , there was hardly any science at all (he more or less created physics out of nothing like MOZART created opera out of nothing), let alone a clear distinction between laws of nature and empirical facts. But in the light of today's knowledge , we certainly can't elevate to law status a premise which can so easily be made to show poor agreement with reality.

However, NEWTON's statement DOES hold true to excellent accuracy when the bodies involved are microscopic rather than macro. Indeed, in 2.3.3 we derived the baby heat equation by assuming that neighbouring particles of the body exchange heat as per NEWTON 's Rule. When that discrete model is made continuous, the conduction principle becomes FOURIER's Law and we get the actual heat equation. It is interesting to note that JOSEPH FOURIER had actually first arrived at his law via this route.

Example 5.4.7

Force-ventilated traction motor

Easier form : Considering the governing equation, perform the variable transformation mentioned in Part (b).

The traction motor 6FRA6068 used in WAP7 is cooled by a blower which forces a stream of atmospheric air past the motor at constant speed v. If the atmospheric temperature is taken as zero, the equation for the temperature $u(x,t)$ of the air stream can be written as

$$\rho C\left(\frac{\partial u}{\partial t}+v\frac{\partial u}{\partial x}\right)-\kappa\frac{\partial^2 u}{\partial x^2}=S(x,t) \quad , \qquad (*)$$

defined in the spatial interval $[-\infty, \infty]$ with BC $u \to 0$ as $|x| \to \infty$. S is the power density being output by the motor. It is constant in the region $-L \le x \le L$ which corresponds to the motor and is zero everywhere outside. The motor is at atmospheric temperature before it is turned ON at $t=0$. Take the parameter values atmospheric temperature 30 ^{o}C, $\rho = 1.2$, $C = 1000$, $\kappa = 0.025$, $v = 5$, $S = 1.3\text{x}10^6$ and $L = 0.5$ SI Units, and let the temperature of the motor be the value of u at $x = 0$.

Consider the case where the blower is inoperative i.e. $v = 0$.

(a) Express $u(0,t)$ in terms of the GREEN's function. Evaluate the integral numerically and determine the temperature of the motor at $t = 2$.

Now consider the case where the blower is operating i.e. $v = 5$.

(b) Perform a variable transformation $\xi = x - vt$ and $\eta = t$ to obtain the GREEN's function for (*).

(c) Determine the temperature of the motor at $t = 2$.

[Level : 3]

(a) After our little tour of Japan and the US we are back to the land of WAP7 and Rajdhani Express. With the blower off i.e. $v=0$, the governing equation reduces to the standard 1+1 heat equation with source term. The question tells us that we should be using the GREEN's function – since the domain is infinite and the BCs are $u\to 0$ as $|x|\to\infty$ the relevant one is the free space GREEN's function (5.3.03) derived in 5.3.2. The integral formula is (5.3.02) which we derived in 5. 3.3 – the solution is, defining $\alpha^2=\kappa/\rho C$ and $P=S/\rho C$,

$$u(x,t)=\int_{\text{interval}} dx'\int_0^t dt'\,P(x',t')G(x,x',t,t') \quad \text{, or}$$

$$u(x,t)=\int_{\text{interval}} dx'\int_0^t dt'\,P(x',t')\frac{1}{2\alpha\sqrt{\pi}\sqrt{t-t'}}\exp\left(-\frac{(x-x')^2}{4\alpha^2(t-t')}\right) \quad . \qquad (01)$$

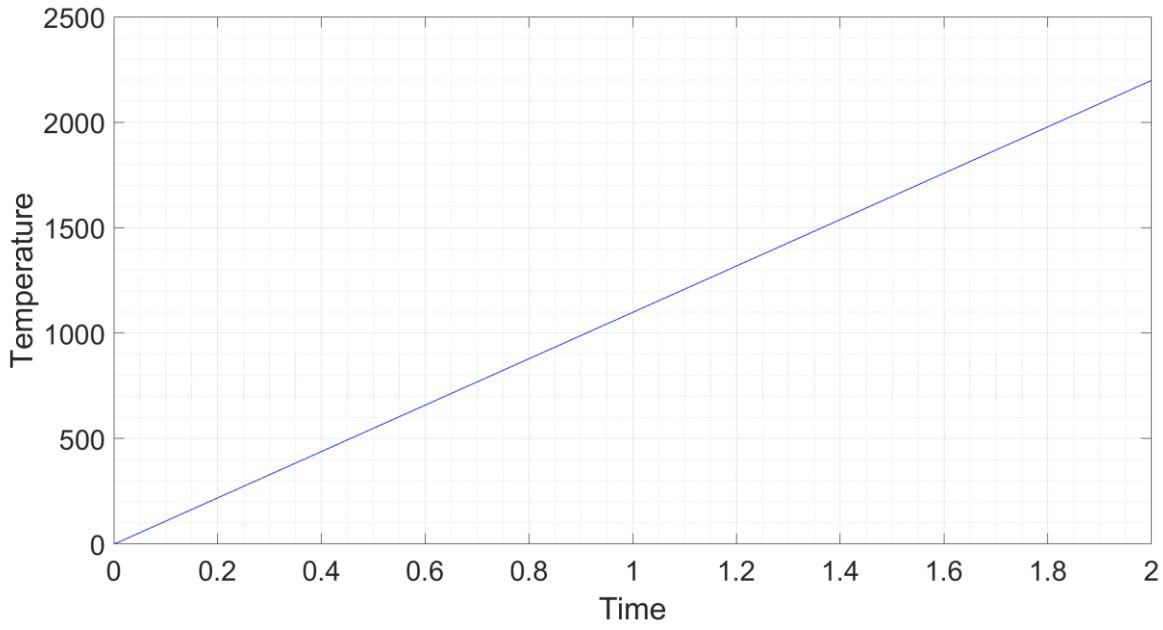
To this we might need to add a homogeneous solution corresponding to the IC (see 5.3.3); however since that is given to be zero, we don't have to bother.

Now $P(x',t')$ is non-zero only in the region corresponding to the motor, hence the x' integral in (01) should run from $-L$ to L. Within this region, P is a constant whose value we call P_0, so it comes outside the integral. Further, we want the temperature at the point $x=0$, so we substitute $x=0$ into (01). Doing this we have

$$u(0,t)=\frac{P_0}{2\alpha\sqrt{\pi}}\int_{-L}^{L} dx'\int_0^t dt'\frac{1}{\sqrt{t-t'}}\exp\left(-\frac{x'^2}{4\alpha^2(t-t')}\right) \quad . \qquad (02)$$

We evaluate this integral numerically using the rectangle method. Because of the huge size of G at $t=t'$, we need a fine grid along the t'-axis. After some hit and trial I have gone with a split of x' interval into 1000 points between –0.5 and 0.5 and t' interval into 2000 points between 0 and 2 (the maximal value of t required in the question). Further refinement of the grid led to significantly higher computational times with no perceptible change in the solution. To ensure the proper limits on the t' integration, I defined G to be automatically zero if t was less than t'.

Plotting $u(0,t)$ for t going from 0 to 2 gives the following picture.



At a time value of 2 s, the temperature of the motor is 2200 °C. The motor is no longer a motor but has long since been reduced to ashes.

(b) Even though the stuff about GREEN's function and all is fancy, variable transformation in a PDE is a popular final exam question so read this carefully. We want to transform u from a function of x and t into a function of two new variables ξ (which is $x-vt$) and η (which is t). Since it's a differential equation we are after, we'll need the DERIVATIVES of u w.r.t these new variables. The variable transformation formula uses the multivariate Chain Rule which you have learnt in the historical past, so here is its statement : to take the partial derivative of a function w.r.t any one of the new variables, you add up all the partials of the function w.r.t all the old variables, multiplying each by the partial of the corresponding old variable w.r.t the concerned new variable. Of course, the statement is much more obscure than the example, which is coming shortly.

But before applying the Rule, let's think about which way we want to go. We are given the governing equation in terms of u_x and u_t – to make the transformation, we would like to express these in terms of u_ξ and u_η and then substitute back into the PDE. So we have

$$\frac{\partial u}{\partial x} = \frac{\partial u}{\partial \xi}\frac{\partial \xi}{\partial x} + \frac{\partial u}{\partial \eta}\frac{\partial \eta}{\partial x}$$
$$= \frac{\partial u}{\partial \xi}\cdot 1 + \frac{\partial u}{\partial \eta}\cdot 0 = \frac{\partial u}{\partial \xi}$$

Evidently, $\partial^2 u/\partial x^2 = \partial^2 u/\partial \xi^2$. As for the time derivative,

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial \xi}\frac{\partial \xi}{\partial t} + \frac{\partial u}{\partial \eta}\frac{\partial \eta}{\partial t}$$
$$= -\nu\frac{\partial u}{\partial \xi} + \frac{\partial u}{\partial \eta}$$

A very common error occurs in the above step – since $t = \eta$, it is tempting to say that u_t must be equal to u_η. As you can see, it is not !

Now plug everything back into the given PDE – that is

$$\frac{\partial u}{\partial t} + \nu\frac{\partial u}{\partial x} - \alpha^2\frac{\partial^2 u}{\partial x^2} = P(x,t) \quad , \qquad (03)$$

and becomes

$$-\nu\frac{\partial u}{\partial \xi} + \frac{\partial u}{\partial \eta} + \nu\frac{\partial u}{\partial \xi} - \alpha^2\frac{\partial^2 u}{\partial \xi^2} = P(\xi,\eta) \quad \text{, or}$$
$$\frac{\partial u}{\partial \eta} - \alpha^2\frac{\partial^2 u}{\partial \xi^2} = P(\xi,\eta) \quad .$$

This is just a plain old 1+1 heat equation for $u(\xi,\eta)$!

The GREEN's function for (03) in terms of ξ and η is of course trivial – it's just (5.3.03) with x,x',t,t' replaced by ξ,ξ',η,η' respectively . Converting back to x and t from ξ and η now gives us

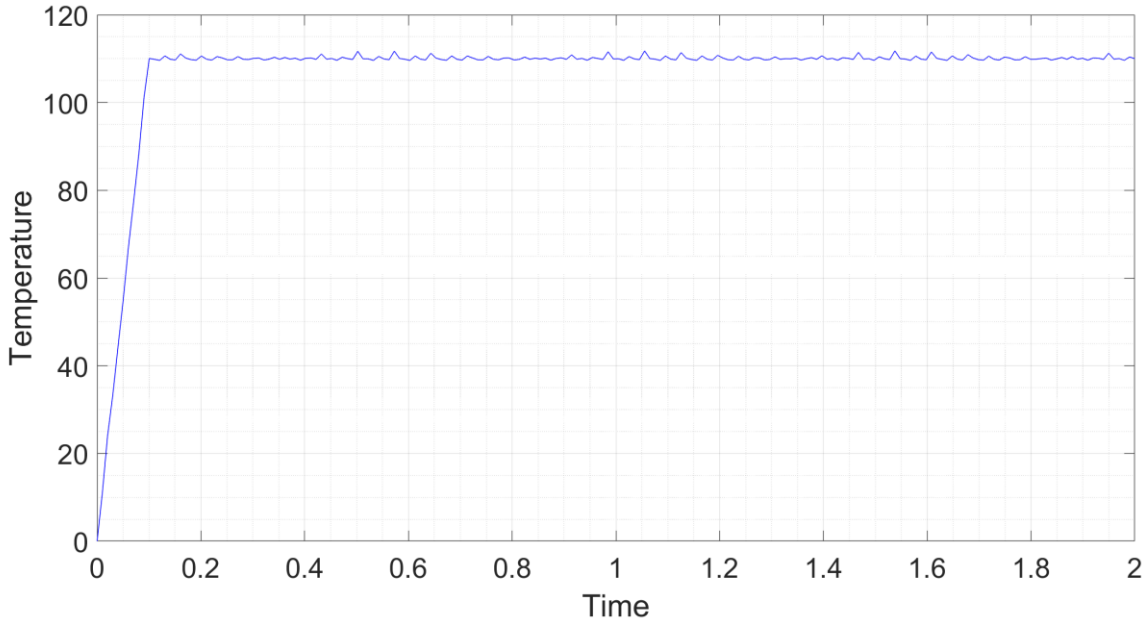
$$* \qquad G(x,x',t,t') = \frac{1}{2\alpha\sqrt{\pi}\sqrt{t-t'}}\exp\left(-\frac{\left[x-x'-\nu(t-t')\right]^2}{4\alpha^2(t-t')}\right) \quad , \qquad *$$

which is the desired answer.

(c) This is just Part (a) again with a different G, so the relevant formula is (02) with the new G replacing the old G. For later reference I'll write it out in full :

$$u(0,t)=\frac{P_0}{2\alpha\sqrt{\pi}}\int_{-L}^{L}\mathrm{d}x'\int_{0}^{t}\mathrm{d}t'\frac{1}{\sqrt{t-t'}}\exp\left(-\frac{\left[-x'-v(t-t')\right]^2}{4\alpha^2(t-t')}\right) \quad . \qquad (04)$$

This time, I have split the t' interval into 6000 to get a better result. The plot is shown below.



This shows temperature on the u-scale; to get in Centigrade scale we add on the constant 30 °C for atmospheric temperature. At a time of 2 seconds, the temperature of the traction motor is 140 °C.

*

COMMENTS : This problem shows the utility of the GREEN's function approach – we are now reaping what we sowed in 5.3.2-3. Analytically solving the governing equation is very likely impossible (at least I have no idea how to do it). But with the help of GREEN's function , we could express the solution as a double integral which was trivial to evaluate on the computer . By contrast , if we didn't know GREEN 's function , we would have had to solve this using matrix method, which is no joke.

A blast of air is of course the way in which WAP7's (or any other loco's) traction motors are kept at operating temperatures (which are typically about 90 °C above ambient). What the setup looks like is as follows – the motor is an elongated cylindrical structure consisting of two concentric cylinders, the rotor

and the stator (look ahead to 6.3.6 for a diagram). These two are separated by an air gap. The air stream is directed parallel to the motor axis. Before hitting the motor it is split up into two parts – one is channelized into the air gap while the second is made to surround the stator. These two streams carry away the heat generated by the two elements.

How did the governing equation come about ? That was essentially by reversing the steps we did in Part (b). Suppose that the air stream is blowing past you at velocity v and carrying heat, and you want to write an equation for its temperature. Consider an observer who is moving with the stream. To him, the stream will appear like static air and the temperature will satisfy the 1+1 heat equation relative to space and time coordinates defined in his reference frame. Time is of course the same for both (at 5 m/s we really don't need to worry about relativistic time dilation). But because the observer is moving, what you call as x, he will call as $x-vt$: imagine a moving metre scale sliding alongside a static one. So ξ and η are actually the spatiotemporal coordinates in the moving observer's frame. In these coordinates, u satisfies the 1+1 heat equation; now transform to x,t coordinates and we get the equation which was the starting point of this problem. This equation is called CONVECTION DIFFUSION equation.

The domain of definition must be the entire real line because it makes no sense to define a place where the stream begins or ends. In a one-dimensional model we cannot account for the air stream spreading out or slowing down. If we artificially set a BC at some point then the model becomes unphysical at that point – if the stream suddenly stops flowing, where does the air go ? As for the parameter values, the peak power output of the motor is about 850 kW or 1150 HP (recall – 1.1.1 says the total power of WAP7 is 7000 HP while 3.3.5 tells you that it has six axles). The rated efficiency of the motor is 95 percent. Assuming all the wasted power to be in the form of heat, we get about 40 kW as the heat output when the motor is at full throttle.

S in the equation is however the power density. The power is being generated in that volume of the air stream which is inside the motor. The length of the motor is about 1 m and the radius is about 20 cm. The thickness of the air gap is less than 1 cm, while the thickness of the airflow round the stator is of the order of 1 cm or a little more. Taking these into account, I have assumed a thickness of 2.5 cm for the air stream. These numbers yield the value of S given in the

question. Of course, S is non-zero only inside the motor since that's where all the heat is being generated. The stream velocity of 5 m/s is also a very plausible figure – industrial fans typically generate air speed of that order, and a significantly higher speed can hardly be enforced in a narrow gap.

The answer of 140 ^{o}C is very plausible. It is slightly above the actual operating temperature of the motor, which is 120-30 ^{o}C. This marginal discrepancy is only to be expected in view of the approximate values of the numbers entering the model. The fact that physically plausible parameter values yielded a very plausible answer shows the robustness of the modeling itself.

The comparison between Parts (a) and (c) shows the importance of keeping the air flowing. With the blower off, the temperature is orders of magnitude higher than with the thing on. This can be explained both mathematically and physically. First we give the mathematical explanation. Considering (02), look at a small and fixed value of x' and a given value of t. At t' slightly less than t, the exponential in the GREEN's function is close to unity because x' is small, while the square root outside is huge. So we have a large contribution to u from this point. As we keep going backwards in t' (remember, all contributions arise from $t'<t$), the exponential gets closer to unity still while the root starts decreasing. This decrease is rather slow (power law decrease), so the contribution is substantial over a long range of t' – we must go to t' in distant past before the contribution becomes negligible. Now consider (04). Again, for small x' and t' slightly less than t, the exponential is close to unity and the radical is big. This time however there is a $(t-t')^2$ in the numerator of the exponential. As we go backwards in t', the argument of the exponential starts increasing as $t-t'$ and the contribution becomes exponentially smaller. So, the significant contribution to the temperature at the current time comes from only a short period in the system's history. This is responsible for u being much smaller in this case.

Physically, air is a lousy conductor so if you pump a few kilowatts into a fistful of static air, you are going to create a blast furnace. With moving air however, you keep replacing the hot air with cooler air. Before the air gets a chance to really become hot, it has already been blown away past the power source and back into the atmosphere. This keeps the temperature down, by more than an order of magnitude as it turns out.

One thing which might surprise you is the almost instantaneous rise of the temperature to 140 °C in the plot of Part (c). Surely you don't expect the motor to attain the maximum temperature after just 0.1 s of operation – the calculation must be wrong. But it isn't. Starting up a power source in a stream of air is exactly analogous to lighting up a gas stove, and the temperature there does indeed take a fraction of a second to attain its steady state value. Still, you aren't convinced. Yes, there IS a simplification here. It lies in the model of the motor as a source. Here we assumed that the motor is dumping 40 kW of heat to the air stream forthwith after it is switched on. That however won't be the case. Initially, a lot of the 40 kW will go into heating the motor itself – the rotor bars and core, and stator conductors and core. Only when the rotor and stator have become hot will they be releasing the full 40 kW to the air. So we have implicitly assumed that the motor and air are in steady state of heat exchange right from the start, while actually there will be some dynamics there. The effect of this transient will be to slow down the temperature rise.

This also means that in the event of a blower failure, the motor won't become ashes instantaneously as it did in Part (a). Rather, there will be enough time for the alarms to go off in the cab and for the loco pilot to cut the throttle and isolate the faulty bogie. Of course, in WAP7, there are automatic temperature sensors which detect overheating and immediately stop current to the motor if a blower failure occurs. But even in older locos like WAP4, there is plenty of time for the loco pilot to take action in the event of a blower malfunction.

After the revelation about the assumption of steady state, you might be tempted to fall upon the PDE itself – surely a 1+1 equation with air coming in from infinity and going out to infinity is a drastic simplification. We all know that the stream velocity is negligible a few metres on either side of the motor. In this case however I will defend my PDE. It is not an intrinsically bad model because it captures very realistically the dynamics of air flow near the motor. In that region, the flow really is one-dimensional along the axis and it takes heat away from the motor. Whether the air is coming in from infinity and going out to infinity at constant speed along its original trajectory or whether it's coming in an accelerating cone and again spreading out into a decelerating cone is not really relevant if all we care is how much heat it carries away. What matters is it's coming from $u = 0$ and going back to $u = 0$. The model is lousy for the air

temperature five metres away from the motor. The model would have been lousy if the actual airflow had been an arc of a circle around the motor. But given the reality, this model is good enough.

Finally, this Tute shows the extreme importance of the blowers in an electric loco. The transformer, rectifier, inverter and traction motors, all are megapower components which have huge heat generation. They are all kept to practical temperature by blowers, which literally blow the heat away. This is why an electric loco constantly makes a loud whirring noise even though the components themselves are relatively quiet – in fact, when a Rajdhani arrives, the howl of the loco's blowers battles with the roar of the 500 HP diesel generator car right behind the loco. The dominant sound inside the loco cab is also that of myriad blowers – only at high speed and fullest throttle do the motors themselves pitch in at the soprano register. But although the sound is unpleasant, one can't imagine a loco without blowers. The power of PDE is that it can show you why.

* * * * *

This page intentionally left blank

Review

VECTOR CALCULUS

In that great subject called multivariate calculus, you have learnt (and possibly forgotten) about gradient, divergence, curl and Laplacian. That course tends to become a huge bore because all the problems take the form of "find the grad of this" and "show that the divergence of that is zero". Here we come back to these operators but this time in the form of differential equations – the thing whose div or grad you want is usually an unknown function. Adding this complication enables us to get meaningful results for meaningful problems. But before that, as usual, we have to get the maths off the ground. The below few pages are essentially a listing of the various vector calculus operators in conventional coordinate systems.

CARTESIAN

In Cartesian coordinates x,y,z the three first order operators are as follows. Grad acts on a scalar (u in the abstract) and produces a vector. It is

$$\nabla u = \frac{\partial u}{\partial x}\hat{\mathbf{x}} + \frac{\partial u}{\partial y}\hat{\mathbf{y}} + \frac{\partial u}{\partial z}\hat{\mathbf{z}} \quad .$$

Div does the opposite – it turns vector ($\mathbf{v}$ in the abstract) to scalar :

$$\nabla \cdot \mathbf{v} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z} \quad .$$

Curl has to do something new – it turns vector to vector :

$$\nabla \times \mathbf{v} = \left(\frac{\partial v_z}{\partial y} - \frac{\partial v_y}{\partial z} \right)\hat{\mathbf{x}} + \left(\frac{\partial v_x}{\partial z} - \frac{\partial v_z}{\partial x} \right)\hat{\mathbf{y}} + \left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y} \right)\hat{\mathbf{z}} \quad .$$

The Laplacian is a second order operator and is the main focus of the upcoming Chapter. It is defined as the divergence of the gradient. Acting on a scalar, it returns a scalar. It is defined as

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \quad .$$

In some books, $\nabla^2 u$ is written as Δu, but we use Δ only to mean small changes or increments, not Laplacian. We can also make Laplacian act on vectors by hitting each component with it; that gives us three components of a new vector. Thus, $\nabla^2 \mathbf{A} = \mathbf{B}$ means that $\nabla^2 A_{x,y,z} = B_{x,y,z}$ where the subscripts correspond. This vector Laplacian however is only possible to do in Cartesian coordinates, and NOT in the other coordinate systems, where curvature gives rise to extra complications.

Finally, there is another operator which we'll need sometimes. It is the operator $(\mathbf{A} \cdot \nabla)$ where **A** is a given vector. Acting on a vector **B**, this operator returns another vector, as follows :

$$(\mathbf{A} \cdot \nabla)\mathbf{B} = \left(A_x \frac{\partial B_x}{\partial x} + A_y \frac{\partial B_x}{\partial y} + A_z \frac{\partial B_x}{\partial z} \right) \hat{\mathbf{x}} + \left(A_x \frac{\partial B_y}{\partial x} + A_y \frac{\partial B_y}{\partial y} + A_z \frac{\partial B_y}{\partial z} \right) \hat{\mathbf{y}} + \left(A_x \frac{\partial B_z}{\partial x} + A_y \frac{\partial B_z}{\partial y} + A_z \frac{\partial B_z}{\partial z} \right) \hat{\mathbf{z}} \quad .$$

This completes the list of the operators in the most well-behaved coordinate system.

CYLINDRICAL POLAR

The coordinates here are the **radius** ρ, the **angle** θ and the **height** z. ρ goes from 0 to ∞, θ from 0 to 2π and z from $-\infty$ to ∞. The x-axis corresponds to $\theta = 0$ and θ is counterclockwise positive. The relations connecting the coordinates are

$$x = \rho \cos\theta \qquad y = \rho \sin\theta \quad ,$$

$$\rho = \left(x^2 + y^2 \right)^{1/2} \qquad \theta = \arctan \frac{y}{x} \quad .$$

The volume element $\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$ is $\rho\,\mathrm{d}\rho\,\mathrm{d}\theta\,\mathrm{d}z$. The unit vectors are related as

$$\hat{\boldsymbol{\rho}} = \cos\theta\hat{\mathbf{x}} + \sin\theta\hat{\mathbf{y}} \qquad \hat{\boldsymbol{\theta}} = -\sin\theta\hat{\mathbf{x}} + \cos\theta\hat{\mathbf{y}} \quad ,$$
$$\hat{\mathbf{x}} = \cos\theta\hat{\boldsymbol{\rho}} - \sin\theta\hat{\boldsymbol{\theta}} \qquad \hat{\mathbf{y}} = \sin\theta\hat{\boldsymbol{\rho}} + \cos\theta\hat{\boldsymbol{\theta}} \quad .$$

Then, to get the first order derivative operators, keep substituting the above into Cartesian expressions and chain-ruling. The results are

$$\nabla u = \frac{\partial u}{\partial \rho}\hat{\boldsymbol{\rho}} + \frac{1}{\rho}\frac{\partial u}{\partial \theta}\hat{\boldsymbol{\theta}} + \frac{\partial u}{\partial z}\hat{\mathbf{z}} \quad ,$$
$$\nabla \cdot \mathbf{v} = \frac{1}{\rho}\frac{\partial}{\partial \rho}\left(\rho v_\rho\right) + \frac{1}{\rho}\frac{\partial v_\theta}{\partial \theta} + \frac{\partial v_z}{\partial z} \quad ,$$
$$\nabla \times \mathbf{v} = \left(\frac{1}{\rho}\frac{\partial v_z}{\partial \theta} - \frac{\partial v_\theta}{\partial z}\right)\hat{\boldsymbol{\rho}} + \left(\frac{\partial v_\rho}{\partial z} - \frac{\partial v_z}{\partial \rho}\right)\hat{\boldsymbol{\theta}} + \frac{1}{\rho}\left(\frac{\partial}{\partial \rho}\left(\rho v_\theta\right) - \frac{\partial v_\rho}{\partial \theta}\right)\hat{\mathbf{z}} \quad .$$

Div of grad now gives Laplacian :

$$\nabla^2 u = \frac{1}{\rho}\frac{\partial}{\partial \rho}\left(\rho\frac{\partial u}{\partial \rho}\right) + \frac{1}{\rho^2}\frac{\partial^2 u}{\partial \theta^2} + \frac{\partial^2 u}{\partial z^2} \quad .$$

Much worse than in Cartesian.

Finally, the nasty vector.grad on vector is

$$\left(\mathbf{A}\cdot\nabla\right)\mathbf{B} = \left(A_\rho\frac{\partial B_\rho}{\partial \rho} + \frac{1}{\rho}A_\theta\frac{\partial B_\rho}{\partial \theta} + A_z\frac{\partial B_\rho}{\partial z} - \frac{A_\theta B_\theta}{\rho}\right)\hat{\boldsymbol{\rho}} +$$
$$\left(A_\rho\frac{\partial B_\theta}{\partial \rho} + \frac{1}{\rho}A_\theta\frac{\partial B_\theta}{\partial \theta} + A_z\frac{\partial B_\theta}{\partial z} + \frac{A_\theta B_\rho}{\rho}\right)\hat{\boldsymbol{\theta}} + \quad .$$
$$\left(A_\rho\frac{\partial B_z}{\partial \rho} + \frac{1}{\rho}A_\theta\frac{\partial B_\theta}{\partial \theta} + A_z\frac{\partial B_z}{\partial z}\right)\hat{\mathbf{z}}$$

I wouldn't have written this down except that we'll need it in one of the Tutes in Chapter 6.

SPHERICAL POLAR

The coordinates are the **radius** *r*, the **angle of depression** or "latitude" θ and the **azimuth** or "longitude" φ. (Latitude and longitude are far easier to visualize than depression and azimuth but if you use those words people

might think you study geography rather than science or engineering. So think latitude, say depression, think longitude, say azimuth.) Radius r goes from 0 to ∞, θ from 0 to π and φ from 0 to 2π. $\theta = 0$ corresponds to the positive z-axis or "North Pole". $\theta = \pi/2$ is the "Equator" and $\theta = \pi$ the "South Pole". $\varphi = 0$ at $\theta = \pi/2$ gives the x-axis and φ is counterclockwise positive. The relations connecting the coordinates are

$$x = r\sin\theta\cos\varphi \qquad y = r\sin\theta\sin\varphi \qquad z = r\cos\theta \quad ,$$

$$r = \left(x^2 + y^2 + z^2\right)^{1/2} \qquad \theta = \arctan\frac{\left(x^2 + y^2\right)^{1/2}}{z} \qquad \varphi = \arctan\frac{y}{x} \quad .$$

The volume element $\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$ is $r^2\sin\theta\,\mathrm{d}r\,\mathrm{d}\theta\,\mathrm{d}\varphi$. The unit vectors are related as

$$\begin{gathered}
\hat{\mathbf{r}} = \sin\theta\cos\varphi\hat{\mathbf{x}} + \sin\theta\sin\varphi\hat{\mathbf{y}} + \cos\theta\hat{\mathbf{z}} \\
\hat{\boldsymbol{\theta}} = \cos\theta\cos\varphi\hat{\mathbf{x}} + \cos\theta\sin\varphi\hat{\mathbf{y}} - \sin\theta\hat{\mathbf{z}} \qquad \hat{\boldsymbol{\varphi}} = -\sin\varphi\hat{\mathbf{x}} + \cos\varphi\hat{\mathbf{y}} \quad , \\
\hat{\mathbf{x}} = \sin\theta\cos\varphi\hat{\mathbf{r}} + \cos\theta\cos\varphi\hat{\boldsymbol{\theta}} - \sin\varphi\hat{\boldsymbol{\varphi}} \\
\hat{\mathbf{y}} = -\sin\theta\sin\varphi\hat{\mathbf{r}} + \cos\theta\sin\varphi\hat{\boldsymbol{\theta}} + \cos\varphi\hat{\boldsymbol{\varphi}} \qquad \hat{\mathbf{z}} = \cos\theta\hat{\mathbf{r}} - \sin\theta\hat{\boldsymbol{\theta}} \quad .
\end{gathered}$$

Now come the first derivatives :

$$\nabla u = \frac{\partial u}{\partial r}\hat{\mathbf{r}} + \frac{1}{r}\frac{\partial u}{\partial\theta}\hat{\boldsymbol{\theta}} + \frac{1}{r\sin\theta}\frac{\partial u}{\partial\varphi}\hat{\boldsymbol{\varphi}}$$

$$\nabla\cdot\mathbf{v} = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 v_r\right) + \frac{1}{r\sin\theta}\frac{\partial}{\partial\theta}\left(v_\theta\sin\theta\right) + \frac{1}{r\sin\theta}\frac{\partial v_\varphi}{\partial\varphi} \quad ,$$

$$\begin{aligned}
\nabla\times\mathbf{v} = {} & \frac{1}{r\sin\theta}\left[\frac{\partial}{\partial\theta}\left(v_\varphi\sin\theta\right) - \frac{\partial v_\theta}{\partial\varphi}\right]\hat{\mathbf{r}} + \\
& \frac{1}{r}\left[\frac{1}{\sin\theta}\frac{\partial v_r}{\partial\varphi} - \frac{\partial}{\partial r}\left(r v_\varphi\right)\right]\hat{\boldsymbol{\theta}} + \frac{1}{r}\left[\frac{\partial}{\partial r}\left(r v_\varphi\right) - \frac{\partial v_r}{\partial\theta}\right]\hat{\boldsymbol{\varphi}}
\end{aligned} \quad .$$

As usual, Laplacian is div of grad, which is

$$\nabla^2 u = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial u}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 u}{\partial\varphi^2} \quad .$$

As for vector.grad on vector, the formula is too messy to quote, and we won't need it anywhere in this book. So I'll let it be. In case you ever come across a

calculation where you need this monster, please look it up on Wikipedia – it's there.

This completes the review. Unlike the trig review, the formulae in cylindrical and spherical are not things which you can be reasonably expected to derive in the exam hall. There is no shortcut like remembering a subset and getting all the rest. So the examiner will very likely provide anything from here which you might need on the test. Rather, this review only acts as a reference, both for this book and elsewhere.

Finally, we have the **divergence theorem** and the **curl theorem**, whose statements (not proofs) you should remember. The divergence theorem states that if V is a volume with S being its bounding surface, and $\mathbf{v}$ is an arbitrary vector field, then the flux of $\mathbf{v}$ over S equals the divergence of $\mathbf{v}$ integrated over V. The curl theorem says that if S is an area with L being its bounding line, and $\mathbf{v}$ is an arbitrary vector field, then the line integral of $\mathbf{v}$ over L equals the curl of $\mathbf{v}$ integrated over S.

*

This page intentionally left blank

6

THE LAPLACIAN OPERATOR AND RELATED EQUATIONS

This Chapter introduces the Laplacian operator and examines partial differential equations built around this operator. PDEs in more than two independent variables are introduced, including the 3-dimensional Laplace equation and the many+1 heat equation. After the theory is developed, it is applied to examples from classical electromagnetism and steady state fluid mechanics.

6.1

INTRODUCTION TO LAPLACE EQUATION

Laplace equation is

$$\nabla^2 u = 0 \quad .$$

Looks really elegant if you write it in this form. Less elegant if you write it explicitly in spherical polars.

This equation and its solutions was investigated by PIERRE SIMON LAPLACE, a French mathematician who also made groundbreaking studies in probability theory. In this Lecture I won't actually solve it in any coordinate system – that I'll do in the Tutes. Rather, I'll give some general properties of solutions of the equation. These solutions are sometimes called "harmonic functions". This term I find confusing especially in a first course since it seems to suggest harmony, music and by extension, wave equation rather than Laplace equation. Because of this, I will use the jargon "**Laplace functions**" even if it is non-standard.

Laplace equation can be defined non-trivially in two or three spatial dimensions. As you can see, in whichever coordinate system you write it, the equation is second order in every independent variable. In 5.1, we saw that as a starting guideline we will need two conditions for every independent variable specified at all possible values of all other independent variables. As in 5.1, a mathematically precise definition of sufficiency of BCs is beyond our course and my knowledge . There is one very famous case however , which needs explicit mention : **if *V* be any volume and *S* be its bounding surface , and if *u* satisfy a Laplace equation inside *V*, then a well-defined PDE problem for *u* can be created by prescribing the value of *u* everywhere on *S***. In physics textbooks this is sometimes called the theorem of uniqueness . Here , since any well-defined PDE problem has a unique solution , we will view it more as a condition for well-definition of the Laplace equation problem . This condition has a striking application in the solution of electromagnetic problems , which we'll see when the topic comes up (6.3).

A second important property of Laplace functions is the **averaging principle**. If u satisfies a Laplace equation in two dimensions, then the average value of u over any arbitrary circle in the plane equals the value of u at the centre of the circle. Likewise if u satisfies Laplace in three dimensions then its average over any arbitrary sphere in space equals its value at the centre of the sphere. This immediately implies that ***u* cannot have a local maxima or minima anywhere** inside its domain of definition. In the 3D case, suppose u has a maximum u_{max} at the internal point P. Then, by definition, $u < u_{max}$ at every neighbouring point, which includes a sufficiently small sphere centred at P. The average of u over this sphere must necessarily be less than u_{max}, which contradicts the averaging principle. So, all maxima and minima of Laplace functions must occur **at the boundaries** of the domain of definition, where it is impossible to draw a sphere including all neighbouring points. Although this property sounds like a technicality, it is actually the bane of the life of the fledgeling magnetic levitation industry. To see why this is so, again we'll have to wait for the electromagnetism Unit.

That's it for general properties of Laplace functions, now let's actually see what they look like.

Example 6.1.1

Lollipop Laplace equation

The function $u(x,y)$ satisfies the Laplace equation inside the box x belongs to $[0, a]$ and y belongs to $[0, b]$. The boundary conditions are $u = \sin(3\pi x/a)$ on the wall $y = b$ and $u = 0$ on the three other walls.

Determine $u(x,y)$.

[Read Only]

Although this is Read Only, if you are feeling up to a challenge, then try extending the concepts of the last Chapter to get to the answer here. It's not particularly easy but it's very good practice for handling an unfamiliar PDE in the exam hall. (Then why is the problem a lollipop? because this is the easiest that Laplace equation can get.)

The equation is

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=0 \quad ,$$

no formula sheets for Cartesian ! As usual we try separation of variables, writing $u = X(x)Y(y)$. Plugging, dividing by XY and invoking the canonical sep var argument,

$$\frac{X_{xx}}{X}=-\frac{Y_{yy}}{Y}=\lambda \quad .$$

For Laplace equation, the subscript notation for derivatives is the most convenient even if it appears cumbersome – it minimizes the chances of error.

As usual, partion into cases. If $\lambda=-k^2$ then the solutions are $X=(\cos,\sin)kx$ and $Y=\mathrm{e}^{\pm ky}$. If $\lambda=k^2$ then X and Y reverse roles – X becomes exponential and Y becomes trig. Finally if $\lambda=0$, X can be 1 and x and Y can be 1 and y. The general solution is an arbitrary linear combination of all these possibilities, so we have

$$\begin{aligned} &u(x,y)\\ &=(A_0+B_0x)(C_0+D_0y)+\sum_k\left[\begin{array}{l}(A_k\cos kx+B_k\sin kx)\left(C_k\,\mathrm{e}^{ky}+D_k\,\mathrm{e}^{-ky}\right)+\\ \left(E_k\,\mathrm{e}^{kx}+F_k\,\mathrm{e}^{-kx}\right)(G_k\cos ky+H_k\sin ky)\end{array}\right] \quad . \end{aligned}$$

This looks infinitely worse than the general solution of the 1+1 heat equation.

Two factors are responsible for this. Firstly, heat equation is first order in time, so the $T(t)$ threw up one arbitrary constant which could be absorbed into the two of X. Here however there's second derivative both ways so two constants in X and two more in Y, nothing gets absorbed into the other. Secondly, for heat equation we could get rid of all positive exponentials because they grew in time. Here since the domain of definition is finite, the positive exponentials are as legitimate as the negative ones. Because of the cumbersome nature of the general solution, I'll not write it again and won't encourage you to write it also.

Rather, this time around we'll see how the BCs whittle things down. Thereafter we'll only keep the form of the general solution in the back of our minds and use the BCs to go straight to the eigenfunction expansion. The thing to note is that in the terms inside the summation, it's the SAME k sitting inside the arguments of the trigs on one variable and the exponentials on the other.

Now invoke the BCs on the messy general solution. In general, homogeneous BCs are easier than non-zero ones, since these come over from the PDE to the separated ODE. So we start from the homogeneous BCs. Clearly, $u = 0$ whenever $x = 0$ and whenever $x = a$. Since Y cannot be the zero function, this means that $X(0) = X(a) = 0$. So X satisfies the BVP $X'' - \lambda X = 0$ with double zero BCs. This is a very familiar situation – the cases $\lambda > 0$ and $\lambda = 0$ both vanish and the third case yields $k = n\pi/a$ with eigenfunctions $\sin kx$. So we have the eigenfunction expansion

$$u(x,y) = \sum_n \sin\frac{n\pi x}{a}\left(C_n e^{n\pi y/a} + D_n e^{-n\pi y/a}\right) \quad , \qquad (01)$$

where I have absorbed the B's into the C's and D's.

There's a third homogeneous BC in the problem – $u = 0$ for $y = 0$. This doesn't immediately simplify the form (01) but works wonders if we rewrite the exponentials as hyperbolics. So let's do that first and get our new eigenfunction expansion :

$$u(x,y) = \sum_n \sin\frac{n\pi x}{a}\left(C_n \cosh\frac{n\pi y}{a} + D_n \sinh\frac{n\pi y}{a}\right) \quad .$$

The C_n and D_n above are different from the ones in (01), so technically I should have used a different letter. But when it's an unknown constant being juggled around, I usually prefer to keep the same letter to reduce clutter. Now set $y = 0$ here; since $\cosh 0 = 1$, C_n must be zero and we have a further reduction

$$u(x,y) = \sum_n \sin\frac{n\pi x}{a}\left(D_n \sinh\frac{n\pi y}{a}\right) \quad .$$

Finally, use the fourth BC – set $y = b$ to get

$$u(x,b) = \sum_n D_n \sinh\frac{n\pi b}{a}\sin\frac{n\pi x}{a} = \sin\frac{3\pi x}{a} \quad .$$

The thing $D_n \sinh(n\pi b/a)$ is a CONSTANT – let's call it E_n – which multiplies the sines in x; this is a standard Fourier series problem for the E_n's. As with the heat equation, the first problem of the Chapter makes it really easy – only the $n = 3$ term is present in the series so $D_3 \sinh(3\pi b/a) = 1$ and $D_n = 0$ for all other n. So we finally have the specific solution

$$u(x,y) = \frac{1}{\sinh(3\pi b/a)} \sin\frac{3\pi x}{a} \sinh\frac{3\pi y}{a} \quad , \qquad *$$

which is the desired answer.

*

COMMENTS : As we can see, the solution process is very similar to that of 5.1.2 (comparing the two solutions line by line might in fact be an instructive exercise). The primary new addition here was the selection of trig vs exponential functions . The trigs came in the x direction because that was the direction with a double zero. If on the other hand the problem had defined $u = 0$ at $y = 0$ and $y = b$ then we would have had to select trigs for y. Barring that, the process is the same – more work because of the second derivative on both variables but no conceptual tricks. But what happens if neither direction has a double zero – say all four walls of the box have non-zero u ? That case can be easy, can be tricky – you'll see both kinds in the next exampel.

Example 6.1.2

Variations on the 2D box

In each of the three questions below, $u(x,y)$ satisfies Laplace equation inside a given box. The box itself and the BCs are different in the three cases. Determine $u(x,y)$ in each case.

(a) x belongs to $[0, a]$, y belongs to $[0, b]$; $u_x(0,y) = u_x(a,y) = 0$, $u_y(x,0) = 0$ and $u_y(x,b) = 2\cos(3\pi x/a) + \cos(4\pi x/a)$

(b) x belongs to $[0, a]$, y belongs to $[0, b]$; $u(0,y) = 0$, $u(a,y) = 1$, $u(x,0) = 0$, $u(x,b) = \sin(3\pi x/a)$. Hint : Exploit the last example and the linearity of Laplace equation to solve this problem.

(c) x, y belong to $[-1, 1]$; $u(-1,y) = 1-y^2$, $u(1,y) = 1-y^2$, $u(x,-1) = x^2-1$, $u(x,1) = x^2-1$. Hint : The answer is a very simple function.

[Level : 3]

(a) This is very similar to the last Tute, only difference is that the BCs involve the derivatives of u. Keeping the general solution in the back of our minds we proceed to the BC fitting step. There are two homogeneous BCs in x, so $X_x(0)$ and $X_x(a)$ must both equal zero. From 4.1.2 Part (a), the eigenfunctions X can be constant or $\cos(n\pi x/a)$. The corresponding Y's are $(1,\ y)$ and $(\cosh,\ \sinh)$ $n\pi y/a$. Attaching these to the X's, we have our eigenfunction expansion

$$u(x,y) = C_0 + D_0 y + \sum_n \cos\frac{n\pi x}{a}\left(C_n \cosh\frac{n\pi y}{a} + D_n \sinh\frac{n\pi y}{a}\right) \quad . \qquad (01)$$

The next two BCs are on u_y, so take the partial derivative of the above w.r.t y. We have

$$u_y(x,y) = D_0 + \sum_n \cos\frac{n\pi x}{a}\left(\frac{n\pi}{a} C_n \sinh\frac{n\pi y}{a} + \frac{n\pi}{a} D_n \cosh\frac{n\pi y}{a}\right) \quad .$$

At $y=0$, $u_y = 0$. Immediately, D_0 and all D_n's go away. Now fit the last BC :

$$u_y(x,b) = \sum_n \cos\frac{n\pi x}{a}\left(\frac{n\pi}{a} C_n \sinh\frac{n\pi b}{a}\right) = 2\cos\frac{3\pi x}{a} + \cos\frac{4\pi x}{a} \quad .$$

Again, the thing inside the brackets, although it looks complicated, is a number and not a function. So we have brought it down to Fourier series problem.

Once again, the question has made it easy – the series consists of two obvious terms. So we have $(3\pi C_3/a)\sinh(3\pi b/a) = 2$, $(4\pi C_4/a)\sinh(4\pi b/a) = 1$ and $C_n = 0$ for all other n. Plugging these C_3 and C_4 into the eigenfunction expansion (01) gives

$$* \qquad u(x,y) = \frac{2a}{3\pi\sinh(3\pi b/a)}\cos\frac{3\pi x}{a}\cosh\frac{3\pi y}{a} + \frac{a}{4\pi\sinh(4\pi b/a)}\cos\frac{4\pi x}{a}\cosh\frac{4\pi y}{a} \quad , \qquad *$$

which is the desired answer. In this last step, please be careful and plug the C_n's into the eigenfunction expansion for u, rather than the expression for u_y !

(b) Here is a problem which has $u \neq 0$ on more than one wall of the box – the top and right walls in this case. The logic by which we so far selected the trigs in one direction now falls through. The hint tells us that we should exploit linearity, and

that the previous Tute is relevant. Indeed, the BC on the top is the same as it was there – the extra is the BC on the right.

The moment you see linearity, think of sum – linearity means that the sum of two solutions is also a solution. This property forms the backbone of this entire 750 page book – we could write the general solution as sums of fundamental solutions, specific solutions as sums of homogeneous and particular solutions, PDE solutions as sums of eigenfunctions etc etc etc. Here, what is the relevant sum ? The only possible choice is, split u into u_1+u_2 where u_1 satisfies the BCs of the previous problem and u_2 satisfies what's left over. So let u_1 go with zero on left, right and bottom and the given BC on the top. What's left for u_2 is zero on the left and bottom, the non-zero thing on the right and again ZERO on the top. So with this partioning, we have managed to generate two Laplace functions, each of which has a double zero BC in one direction.

We already know what u_1 is so let's hack it on u_2. Keeping the general solution in the back of our minds, we proceed to the BC fitting step. Two zeroes in y, so $Y(y)$ must be $\sin(n\pi y/b)$ and the eigenfunction expansion is

$$u_2(x,y)=\sum_n \sin\frac{n\pi y}{b}\left(C_n\cosh\frac{n\pi x}{b}+D_n\sinh\frac{n\pi x}{b}\right) \quad . \qquad (01)$$

Now fit the BCs in x. At $x=0$, $u=0$ so C_n must be zero for all n. Finally, setting $x=a$,

$$u_2(a,y)=\sum_n D_n\sinh\frac{n\pi a}{b}\sin\frac{n\pi y}{b}=1 \quad .$$

Letting $E_n=D_n\sinh(n\pi a/b)$, we recognize the Fourier series problem for E_n. We can either do it by exploiting eigenfunction orthonormality or by performing an antisymmetric extension in y and replicating periodically. The main point is that the eigenfunctions are the SAME as for the 1+1 heat equation, only with y in place of x, so the procedure carries over completely. The specific function of the BC, we have already handled in 5.4.6, so now we copy the coefficient formulae from there – we have $E_n=4/n\pi$ if n is odd and zero if n is even. So, D_n becomes

$$D_n=\frac{4}{n\pi\sinh(n\pi a/b)} \quad ,$$

if n is odd and zero if n is even. Plugging this into (01),

$$u_2(x,y) = \sum_{\text{odd } n} \frac{4}{n\pi \sinh(n\pi a/b)} \sinh\frac{n\pi x}{b} \sin\frac{n\pi y}{b} \quad .$$

Finally, to get $u(x,y)$ we add to this function the u_1 we found in the last example. Doing this we get

$$u(x,y) = \frac{1}{\sinh(3\pi b/a)} \sin\frac{3\pi x}{a} \sinh\frac{3\pi y}{a} + \sum_{\text{odd } n} \frac{4}{n\pi \sinh(n\pi a/b)} \sinh\frac{n\pi x}{b} \sin\frac{n\pi y}{b}$$

which is the desired answer.

(c) Although this looks like a tedious extension of the above part, where we now have to split u into four functions each satisfying one non-zero BC, the hint gives us pause. For one, if that was the path to the solution, the question would not be giving a hint. Secondly, an eigenfunction expansion of $1-y^2$ on one wall does not sound like a simple proposition; four such serieses, one corresponding to each BC, will hardly add up to a "very simple function".

Instead, we think outside the box. $1-y^2$ on $x=\pm 1$ and x^2-1 on $y=\pm 1$ seem to suggest the function x^2-y^2. Take its Laplacian : $2-2=0$. So this function satisfies Laplace equation and also satisfies the BCs on all the walls. Since the problem is well-defined, if x^2-y^2 is A solution, it must also be THE solution. Hence $\underline{u(x,y)=x^2-y^2}$.

*

COMMENTS : The three parts of this problem show you three different things. Part (a) handles BCs specified in terms of derivatives. Part (b) shows how you can exploit superposition to crack boxes with more than one non-zero walls. In these problems, the locations of the zero walls were such that either the cosh or sinh terms dropped out on inspection. If the walls are located elsewhere, then we get a pair of simultaneous equations for C_n and D_n which can be solved quite easily.

Finally, Part (c) shows a case where sep var doesn't go too far. It's not that the method fails – but if you can actually get to $u = x^2 - y^2$ from the eigenfunction expansions, then you are probably headed for a career in mathematics. The guessing method is non-generic, but is extremely common and useful in physics and engineering. Sometimes when faced with a real life problem, the only practical option is to first guess the solution and then show that it is true. This question also shows the complications unique to PDEs – a systematic solution method is given, you crank it diligently and fail to recover what really is a very simple function. It turns out that sep var in cylindrical polar yields this as a fundamental solution, but who would guess that ?

Example 6.1.3

Laplace in 3D

The function $u(x,y,z)$ satisfies Laplace equation in the region x belongs to $[0, L_1]$, y belongs to $[0, L_2]$ and z belongs to $[0, \infty]$. The BCs are $u(0,y,z) = u(L_1,y,z) = 0$, $u(x,0,z) = u(x,L_2,z) = 0$, $u(x,y,0) = 1$ and $u(x,y,\infty) = 0$.

Determine $u(x,y,z)$.

[Read Only]

Our first PDE with three independent variables – technically, this is one of the harder problems you can expect to face in the final exam. This is not CONCEPTUALLY difficult though – it's the same old sep var and eig exp now expanded into three dimensions. Since this is the archetypal problem in many-variable PDEs, I will recommend that you read it very carefully and understand the concepts thoroughly.

The equation is of course

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}+\frac{\partial^2 u}{\partial z^2}=0 \quad .$$

As a first step we can see that u has been prescribed everywhere on the boundary of the volume where it is defined. The point $z=\infty$ is a perfectly legitimate boundary for this purpose, so much so that we'll use the notation $u(x,y,\infty)=0$ rather than the more pedantic $u(x,y,z)\to 0$ as $z\to\infty$. This might give mathematicians apoplexy – let it. So the problem we have is well-defined. Now separate variables, writing $u=X(x)Y(y)Z(z)$. Plugging and dividing by XYZ,

$$\frac{X_{xx}}{X}+\frac{Y_{yy}}{Y}+\frac{Z_{zz}}{Z}=0 \quad .$$

Transferring the Z terms over to the RHS, a function of z alone equals a function of x and y alone. So Z_{zz}/Z must equal a constant, say μ. Then, a function of x alone plus a function of y alone equals μ; if we carry the y-term over to the RHS then a function of x alone becomes equal to a function of y alone (the μ can be a function of either x or y, as per our convenience). So X_{xx}/X and Y_{yy}/Y must be constants also [we already saw how to start separating a 3-variable PDE in T076 Part (b)]. This immediately suggests the general rule – in a many-variable PDE, the moment you bring it to the form that a function of the first independent variable alone plus a function of the second alone, plus one of the third alone etc. equals zero or a constant, then you can say that each individual function itself equals a constant.

This time however there are TWO independent constants, not just one. X_{xx}/X can be some λ_1, Y_{yy}/Y can be some other λ_2 and that will force Z_{zz}/Z to equal $-\lambda_1-\lambda_2$. Corresponding to every possible sign of λ_1 and λ_2 we'll get either trigs or exponentials or linear functions in each of x, y, and z – counting all the possibilities makes for an excellent exercise in combinatorics. More relevant for

us is that we shouldn't even try to write the general solution. Rather, keeping the concepts trig, exponential and linear in the back of our minds, we proceed to the BC fitting step.

The first two BCs imply $X(0)YZ$ and $X(L_1)YZ$ both equal zero – since YZ isn't the zero function, we have a double zero in X. This function now satisfies 4.1.1 . [4.1.1 surely takes the crown for example with maximum number of internal citations]. We know what this means : $\lambda_1 = -(n_1\pi/L_1)_2$ and X eigenfunctions are $\sin(n_1\pi x/L_1)$. Similarly, $Y(0)$ and $Y(L_2)$ are again both zero : $\lambda_2 = -(n_2\pi/L_2)_2$ and Y eigenfunctions are $\sin(n_2\pi x/L_2)$. Here n_1 and n_2 are both natural numbers but NOT the same natural number – the two λ's are completely independent of each other ! Writing the same n for both cases is the biggest source of error among students on such a problem , so be very careful here. $X = \sin(2\pi x/L_1)$ does not in any way force $Y = \sin(2\pi y/L_2)$: Y can be $\sin(8\pi y/L_2)$, $\sin(24\pi y/L_2)$ whatever it wishes.

Given these, we now have the ODE for $Z(z)$

$$Z_{zz} = \pi^2\left(\frac{n_1^2}{L_1^2} + \frac{n_2^2}{L_2^2}\right)Z \quad ,$$

which has the solutions

$$Z = \exp\left(\pm\pi\left[\frac{n_1^2}{L_1^2} + \frac{n_2^2}{L_2^2}\right]^{1/2} z\right) \quad . \qquad (01)$$

The BCs are $Z(0)$ equals something while $Z(\infty) = 0$. (Another potential trap : going with the flow, you might be tempted to write the first BC as $Z(0) = 1$. That's completely false – $u(x,y,0) = 1$ means $X(x)Y(y)Z(0) = 1$, and there's nothing to suggest that the product XY equals 1 on the plane $z = 0$. Transferring a BC from the PDE to the separated ODE only works if the RHS of the BC has the value zero !)

Long back, in 4.1.2 part (c), we have already seen how to handle such a BC (this setup by the way is what motivated that problem). The exponentially growing term goes away and the decaying one survives. So the eligible Z's are (01) with the negative sign inside the exponential . Now we cobble all this together into our eigenfunction expansion . Since there are two independent sets of eigenvalues n_1 and n_2, this expansion now becomes a double summation, and

the arbitrary constants in the expansion become functions of both n_1 and n_2. We have

$$u(x,y,z) = \sum_{n_1}\sum_{n_2} D_{n_1,n_2}\left(\sin\frac{n_1\pi x}{L_1}\right)\left(\sin\frac{n_2\pi y}{L^2}\right)\exp\left(-\pi\left[\frac{n_1^2}{L_1^2}+\frac{n_2^2}{L_2^2}\right]z\right) \quad , \qquad (02)$$

and the most important waypoint in the path to the solution has been reached. As I said while doing heat equation, in an exam it is essential that you clearly write and label the eigenfunction expansion – a correct eig exp is bound to fetch significant partial credit provided that the examiner can see it and recognize it.

The last step is the determination of the coefficients D_{n_1,n_2}. For this we use the BC which hasn't entered into the picture so far – $u(x,y,0)=1$. Setting $z=0$ into (02) we have

$$u(x,y,0) = \sum_{n_1}\sum_{n_2} D_{n_1,n_2}\left(\sin\frac{n_1\pi x}{L_1}\right)\left(\sin\frac{n_2\pi y}{L_2}\right)=1 \quad . \qquad (03)$$

This is a double Fourier series problem; although we can solve it by a two-dimensional extension, it is easier in this case to exploit orthonormality of eigenfunctions. We have already seen that the functions $v_{n_1}\sin(n_1\pi x/L_1)$ with $v_{n_1}=\sqrt{(2/L_1)}$ are orthonormal over $[0, L_1]$. Likewise, $\kappa_{n_2}\sin(n_2\pi x/L_2)$ with $\kappa_{n_2}=\sqrt{(2/L_2)}$ are orthonormal over $[0, L_2]$. So let's rewrite (03) as

$$\sum_{n_1}\sum_{n_2} E_{n_1,n_2}\left(v_{n_1}\sin\frac{n_1\pi x}{L_1}\right)\left(\kappa_{n_2}\sin\frac{n_2\pi y}{L_2}\right)=1 \quad , \qquad (04)$$

and then take dot product with the orthonormal eigenfunctions on the both sides.

Doing it first for y, we consider the dot product of (04) with $\kappa_m\sin(m\pi y/L_2)$ for some arbitrary integer m :

$$\int_0^{L_2} dy\left(\kappa_m\sin\frac{m\pi y}{L_2}\right)\sum_{n_1}\sum_{n_2} E_{n_1,n_2}\left(v_{n_1}\sin\frac{n_1\pi x}{L_1}\right)\left(\kappa_{n_2}\sin\frac{n_2\pi y}{L_2}\right)$$
$$= \int_0^{L_2} dy\; 1\cdot\left(\kappa_m\sin\frac{m\pi y}{L_2}\right)$$

In the above LHS we can take the summation on x outside the integration on y; that gives

$$\sum_{n_1} v_{n_1} \sin\frac{n_1\pi x}{L_1}\left[\int_0^{L_2} \mathrm{d}y\ \left(\kappa_m \sin\frac{m\pi y}{L_2}\sum_{n_2} E_{n_1,n_2}\kappa_{n_2}\sin\frac{n_2\pi y}{L_2}\right)\right]$$
$$= \int_0^{L_2} \kappa_m \sin\frac{m\pi y}{L_2}\ \mathrm{d}y$$

In this LHS, the thing inside the first sum is now in standard form. The integral evaluates to E_{n_1,n_2} for $m = n_2$ and zero for all m other than n_2. So

$$\sum_{n_1} E_{n_1,n_2} v_{n_1} \sin\frac{n_1\pi x}{L_1} = \int_0^{L_2} \kappa_{n_2} \sin\frac{n_2\pi y}{L_2}\ \mathrm{d}y \quad .$$

After this, the next step is obvious – take dot product with $v_m \sin(m\pi x/L_1)$ on the both sides (m again arbitrary). That gives

$$\sum_{n_1} E_{n_1,n_2} \int_0^{L_1} \mathrm{d}x\ \left(v_m \sin\frac{m\pi x}{L_1}\right) v_{n_1} \sin\frac{n_1\pi x}{L_1}$$
$$= \int_0^{L_1} \mathrm{d}x \int_0^{L_2} \mathrm{d}y\ v_m \kappa_{n_2} \sin\frac{m\pi x}{L_1}\sin\frac{n_2\pi y}{L_2}$$

Standard setup on the LHS – m equals n_1 and sum equals E_{n_1,n_2}. Hence we have

$$E_{n_1,n_2} = \int_0^{L_1} \mathrm{d}x \int_0^{L_2} \mathrm{d}y\ v_{n_1}\kappa_{n_2} \sin\frac{n_1\pi x}{L_1}\sin\frac{n_2\pi y}{L_2} \quad .$$

Now evaluate the integrals – stripped of the factors v and κ it's a two-dimensional square wave which gives $4/\pi^2 n_1 n_2$ if both n_1 and n_2 are odd and zero in all other cases. So $E_{n_1,n_2} = 4v_{n_1}\kappa_{n_2}/\pi^2 n_1 n_2$ if n_1 and n_2 are both odd.

Finally, since the eigenfunction expansion is in terms of the D's, we need the D's rather than the E's. Comparing (03) and (04) we see that $D_{n_1,n_2} = v_{n_1}\kappa_{n_2}E_{n_1,n_2}$. Using the values of v and κ we got long back, $D_{n_1,n_2} = 16/\pi^2 n_1 n_2 L_1 L_2$ if both n's are odd and zero otherwise. Plugging this into (02) now gives the specific solution

$$u(x,y,z) =$$
$$\sum_{\text{odd } n_1}\ \sum_{\text{odd } n_2} \frac{16}{\pi^2 n_1 n_2 L_1 L_2}\left(\sin\frac{n_1\pi x}{L_1}\right)\left(\sin\frac{n_2\pi y}{L_2}\right)\exp\left(-\pi\left[\frac{n_1^2}{L_1^2}+\frac{n_2^2}{L_2^2}\right]z\right) \quad *$$

which is the desired answer.

The calculation after the eigenfunction expansion (02) became something of a tour de force so let's step back and re-assess what exactly happened. The final result was the natural and intuitive extension of what happens in one less dimension. There, if we work in terms of an arbitrary BC/IC function $f(x)$, the coefficient works out to

$$D_n = \frac{2}{L}\int_0^L f(x)\sin\frac{n\pi x}{L}\,\mathrm{d}x \quad .$$

This is the integral of f and the eigenfunction over the appropriate region, times the square of the normalization constant (recall that $v_n = \sqrt{(2/L)}$ for all n). Here, it's the same thing : generalizing the BC function '1' to $f(x,y)$ we have

$$D_{n_1,n_2} = \frac{2}{L_1}\frac{2}{L_2}\int_0^{L_1}\mathrm{d}x\int_0^{L_2}\mathrm{d}y\, f(x,y)\sin\frac{n_1\pi x}{L_1}\sin\frac{n_2\pi x}{L_2} \quad .$$

This is again the integral of f and the eigenfunctions over the appropriate region times the squares of the normalization constants involved. Although I could have skipped two pages of calculation and written down this formula as "an intuitive extension of the easier case", you could possibly have challenged it. Now you see once and for all how the thing works; then, when you face it next time, just write down the expression which looks and feels – and is – right.

Now we'll turn to the linear algebraic representation of 5.1.3 to gain some more insight into the eigenfunction expansion. Writing functions as vectors is more abstract than writing them explicitly but the notation is very compact, and the concept comes through with a minimum of fuss. The general coefficient - determination problem has the following form : we are given an expression like

$$f(x,y) = \sum_{n_1,n_2} C_{n_1,n_2} X_{n_1}(x) Y_{n_2}(y) \quad , \qquad (05)$$

where f is an arbitrary function, X's and Y's are orthogonal eigenfunctions and we want to find the coefficients C_{n_1,n_2}.

The first thing is, what does a 2-dimensional function look like in linear algebraic notation ? The 1D case was simple – a function is just a vector of its values defined at certain lattice sites. In 2D, there are lattice sites in both x and y, so

the discretization of the space becomes a mesh or matrix, looking something like the following :

$$\begin{bmatrix} (x_1,y_1) & (x_1,y_2) & (x_1,y_3) & (x_1,y_4) \\ (x_2,y_1) & (x_2,y_2) & (x_2,y_3) & (x_2,y_4) \\ (x_3,y_1) & (x_3,y_2) & (x_3,y_3) & (x_3,y_4) \\ (x_4,y_1) & (x_4,y_2) & (x_4,y_3) & (x_4,y_4) \end{bmatrix} .$$

A two-variate function $f(x,y)$ will also be a matrix $\mathbf{F}$ with entries corresponding to the values of f at the appropriate lattice sites. The individual eigenfunctions $X_{n_1}(x)$ and $Y_{n_2}(y)$ are of course orthogonal vectors $\mathbf{u}_{n_1}$ and $\mathbf{v}_{n_2}$. The product $X_{n_1}Y_{n_2}$ is their MATRIX PRODUCT $\mathbf{u}_{n_1}\mathbf{v}^T_{n_2}$. To see this, lets consider a toy example. We have, for two vectors $\mathbf{u}$ and $\mathbf{v}$,

$$\mathbf{u}\mathbf{v}^T = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} \begin{bmatrix} v_1 & v_2 & v_3 & v_4 \end{bmatrix} = \begin{bmatrix} u_1v_1 & u_1v_2 & u_1v_3 & u_1v_4 \\ u_2v_1 & u_2v_2 & u_2v_3 & u_2v_4 \\ u_3v_1 & u_3v_2 & u_3v_3 & u_3v_4 \\ u_4v_1 & u_4v_2 & u_4v_3 & u_4v_4 \end{bmatrix} .$$

This is exactly the product XY defined at the lattice sites which I listed above. Note that this product is like the 'reverse' of the dot product – that has the first vector transposed while this has the second.

With this in our hand, we can express (05) in linear algebra form as follows : given

$$\mathbf{F} = \sum_{n_1,n_2} c_{n_1,n_2}\mathbf{u}_{n_1}\mathbf{v}^T_{n_2} \quad , \tag{06}$$

with $\mathbf{F}$ arbitrary and $\mathbf{u}$'s, $\mathbf{v}$'s orthogonal, determine c_{n_1,n_2}.

As usual, start from the orthogonality relations. Let v_{n_1} and κ_{n_2} be the usual normalization constants for $\mathbf{u}_{n_1}$ and $\mathbf{v}_{n_2}$. Then define $\mathbf{w}_{n_1} = v_{n_1}\mathbf{u}_{n_1}$ and $\mathbf{z}_{n_2} = \kappa_{n_2}\mathbf{v}_{n_2}$ to be unit vectors so that (06) becomes

$$\mathbf{F} = \sum_{n_1,n_2} \frac{c_{n_1,n_2}}{v_{n_1}\kappa_{n_2}} \mathbf{w}_{n_1}\mathbf{z}^T_{n_2} \quad .$$

Now consider arbitrary natural numbers m and n and multiply the both sides above with $\mathbf{u}^T_m$ on the left and $\mathbf{v}_n$ on the right :

$$\mathbf{u}_m^T \mathbf{F} \mathbf{v}_n = \sum_{n_1,n_2} \frac{c_{n_1,n_2}}{v_{n_1}\kappa_{n_2}} \mathbf{u}_m^T \mathbf{w}_{n_1} \mathbf{z}_{n_2}^T \mathbf{v}_n$$
$$= \sum_{n_1,n_2} \frac{c_{n_1,n_2}}{v_{n_1}^2\kappa_{n_2}^2} \mathbf{w}_m^T \mathbf{w}_{n_1} \mathbf{z}_{n_2}^T \mathbf{z}_n \quad .$$

On the RHS, start evaluating the 4-term product from the right – the last two terms are the dot product of $\mathbf{z}_n$ and $\mathbf{z}_{n_2}$ which is unity if $n = n_2$ and zero otherwise. The first two terms are the dot product of $\mathbf{w}_m$ and $\mathbf{w}_{n_1}$ which is unity if $m = n_1$ and zero otherwise. The whole summation on the RHS therefore reduces to a single non-trivial term; equating this to the LHS gives

$$\mathbf{u}_m^T \mathbf{F} \mathbf{v}_n = \frac{c_{mn}}{v_m^2 \kappa_n^2} \quad .$$

In other words, the coefficients c_{n_1,n_2} in (06) are given by

$$c_{n_1,n_2} = v_{n_1}^2 \kappa_{n_2}^2 \mathbf{u}_{n_1}^T \mathbf{F} \mathbf{v}_{n_2} \quad .$$

Finally, we come back from (06) to (05). As we increase the grid refinement, the product $\mathbf{u}^T\mathbf{F}\mathbf{v}$ (I'm letting go of the subscripts n_1, n_2 for clarity) increases quadratically with the number of lattice sites. To make the limiting process tractable, we must multiply by h^2 where h is the step size. But $h^2\,\mathbf{u}^T\mathbf{F}\mathbf{v}$ is nothing but the discrete form of the double integral of $f(x,y)X(x)Y(y)$ over the interval – as we take the limit, it becomes the integral itself. Thus, the coefficients in the formula (05) are

$$C_{n_1,n_2} = v_{n_1}^2 \kappa_{n_2}^2 \int_0^{L_1} \mathrm{d}x \int_0^{L_2} \mathrm{d}y\, f(x,y) X_{n_1}(x) Y_{n_2}(y) \quad \text{, where}$$

$$v_{n_1} = \frac{1}{\left(\int_0^{L_1} \left[X_{n_1}(x) \right]^2 \mathrm{d}x \right)^{1/2}} \qquad \kappa_{n_2} = \frac{1}{\left(\int_0^{L_2} \left[Y_{n_1}(y) \right]^2 \mathrm{d}y \right)^{1/2}} \quad .$$

This formula takes us from any arbitrary two-dimensional orthogonal eigenfunction expansion to BC/IC fit and completes the solution.

This is the natural extension of what happened in 5.1.3, and the extension of this formula to still higher dimensions should by now be obvious (its proof however isn't – to handle 3D for instance using this linear algebra approach, we'll have to

deal with tensors of rank 3!). Technique-oriented books on PDEs – such as anything written by PHILIP MORSE – go blue in the face writing arrays of double, triple and sextuple sums and integrals. But behind every devilish expression lies this kind of reasoning – eigenfunction expansion followed by exploitation of orthonormality to evaluate the constants.

What guarantees orthonormality (and for that matter completeness) of eigenfunctions ? For 1D we could use the matrix method to get at it – for higher dimensions, the matrices are not at all simple to construct. Fortunately, in the classes of problems we'll see here, the eigenfunctions will be the same as they were in the 1D case, and we can carry over the results from there.

Finally, you might wonder about the preferential selection of the exponentially decaying solution in (01) – wasn't this just the sort of thing I was railing at in all of Chapter 2 – that even one growing solution makes the whole system unstable or the model unacceptable ? Yes I was, and I still will. Dynamical systems are INITIAL value problems where you really need a very special IC as in 2.1.5 to preferentially select the decaying solution over the intractable one. You can't generate such an IC except on a sheet of paper . These however are BOUNDARY value problems where you CAN (or rather Nature can) select one of them to satisfy the BC at infinity . So there 's no contradiction here – dynamical systems are fundamentally different from problems defined in space.

Example 6.1.4

Cylindrical polar

The function $u(\rho,\theta,z)$ satisfies Laplace equation.

(a) Separate $u = R(\rho)\,\Theta(\theta)\,Z(z)$ and write ODEs satisfied by R, Θ and Z. These should feature arbitrary separation constants.

(b) Obtain the eigenfunction expansion in the case that u is independent of z and there are no boundaries in the θ-direction.

(c) Consider the function $u(\rho,\theta)$ defined in the unit disk ρ belongs to [0, 1] and θ periodic with period 2π. The BC is $u(1,\theta) = \delta(\theta)$ [Dirac delta function]. Obtain $u(\rho,\theta)$ in the form of a series.

(d) Sum the series obtained above and hence determine $u(\rho,\theta)$ explicitly.

[Level : 3]

(a) The review gives us the equation :

$$\frac{1}{\rho}\frac{\partial}{\partial\rho}\left(\rho\frac{\partial u}{\partial\rho}\right)+\frac{1}{\rho^2}\frac{\partial^2 u}{\partial\theta^2}+\frac{\partial^2 u}{\partial z^2}=0 \quad .$$

Opening out the ρ terms,

$$\frac{\partial^2 u}{\partial\rho^2}+\frac{1}{\rho}\frac{\partial u}{\partial\rho}+\frac{1}{\rho^2}\frac{\partial^2 u}{\partial\theta^2}+\frac{\partial^2 u}{\partial z^2}=0 \quad .$$

Now plug in $u=R\Theta Z$ and divide by $R\Theta Z$ to get

$$\frac{R_{\rho\rho}}{R}+\frac{1}{\rho}\frac{R_\rho}{R}+\frac{1}{\rho^2}\frac{\Theta_{\theta\theta}}{\Theta}+\frac{Z_{zz}}{Z}=0 \quad .$$

The first three terms together are some function of ρ and θ alone; the fourth is a function of z alone. Hence these two entities must be constants and we have

$$\frac{Z_{zz}}{Z}=\lambda_1 \quad ,$$

$$\frac{R_{\rho\rho}}{R}+\frac{1}{\rho}\frac{R_\rho}{R}+\frac{1}{\rho^2}\frac{\Theta_{\theta\theta}}{\Theta}+\lambda_1=0 \quad ,$$

for some unknown constant λ_1.

For the next step, we would like a function of ρ alone plus a function of θ alone to equal zero; that way we can pull out the two parts. Multiplying out by ρ^2 we have

$$\rho^2\frac{R_{\rho\rho}}{R}+\rho\frac{R_\rho}{R}+\lambda_1\rho^2+\frac{\Theta_{\theta\theta}}{\Theta}=0 \quad .$$

The θ term has now separated away from the ρ terms, so we set $\Theta_{\theta\theta}/\Theta$ equal to some second unknown constant λ_2 (Laplace in 3D Cartesian had two independent separation constants, so this had better have two also – the equations can change but the number of constants can't). That gives

$$\frac{\Theta_{\theta\theta}}{\Theta}=\lambda_2 \quad ,$$

$$\rho^2\frac{R_{\rho\rho}}{R}+\rho\frac{R_\rho}{R}+\lambda_1\rho^2+\lambda_2=0 \quad .$$

The equation has now separated fully.

Collecting what we have got and writing it neatly,

$$\rho^2\frac{\mathrm{d}^2R}{\mathrm{d}\rho^2}+\rho\frac{\mathrm{d}R}{\mathrm{d}\rho}+\left(\lambda_1\rho^2+\lambda_2\right)R=0 \quad , \qquad (01a)$$

$$\frac{\mathrm{d}^2\Theta}{\mathrm{d}\theta^2}-\lambda_2\Theta=0 \quad , \qquad (01b)$$

$$\frac{\mathrm{d}^2Z}{\mathrm{d}z^2}-\lambda_1 Z=0 \quad , \qquad (01c)$$

which are the ODEs asked for.

(b) In the case that u is independent of z, the function $Z(z)$ is actually a constant (it's not ZERO, that would make u zero). So $Z_{zz}=0$, λ_1 evaluates to zero, and (01c) goes away. Now, take on the Θ equation (01b). Since there are no boundaries in θ, physical plausibility dictates that u must be periodic in θ with period 2π. (Note that this assertion would NOT remain true if we had BCs like say $u(\rho,0)=0$ and $u(\rho,\pi/2)=0$ and wanted to find u inside the 90° sector – it's a complete mistake to apply this assumption, and the resulting eigenfunction expansion, to such situations.) Seeing the future in a crystal ball, we'd already anticipated this situation in 4.1.2 Part (b) – the answer is $\lambda_2=-n_2$ with eigs $\Theta=(\cos,\sin)n\theta$ and $\lambda_2=0$ with eig $\Theta=\text{const}$.

Now substitute these values of λ_1 and λ_2 into (01a). We have

$$\rho^2\frac{\mathrm{d}^2R}{\mathrm{d}\rho^2}+\rho\frac{\mathrm{d}R}{\mathrm{d}\rho}-n^2R=0 \quad \text{, for } \lambda_2\neq 0 \text{ and} \qquad (02a)$$

$$\rho^2\frac{\mathrm{d}^2R}{\mathrm{d}\rho^2}+\rho\frac{\mathrm{d}R}{\mathrm{d}\rho}=0 \quad \text{, for } \lambda_2=0. \qquad (02b)$$

We recognize the Euler Cauchy equations from N°. 10 and try $R=\rho^\alpha$ (using either r or λ for the exponent would cause enormous confusion). For (02a) this yields $\alpha=n$ and $\alpha=-n$, giving the radial eigenfunctions $R=\rho^{\pm n}$. For (02b), the root $\alpha=0$ is repeated twice, so the solutions are $R=1$ and $R=\log\rho$. Multiplying the R's by the corresponding Θ's, attaching an arbitrary constant to every possible combination and adding them all up gives the eigenfunction expansion

$$u(\rho,\theta)=A_0+B_0\log\rho+\sum_n\left(A_n\rho^n+\frac{B_n}{\rho^n}\right)\left(C_n\cos n\theta+D_n\sin n\theta\right) \quad ,$$

which is the desired answer.

(c) We start from the eigenfunction expansion above. Since u is defined inside the unit disk, it must be well-behaved at the origin. This kills B_0 and all the B_n's. What remains is

$$u(\rho,\theta) = A_0 + \sum_n \left(\rho^n\right)\left(C_n \cos n\theta + D_n \sin n\theta\right) \quad , \tag{03}$$

where we've absorbed A_n into C_n and D_n. Setting $\rho = 1$ into this expression evidently identifies C_n and D_n as the Fourier coefficients of the delta function.

Now let's do the Fourier expansion. The limits of integration in the coefficient formulas can be anything other than 0 since the delta is right there. Let's take them to be $-\pi$ and π as is conventional. We then have

$$A_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} \delta(\theta)\ \mathrm{d}\theta = \frac{1}{2\pi} \quad ,$$

$$C_n = \frac{1}{\pi}\int_{-\pi}^{\pi} \delta(\theta)\cos n\theta\ \mathrm{d}\theta = \frac{1}{\pi} \quad ,$$

$$D_n = \frac{1}{\pi}\int_{-\pi}^{\pi} \delta(\theta)\sin n\theta\ \mathrm{d}\theta = 0 \quad .$$

We could also have got this last line from the fact that the delta is a limit of the symmetric rectangle functions as we saw in L03. Note that all the Fourier coefficients of delta are of the same size – a pathological extension of the relation between continuity properties and asymptotic size of the Fourier coefficients.

Plugging these relations into (03) yields the series expression

* $$u(\rho,\theta) = \frac{1}{2\pi} + \frac{1}{\pi}\sum_n \rho^n \cos n\theta \quad ,$$ *

which is the desired answer.

(d) For the first time in this book, the series is simple enough for us to sum. We write $\cos n\theta$ as $\mathrm{Re}\, \mathrm{e}^{\mathrm{j}n\theta}$; then the series becomes

$$u(\rho,\theta) = \frac{1}{2\pi} + \frac{1}{\pi}\mathrm{Re}\sum_n \rho^n\, \mathrm{e}^{\mathrm{j}n\theta} \quad .$$

The thing inside the summation is a GP; the common ratio is $\rho e^{j\theta}$ which has a magnitude strictly less than unity since we are working inside the disk where $\rho < 1$. Hence this series can be summed. A slight convenience comes about if we split the $1/2\pi$ outside as $-1/2\pi + 1/\pi$; the latter then becomes the first term of the GP which looks like

$$u(\rho,\theta) = -\frac{1}{2\pi} + \frac{1}{\pi}\mathrm{Re}\left(1 + \sum_n \left(\rho e^{j\theta}\right)^n\right) \quad .$$

Now we sum the series and get

$$u(\rho,\theta) = -\frac{1}{2\pi} + \frac{1}{\pi}\mathrm{Re}\frac{1}{1-\rho\cos\theta - j\rho\sin\theta} \quad .$$

Finally, we use the fact that $\mathrm{Re}\,(1/[A+jB])$ is $A/(A^2 + B^2)$; this followed by elementary trig gives

$$u(\rho,\theta) = \frac{1}{\pi}\left(-\frac{1}{2} + \frac{1-\rho\cos\theta}{1-2\rho\cos\theta+\rho^2}\right) \quad , \qquad (04)^*$$

which is the desired answer.

*

COMMENTS : The formula (04) is called POISSON's solution to the Laplace equation. If we locate the delta function at some arbitrary θ' instead of 0 and replace θ with $\theta - \theta'$ in (04) then we get a GREEN's function which can yield an integral for u inside the disk for arbitrary BC on the periphery. It's also noteworthy that (04) is one of the few problems where the series bunches up into a closed form expression without our having to do anything fancy. This final result doesn't look anything like a separable function $R(\rho)\,\Theta(\theta)$ or a sum of separable functions and hence acts as another example (after T085) showing that sep var can yield solutions which are more complex than the starting ansatz.

Example 6.1.5

Spherical polar

The function $u(r,\theta,\varphi)$ satisfies Laplace equation.

(a) Separate $u = R(r)\,\Theta(\theta)\,\Phi(\varphi)$ and write ODEs satisfied by R, Θ and Φ. These should feature arbitrary separation constants.

(b) Obtain the solutions in the cases where u has the following dependences on the angles of depression and azimuth : (i) independent of both, (ii) $\cos\theta$, (iii) $\sin\theta\cos\varphi$ or $\sin\theta\sin\varphi$.

[Level : 2]

(a) The review gives us :

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right)+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial u}{\partial\theta}\right)+\frac{1}{r^2\sin^2\theta}\frac{\partial^2 u}{\partial\varphi^2}=0 \quad .$$

$1/r^2$ is a factor, so let's pull it out and open out the derivatives :

$$r^2\frac{\partial^2 u}{\partial r^2}+2r\frac{\partial u}{\partial r}+\frac{\partial^2 u}{\partial\theta^2}+\frac{\cos\theta}{\sin\theta}\frac{\partial u}{\partial\theta}+\frac{1}{\sin^2\theta}\frac{\partial^2 u}{\partial\varphi^2}=0 \quad .$$

The last term looks like it's hanging onto the rest of the stuff by a thin thread – to snap that link we multiply by $\sin^2\theta$ throughout. Then of course substitute in $u=R(r)\,\Theta(\theta)\,\Phi(\varphi)$ and divide by $R\Theta\Phi$. This gives

$$r^2\sin^2\theta\frac{R_{rr}}{R}+2r\sin^2\theta\frac{R_r}{R}+\sin^2\theta\frac{\Theta_{\theta\theta}}{\Theta}+\cos\theta\sin\theta\frac{\Theta_\theta}{\Theta}+\frac{\Phi_{\varphi\varphi}}{\Phi}=0 \quad .$$

The φ term indeed separates away from the rest, and we have

$$\frac{\Phi_{\varphi\varphi}}{\Phi}=\lambda_1 \quad ,$$

$$r^2\sin^2\theta\frac{R_{rr}}{R}+2r\sin^2\theta\frac{R_r}{R}+\sin^2\theta\frac{\Theta_{\theta\theta}}{\Theta}+\cos\theta\sin\theta\frac{\Theta_\theta}{\Theta}+\lambda_1=0 \quad .$$

λ_1 is of course an arbitrary separation constant.

To shake loose the r terms, we divide out the $\sin^2\theta$ we'd introduced before :

$$r^2\frac{R_{rr}}{R}+2r\frac{R_r}{R}+\frac{\Theta_{\theta\theta}}{\Theta}+\frac{\cos\theta}{\sin\theta}\frac{\Theta_\theta}{\Theta}+\frac{\lambda_1}{\sin^2\theta}=0 \quad .$$

Objective achieved : the first two terms are function of r alone and last three terms are function of θ alone. So let them equal $\pm\lambda_2$. This done, we have our separated equations

* $$r^2\frac{\mathrm{d}^2R}{\mathrm{d}r^2}+2r\frac{\mathrm{d}R}{\mathrm{d}r}+\lambda_2 R=0 \quad , \qquad (01a)^*$$

* $$\frac{\mathrm{d}^2\Theta}{\mathrm{d}\theta^2}+\frac{\cos\theta}{\sin\theta}\frac{\mathrm{d}\Theta}{\mathrm{d}\theta}+\left(\frac{\lambda_1}{\sin^2\theta}-\lambda_2\right)\Theta=0 \quad , \qquad (01b)^*$$

* $$\frac{\mathrm{d}^2\Phi}{\mathrm{d}\varphi^2}-\lambda_1\Phi=0 \quad , \qquad (01c)^*$$

which are the ODEs asked for.

(b) When u is independent of both θ and φ, the functions Θ and Φ are constants and λ_1 and λ_2 both equal zero. In this case, (01a) becomes

$$r^2 \frac{\mathrm{d}^2 R}{\mathrm{d}r^2} + 2r \frac{\mathrm{d}R}{\mathrm{d}r} = 0 \quad .$$

This is Euler Cauchy; $R = r^\alpha$ yields $\alpha = 0$ and $\alpha = -1$. Hence the solutions for u are $u = 1/r$ and $u = \text{const}$.

When u goes as $\cos\theta$, the function Φ is constant and $\lambda_1 = 0$. To obtain λ_2, we plug in $\Theta = \cos\theta$ into (01b); this reduces it to

$$-\cos\theta + \frac{\cos\theta}{\sin\theta}\left(-\sin\theta\right) - \lambda_2 \cos\theta = 0 \quad ,$$

or $\lambda_2 = -2$. Setting this into (01a),

$$r^2 \frac{\mathrm{d}^2 R}{\mathrm{d}r^2} + 2r \frac{\mathrm{d}R}{\mathrm{d}r} + -2R = 0 \quad ,$$

which is Euler Cauchy with solutions r and $1/r^2$. Hence the solutions for u are $u = r\cos\theta$ and $u = (1/r^2)\cos\theta$ in this case.

When u goes as $\sin\theta\cos\varphi$, $\Phi = \cos\varphi$. Plugging into (01c) shows $\lambda_1 = -1$. Evidently the same holds true if $\Phi = \sin\varphi$. Setting this into (01b) gives

$$-\sin\theta + \frac{\cos^2\theta}{\sin\theta} - \frac{1}{\sin\theta} - \lambda_2 \sin\theta = 0 \quad ,$$

and the most basic trig identity now gives $\lambda_2 = -2$ again. Hence R is again r or $1/r^2$ and the solutions are $u = r\sin\theta\cos\varphi$ and $u = (1/r^2)\sin\theta\cos\varphi$ (and their equivalents with $\sin\varphi$ in place of $\cos\varphi$) in this case.

*

COMMENTS : This and last example literally rammed into the boundary wall of DELTA – the limit of technical knowledge which we won't cross . The simplifying assumptions – two -dimensionalization in one case and near - trivial angular dependences in the other – were only so that the ODEs remained within the scope of our syllabus . It's not that the more general cases produce ill-behaved solutions or anything . Rather , those solutions feature special functions – Bessel functions of ρ in the cylindrical setup and spherical harmonics $Y_l^m(\theta,\varphi)$ in the spherical.

Nevertheless, these special cases convey a lot of information. In both cylindrical and spherical cases we see that the radial equation admits two kinds of solutions – one a positive power and one a negative power. The former is zero at zero radius and blows up at large radius, while the latter goes the other way. In physical problems, we can't have any blow up. So, to fit a region which includes the origin but doesn't go to infinity, we need the positive powers [as in 6.1.4 (c)], while to fit a region which excludes the origin but extends to large distances, we need the negative powers. These basic properties hold true for the fancier special function solutions also. An exception however is the 2D cylindrical solution corresponding to $\lambda_2 = 0$; the two options here are constant and $\log \rho$. This latter term is trouble at both zero and infinity – it turns out that although it is forbidden at $\rho = 0$, it is in fact permissible at $\rho \to \infty$.

So when do these simple solutions work ? For the cylinder, it works if the height is substantially greater than the radius – then the variation of physical quantities along the height is negligible. Theoretically, we would need a height which is order of magnitude bigger than radius – practically, we can relax the constraint a lot and still get away with it. For the sphere, it works only if the physical situation features such limited dependence on the angles. In fact, the three angular distributions $\text{const.}, \cos\theta$ and $\sin\theta (\cos, \sin)\varphi$ correspond to the mono-pole, electric dipole and magnetic dipole respectively, as you'll see a little later. You might ask, how did I even get these three simple solutions ? I got them from the special functions theory which is outside our scope. However, to just understand these solutions we don't really need to go into that theory – the fancy stuff becomes necessary when we have a complicated situation and need to exploit orthogonality etc. But it turns out that we can go quite some distance with these simple solutions alone.

Finally, I had said in the comments to 6.1.2 that the function $x_2 - y_2$ comes out naturally in cylindrical coordinates. How ? It's the eigenfunction $\rho^2 \cos 2\theta$; we can write $\rho^2 \cos 2\theta = \text{Re}\left(\rho^2 \, \text{e}^{\text{j}2\theta}\right) = \text{Re}\left(\left[\rho \, \text{e}^{\text{j}\theta}\right]^2\right) = \text{Re}\left(\left[x + \text{j}y\right]^2\right) = x^2 - y^2$. All the 'internal' eigenfunctions $\rho^n \cos n\theta$ and $\rho^n \sin n\theta$ are actually simple polynomials in x and y. But they aren't even close to eig in Cartesian – a rather odd fact.

6.2

MANY PLUS ONE HEAT EQUATION

Here is the generalization of the heat equation to include more than one spatial dimension. This time I won't do the derivation in detail but more or less state the changes which occur. FOURIER's law now states that the heat current is proportional to the gradient of the temperature, thus $\mathbf{q} = -\kappa \nabla u$. Taking into account the effect of sources inside the region, the heat equation becomes

$$\rho C \frac{\partial u}{\partial t} - \kappa \nabla^2 u = S(x, y, z, t) \quad ,$$

where S, we recall, is the power per unit volume developed by the source. Here the LHS is in coordinate free form but the RHS makes explicit references to the coordinates of the source. To remove this nuisance feature, we specify the location of the source point in terms of the vector from the origin to the point in question. This **position vector** is typically denoted by **r**. So we can write

$$\rho C \frac{\partial u}{\partial t} - \kappa \nabla^2 u = S(\mathbf{r}, t) \quad ,$$

which is the final form of the heat equation. This formula is the same in two and three space dimensions.

In steady state problems, $\partial u / \partial t = 0$ and S is also a function of **r** alone. In source-free regions, u satisfies a Laplace equation. Indeed, steady state temperature is often used as an example of a variable which satisfies Laplace equation. Transient problems of course require the full dynamic equation for their solution.

Example 6.2.1

2+1 Heat equation

The temperature $u(x,y,t)$ of the box x,y belong to $[0, \pi]$ satisfies the 2+1 heat equation

$$4\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}\right)=\frac{\partial u}{\partial t} \quad .$$

(a) If the BCs are that all four walls are maintained at $u=0$ and the IC is $u(x,y,0)=10$, determine $u(x,y,t)$.

(b) In the above solution, describe the shape of $u(x,y,t)$ at moderately large times.

(c) The BCs now are that the four walls are insulated. Write these conditions mathematically.

(d) If the IC is $u(x,y,0)=10$, again determine $u(x,y,t)$.

Suppose now that Part (a) is used to model the temperature of an infinitely long rod whose cross-section is the square of size π. The rod is immersed in water at temperature $u=0$. A simpler model of the rod can use the 1+1 heat equation

$$\frac{\partial u}{\partial t}-\frac{\partial^2 u}{\partial x^2}=-\beta u \quad ,$$

where β represents the lateral leakage term.

(e) Determine the value of β which is most appropriate and comment of the accuracy of the simplified model.

[Level : 3]

(a) The starting step is no surprise – separate the variables. Substituting $u = X(x)Y(y)T(t)$ and dividing by XYT we have

$$\frac{X_{xx}}{X} + \frac{Y_{yy}}{y} - \frac{1}{4}\frac{\dot{T}}{T} = 0 \quad .$$

Each of the three terms here is a function of a different variable so they are all equal to constants. Let $X_{xx}/X = \lambda_1$, $Y_{yy}/Y = \lambda_2$, then $\dot{T}/T$ becomes $4(\lambda_1+\lambda_2)$. The physical plausibility condition (no temporally growing solutions) forces $\lambda_1+\lambda_2$ to be negative but it doesn't reveal anything about the signs of each of them separately. So the general solution can admit quite a lot of possibilities as regards X and Y – trig trig, trig exponential, trig linear etc. Once again we keep it in the back of our minds and proceed to the BC fitting step.

The BCs are the same as in 6.1.3 – double zero in both directions. So both X and Y must be trigs. We can easily see that $X = \sin n_1 x$ and $Y = \sin n_2 y$ for natural numbers n_1 and n_2. This forces $\lambda_1 = -n_1^2$ and $\lambda_2 = -n_2^2$, whereby the T equation becomes

$$\frac{\mathrm{d}T}{\mathrm{d}t} = -4\left(n_1^2 + n_2^2\right)T \quad ,$$

with solution

$$T = \exp\left(-4\left[n_1^2 + n_2^2\right]t\right) \quad .$$

This now leads to the eigenfunction expansion

$$u(x,y,t) = \sum_{n_1}\sum_{n_2} D_{n_1,n_2}(\sin n_1 x)(\sin n_2 y)\exp\left(-4\left[n_1^2 + n_2^2\right]t\right) \quad .$$

The last step is the fitting of the IC. Setting $t = 0$ we get

$$u(x,y,0) = \sum_{n_1}\sum_{n_2} D_{n_1,n_2}\sin n_1 x \sin n_2 y = 10 \quad .$$

This is the same double Fourier series as we got in 6.1.3; here $L_1 = L_2 = \pi$ and $f(x,y) = 10$ instead of 1. Copying the results of that problem we have $D_{n_1,n_2} = 160/\pi^4 n_1 n_2$ if both n's are odd and zero otherwise. Hence

$$u(x,y,t)$$

$$= \sum_{\text{odd } n_1} \sum_{\text{odd } n_2} \frac{160}{\pi^2 n_1 n_2} (\sin n_1 x)(\sin n_2 y) \exp\left(-4\left[n_1^2 + n_2^2\right]t\right) \quad , \qquad *$$

which is the desired answer.

(b) The maxim is smallest k slowest decay, and we have to extend it here to one extra dimension. Looking at the exponential, the slowest decay term corresponds to $n_1 = n_2 = 1$. So at moderately large time, u looks like

$$u(x,y,t) \sim \sin x \sin y \, \mathrm{e}^{-8t} \quad , \qquad *$$

which is the desired answer. In this type of problem, be careful not to let even one of the n's be zero – that corresponds to the trivial solution !

(c) An insulated wall means that the heat current cannot cross it. By FOURIER's law, $\mathbf{q} = -\nabla u$, and it is tempting to write $\nabla u = 0$ on all four walls of the box. This however is wrong ! The insulation only blocks the current from flowing ACROSS the wall – it does nothing to a current flowing ALONG the wall. So, only the x-component of current is zero on the walls $x = 0$ and $x = \pi$. Only the y-component is zero on the walls $y = 0$ and $y = \pi$. This gives $\underline{u_x(0,y) = u_x(\pi,y) = 0}$ and $\underline{u_y(x,0) = u_y(x,\pi) = 0}$.

(d) We can of course embark on another eigenfunction expansion. But in this case, physical argument gives the answer in one line. We know that heat diffusion tends to spread out the temperature all over the body. Since the box is completely insulated, the initial heat cannot escape from it but can only redistribute itself inside. Just as in 5.1.4, the steady state must be $u = \text{const.}$ all over the box with the constant being equal to the average of the IC. But in this case, the IC itself is a constant – there is no redistribution to take place and system is in steady state right from the start. Hence the answer is $\underline{u(x,y,t) = 10}$.

(e) The 2+1 heat equation is a good model for an infinitely long rod since we expect its temperature to be independent along the length. The comparison model is 1+1 with leakage – let's solve that for a rod of length L first and extend it to infinity at the end. Since the entire rod is immersed in water, the natural BCs to impose on the 1+1 problem are $u(0) = u(L) = 0$. Separating variables we have

$$X'' = \lambda X \qquad \dot{T} = T(\lambda - \beta) \quad .$$

The hackneyed double zero BCs give the same old eigenvalues $\lambda = -n^2\pi^2/L^2$; plugging this into the T equation yields that all T's are exponentially decaying with rate constant $\lambda - \beta$. Now we take the limit $L \to \infty$; in this limit the smallest eigenvalue becomes zero and the smallest decay constant becomes β. We have to match this to the 2+1 model; for a good fit, the asymptotic decay rates must be the same so that the long time solutions agree. We found in Part (b) that the slowest decay constant for the 2+1 setup is 8, so we must have $\underline{\beta = 8}$.

The approximation here is very similar to what we did in T092 (NEWTON's Rule of cooling). Here we threw one dimension away from a 2+1 PDE while there we converted a 1+1 PDE into an ODE. Just as in there, we can't expect the simpler model to capture everything – there's no way the 1+1 model can give us the higher modes of the 2+1 model. Nevertheless, the leading order behaviour is often the most important and in a realistic situation the reduction of dimension (especially if the rod has a weird cross-sectional shape) can make the difference between getting an approximate solution and making no headway at all.

*

COMMENTS : Looking at the answer to Part (a), the primary difference with the corresponding solutions of 1+1 problems is the presence of DEGENERACY – the phenomenon that more than one spatial eigenfunction can show the same temporal behaviour. In the 1+1 case, every spatial eigenfunction had its own characteristic time evolution but here that is no longer true. For example, $\sin x \sin 2y$ and $\sin 2x \sin y$ both evolve as e^{-20t}. At larger n_1 and n_2 there are more space functions sharing the same temporal function – $n_1 = 5$ $n_2 = 5$, $n_1 = 1$ $n_2 = 7$, and $n_1 = 7$ $n_2 = 1$ all evolve as e^{-200t}. For the heat equation, this degeneracy has very little physical significance since the modes are all decaying. However, in wave equation and Schroedinger equation, where the modes are oscillatory rather than decaying, these degeneracies are often very important. We also saw that the lowest mode (the "smallest k") was unique, no degeneracy there – that too is a more or less universal feature of all multidimensional PDE systems.

With this problem, this Unit is over. Although this was just one example, after Chapter 5 I really feel that we have had enough of problems on heat conduction.

6.3

STATIC AND DYNAMIC ELECTRICITY

The Unit title draws its name from the book by WILLIAM RALPH SMYTHE – a PDE tour de force. Majority of the problems there feature functions I have barely heard of – Bessels of complex modulus and complex argument, parabolic cylinders of the n^{th} kind and the like. Another special feature of the book is that the double triple quadruple eigenfunction expansions are often summed explicitly to get closed form final answers. Viewed from one perspective, SMYTHE starts off where we leave off – our problems will for him be child's play. But only in a way. With a little imagination, our simple functions will carry us a long distance; our tag-rag band of cosines, sines, powers and exponentials will not exactly run out of fuel before reaching top of climb.

Of course, this Unit is not a whole book on electromagnetism, so it will not teach you everything about electric and magnetic fields ab initio. Rather, the way we'll approach the topic is as follows. We'll treat the PDEs governing these fields as the core of the subject. Then, we'll accept certain physical facts as given and use those together with our PDE skills to solve problems.

Electromagnetism deals with two field variables – the electric field **E** and the magnetic field **B**. A field is defined as any quantity which permeates space (and optionally varies in time) – it can be a scalar or a vector. For example, temperature u is a scalar field. **E** and **B** are vector fields. **E** has as its signature the force it exerts on a point charge : $\mathbf{F} = q\mathbf{E}$ where q is the charge. **B** has as signature the force it exerts on a moving charge : $\mathbf{F} = q(\mathbf{v} \times \mathbf{B})$ where **v** is the velocity of the charge. Electric fields are created by static charges q (and more generally charge densities ρ) and by time-dependent magnetic fields. Magnetic fields are created by electric currents i (and more generally current densities **J**) and by time-dependent electric fields. The PDEs governing this creation process are :

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon_0} \quad , \qquad \textbf{(01a)*}$$

$$\nabla \times \mathbf{E} = -\frac{\partial}{\partial t}\mathbf{B} \quad , \qquad \textbf{(01b)*}$$

$$\nabla \cdot \mathbf{B} = 0 \quad , \qquad \textbf{(01c)*}$$

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \varepsilon_0 \frac{\partial}{\partial t}\mathbf{E} \quad . \qquad \textbf{(01d)*}$$

Here $\varepsilon_0 = 8.85\text{x}10^{-12}$ and $\mu_0 = 1.2\text{x}10^{-6}$ SI Units are the permittivity and permeability of free space – just two pesky constants. These values make the equations hold true in vacuum (or air, or any other medium which is insensitive to electric and magnetic fields). If there are materials present which respond to these fields, a few complications arise which we'll see later.

The four equations (01) are fundamental and inviolable laws of Nature. In this form, they are due to Sir **JAMES CLERK MAXWELL** and are known as MAXWELL's equations (technically one should have called them CLERK MAXWELL's equations since Sir JAMES apparently treated the whole thing as his last name – anyway we'll let that be). In other forms however, most of these principles were discovered earlier by other people. Even among laws of Nature, these are among the most inflexible – there's no regime of operation in the known universe (small size, high speed etc.) where they don't hold true. Let's now get a little more into the details.

ELECTROSTATICS

This subfield concerns itself with (01a,b) in the case that the fields are independent of time. $\nabla \times \mathbf{E}$ then becomes zero. Since curl of a gradient is zero, we can write **E** as the gradient of some scalar function, which we call the **potential** V. Convention has it to attach a minus sign also; thus, we define $\mathbf{E} = -\nabla V$. Now substitute this into (01a) and recall that divergence of grad is Laplacian; this gives

$$\nabla^2 V = -\frac{\rho}{\varepsilon_0} \quad ,$$

which is an inhomogeneous Laplace equation for V. (This equation is also called POISSON's equation – the use of different names for the homogeneous and inhomogeneous forms of the same equation might create confusion so

I'll go with "inhomogeneous Laplace".) In regions where there are no charges, potential satisfies the (homogeneous) Laplace equation. If we can find the potential using PDE techniques, we can then easily find **E** by taking its gradient.

Physical plausibility implies that V is continuous. If V has a discontinuity then **E** will become delta function, which is unacceptable. We note that the electric field doesn't change if we make a transformation $V \rightarrow V +$ const. since the gradient of any constant is zero. The potential energy of a point charge q in an electric field V is qV. To make physical sense of this, the potential far away from a finite-sized charge distribution is taken to be $V = 0$ – the charge has no potential energy when far away from all other charges. This assumption doesn't always hold however for infinite charge distributions.

A perfect electric **conductor** is a material which doesn't allow an electric field to develop inside. If exposed to an external **E**, charges redistribute themselves inside the conductor in such a way that the net **E** becomes zero everywhere. Since **E** is grad of V, it implies that V inside a conductor is constant. A **grounded** conductor is one where the potential is zero. It often enforces BCs on electrostatic problems. A **dielectric** is a material which strengthens the electric field inside. If exposed to an external **E**, charges redistribute themselves inside the dielectric in such a way that the net **E** is stronger than the originally applied value. If the dielectric is linear, the strengthening is by a constant factor.

A grounded conductor is an obvious BC on an electrostatic PDE – a more subtle one comes from a sheet of charge carrying some surface charge density σ. (In an unfortunate degeneracy, σ is used both for surface charge density and material conductivity. The convention is so deep-rooted that a change will only confuse more; the saving grace is that we don't have to deal with conductivities in electrostatic problems.) This σ has the dimensions of charge per unit area and in general is a function of position on the sheet. Then, at every point on the sheet, the components of electric field parallel and perpendicular to the sheet ($\mathbf{E}_{\parallel}$ and $E_{\perp}$ respectively – note that $\mathbf{E}_{\parallel}$ has two components so is a vector while $E_{\perp}$ has only the one component and is a scalar) satisfy certain BCs which we derive below.

Applying the divergence theorem on **E**, the flux of **E** over a closed surface is the integral of its divergence over the enclosed volume. Then from (01a), this

integral is proportional to the total charge contained in the volume. This is called GAUSS' law – the flux of **E** over the bounding surface is $1/\varepsilon_0$ times the total charge contained in the enclosed volume. (A side application of GAUSS' law is in predicting the electric fields of some singularly easy configurations.) Now consider a rectangular parallelopiped straddling the charge sheet – the top surface is above the sheet and parallel to it, the bottom surface is below the sheet and the lateral surfaces pass through the sheet. The dimensions of this box are small (infinitesimal) so that it covers 'only a point' on the sheet (i.e. a region where **E** is constant for all practical purposes). But among them, the dimension perpendicular to the sheet is smaller still, by orders of magnitude compared to the other two dimensions. Then, the flux of **E** through the box comes only from the top and bottom surfaces – it is $E_{\perp,\text{top}}A$ through the top and $-E_{\perp,\text{bot}}A$ through the bottom, where A is the cross-sectional area of the top and bottom of the parallelopiped, and the signs account for the fact that the area vector is always defined outward to the surface. From GAUSS' law this equals $1/\varepsilon_0$ times the total charge enclosed in the parallelopiped, which is σA. Thus, we get the BC

$$E_{\perp,\text{top}} - E_{\perp,\text{bot}} = \frac{\sigma}{\varepsilon_0} \quad .$$

For the BC on the parallel component, we start from the curl theorem. Since $\nabla \times \mathbf{E} = 0$, the line integral of **E** over any closed curve is zero. Now consider a rectangle straddling the sheet – on the top side it runs along whatever is $\mathbf{E}_{||}$ there; then it turns 90°, penetrates the sheet and comes over to the bottom, then again turns 90° and runs in the reverse of what it did on the top, and finally penetrates the sheet to return to its starting point. Like the box, all dimensions of this rectangle are small but the height is much smaller than the length. Then, the contributions to $\oint \mathbf{E} \cdot \mathrm{d}\mathbf{l}$ come only from the top and bottom. Their signs are clearly opposite; since their resultant is zero, $\mathbf{E}_{||}$ must be the same on both sides of the sheet.

I can hear an objection – $\mathbf{E}_{||}$ is a two-component vector, so suppose on the bottom it has a component perpendicular to the side of the rectangle ? This component would not get picked up during our cute little argument. To quash this objection, consider a second rectangle with the long side oriented along

this new component i.e. perpendicular to the original rectangle, and calculate $\oint \mathbf{E} \cdot d\mathbf{l}$ here. The top of this rectangle won't contribute – we chose the original rectangle to be oriented along $\mathbf{E}_{||}$ there, so the dot product will be zero by definition. The height won't contribute also because it's small. The only contribution will come from the bottom – since the total is zero, this mystery component of $\mathbf{E}_{||}$ must be zero also.

The rather fancy argumentation above has a simple bottom-line. Whenever we have a sheet carrying surface charge, the following hold :

* $$E_{\perp,\text{top}} - E_{\perp,\text{bot}} = \frac{\sigma}{\varepsilon_0} \quad , \qquad \textbf{(02a)*}$$

$$\mathbf{E}_{||,\text{top}} - \mathbf{E}_{||,\text{bot}} = 0 \quad . \qquad \text{(02b)}$$

In words, the normal component of **E** jumps by the surface charge density while the parallel component is continuous across the sheet. Generally, the second of these conditions is satisfied automatically by the continuity of *V*; it is the first one which is a new condition and is of great utility. Since we derived the BC (02a) basis of GAUSS' law (01a) alone, it is equally valid when the situation is dynamic. On the other hand, (02b) appears to change if there's a time-dependent magnetic field around. However, unless there's an infinite **B** involved (which is absurd), an infinitesimal loop can only enclose a negligible amount of $-\partial\mathbf{B}/\partial t$ so (02b) holds true also. Thus, these BCs are valid not just for static situations but **dynamic situations also**, a very important fact.

An easy calculation (one of the Tutes) shows that the field of a point charge *q* located at the origin of coordinates is

$$\mathbf{E} = \frac{1}{4\pi\varepsilon_0}\frac{q}{r^2}\hat{\mathbf{r}} \quad .$$

This is called COULOMB's law. An **electric dipole** consists of two point charges, equal in magnitude and opposite in sign, separated by a small distance. An ideal dipole has the strength of each charge going to infinity and the separation between them to zero, in such a way that their product is finite. This product is called the electric dipole moment **p**. It is a vector, being oriented from the negative to the positive charge. The potential of a dipole located at the origin is

$$V = \frac{1}{4\pi\varepsilon_0}\frac{\mathbf{p}\cdot\hat{\mathbf{r}}}{r^2} \quad ,$$

and its electric field is

$$\mathbf{E} = \frac{1}{4\pi\varepsilon_0}\frac{3(\mathbf{p}\cdot\hat{\mathbf{r}})\hat{\mathbf{r}} - \mathbf{p}}{r^3} \quad . \qquad \textbf{(03)*}$$

Although the point dipole may seem like a completely artificial construct, it turns out that the electric field of an arbitrary charge distribution can be expressed in a series form as a sum of a field due to an equivalent point charge, plus a field due to an equivalent dipole plus higher order terms. This is called **multipolar expansion**. It is also a surprising fact that a sphere carrying a surface charge density varying as $\cos\theta$ produces an external field which is identical to that of a point dipole placed at its centre.

The force exerted on a point dipole by an electric field $\mathbf{E}$ is $\mathbf{F} = (\mathbf{p}\cdot\nabla)\mathbf{E}$. This means that $F_{x,y,z} = p_x(\partial E_{x,y,z}/\partial x) + p_y(\partial E_{x,y,z}/\partial y) + p_z(\partial E_{x,y,z}/\partial z)$ where the comma-separated subscripts correspond. The potential energy of dipole in electric field is $\mathbf{p}\cdot\mathbf{E}$. A dipole in an electric field also experiences a torque given by $\mathbf{T} = \mathbf{p}\times\mathbf{E}$.

MAGNETOSTATICS IN AIR

We now concern ourselves with (01c,d). But before that, we must discuss two very important properties of electric currents. The first one – is current a scalar or a vector ? The wire current i is a scalar, while the current density $\mathbf{J}$ is a vector. $\mathbf{J}$ is obvious – it must point in the direction in which is flows. But how can i be a scalar ? This is because i is defined as $\int \mathbf{J}\cdot\mathrm{d}\mathbf{A}$ over the cross-section of the wire. To get a vector from i, we tack on the infinitesimal length $\mathrm{d}\mathbf{l}$ of the wire onto the current to get the vector $i\mathrm{d}\mathbf{l}$. The second property – the vast majority of currents flow in such a way that there is no accumulation or depletion of charge anywhere (i.e. currents flow in closed circuits). How can we express this more mathematically ? We do this by considering a volume of space into which currents are flowing in and out. A current $\mathbf{J}$ over an infinitesimal area on the bounding surface takes out an amount of charge $\mathbf{J}\cdot\mathrm{d}\mathbf{A}$ (the normal being defined positive outwards). So the total charge flowing out of the volume is $\int \mathbf{J}\cdot\mathrm{d}\mathbf{A}$ all over the bounding surface. By the

divergence theorem, this is $\nabla \cdot \mathbf{J}$ integrated over the volume. No accumulation or depletion of charge from the volume means that this integral is zero. Since this holds true for any arbitrary volume, we must have

$$\nabla \cdot \mathbf{J} = 0 \quad .$$

This property must be satisfied by all currents which flow in closed circuits. In the more general case, when there are regions where charge can accumulate or deplete, we have the conservation law

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0 \quad .$$

This is the same kind of principle as satisfied by salt in a chemical tank (example 1.2.3) and heat in a bar (5.4).

Now let's focus on the magnetic field created by a time-independent current. Since the configuration is static, we can throw out the $\partial \mathbf{E}/\partial t$ term from (01d). This time, it is the divergence rather than the curl which vanishes. Since the divergence of a curl is always zero, we can write **B** as the curl of some vector, which we call the **vector potential A**. We thus define

$$\mathbf{B} = \nabla \times \mathbf{A} \quad .$$

Now plug this into (01d) – using one of those vector identities for curl of curl we have

$$\nabla\left(\nabla \cdot \mathbf{A}\right) - \nabla^2 \mathbf{A} = \mu_0 \mathbf{J} \quad .$$

Not much help unless we can kick out the first term in the LHS. (By setting the permeability to μ_0 we are using the fact that everything is taking place in air. See the next section of this Lecture for magnetism in materials.)

But it so happens that we can always do it. The physically measurable quantity is **B** and not **A**, so if we have two different **A**'s which yield the same **B**, both are equally acceptable. But since **B** is curl of **A** and curl of gradient is zero, we can always add to **A** the gradient of some scalar function ψ and leave **B** unchanged. This is similar to the way we could add an arbitrary constant to V, only that the flexibility is much greater this time. We can exploit this freedom to choose ψ in such a way that the divergence of **A** becomes zero. (The choice of ψ is called a gauge and is the starting point for a great deal of very sophisticated physics and maths. True to form we bypass all that and focus on the fact that we can do it.) Then the above simplifies to

$$\nabla^2 \mathbf{A} = -\mu_0 \mathbf{J} \quad .$$

This is an inhomogeneous Laplace equation for every component of **A**. In **Cartesian coordinates** we can break this up into three problems $\nabla^2 A_x = -\mu_0 J_x$ etc. and solve each component separately. It is dangerous however to do this in any other coordinate system – for example, writing $\nabla^2 A_\theta = -\mu_0 J_\theta$ in cylindrical polars and solving this leads to complete garbage except probably in some fortuitious cases which are of no interest. It does generally hold true however, that if **J** is one direction (say $\hat{\boldsymbol{\theta}}$) then **A** is also in the same direction, even in non-Cartesian coordinate systems. Like *V*, **A** must be continuous to avoid unphysical delta functions in its curl.

Another way of expressing the magnetic field of a current is BIOT-SAVART's law which holds for wire currents *i*. This law states that the magnetic field at point **r** due to an infinitesimal section d**l** of wire located at the origin and carrying current *i* is

$$\mathrm{d}\mathbf{B} = \frac{\mu_0}{4\pi} \frac{i\mathrm{d}\mathbf{l} \times \mathbf{r}}{r^3} \quad . \tag{04}$$

To find the magnetic field at a point due to an entire circuit, we can integrate this law all over the circuit. Like COULOMB's law, this is not a new law but can be derived from the PDE for magnetic field. Although the integrals over the circuit are often impossible to evaluate, there are a couple of special occasions where this formulation rather than the PDE formulation is the quickest path to exact results.

The magnetic equivalent of the electrostatic charge sheet is a sheet carrying a surface current **K**. This has the dimension of current per unit length (recall : ρ is charge per volume and σ is charge per area; likewise **J** is current per area so **K** is current per length). Arguments involving boxes and rectangles straddling the sheet yield the BCs on **B** at a current sheet. This time I won't go into the details. I will however mention AMPERE's law which is the magnetic equivalent of GAUSS' – this states that the line integral of **B** over a closed curve is μ_0 times the total current enclosed within the curve. (Like GAUSS' law, AMPERE's law can be used to explicitly find magnetic fields in some near-trivial situations.) Some manipulations from this starting point lead to :

$$B_{\perp,\text{top}} - B_{\perp,\text{bot}} = 0 \quad , \tag{05a}$$

$$* \qquad \mathbf{B}_{||,\text{top}} - \mathbf{B}_{||,\text{bot}} = \mu_0 \left(\mathbf{K} \times \hat{\mathbf{n}} \right) \quad . \tag{05b)*}$$

Here, $\hat{\mathbf{n}}$ is the unit vector normal to the sheet, directed from bot to top. In words, the normal component of **B** is continuous while the parallel component jumps by an amount proportional to the surface current in the sheet. Practically, the continuity of **A** automatically ensures satisfaction of (04a); it is the second one which acts as an additional BC. Just as for (02), these BCs also remain valid if the situation is dynamic rather than static.

Magnetostatics has no 'point magnetic charge' (which is called magnetic monopole) – indeed, a whole collection of physics research exists on "what if monopoles were real". The basic unit of magnetic field generator is the **magnetic dipole**, which is a circular wire loop carrying a current. For the ideal dipole, the strength of the current goes to infinity and the area of the loop to zero in such a way that their product is finite. This product is called the magnetic dipole moment **m**. The orientation of **m** is perpendicular to the loop – positive upwards if current flow is counter-clockwise. It turns out that the magnetic field of an ideal magnetic dipole is identical to that of an ideal electric dipole (03) – we have

$$* \qquad \mathbf{B} = \frac{\mu_0}{4\pi} \frac{3\left(\mathbf{m} \cdot \hat{\mathbf{r}}\right)\hat{\mathbf{r}} - \mathbf{m}}{r^3} \quad . \tag{06)*}$$

Once again, a sphere with sinusoidal current distribution recreates perfectly the field of an ideal magnetic dipole at its centre. This derivation we'll see in one of the Tutes. A dipole in an external magnetic field experiences a potential energy of $\mathbf{m} \cdot \mathbf{B}$, a force of $\left(\mathbf{m} \cdot \nabla\right)\mathbf{B}$, and a torque of $\mathbf{m} \times \mathbf{B}$. A current wire in a magnetic field experiences a force $\mathrm{d}\mathbf{F} = i\mathrm{d}\mathbf{l} \times \mathbf{B}$. This force is the cause of the torque on a motor and the lift on a Maglev train, as we'll see in the Tutes.

MAGNETOSTATICS IN MATERIALS

Like dielectrics, there are also special magnetic materials. Paramagnets tend to make **B** stronger inside. A very strong paramagnet is called a **ferromagnet**. Diamagnets make **B** weaker inside. The limiting case of a diamagnet is a **superconductor**, which completely expels a magnetic field from inside itself. Unlike in the electrical case, I'll go into some of the details here. Skipping a

lot of the physics (see "Introduction to Electrodynamics" by DAVID GRIFFITHS for that), I will present some quick mathematical equations which we can solve for a given situation to get answers.

When a magnetic medium is kept inside an external magnetic field, currents start flowing spontaneously inside the material. These are called **bound currents**. In the most general case, there is a volume bound current $\mathbf{J}_b$ inside the bulk of the material and a surface bound current $\mathbf{K}_b$ on its surface. On the other hand, the currents which we are driving externally to create the original magnetic field are called free currents – we denote these by $\mathbf{J}_f$ and $\mathbf{K}_f$. The magnetic field created by the free currents alone is denoted by $\mathbf{H}$. Well, that was a little sloppy – $\mathbf{H}$ is something like the magnetic field created by the free currents alone, but not quite. To make things precise, I'll have to introduce the magnetization $\mathbf{M}$. The easiest way to define this is, $\mathbf{M}$ is a vector field defined inside the magnetic material such that

$$\mathbf{J}_b = \nabla \times \mathbf{M} \quad ,$$
$$\mathbf{K}_b = \mathbf{M} \times \hat{\mathbf{n}} \quad ,$$

where $\mathbf{n}$ denotes the outward-pointing unit normal vector. Evidently, $\mathbf{M}$ is zero outside the magnetic material. Then, we define the vector field $\mathbf{H}$ (which by the way has no name in words) as the solution of the equations

$$\nabla \cdot \mathbf{H} = -\nabla \cdot \mathbf{M} \quad ,$$
$$\nabla \times \mathbf{H} = \mathbf{J}_f \quad .$$

The second equation above looks a lot like (01d) except that there is no permeability in it – the first one is different however. Unlike $\mathbf{M}$, $\mathbf{H}$ is defined everywhere – both inside magnetic materials and in air.

I have introduced two new fields $\mathbf{H}$ and $\mathbf{M}$ which are primarily mathematical constructs (at least in the way we are viewing it); now I will have to relate them to the actual magnetic field $\mathbf{B}$, which is what we can actually measure. The defining relation is

$$\mathbf{B} = \mu_0 \left(\mathbf{H} + \mathbf{M} \right) \quad . \tag{07}$$

We are almost done with variable introduction. There's only one more and that one enters the picture if the material is linear. Since this book is all about linearity, it's no surprise that all materials we'll work with shall be linear. The definition of a linear material is

$$\mathbf{M} = \chi \mathbf{H} \quad ,$$

everywhere inside the material, where χ is a number, called the **magnetic susceptibility**. The value of susceptibility is a property of the material. Using this relation in (07), we have

$$\mathbf{B} = \mu_0 (1 + \chi) \mathbf{H} \quad ,$$

as the constitutive relation for a linear medium. The value of χ for air is zero – this implies that the magnetization in air is zero and $\mathbf{B} = \mu_0 \mathbf{H}$. Since $\mathbf{J}_b = \mathbf{K}_b = 0$ in this case, this is consistent with (01d). Paramagnetic materials have χ positive – the value for electric steel (ferromagnet) is about 100 or more. Diamagnetic materials have χ negative – while ordinary diamagnets have $|\chi|$ of about 10^{-5} or smaller, the superconductor has $\chi = -1$. Evidently, this implies that whatever the value of **H** inside the superconductor, **B** inside is always identically zero.

A trivial class of problem is where a current configuration is mounted in an infinite space entirely filled with magnetic medium. In this case, the magnetic field simply $1+\chi$ times what it would have been if the setup had been mounted in air. Realistic problems feature both magnetic materials and air – for example we might have a current wire in air and an iron block close by. In this case, we need the BCs at the interface between material and air. The derivation of the BCs is similar to that for the current sheet which I skipped, so I'll skip this also. The final answers are

* $$H_{\perp,\text{top}} - H_{\perp,\text{bot}} = -\left(M_{\perp,\text{top}} - M_{\perp,\text{bot}} \right) \quad , \qquad \textbf{(08a)*}$$

* $$\mathbf{H}_{||,\text{top}} - \mathbf{H}_{||,\text{bot}} = \mathbf{K}_f \times \hat{\mathbf{n}} \quad . \qquad \textbf{(08b)*}$$

These BCs are equivalent to (05), which also continue to hold in this case. Whether (05) or (08) prove more useful depends on the problem at hand. The equations and BCs we developed here are sufficient to solve any problem involving magnetic media, as we'll see in the Tutes.

Before that, a quick summary of several jargons which appear while dealing with magnetic media. The quantity $\mu_0(1+\chi)$ is called the permeability and denoted by μ. The quantity $1+\chi$ on the other hand is called the relative permeability μ_r. "An orgy of unnecessary terminology" is what GRIFFITHS calls this, but so long as there are authors who revel in it, you'll have to be familiar with it.

SLOW ELECTRODYNAMICS

There are two $\partial/\partial t$ terms in MAXWELL's equations. Consider a region which is free of sources i.e. ρ=**J**=0. Taking the curl of (01b) and plugging in (01d) we have

$$\nabla\times\left(\nabla\times\mathbf{E}\right)=-\frac{\partial}{\partial t}\left(\nabla\times\mathbf{B}\right)=-\frac{\partial}{\partial t}\left(\mu_0\varepsilon_0\frac{\partial}{\partial t}\mathbf{E}\right) \quad .$$

We just used curl of curl a while back; since the region is source-free, **E** is divergenceless and it reduces to the negative of the Laplacian. This implies

$$\nabla^2\mathbf{E}=\mu_0\varepsilon_0\frac{\partial^2\mathbf{E}}{\partial t^2} \quad .$$

This is a wave equation for **E**, which we are not yet prepared to analyse (that's the next Chapter). So we must beat a retreat from here for now. (The product $\mu_0\varepsilon_0$ by the way is $1/c^2$ where c = 29,97,92,458 m/s – 3x10^8 m/s for all practical purposes – is the velocity of light, the ultimate MPS.)

In a different regime however, we get heat equation like structures. This starts from the fact that the two $\partial/\partial t$ terms in the equations have coefficients of vastly different sizes. The one in front of **B** in (01b) is unity while the one in front of **E** in (01d) is 10^{-17} SI Units ! So if we have slow variation of **E** and **B** in time, then the term in (01b) is sizeable while the other one is negligible and can be dropped. (Since it's 10^{17} which is the comparison scale, 'slow' here can mean a few gigahertz !) Then, slowly varying **J**'s create **B**'s which are identical to their static counterparts – these **B**'s then dynamically create **E**'s. Now if the material involved is a conductor rather than air, then an electric field gives rise to a current inside – for a linear conductor, **J**=σ**E**, which is OHM's law. (A couple of pages back I said **E**=0 inside a conductor, now I'm saying **J**=σ**E** – why no contradiction ? This is because, the zero electric field is for a perfect conductor, for which $\sigma=\infty$; in this case a finite **J** is indeed supported at zero **E**. But practical conductors have only finitely high σ so they can sustain non-zero **E**'s inside.) So the dynamically created **E**'s produce more **J**'s, and the current distribution in the conductor gets substantially modified from what it would have been in the absence of the time variation. This phenomenon is called **electromagnetic induction** (of induction motor fame), and the currents created by time-varying magnetic field are called **induced currents** or **eddy currents**.

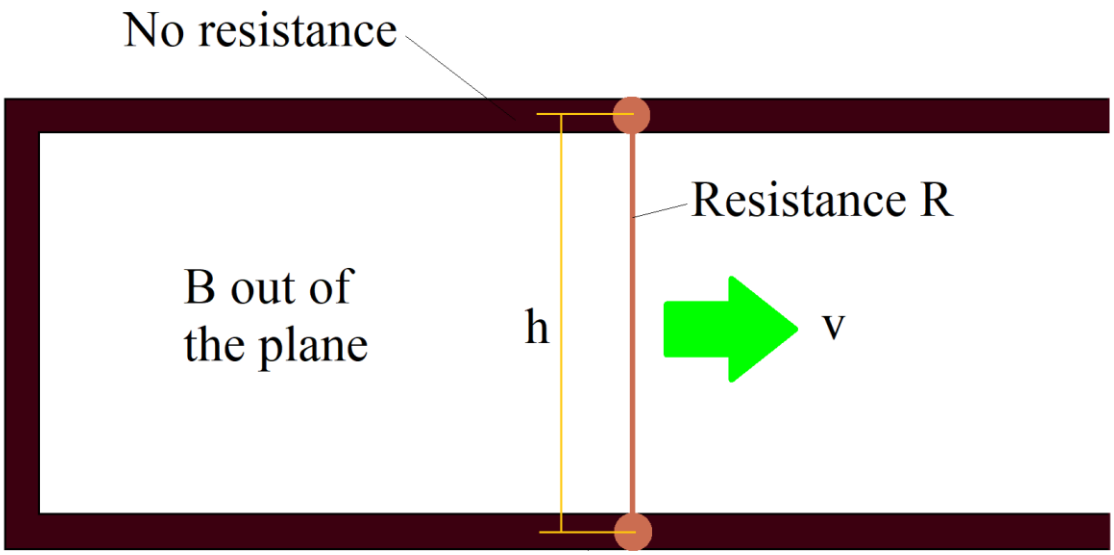
A second source of induced electric fields is the motion of a conductor through a magnetic field. Since **B** exerts force on all moving charges, it does so on the ones inside the conductor. It is as if the charges are acted on by an electric field, whose value is $\mathbf{E} = \mathbf{v} \times \mathbf{B}$. This is called **motional EMF** effect (the EMF standing for electromotive force, an outdated word for electric field which remains active today in this one context). On the other hand, if we look at a moving conductor from a frame which is moving with it, then we won't see motional effect but will instead see a time-dependent **B**, giving rise to an **E** via the second of MAXWELL's equations. ALBERT EINSTEIN assured us in his groundbreaking 1905 paper that the two descriptions of this phenomenon are entirely equivalent and equally legitimate; of course they produce the same answer.

To get a PDE for induction, we start from (01b). $\nabla \times \mathbf{E} = -\partial\left(\nabla \times \mathbf{A}\right)/\partial t$; taking the curl outside the $\partial/\partial t$ on the RHS, we have $\mathbf{E}=-\partial\mathbf{A}/\partial t$. On the other hand, in the magnetostatic situation we have $\nabla^2 \mathbf{A} = -\mu_0 \mathbf{J}$; since we are dropping time dependence from (01d), this continues to hold. Sometimes, ferromagnetic materials are used to maximize induction effects and the permeability μ_0 gets replaced by some value μ which is typically 100 times greater than μ_0. Now $\mathbf{J} = \sigma\mathbf{E}$, and then $\mathbf{E} = -\partial\mathbf{A}/\partial t$ so

$$\nabla^2 \mathbf{A} = \mu\sigma \frac{\partial \mathbf{A}}{\partial t} \quad ,$$

and we have a many+1 heat equation for every component of **A**. The physical implications of this equation in this case are striking – at least in my own opinion, much more so than the simple diffusion of heat within a body.

A trivial application of induced currents is the following (it's too easy to find a place in the Tutes). A rectangular circuit consists of three resistanceless bars connected by a sliding bar of resistance R which moves with speed v. The contraption is mounted in a region where there's a uniform magnetic field **B** coming out of the plane. This setup is shown in the Figure below.



To solve this system, we calculate the voltage induced across the movable bar. Since it's moving at velocity **v** in a magnetic field **B** with **v** and **B** perpendicular, the electric field has magnitude vB. Its direction is evidently from top to bottom. The voltage across the bar, and hence across the circuit, is vBh, and the current is $i = vBh/R$. The current wire in a magnetic field experiences a force ihB. Since the current flows from top to bottom and the magnetic field is out of the plane, the force is towards the left i.e. opposite to the bar's motion. This gives the force as $F = -vB^2h^2/R$. Another way we can do this calculation is to use the integral form of (01b), which is called FARADAY and LENZ's law. This states that the voltage induced across a closed circuit is the time derivative of the magnetic flux through the circuit, and the induced EMF always opposes the cause which produces it. Using this principle also, we can easily get to the voltage vBh across the circuit, and the rest of the steps follow.

The thing to note in this example is that the induced currents exert a **retarding force** on the moving bar. This braking effect of eddy currents is something which gets included in many physics courses – a popular, more meaningful and much more difficult example features a magnetic dipole falling through a metallic pipe. Here however, we'll see some other applications of induced currents – in one case they'll make an object rotate while in another they'll

make it levitate. Let's now head over to the Tutes and see these phenomena for ourselves.

Example 6.3.1

EARNSHAW's theorem

Eight equal point charges are placed at the vertices of a cube. A test charge is now placed at its centre.

(a) Suppose the test charge were to experience a restoring force upon small displacement in any arbitrary direction. Show that this would imply that the potential has a local minimum somewhere in the region near the centre of the cube.

(b) Basis the above result determine whether it is possible to achieve stable confinement of the test charge at the centre of the cube.

[Level : 2]

(a) Consider an arbitrary displacement of the test charge $x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}}$ from the centre of the cube, where x, y and z are small (I could have written Δx, Δy etc but this is clear enough). Then, by the hypothesis of the problem, the test charge experiences a restoring force i.e. a force which is directed along the displacement but opposite to it in sign. This can be written as

$$\mathbf{F} = -k(x,y,z)\left(x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}}\right) \quad ,$$

where k is a function of x, y and z since the spring constant need not be the same in all directions. But, we can expand this $k(x,y,z)$ in a Taylor series $k = k_0 + \alpha x + \beta y + \gamma z + \ldots..$; when the displacements are small, the only term which survives after multiplication by the displacement vector is the constant term. Thus, $\mathbf{F}$ must equal $-k_0\left(x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}}\right)$, which implies that $\mathbf{E}$ in the region is

$$\mathbf{E} = -A\left(x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}}\right) \quad , \qquad (01)$$

where $A = k_0/q$.

Now, consider V. Since we can always add an arbitrary constant to V, let $V = 0$ at the centre of the cube without loss of generality (this flouts the BC that $V \rightarrow 0$ at large distances from the cube, but since we aren't interested in those regions, we don't bother). By inspection, the potential which produces the $\mathbf{E}$ of (01) is

$$V = \frac{A}{2}\left(x^2 + y^2 + z^2\right) \quad . \qquad (02)$$

This potential is minimum at the centre of the cube and increases everywhere as one moves away from it.

(b) In source-free region, which the centre of the cube clearly is, V satisfies Laplace equation. But we saw in 6.1 that Laplace function cannot have maxima and minima at any internal point. Hence (02) does not describe a genuine electric potential. This implies that the assumption leading to (02) is false i.e. the particle cannot be experiencing a restoring force for all possible displacements. Hence it is impossible to achieve stable confinement of the test charge at the centre of the cube.

*

COMMENTS : This is one example of EARNSHAW's theorem (obviously, since that's the Tute title), which states that it is impossible to achieve statically stable spatial confinement of any configuration of time-independent charges and/or currents using any configuration of time-independent charges and/or currents and linear materials of positive permittivity and permeability. This theorem is the bane of life of magnetic levitation technology, where it is essential for the levitating object (either a train or a rotor) to be stable at its equilibrium position. Of course, using charges isn't practical anyway because it's very difficult to isolate and maintain charges. Generating large currents and large magnetic fields is practical however but EARNSHAW's theorem says that we won't go anywhere with a static configuration. In fact, amateur maglev designers often propose elaborate configurations of static magnets which will confine a test magnet inside – they claim that the second pole pair will overcome the instability of the first, the third will overcome that of the second and so on. However, EARNSHAW's theorem shows that such designs are impossible by Laws of Nature, like perpetual motion machines.

Then how does Maglev work ? It aims at the caveats in the theorem – either it uses time-dependent configurations or negative permeability materials (superconductors). But superconductors are very expensive and show their special property only at very low temperatures ($-200\ ^{o}C$ give or take). Dynamic configurations which show the requisite stability while remaining sufficiently simple and inexpensive are very difficult to achieve. This is why Maglev technology is still a long way from mature. We'll see an example of bypassing EARNSHAW in the last Tute of this Unit, and one more in the next Chapter.

Example 6.3.2

Free space GREEN's function

(a) Consider a spherical shell of radius R centred at the origin which carries a constant surface charge density σ. The BC is given that $V = 0$ at $r = \infty$. By solving Laplace equation, determine V and $\mathbf{E}$ everywhere.

(b) In the above solution, let R decrease and σ increase in such a way that the total charge on the sphere is a constant. Take the limit $R \rightarrow 0$ to obtain the GREEN's function for electrostatics with homogeneous BC at infinite distance.

(c) A disk of radius b carries a constant surface charge density σ_0. Using the above GREEN's function, determine V at a point on the disk's axis located a distance z_0 above the disk. Check that the limit $z_0 \rightarrow \infty$ yields a physically plausible result.

[Level : 3]

(a) We use spherical polar coordinates (of course) with the sphere centred at the origin. Evidently, the potential satisfies the homogeneous Laplace equation in the two regions inside and outside the sphere. Let these two potentials be V_1 and V_2 (we'll use the convention 1 for inside and 2 for outside in all problems – and there are quite a few – which feature this kind of geometry). The connection between V_1 and V_2 will come from the boundary matching conditions on V and $\mathbf{E}$ at the surface of the sphere [recall the similar procedure in 4.2.2 Part (a)].

Since the charge density is constant on the sphere surface, we expect on physical grounds that V will also be independent of θ and φ. From 6.1.5, the solutions in this case are $V = \text{const.}$ and $V = 1/r$. Inside the sphere, the latter solution is a no-no. Outside, the former must be zero to satisfy the BC at infinity. So we have

$$V_1 = A_1 \quad ,$$

$$V_2 = \frac{A_2}{r} \quad ,$$

as the "eigenfunction expansion" for V.

Now impose the BCs at $r = R$. First is that V is continuous so $A_1 = A_2/R$. The second BC comes from the matching conditions on $\mathbf{E}$ (6.3.02a). So let's find $\mathbf{E}$ first by taking the grad of V. Inside the sphere, V is constant so $\mathbf{E} = 0$. Outside, $\mathbf{E} = \left(A_2 / r^2\right)\hat{\mathbf{r}}$. Evidently, at the surface of the sphere, $\mathbf{E}$ only has a normal component and no parallel one. So the matching condition yields

$$\frac{A_2}{R^2} - 0 = \frac{\sigma}{\varepsilon_0} \quad ,$$

or, $A_2 = \sigma R^2/\varepsilon_0$. This gives $A_1 = \sigma R/\varepsilon_0$.

Combining all this we have our solution

$$V = \begin{cases} \dfrac{\sigma R}{\varepsilon_0} & \text{if } r \le R \\ \dfrac{\sigma R^2}{\varepsilon_0}\dfrac{1}{r} & \text{if } r \ge R \end{cases} \quad , \qquad \text{(01a)*}$$

$$\mathbf{E} = \begin{cases} 0 & \text{if } r < R \\ \dfrac{\sigma R^2}{\varepsilon_0}\dfrac{1}{r^2}\hat{\mathbf{r}} & \text{if } r > R \end{cases} \quad , \qquad \text{(01b)*}$$

which is the desired answer. Since V is continuous, we could use $\leq$ and $\geq$ when comparing r and R in (01a); on the other hand $\mathbf{E}$ is different inside and outside the shell so the only correct symbols are $<$ and $>$ in (01b).

(b) We start from what the free space GREEN's function of electromagnetism is. $G(\mathbf{r},\mathbf{r}')$ is the solution of $\nabla^2 G(\mathbf{r},\mathbf{r}') = -(1/\varepsilon_0)\delta(x-x')\delta(y-y')\delta(z-z')$ with BC $G=0$ at $|\mathbf{r}|=\infty$. In words, G is the potential created at point with position vector $\mathbf{r}$, on account of a point charge of unit strength at position $\mathbf{r}'$, satisfying homogeneous BC at infinite distance. Since the Laplace equation is translation-invariant, it suffices to find $G(\mathbf{r})$ with the point charge located at $\mathbf{r}'=0$; we can then introduce $\mathbf{r}-\mathbf{r}'$ in place of $\mathbf{r}$ to extend to the general case (recall the derivation in 5.3.2).

The question tells us how to move forward from here. We have just found V for a sphere of charge centred at $\mathbf{r}'=0$; now shrink its radius to zero while keeping the total charge constant. If q be the total charge on the sphere, then $\sigma = q/4\pi R^2$; substitute this into (01a), and the R's cancel away giving $V = q/4\pi\varepsilon_0 r$ outside the sphere. V inside obviously makes no sense in the limit $R\to 0$. Evidently, r in spherical coordinates is nothing but the distance from the origin to the target point, so $V = q/(4\pi\varepsilon_0|\mathbf{r}|)$. Finally, setting $q=1$ and substituting $\mathbf{r}-\mathbf{r}'$ in place of $\mathbf{r}$ gives the GREEN's function

$$* \qquad G(\mathbf{r},\mathbf{r}') = \frac{1}{4\pi\varepsilon_0|\mathbf{r}-\mathbf{r}'|} \quad , \qquad *$$

which is the desired answer.

(c) We have obtained the potential due to a point charge at arbitrary location, and we want the potential due to a given charge density function ρ. Imagining the given distribution to be an infinite collection of point charges at locations $\mathbf{r}'$, we get the usual GREEN's function integral formula :

$$V(\mathbf{r}) = \int \mathrm{d}^3\mathbf{r}'\,\rho(\mathbf{r}')G(\mathbf{r},\mathbf{r}') \quad .$$

Here, $\mathrm{d}^3\mathbf{r}'$ means that the entire volume in $\mathbf{r}'$ is to be integrated over, in whatever coordinates may be appropriate.

The given charge distribution is a surface charge rather than a volume charge because the disk is very thin. Putting the origin of the primed system at the centre of the disk, the z' dependence of the distribution is $\delta(z')$. In x' and y', the

distribution has the value σ_0 if $(x'^2+y'^2) \le b^2$ and zero otherwise. Hence the charge configuration is

$$\rho(x',y',z') = \begin{cases} \sigma_0 \delta(z') & \text{if } x'^2+y'^2 \le b^2 \\ 0 & \text{otherwise} \end{cases} .$$

What is desired is the potential at the point $(0,0,z_0)$. Using the integral formula,

$$V(0,0,z_0) = \int dx' \int dy' \int dz' \sigma_0 \delta(z') \frac{1}{4\pi\varepsilon_0 \left(x'^2+y'^2+z_0^2\right)^{1/2}}$$

.

The integration on z' clearly produces a factor of σ_0; for the rest of the integral we transform to cylindrical polar coordinates $\tilde{\rho}'$ and θ' (the tilda on ρ differentiates between cylindrical coordinate and charge density). We have

$$V(0,0,z_0) = \frac{\sigma_0}{4\pi\varepsilon_0} \int_0^{2\pi} d\theta' \int_0^R \frac{\tilde{\rho}' d\tilde{\rho}'}{\left(\tilde{\rho}'^2+z_0^2\right)^{1/2}} .$$

The angular integration is trivial; the radial one is easy because we recognize the derivative of $\left(\tilde{\rho}'^2+z_0^2\right)^{1/2}$; doing these integrations we get

$$* \qquad V(0,0,z_0) = \frac{\sigma}{2\varepsilon_0}\left[\left(R^2+z_0^2\right)^{1/2} - z_0\right] , \qquad *$$

which is the desired answer.

In the limit $z_0 \to \infty$, we Taylor expand the surd, treating R/z_0 as a small parameter; this gives $V = \sigma_0 R^2 / 4\varepsilon_0 z_0$. This is like the potential of a single point charge at the origin; this equivalent point charge has the value $\pi R^2 \sigma_0$, which is the total charge on the disk. Far away from the disk, we indeed expect that such a thing will happen – since the whole disk is so far off, the microscopic charge distribution won't matter and all we'll see is the total. Hence this is a physically plausible result.

*

COMMENTS : I can understand your reaction. "COULOMB 's law and the simple integration which followed are stuffs I know

since 11th standard – minus your mathematical high-falutions." This reaction is probably justified. But that's reality. Like babies who are given medicine mixed in food, we are introduced to electromagnetism through what is actually the free space GREEN's function of a Laplace equation. At that time, we can barely spell the word "differential equation" but see COULOMB's law as something which holds true, and the summation or integration over charges to find potentials as something completely intuitive. Then, we do electromagnetism a second time, now using the full-blown PDE machinery, and it appears like a monster.

The primary point of this problem – more important that the pedagogic demonstration of COULOMB's law from MAXWELL's equations – is that there is absolutely nothing difficult about GREEN's function. Somehow, that word always has an aura – perhaps on account of the mind-boggling eigenfunction expansions for GREEN's functions which are present in works like JACKSON and SMYTHE. GRIFFITHS carefully avoids using the word, although like everyone else he starts from COULOMB's law. But the concept is among the simplest possible – literally high school material. Because of this, I was tempted to introduce G through this example rather than the cantilever of T065-6. But mathematical rationality dictates that we see the thing first in ODE, and then in an easy PDE (1+1 heat) before coming over to this more difficult context.

You may also object that Part (a) could have been done much more easily using GAUSS' law – this is one of the isolated handful of situations where that method actually works. Yes, that's true. But the PDE formalism is a far more powerful tool than the integrated equations, as you'll soon see for yourself. This problem gives you a very simple situation where you can acquire familiarity with the new mathematical machine before letting it loose in the field.

Example 6.3.3

Method of images

(a) The plane $z=0$ is grounded and a point charge q kept at a height h above it. Determine V everywhere in the upper half-space.

(b) The planes $x=0$ and $z=0$ are grounded and a point charge q kept at distances x_0 and z_0 respectively away from the two planes. Obtain a fictitious charge configuration which can enable the determination of V everywhere in $x,z>0$.

(c) A spherical shell of radius R is grounded and a point charge q is placed a distance d ($>R$) away from its centre. Given that a single fictitious point charge can replicate the BC on the sphere surface in the absence of external grounding, determine its strength and location.

[Read Only]

(a) This is 6.1.2 Part (c) taken to an extreme. The problem we want to solve is : V satisfies the inhomogeneous Laplace equation $\nabla^2 V = q\delta(\mathbf{r}-\mathbf{r}')$ [the delta function is three-dimensional and $\mathbf{r}'$ is the location of the charge] everywhere in the upper half-space (i.e. $z>0$) with the BCs $V=0$ at $z=0$ and at $|x^2+y^2+z^2|=\infty$. The plane and the hemisphere together enclose the region where we are seeking V so the problem is well-defined and has a unique solution.

What we do is, we remove the externally imposed BC at $z=0$ and instead try to create a fictitious charge configuration which can somehow reproduce this BC automatically. In creating this configuration, we CANNOT tamper with the upper half-space – introducing a charge there would make it a different problem from the one we are trying to solve. But we CAN muck about with the lower half-space – nobody has asked for the potential there.

Let the coordinates of the given charge q be $(0,0,h)$ and let's remove the grounded plane. We now guess (and there's NO substitute!) that if we bring in a second charge $-q$ at $(0,0,-h)$, then we'll satisfy $V=0$ at $z=0$ even without the grounded conductor. To prove the guess : at any point $(x,y,0)$, the potential is

$$V(x,y,0)=\frac{1}{4\pi\varepsilon_0}\left[\frac{q}{\left(x^2+y^2+h^2\right)^{1/2}}-\frac{q}{\left(x^2+y^2+h^2\right)^{1/2}}\right] ,$$

which is obviously zero. Hence this configuration indeed recreates the BC on the plane. The additional charge $-q$ is called the IMAGE charge. (The word "image" is motivated by the fictitious charge appearing like the reflection of the given charge in the plane in this particular problem; it is also used in more general situations as it sounds better than "fake".)

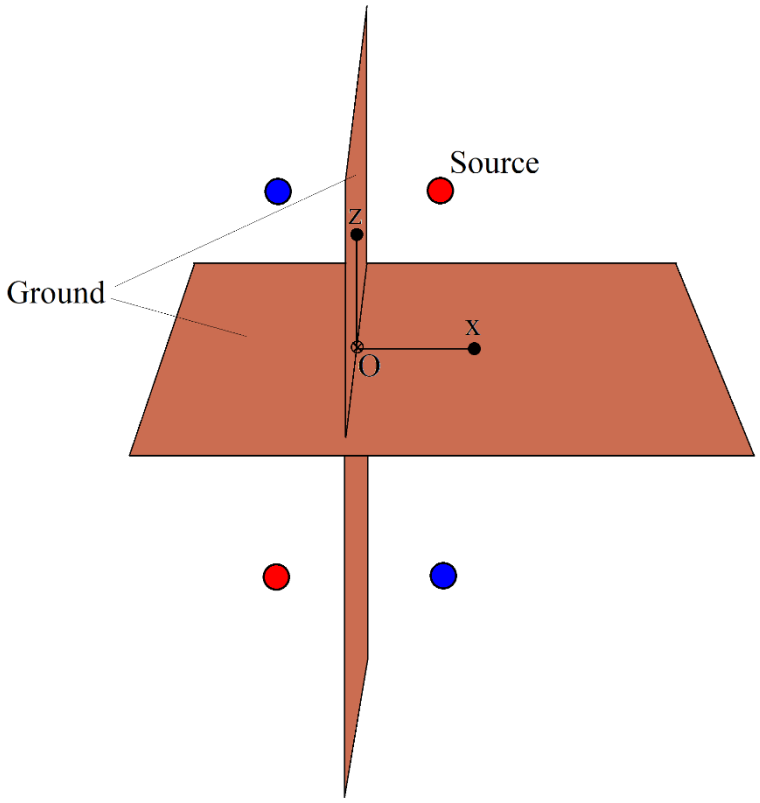
From this point it's another one-liner – to find V at x,y,z for $z>0$, simply add up the potentials coming from the original charge and the image. We have

$$* \qquad V(x,y,z)=\frac{1}{4\pi\varepsilon_0}\left[\frac{q}{\left(x^2+y^2+(z-h)^2\right)^{1/2}}-\frac{q}{\left(x^2+y^2+(z+h)^2\right)^{1/2}}\right] , \qquad *$$

which is the desired answer.

(b) This time, we have grounded planes at both $x=0$ and $z=0$, and the charge q is at $(x_0,0,z_0)$. Corresponding to the first one, we can install the image $-q$ at

$(-x_0,0,z_0)$. Corresponding to the second, we install another image $-q$ at $(x_0,0,-z_0)$. However the two images are at war with each other – each violates the other's BC. What we now do is, install a third image $+q$ at $(-x_0,0,-z_0)$. Now, there are two equidistant $+q$ and $-q$'s from $z=0$, as well as two equidistant $\pm q$'s from $x=0$, so both the BCs are satisfied. And since we didn't introduce anything in $x,z>0$, the original problem hasn't been tampered with. Bingo ! This configuration is shown in the Figure below.



A red charge is positive and a blue one is negative. The source charge is labelled while the other three charges are images. Evidently, we can see how both pairs act together to satisfy the grounded BCs on both the planes (copper). The potential at any point for $x,z>0$ is of course the sums of the contributions from the original charge as well as the three images.

(c) Let the sphere be centred at the origin and let the given point charge q be at $(d,0,0)$. The question contains an enormous hint – it actually tells us that there at all exists a point charge of some suitable strength and location which can reproduce the BC at the surface of the sphere. We'll exploit this hint to be slick in the calculation. First thing, physical argument tells us that the fictitious charge must lie on the x-axis. The external charge is on that axis – if the image is not,

then which way is it off ? So we can let the strength of the image charge be Q and its coordinates be $(D,0,0)$.

Now is the sneaky step – instead of even trying to write the potential everywhere on the sphere surface and find Q and D such that it vanishes, we'll only make V equal zero at the two easiest points $(R,0,0)$ and $(-R,0,0)$. These two constraints will automatically determine the two unknowns – then, since the question says that such a configuration does make V disappear on the sphere surface, the one we've found must be the answer.

The potential of a point charge is of course $(1/4\pi\varepsilon_0)\,(q/r)$ as you likely know already and as we calculated in the last Tute. The potential from the source and image at point $(R,0,0)$ is

$$V(R,0,0)=\frac{1}{4\pi\varepsilon_0}\left(\frac{q}{d-R}+\frac{Q}{R-D}\right)=0 \quad ,$$

while at the point $(-R,0,0)$ it is

$$V(-R,0,0)=\frac{1}{4\pi\varepsilon_0}\left(\frac{q}{d+R}+\frac{Q}{d+R}\right)=0 \quad .$$

Clearing the fractions off the last two equations gives us a linear system :

$$\begin{aligned}(d-R)Q-qD&=-qR\\(d+R)Q+QD&=-qR\end{aligned} \quad .$$

We solve this using KRAMER's Rule – it yields

$$\begin{bmatrix}Q\\D\end{bmatrix}=\frac{1}{2qd}\begin{bmatrix}q & q\\-d-R & d-R\end{bmatrix}\begin{bmatrix}-qR\\-qR\end{bmatrix} \quad ,$$

i.e.

* $$Q=-\frac{qR}{d} \quad ,$$ *

* $$D=\frac{R^2}{d} \quad ,$$ *

which is the desired answer. We can see that the sign of the image charge is opposite that of the original, and that the image always lies inside the sphere, both of which are important consistency checks. *****

COMMENTS : As in 6.1.2, it will take a genius in maths to get these results from eigenfunction expansion on the original problems. It took a genius in physics (this technique is due to Sir WILLIAM THOMSON the Lord KELVIN , and MAXWELL himself) to get them this way for the first time. But once shown the method, every fool can do it.

The only thing to be careful of when doing an image problem is to not insert a charge in the region of interest. As an example, in Part (a) let's calculate the potential in the lower half-space for the original configuration. This satisfies the homogeneous Laplace equation with BCs $V=0$ on the plane and at infinity. Once again, the BC region completely encloses the space so the problem is well-defined and the solution unique. One solution which satisfies the equation and the BCs is the trivial one. Hence $V=0$ everywhere in the lower half-space for the given configuration. On the other hand, V of the charge and its image is clearly not zero everywhere in the lower half-space. In the same way, the image configuration of Part (c) gives only the potential outside the sphere – inside the actual sphere V is again trivial.

In Part (c), I didn't do the derivation that the configuration does indeed cause V to be zero everywhere on the sphere surface. If you want an exercise in coordinate geometry, you can check that fact yourself. Else you can just take it for granted – the image method is so non-generic that there's no way one can know the answer unless somebody discovers it first. Carrying through the lengthy derivation is somewhat pointless since there's no question of systematically applying the method to general configurations. Another special geometry which does work this way is a line charge outside a grounded cylinder – using the fact that the potential of line charge is $\log\rho$, you can calculate two messy expressions for the strength and location of an image line which enforces the BC. In both these configurations, you can also invert the setup – to get potential due to a charge INSIDE a grounded sphere or cylinder, you can introduce an image outside.

A popular follow-up to an image question in an exam is to calculate the charge density induced on the grounded conductor, for example on the plane of Part (a). To do this, you must first take the grad of the potential to find the electric field above the plane (it is zero below). Then use the BC (6.3.02a) to find σ on the plane as a function of x and y. I would have done more details if there had

been practical uses of such charges and grounded conductors. But as I said in 6 .3.1, it's very difficult to isolate and maintain charges. So we'll be more elaborate when we get to currents and magnetic fields.

Note also how we cleverly exploited the info given in the question to minimize work in Part (c). Such short cuts are very useful in the exam hall. The flip side is that their application requires some care. For example, our two-point calculation would have been rubbish if the question had not mentioned that the image is a point – the actual image configuration might well have been a ring or sphere or some weird axisymmetric cloud of charges (or an image could have not existed at all). If simply guessing the answer without the hint, then it would have made no sense unless we showed explicitly that the guessed configuration does make V become zero everywhere on the surface of the sphere.

Since the potential of a point charge in the presence of homogeneous BCs is the GREEN's function for those BCs, the results of this problem can easily be extended to find potential for arbitrary charge configurations in the presence of grounded plates or spheres. Indeed, the image method is probably one of the most beautiful PDE calculations in Literature. But it's more than a physico-mathematical curiosity and a great exam problem – in the material which follows, we'll actually use it to telling effect, more than once. So please master the art of image-guessing before you proceed with the further material.

Example 6.3.4

Currents on a cylinder

Each of the three situations below features an infinitely long cylindrical shell of radius R carrying a surface current $\mathbf{K}$. The profile of the current is different in the three problems. Using the vector potential approach, determine the magnetic field everywhere in each case.

(a) $\mathbf{K} = K_0 \hat{\mathbf{z}}$

(b) $\mathbf{K} = \left(K_d \cos\theta + K_q \sin\theta \right) \hat{\mathbf{z}}$

(c) $\mathbf{K} = K_0 \hat{\boldsymbol{\theta}}$. In this problem, write and solve PDEs for the Cartesian components of $\mathbf{A}$ in cylindrical coordinates.

[Level : 3]

(a) The current is along the z-direction, so we expect that $\mathbf{A}$ will also be along z. Further, z is a Cartesian direction even though the coordinate system is cylindrical. So we can write

$$\nabla^2 A_z = -\mu_0 J_z \quad .$$

Hereafter I will throw away the subscripts z on A and J and write them as scalars.

J is non-zero only at the surface of the cylinder, similar to the charged spherical shell of 6.3.2. So once again we have two homogeneous Laplace equations inside and outside, and will have to match A and $\mathbf{B}$ at the cylinder to complete the solution. Now we obtained the general solution of Laplace equation in 2D cylindrical polars in 6.1.4. This consists of a series with terms independent of θ, terms going as $(\cos, \sin)\,\theta$, terms as $(\cos, \sin)\,2\theta$ etc. Since the source term J is independent of θ, we expect these solutions to be the ones present in this case . These are $A = \text{const.}$ and $A = \log\rho$.

Ordinary logic would say to throw away the $\log\rho$ term both inside and outside, since it is divergent both at $\rho = 0$ and $\rho = \infty$. However, $A = \text{const.}$ both inside and outside, followed by A is continuous would imply that A is the same constant everywhere. This leads to $\mathbf{B} = 0$ everywhere, a patently nonsensical result. What we have to do is allow the $\log\rho$ to be present outside. After all, $\log\rho$ at $\rho = 0$ diverges sharply but does so only marginally at $\rho = \infty$; $\log 10$ is 2.3, $\log 10000$ is 9.2 and so on. So we write

$$A_1 = C_1 \quad ,$$
$$A_2 = C_2 + D_2 \log\rho \quad ,$$

as our eigenfunction expansion and proceed to the BC fitting step.

Continuity of A gives $C_1 = C_2 + D_2 \log R$. To fit the BC on $\mathbf{B}$, we need to find it first. This we do by taking the curl of $\mathbf{A}$ (which we now treat as vector – nothing along ρ and θ and our scalar A along z). Using the formulae from the review, we have

$$\mathbf{B}_1 = 0 \quad ,$$
$$\mathbf{B}_2 = -\frac{D_2}{\rho}\hat{\boldsymbol{\theta}} \quad .$$

Since this $\mathbf{B}$ has only a tangential component, we now employ (6.3.05b) at $\rho = R$; this says $\mathbf{B}_2 - \mathbf{B}_1 = \mu_0 (\mathbf{K} \times \mathbf{n})$. Here, $\mathbf{K}$ is along $\hat{\mathbf{z}}$ while $\hat{\mathbf{n}}$ is along $\hat{\boldsymbol{\rho}}$; their cross product is positive $\hat{\boldsymbol{\theta}}$. This implies that

$$-\frac{D_2}{R} = \mu_0 K_0 \quad ,$$

whereby we get our final expressions for $\mathbf{B}$:

$$* \qquad \mathbf{B} = \begin{cases} 0 \ \text{ if } \ \rho < R \\ \dfrac{\mu_0 K_0 R}{\rho} \hat{\boldsymbol{\theta}} \ \text{ if } \ \rho > R \end{cases} \quad , \qquad *$$

which is the desired answer.

(b) We consider the $K_d \cos\theta$ component of $\mathbf{K}$ alone – we can easily extend the result at the end to solve for the $K_q \sin\theta$ component also. Once again, A_z is the only component of $\mathbf{A}$, and it satisfies Laplace equation inside and outside the cylinder. The matching condition comes from K; since this has a $\cos\theta$ component, we expect the $\rho\,(\cos\theta,\ \sin\theta)$ and $(1/\rho)\,(\cos\theta,\ \sin\theta)$ terms of the general solution to be the relevant ones. (Although the current is $\cos\theta$ only, throwing out the $\sin\theta$ component also might be unwise, just as in undet coeff with sinusoidal excitation – after all a curl can convert a cos to sin. But it can't convert θ to 2θ or 5θ.) Inside the cylinder, $1/\rho$ blows up while outside the cylinder ρ blows up. So we reject these two terms and have our eigenfunction expansion

$$A_1 = \rho\left(C_1 \cos\theta + D_1 \sin\theta\right) \quad ,$$

$$A_2 = \frac{1}{\rho}\left(C_2 \cos\theta + D_2 \sin\theta\right) \quad .$$

Now we proceed to the BC fitting step.

Continuity of A at $\rho = R$ gives $(C,D)_1 = (C,D)_2 / R^2$. $\mathbf{B}$ has two components – radial and tangential. The former is normal to the surface, and is continuous; we expect that it is taken care of by imposing continuity of A. The latter is parallel to the surface and has to be obtained by imposing (6.3.05b). Taking the θ-component of curl $\mathbf{A}$,

$$B_{\theta 1} = -C_1 \cos\theta - D_1 \sin\theta \quad ,$$
$$B_{\theta 2} = \frac{C_2}{\rho^2}\cos\theta + \frac{D_2}{\rho^2}\sin\theta \quad ,$$

and then imposing the BC,

$$\frac{C_2}{R^2} + C_1 = \mu_0 K_d \quad ,$$
$$\frac{D_2}{R^2} + D_1 = 0 \quad .$$

Finally, using the relations between C and D obtained from continuity of A, we have $C_1 = \mu_0 K_d/2$, $C_2 = \mu_0 K_d R^2/2$ and $D_1 = D_2 = 0$.

Thus our vector potential is

$$A_z = \begin{cases} \dfrac{\mu_0 K_d}{2}\rho\cos\theta & \text{if } \rho \le R \\ \dfrac{\mu_0 K_d R^2}{2\rho}\cos\theta & \text{if } \rho \ge R \end{cases} .$$

Now we take the curl of this $\mathbf{A}$ to get $\mathbf{B}$ inside and outside. We have

$$\mathbf{B} = \begin{cases} \dfrac{\mu_0 K_d}{2}\left(-\sin\theta\hat{\boldsymbol{\rho}} - \cos\theta\hat{\boldsymbol{\theta}}\right) & \text{if } \rho < R \\ \dfrac{\mu_0 K_d R^2}{2\rho^2}\left(-\sin\theta\hat{\boldsymbol{\rho}} + \cos\theta\hat{\boldsymbol{\theta}}\right) & \text{if } \rho > R \end{cases} . \qquad (01)$$

We are almost done except for that $K_q \sin\theta$ also given as part of the current distribution. Since $\sin\theta = \cos(\theta - \pi/2)$, the result for this component should be the same as (01) with $K_d \rightarrow K_q$ and $\theta \rightarrow \theta - \pi/2$ (which changes cos to sin and sin to −cos). Implementing this, we get our final expressions

$$A_z = \begin{cases} \dfrac{\mu_0 K_d}{2}\rho\cos\theta + \dfrac{\mu_0 K_q}{2}\rho\sin\theta & \text{if } \rho \le R \\ \dfrac{\mu_0 K_d R^2}{2\rho}\cos\theta + \dfrac{\mu_0 K_q R^2}{2\rho}\sin\theta & \text{if } \rho \ge R \end{cases} , \qquad (02)$$

and

$$* \qquad \mathbf{B} = \begin{cases} \frac{\mu_0}{2}\left[\begin{array}{l}\left(-K_d\sin\theta + K_q\cos\theta\right)\hat{\boldsymbol{\rho}} \\ +\left(-K_d\cos\theta - K_q\sin\theta\right)\hat{\boldsymbol{\theta}}\end{array}\right] & \text{if } \rho < R \\ \frac{\mu_0 R^2}{2\rho^2}\left[\begin{array}{l}\left(-K_d\sin\theta + K_q\cos\theta\right)\hat{\boldsymbol{\rho}} \\ +\left(K_d\cos\theta + K_q\sin\theta\right)\hat{\boldsymbol{\theta}}\end{array}\right] & \text{if } \rho > R \end{cases} , \qquad (03)*$$

which latter is the desired answer.

(c) The problem explicitly tells us what to do – we are to split $\mathbf{A}$ and hence $\mathbf{K}$ into CARTESIAN components rather than cylindrical but solve the problem in cylindrical. Curious, but let's go ahead. (Exam tip : If a problem gives an instruction like this which seems counter-intuitive or odd, don't ask why – just DO IT. Ask the why only after the exam is over.) Since $\mathbf{K}$ is directed along $\hat{\boldsymbol{\theta}}$, we can write it as $\mathbf{K} = K_0\left(-\sin\theta\hat{\mathbf{x}} + \cos\theta\hat{\mathbf{y}}\right)$. Then, we write inhomogeneous Laplace equations for $\mathbf{A}$ component-by-component : $\nabla^2 A_{x,y} = -\mu_0 J_{x,y}$, where $\mathbf{J}$ of course exists only on the surface of the cylinder.

Now, each of A_x and A_y satisfies a PDE very similar to the one in the above part. Instead of scribbling eigenfunction expansions, we think about the results a bit. We saw that the angular dependence of a scalar component of $\mathbf{A}$ mimics that of the corresponding K everywhere i.e. a $\cos\theta$ K gives rise to a $\cos\theta$ A while $\sin\theta$ K creates $\sin\theta$ A. So A_x here will be some function of ρ times $-\sin\theta$ everywhere. Since A_y has identical BC with $\cos\theta$ in place of $-\sin\theta$, it will the SAME function of ρ times $\cos\theta$ everywhere. This implies that $\mathbf{A}$ everywhere will be in the $\hat{\boldsymbol{\theta}}$ direction and will be some function of ρ. The ρ-dependences of A_x and A_y will of course be the ones from Part (b) – ρ inside and $1/\rho$ outside. Hence we can come back from Cartesian to cylindrical to write

$$\mathbf{A}_1 = C_1\rho\hat{\boldsymbol{\theta}} \quad ,$$

$$\mathbf{A}_2 = \frac{C_2}{\rho}\hat{\boldsymbol{\theta}} \quad .$$

Now of course we proceed to the BC fitting step.

Continuity of A at $\rho = R$ forces $C_1 = C_2/R^2$. The only component of curl $\mathbf{A}$ which is not obviously zero is B_z; it equals $2C_1$ inside and zero outside. Finally, use (L2205b), noting that $\hat{\boldsymbol{\theta}}\times\hat{\boldsymbol{\rho}} = -\hat{\mathbf{z}}$; this gives

$$0 - 2C_1 = -\mu_0 K_0 \quad ,$$

or $2C_1 = \mu_0 K_0$. We now have our final expressions

$$* \qquad \mathbf{B} = \begin{cases} \mu_0 K_0 \hat{\mathbf{z}} & \text{if } \rho < R \\ 0 & \text{if } \rho > R \end{cases} \quad , \qquad *$$

which is the desired answer.

*

COMMENTS : Part (a) is essentially a current-carrying wire – shrink down R to zero keeping $KR = \text{const.}$ and that's what it becomes. The answer agrees with what is conventionally obtained from other methods. Part (b) is a typical example of the PDE method. What is interesting is that the magnetic field inside the cylinder is uniform – draw (01) at a few internal points and you'll see that. Another interesting feature is the subscripts d and q on K – the only time these funky subscripts appeared before was in 3.3.1 and 4.4.4, both on induction motors. This is not a coincidence as you'll see soon enough.

Part (c) is the SOLENOID – an archetypal current configuration which nevertheless is not particularly easy to solve by 'elementary' methods. The usual solution is to arrive at the direction of $\mathbf{B}$ from symmetry arguments and then consider $\oint \mathbf{B} \cdot \mathrm{d}\mathbf{l}$ along several rectangular paths, inside, outside and crossing the solenoid. The tricky step in this approach is to justify that $\mathbf{B}(\infty, \theta) = 0$; this is not absolutely self-evident because the current configuration is infinite also. However, the vector potential approach bypasses all this and throws up the answer as the inevitable consequence of the solution of the governing PDEs.

In this part, we also did something weird – even though both $\mathbf{J}$ and as it turned out $\mathbf{A}$ also were so nice and tractable along $\hat{\boldsymbol{\theta}}$, we solved for Cartesian components rather than the "θ-component". So now let's do what feels intuitive – start from $\nabla^2 A_\theta = -\mu_0 J_\theta$. This J_θ is of course the K of this problem; since it is independent of θ we have the situation of Part (a) and have to try the functions $A_\theta = \text{const.}$ and $A_\theta = \log \rho$ inside and outside. Throwing the latter away inside, we get $\mathbf{B} = 0$ inside and $\mathbf{B} = (1/\rho)\hat{\mathbf{z}}$ outside. This is garbage. This apparently intuitive procedure broke down because the cylindrical and spherical coordinate systems are curved – a Cartesian unit vector like $\hat{\mathbf{x}}$ points in the same direction

everywhere but something like $\hat{\boldsymbol{\rho}}$ does not. $\hat{\mathbf{z}}$ of cylindrical does, which is why we didn't run into problems in Parts (a) and (b). So, when solving Laplace equation for vectors, we must always work in terms of Cartesian components alone.

Example 6.3.5

Current on a sphere

A spherical shell of radius R carries a current $\mathbf{K} = K_0 \sin\theta \hat{\boldsymbol{\varphi}}$ on its surface.

(a) Drawing reference to 6.1.5, determine **B** everywhere.

(b) Show that the field outside the sphere is identical to that produced by an ideal dipole **m** at the centre, and determine the value of **m**.

[Level : 2]

(a) Having learnt our lesson in the last Tute, we break up $\mathbf{A}$ and $\mathbf{J}$ into Cartesian components and apply spherical Laplace equation on each. Our current distribution is $\mathbf{K} = K_0 \sin\theta\left(-\sin\varphi\hat{\mathbf{x}} + \cos\varphi\hat{\mathbf{y}}\right)$ and A_x and A_y both satisfy Laplace inside and outside the sphere with this kind of BC at the surface. Now, look at 6 .1.5. $\sin\theta\cos\varphi$ and $\sin\theta\sin\varphi$ are exactly the angular dependences we handled in Part (b) there; we found that the corresponding radial solutions are r and $1/r^2$. The former is implausible outside the sphere and the latter implausible inside.

What we could have done is, gone with A_x and A_y, written eigenfunction expansions for each, converted spherical to Cartesian and Cartesian to spherical, created a giant mess and somehow or other arrived at A_x and A_y everywhere. Instead we'll take a leap of faith. In the cylindrical case we just saw that $\mathbf{A}$ runs in circles like $\mathbf{J}$ and is independent of θ – there we proved it rigorously. Here we'll ASSUME that the thing holds true again i.e. we'll take it for granted that the $\cos\varphi$ and $\sin\varphi$ components of $A_x\hat{\boldsymbol{\varphi}}$ and A_y work out in such as way to generate a resultant $\mathbf{A}$ which is entirely in the direction and independent of φ. Further, since $\sin\theta\cos\varphi$ and $\sin\theta\sin\varphi$ are components of fundamental solutions of Laplace in spherical, we will also assume that the $\sin\theta$ dependence of $\mathbf{A}$ remains untouched. If the calculation goes through without contradiction, our assumption will get justified retroactively; if it reaches a hitch then we'll see about it.

So we claim

$$\mathbf{A}_1 = C_1 r\sin\theta\hat{\boldsymbol{\varphi}} \quad , \tag{01a}$$

$$\mathbf{A}_2 = \frac{C_2}{r^2}\sin\theta\hat{\boldsymbol{\varphi}} \quad , \tag{01b}$$

and proceed tentatively to the BC fitting step. Continuity of $\mathbf{A}$ at $r = R$ forces $C_1 = C_2/R^3$. This also takes care of the continuity of the normal component of $\mathbf{B}$, which is B_ρ in this case. The tangential component B_θ has the usual discontinuity. Taking the curl (assisted by the review material), we have

$$B_{\theta 1} = -2C_1\sin\theta \quad ,$$

$$B_{\theta 2} = \frac{C_2}{r^3}\sin\theta \quad .$$

Now the matching step says that their difference is $\mu_0 K_0\sin\theta$ (positive sign since $\mathbf{K}$ is $\hat{\boldsymbol{\varphi}}$ and $\hat{\mathbf{n}}$ is $\hat{\boldsymbol{\rho}}$). This is consistent – the B_θ's and the current indeed have the

same angular dependence, making a match possible and validating our assumptions on the character of **A**. So we have

$$\frac{C_2}{R^3} + 2C_1 = \mu_0 K_0 \quad ,$$

whereby $C_1 = \mu_0 K_0/3$ and $C_2 = \mu_0 K_0 R^3/3$.

Plugging this into (01) and taking the curl,

$$\mathbf{B} = \begin{cases} \dfrac{2\mu_0 K_0}{3}\left(\cos\theta\hat{\mathbf{r}} - \sin\theta\hat{\boldsymbol{\theta}}\right) & \text{if } r < R \\ \dfrac{\mu_0 K_0 R^3}{3r^3}\left(2\cos\theta\hat{\mathbf{r}} + \sin\theta\hat{\boldsymbol{\theta}}\right) & \text{if } r > R \end{cases} \quad , \qquad (02)^*$$

which is the desired answer.

(b) We are told that an ideal dipole of strength **m** at the origin has a magnetic field

$$\mathbf{B} = \frac{\mu_0}{4\pi}\frac{3(\mathbf{m}\cdot\hat{\mathbf{r}})\hat{\mathbf{r}} - \mathbf{m}}{r^3} \quad . \qquad (03)$$

The only tricky step before a routine checking calculation is to find the direction of the effective **m** which stands in for the sphere. Recall that an ideal dipole is a current loop and its dipole moment is directed perpendicular to the loop – positive if flow is counter-clockwise. Since all the currents in the sphere are flowing in the $+\hat{\boldsymbol{\varphi}}$ direction, the effective dipole must also be a loop which carries current in this direction. Its dipole moment points in the z-direction. Hence (at least we suspect) $\mathbf{m} = m\hat{\mathbf{z}}$.

Now plug in this as well as the formulae for unit vectors into (03). The denominator contains the $1/r^3$, same as the external field in (02) so the match looks promising. The dot product in the first term in the numerator reduces to $m\cos\theta$ while **m** itself becomes $m\hat{\mathbf{z}} = m\left(\cos\theta\hat{\mathbf{r}} + \sin\theta\hat{\boldsymbol{\theta}}\right)$. Then, (03) easily reduces to

$$\mathbf{B} = \frac{\mu_0}{4\pi}\frac{2m\cos\theta\hat{\mathbf{r}} + m\sin\theta\hat{\boldsymbol{\theta}}}{r^3} \quad ,$$

which agrees with the external field of (02) if we set $\underline{m = 4\pi K_0 R^3/3}$. Our assumption about the direction of **m** gets validated at this step. *

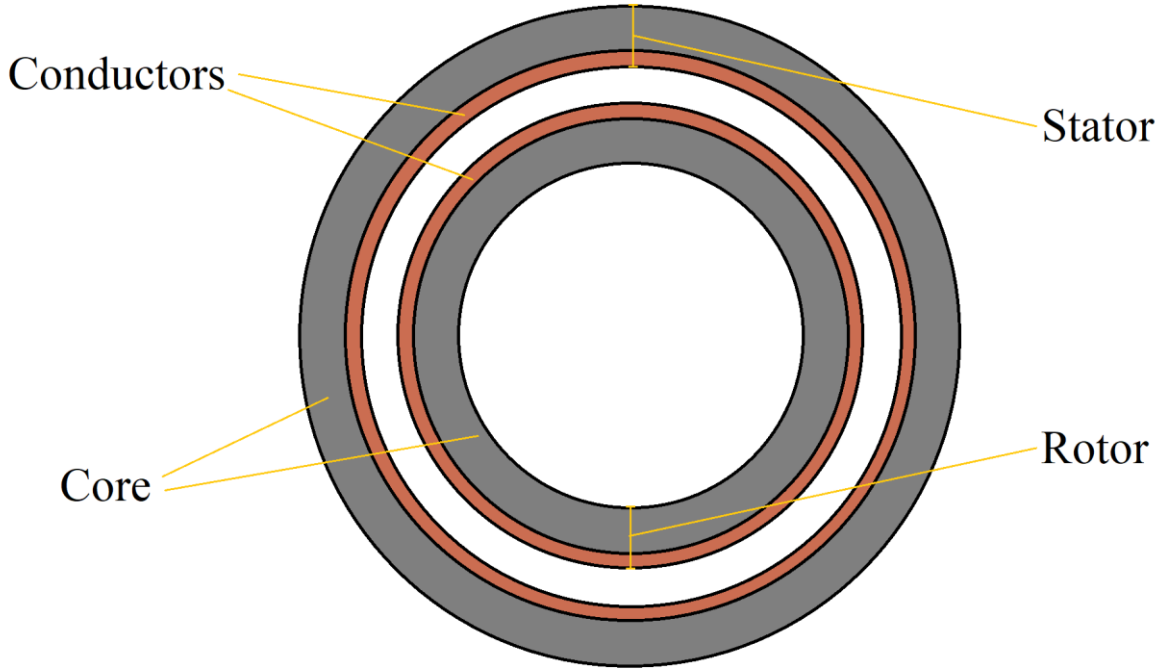
COMMENTS : Another surprising result to say the least. A vanishingly small circular loop carrying infinite current and a hulking sphere with current distribution $K_0 \sin\theta$ can hardly be expected to show the same magnetic field outside. Yet they do. This is unbelievable unless we actually do the calculation and show that it really is true. This means that we can create an "ideal" dipole inside a laboratory to excellent accuracy – a current configuration with no zeroes or infinities anywhere can reproduce its field as a mathematical identity. Inside the sphere by the way, the field is uniform, as it was for the cylinder of the last Tute.

It is also interesting that the dipole moment of the sphere is exactly the surface current times the volume of the sphere – the closest possible analogue of the more common m equals current times area of loop. A macroscopic ELECTRIC dipole on the other hand features a sphere with charge distribution $\sigma\cos\theta$; as we saw in 6.1.5, that angular dependence also produces the same radial eigenfunctions as this one. And the angular dependence of const. corresponds of course to the charged shell of 6.3.2. This explains the names I'd mentioned in the comments to 6.1.5. Although the general solution of spherical Laplace is beyond our scope, I'll nevertheless recommend you strongly to be familiar with these three special cases – they are very easy and more useful than you may think.

Example 6.3.6

Induction motor dynamic model

An induction motor consists of two coaxial cylindrical shells, the stator and the rotor, with the rotor being internal to the stator. An idealized motor is shown in cross-sectional view in the Figure below.



All currents flow along the axial direction only and have $\cos\theta$ or $\sin\theta$ dependence on the angle. The former is called the direct component and the latter is called the quadrature component. The rotor and stator can be treated as sufficiently long so that the two-dimensional formulae are applicable. Further, the thickness of the conducting region of the rotor and stator is assumed small compared to their radius, so that the currents in the elements may be treated as surface currents where appropriate.

External voltage $V_d \cos\theta + V_q \sin\theta$ is applied on the stator, with V_d and V_q being functions of time. No voltage is externally applied on the rotor. Define the rotor current to be $K_{rd} \cos\theta + K_{rq} \sin\theta$ and the stator current to be $K_{sd} \cos\theta + K_{sq} \sin\theta$, with K_{rd}, K_{rq}, K_{sd} and K_{sq} all being

functions of time. Let the motor parameters be height h, rotor radius r, stator radius R ($R > r$), rotor thickness b, stator thickness b', rotor material conductivity σ and stator material conductivity σ'. Assume that magnetic material of susceptibility χ occupies all space so that μ_0 in all expressions may be replaced by $\mu = 1+\chi$. The rotor rotates at angular velocity ω relative to the laboratory while the stator is static.

(a) Working in the laboratory frame, determine the electric field induced inside the rotor and stator in terms of the direct and quadrature components of the rotor and stator currents. Be sure to include all dynamic effects as well as motional EMF effect on the rotor.

Since the rotor and stator are linear conductors, the net currents flowing through them must be proportional to the electric fields inside.

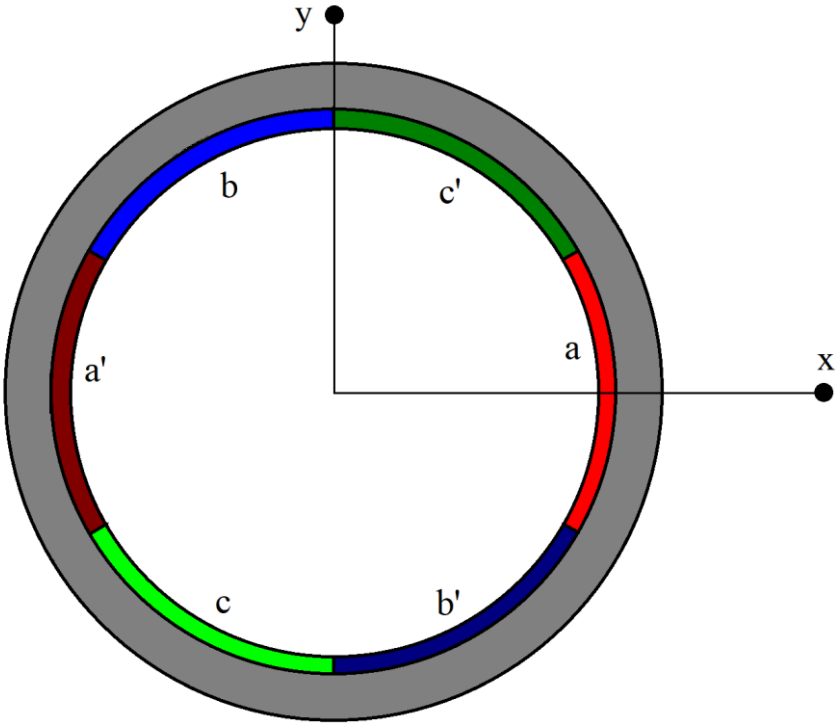
(b) Writing the total electric field as the sum of the externally applied and dynamically induced fields, exploit this proportionality relationship to arrive at dynamical equations for K_{rd}, K_{rq}, K_{sd} and K_{sq} in terms of V_d and V_q.

(c) Determine the torque on the motor in terms of the current components and relevant motor parameters.

(d) Defining the space phasors $\mathbf{X} = X_d + \mathrm{j}X_q$ where X can be voltage or current, express the equations of Part (b) in terms of the space phasors.

The stator of a realistic three-phase motor is shown in the Figure on the next page. The three phases of windings are denoted by a, b and c. Each phase has two branches, the proper branch which is labelled without a prime and the primed branch which is labelled by a prime and acts as the return path for the current. Voltages V_a, V_b and V_c are applied to the three phases. By definition, if V_a is positive then the electric field is directed upwards in the branch a and downwards in the branch a', and similarly in the other phases. Each branch spans a 60° angular sector and carries plus or minus the phase voltage everywhere, i.e. voltage is independent of θ

within each branch. The x-axis passes through the centre of the a windings as shown.



(e) If only V_a is non-zero, obtain the $\cos\theta$ and $\sin\theta$ components i.e. the direct and quadrature components of the applied voltage. Hence, given arbitrary V_a, V_b and V_c, determine the voltage phasor **V**.

[Level : 4]

The very length of the problem statement can throw you off balance in an exam. As always, don't lose your head if faced with such a thing. For one, there are two large pictures. Second, if you read the text even semi-carefully, you'll see that some parts are independent of the other parts. Part (a) is a lead-up to (b), (c) is very likely independent of (b), (d) is a simple manipulation on (b) [express real equation in terms of complex variable] and (e) is decoupled from the rest. This overview at once makes the problem a lot easier.

(a) We have to find E induced inside the stator and rotor on account of time-dependent currents and the rotor's motion. The currents of course are K_{rd}, K_{rq}, K_{sd} and K_{sq}, and everything can be treated as scalars since they are all in z-direction.

Let's first find E in the stator, since that's easier (no rotation). Recall that $E = -\partial A/\partial t$. In 6.3.4, we found A everywhere on account of a single cylinder with direct and quadrature current components . Due to linearity of electromagnetism , the resultant A at any point is the sum of those created by the rotor and stator alone. At the stator surface, the rotor current contributes an A corresponding to the second line of (6.3.4.02) since stator is outside . Plugging r in place of R and R in place of ρ into the formula, we have

$$A_{s,r} = \frac{\mu r^2}{2R}\left(K_{rd}\cos\theta + K_{rq}\sin\theta\right) \quad ,$$

where the double subscript indicates the vector potential created AT the stator surface BY the rotor current. The stator's contribution at its own surface can be got from either line of (6.3.4.02); since A is continuous at the stator surface, the two lines yield the same result. We have

$$A_{s,s} = \frac{\mu R}{2}\left(K_{sd}\cos\theta + K_{sq}\sin\theta\right) \quad .$$

Adding the two contributions and taking the $-\partial/\partial t$ gives E induced inside the stator

$$* \qquad E_{s,\text{ind}} = -\frac{\mu R}{2}\left[\dot{K}_{sd}\cos\theta + \dot{K}_{sq}\sin\theta + \frac{r^2}{R^2}\left(\dot{K}_{rd}\cos\theta + \dot{K}_{rq}\sin\theta\right)\right] \quad . \qquad (01)*$$

In a like manner we can calculate E inside the rotor due to the $-\partial A/\partial t$ effects. However, over and above that there will be a motional EMF term. This is given by $\mathbf{E}=\mathbf{v}\times\mathbf{B}$. Here, $\mathbf{v}$ is due to the rotation of the rotor and equals $\omega r\hat{\boldsymbol{\theta}}$. Then, the only component of $\mathbf{B}$ which is useful is the radial one; the contribution comes out in the $-z$-direction. We use the formulae (6.3.4.03); the top line is relevant for the stator field since the rotor lies inside while either line works for the rotor field since $B_\perp$ is continuous. Putting all these together we have

$$* \qquad E_{r,\text{ind}}=\frac{\mu r}{2}\begin{pmatrix}-\dot{K}_{rd}\cos\theta-K_{rq}\sin\theta-\dot{K}_{sd}\cos\theta-\dot{K}_{sq}\sin\theta+\\ \omega K_{rd}\sin\theta-\omega K_{rq}\cos\theta+\omega K_{sd}\sin\theta-\omega K_{sq}\cos\theta\end{pmatrix} \quad . \qquad (02)*$$

Equations (01) and (02) are the electric fields asked for.

(b) The total electric field in rotor and stator is the sum of the applied and induced fields. Voltage is applied externally on the stator; since the height is h, the applied E is $(V_d\cos\theta+V_q\sin\theta)/h$. There is no external voltage on the rotor so applied E there is zero. So we have

$$E_{r,\text{app}}=0 \quad , \qquad (03a)$$

$$E_{s,\text{app}}=\frac{V_d\cos\theta+V_q\sin\theta}{h} \quad . \qquad (03b)$$

The total electric fields E_r and E_s are of course $E_{r,\text{ind}}+E_{r,\text{app}}$ and $E_{s,\text{ind}}+E_{s,\text{app}}$.

Now is the final step. By definition of a linear conductor, $J=\sigma E$. So far we have been using K; now we have to make it a J. For this step, we use the finite thickness of the rotor and stator. Since the thickness is small, J is radially uniform – K is evidently J multiplied by this thickness. Hence we have $J_r=K_r/b$ for the rotor and $J_s=K_s/b$' for the stator. Then, we easily get to

$$E_r=\frac{1}{\sigma b}\left(K_{rd}\cos\theta+K_{rq}\sin\theta\right) \quad , \qquad (04a)$$

$$E_s=\frac{1}{\sigma' b'}\left(K_{sd}\cos\theta+K_{sq}\sin\theta\right) \quad . \qquad (04b)$$

Since the K's are the total currents in the rotor and stator, these electric fields are also the total E's in the two elements.

But we have just found E_r and E_s as the sum of the RHSes of (02) and (03a) and (01) and (03b) respectively. Equating them to the RHSes of (04a) and (04b) and collecting $\cos\theta$ and $\sin\theta$ components from each resulting equation gives the

dynamic model of the induction motor. We define $\tau_r = \mu\sigma rb/2$ and $\tau_s = \mu\sigma' Rb'/2$, and get

* $$\tau_r \dot{K}_{rd} + \omega\tau_r K_{rq} + K_{rd} + \tau_r \dot{K}_{sd} + \omega\tau_r K_{sq} = 0 \quad , \qquad (05a)^*$$

* $$\tau_r \dot{K}_{rq} - \omega\tau_r K_{rd} + K_{rq} + \tau_r \dot{K}_{sq} - \omega\tau_r K_{sd} = 0 \quad , \qquad (05b)^*$$

* $$\tau_s \dot{K}_{sd} + K_{sd} + \frac{r^2}{R^2}\tau_s \dot{K}_{rd} = \frac{\sigma' b'}{h} V_d \quad , \qquad (05c)^*$$

* $$\tau_s \dot{K}_{sq} + K_{sq} + \frac{r^2}{R^2}\tau_s \dot{K}_{rq} = \frac{\sigma' b'}{h} V_q \quad , \qquad (05d)^*$$

which are the equations asked for.

(c) The torque on the rotor arises because the rotor carries a current and is inside a magnetic field (recall $d\mathbf{F} = i d\mathbf{l} \times \mathbf{B}$). Every infinitesimal element $r\,d\theta$ of the rotor experiences a force $\left(K_r r d\theta h \hat{\mathbf{z}}\right) \times \mathbf{B}$. Now, torque is produced only by the tangential component of the force, and equals r times this force – the radial component has no effect. The tangential force in turn comes from the radial component of $\mathbf{B}$, which we got from (6.3.4.03) and already used once in arriving at (02). But before plugging things into a formula, we think a bit. The motion of the rotor through its own field cannot be creating a torque on it – if that were true, then a single current-carrying cylinder set into rotation would spontaneously accelerate or decelerate, which is absurd. Hence, only the field created by the stator is relevant for generation of the torque; this field must of course be evaluated at the rotor surface.

Putting all this together we have

$$\Gamma = \int_0^{2\pi} d\theta \; r^2 h \left(K_{rd} \cos\theta + K_{rq} \sin\theta \right) B_s \quad ,$$

where the scalar B_s refers to the radial component. Now from (6.3.4.03) this equals $(\mu/2)(-K_{sd}\sin\theta + K_{sq}\cos\theta)$. So we have

$$\Gamma = \frac{\mu r^2 h}{2} \int_0^{2\pi} d\theta \; \left(K_{rd} \cos\theta + K_{rq} \sin\theta \right)\left(-K_{sd} \sin\theta + K_{sq} \cos\theta \right) \quad .$$

In the integral, the $\cos^2\theta$ and $\sin^2\theta$ terms integrate to π; the $\cos\theta \sin\theta$ gives zero. Hence we have

* $$\Gamma = \frac{\pi \mu r^2 h}{2}\left(K_{rd}K_{sq} - K_{sd}K_{rq}\right) \quad , \qquad (06)^*$$

which is the desired answer.

(d) We already have the motor equations in terms of d and q; now we just have to bunch them up into space phasors. This is very simple and I'll give the answer without showing steps :

* $$\tau_r \dot{\mathbf{K}}_r + (1 - \mathrm{j}\tau_r\omega)\mathbf{K}_r + \tau_r \dot{\mathbf{K}}_s - \mathrm{j}\tau_r\omega\mathbf{K}_s = 0 \quad , \qquad (07a)^*$$

* $$\tau_s \dot{\mathbf{K}}_s + \mathbf{K}_s + \frac{r^2}{R^2}\tau_s \dot{\mathbf{K}}_r = \frac{\sigma' b'}{h}\mathbf{V} \quad , \qquad (07b)^*$$

which is the space phasor formulation of the motor dynamic model.

(e) When V_a is nonzero and $V_b = V_c = 0$, the voltage as a function of θ is

$$V(\theta) = \begin{cases} V_a & \text{if} \quad -\pi/6 \le \theta \le \pi/6 \\ -V_a & \text{if} \quad -\pi \le \theta \le -5\pi/6 \quad \text{and} \quad 5\pi/6 \le \theta \le \pi \\ 0 & \text{otherwise} \end{cases} .$$

The d and q components are of course the first harmonics when we expand this function in a Fourier series. The function is symmetric so the quadrature component is obviously zero. As regards the direct component, the integral of $V\cos\theta$ has equal contributions from the a and a' windings. Now the contribution of the a windings is just what we calculated in 4.3.3, with d of that problem being 1/6 and n being 1 in our case. Copying the result of that Tute, we have $V_d = 2V_a/\pi$.

Although we can easily Fourier expand again, with arbitrary voltages V_a, V_b and V_c, a trick saves a lot of work. We see that when phase a alone is energized, the voltage phasor is $\mathbf{V} = 2V_a/\pi + \mathrm{j}0$. Imagining the d-q plane on top of the x-y plane with d lying along x, the phasor points from the centre of the a' branch to the centre of the a branch. This geometrical relationship must hold true for all phases i.e. if V_b alone is energized then the phasor must have a magnitude $2V_b/\pi$ and point from the centre of the b' windings to that of the b windings. The phasor which does so is $(2V_b/\pi)\,\mathrm{e}^{\mathrm{j}2\pi/3}$. Analogously, energizing phase c alone produces $(2V_c/\pi)\,\mathrm{e}^{\mathrm{j}4\pi/3}$. Finally, since everything is nice and linear, the contribution of the three phases together must be the sum of the contributions of each phase separately, i.e.

* $$\mathbf{V} = \frac{2}{\pi}\left(V_a + V_b\, e^{j2\pi/3} + V_c\, e^{j4\pi/3}\right) \quad , \qquad (08)^*$$

which is the desired answer.

*

COMMENTS : We have already seen the induction motor dynamic model in (4.4.4.11). There we had cast it in traditional engineering language i.e. terms of wire currents i as against surface currents K, and resistances and inductances as against time constants. To get the parameters of the engineering model, we start from the fact that the surface current in the stator equals the wire current divided by the stator circumference times some dimensionless factor α arising from the winding structure etc. Hence we can write $\mathbf{K}_s = \mathbf{i}_s/\alpha R$. The rotor is generally a metal cage with bars around the circumference instead of windings; convention has it to treat its α as the same as that of the stator. This enables us to introduce $\mathbf{i}_r$ and $\mathbf{i}_s$ instead of $\mathbf{K}_r$ and $\mathbf{K}_s$ in (07), giving

$$\frac{\tau_r}{\alpha r}\mathbf{i}_r' + \left(\frac{1 - j\tau_r\omega}{\alpha r}\right)\mathbf{i}_r + \frac{\tau_r}{\alpha' R}\left(\mathbf{i}_s' - j\omega\mathbf{i}_s\right) = 0 \quad ,$$

$$\frac{\tau_s}{\alpha' R}\mathbf{i}_s' + \frac{1}{\alpha' R}\mathbf{i}_s + \frac{r^2}{R^2}\frac{\tau_s}{\alpha r}\mathbf{i}_r' = \frac{\sigma' b'}{h}\mathbf{V} \quad .$$

Then, we realize that the quantities on the RHS of the engineering equations are pure voltages – no extra constants. This gives us a handle on the stator, where the voltage is not zero. If we multiply the RHS, and hence the both sides, by $h/\sigma' b'$, then we'll have the equation in the requisite dimension. In like manner, we should multiply the rotor equation by $h/\sigma b$ so as to get it into the same dimension as the stator one. This yields

$$\frac{\mu h}{2\alpha}\mathbf{i}_r' + \left(\frac{h}{\alpha\sigma r b} - j\frac{\mu h\omega}{2\alpha}\right)\mathbf{i}_r + \frac{\mu h r}{2\alpha R}\left(\mathbf{i}_s' - j\omega\mathbf{i}_s\right) = 0 \quad ,$$

$$\frac{\mu h}{2\alpha}\mathbf{i}_s' + \frac{h}{\alpha\sigma' R b'}\mathbf{i}_s + \frac{\mu h r}{2\alpha R}\mathbf{i}_r' = \mathbf{V} \quad ,$$

which agrees with (4.4.4.11) if we make the identification

$$L_r = L_s = \frac{\mu h}{2\alpha} \qquad \Re_r = \frac{h}{\alpha\sigma r b} \qquad \Re_s = \frac{h}{\alpha\sigma' R b'} \qquad M = \frac{\mu h r}{2\alpha R} \quad .$$

The torque equation (06) we got is the same as what was stated in example 4.4 .4 (without the proportionality constants). So at least we have recovered the engineering model, a dull but important consistency check.

Part (e) shows how one actually generates $V_d \cos\theta + V_q \sin\theta$. The stator phase windings are designed to generate constant voltages across 60° angular sectors, and then taking the first harmonic gives V_d and V_q. In reality, other angular harmonics are also present; they have no utility, only nuisance value. Equation (08), which relates the voltage phasor to the individual phase voltages, is called Clarke transformation after EDITH CLARKE who invented it. In a line supplied motor, each phase is hooked up to a supply line which produces (in our country) 50 Hz ac voltage – the three phases are $2\pi/3$ radians out of phase with each other. Setting the three phase voltages in (08) equal to $V\cos(\Omega t - k\pi/3)$ for $k = 0,1,2$ shows at once that the voltage phasor rotates counter-clockwise at constant angular velocity Ω. In a controlled motor, each phase connects to an inverter. The derivation we did above is for a two-pole motor; if the polarity is $2n$ ($n = 2,3,4.....$) then the fundamental angular harmonics become $\cos n\theta$ and $\sin n\theta$ and ω in (07) gets replaced by $n\omega$.

What are the typical parameter values for a motor ? For a large motor, the height is $O(1)$ m while radius is $O(0.1)$ m. The magnetic permeability is typically about a hundred times greater than μ_0 – an iron or equivalent ferromagnetic core achieves this. A 1 cm diameter wire carries about 200 Amperes; wound multiple times, this can amount to an equivalent surface current of $O(10^5)$ A/m. Now plugging this into (06) gives a torque of about $O(10^4)$ Nm, which is very reasonable. Note the very important role played by the iron core – without it, the torque would get reduced by a factor of about 100 !

Induction motor is one of the two biggest everyday applications of MAXWELL's theories, the other being communication via signal transmission. Unlike the latter one, this has never got its due in the basic electromagnetism curriculum, probably because the literature derives the model using not this approach but equivalent circuit concept. Induction motor was invented simultaneously by NIKOLA TESLA and GALILEO FERRARI and is the first choice in almost all motor applications in industry and technology. This is because of three features : (a) the rotor has no external electrical contacts which are prone to wear and tear, (b) the power density is very high, and (c) when excited with constant frequency

ac, the motor settles into a fixed point with ω constant and torque balancing load. This fp is stable and has a practically infinite basin of attraction, so the motor always comes back to it after any kind of disturbance. At the fp, ω is generally slightly short of the line frequency divided by the semi-polarity n. Since 50 Hz corresponds to 3000 rpm, line supplied motors usually have operating speeds of about 2900, 1450, 950 and similar rpms corresponding to 2, 4, 6 etc pole motors. If you come across a random motor, an rpm in this range is often a hallmark of its being induction.

Almost all motors in factories, workshops and warehouses are induction motors. Non-factory uses include the fancy washing machines and air conditioners which are typically advertised as direct drive or inverter drive. Even the household ceiling fan is a highly watered down (and inefficient) variant of this, called shedded pole motor. And of course, there is our friend WAP7 powered by six hulking induction motors, driven by TAKAHASHI and NOGUCHI's control. In this case, the desirable feature is not the stability of the fixed point but the absence of electrical contacts and the high power at light weight (WAP7's motors are 1150 HP max, 2.1 tons while WAP4's dc motors are 1000+ HP max, 3.5 tons).

Example 6.3.7

Magnetic fridge ornament

Easier form : Parts (b), (d) and (e) only, treating the results of (a) and (c) as given.

A magnetic material of susceptibility χ occupies the lower half-space $z<0$ while the upper half-space $z>0$ is filled with air. A current configuration of finite size is introduced into the air. An image method can be used to determine the magnetic field in the air, as outlined below.

Consider the special case where $\chi=-1$. In this case the material is a perfect flux-expeller (superconductor) i.e. $\mathbf{B}=0$ for $z<0$.

(a) Argue that the material can be removed and replaced with a fictitious current configuration, located entirely in the lower half-space, such that the resultant magnetic field satisfies the BC $B_z=0$ at the surface of the superconductor.

(b) Use the above result to show that for an infinitesimal current element $\mathrm{d}l\left(i_1\hat{\mathbf{x}}+i_2\hat{\mathbf{y}}+i_3\hat{\mathbf{z}}\right)$ of the original configuration located at (0,0,h), the corresponding image current element is $-\mathrm{d}l\left(i_1\hat{\mathbf{x}}+i_2\hat{\mathbf{y}}-i_3\hat{\mathbf{z}}\right)$ located at (0,0,$-h$).

Now consider the general case where χ is arbitrary.

(c) Use the preceding results to show that for an infinitesimal current element $\mathrm{d}l\left(i_1\hat{\mathbf{x}}+i_2\hat{\mathbf{y}}+i_3\hat{\mathbf{z}}\right)$ of the original configuration located at (0,0,h), the corresponding image current element is $\alpha\ \mathrm{d}l\left(i_1\hat{\mathbf{x}}+i_2\hat{\mathbf{y}}-i_3\hat{\mathbf{z}}\right)$ located at (0,0,$-h$) for some number α, and determine α.

(d) If the current configuration be an ideal dipole of strength m pointing in the z-direction, determine the force on it. Discuss such limiting values of χ as you find suitable.

(e) A magnetic ornament is to be attached to a refrigerator door of very large susceptibility χ. The door is assumed to be extremely

thick and the ornament may be treated as a dipole of strength m directed perpendicular to the surface of the door and located a distance d away from it. If the friction coefficient between the ornament and the door is v, determine the minimum value of m necessary for the ornament not to fall off.

[Level : 4]

(a) We have some current configuration in the upper half-space and $\mathbf{B}=0$ in the lower half-space. Evidently, $\mathbf{H}$ and $\mathbf{M}$ are also zero in the lower half, and so is the volume bound current which is $\nabla\times\mathbf{M}$. Hence the only bound current is a surface current $\mathbf{K}_b$ at $z=0$ which somehow manages to cancel the field of the configuration everywhere below it. So the actual problem features the original current configuration and a surface current at $z=0$. For the image problem, we are told to remove the material i.e. the surface current and install a fictitious current configuration such that $B_z=0$ at $z=0$. Since the image is entirely in the lower half-space and we want $\mathbf{B}$ only in the upper half-space, the image is not trivially invalid. We now claim that the solution to the actual problem is $\mathbf{B}$ equal to that of the source plus image above the superconductor and $\mathbf{B}=0$ below.

To prove this claim, we use the BCs (6.3.05) at the surface current i.e. B_z is continuous while B_x and B_y suffer a jump. Evidently, the assumption of $B_z=0$ at $z=0$ for the clamed solution is necessary for consistency with the BC – if we didn't have it then the solution wouldn't work. But then, we can always CHOOSE the surface bound current to satisfy the remaining BCs – given B_x and B_y above (source-plus-image) and below (zero), simply design $\mathbf{K}_b$ to achieve compliance with (6.3.05b). This surface current thus creates zero B_x and B_y just below the sheet.

Now enter the well-definition argument. In any region which is source-free, every Cartesian component of the vector potential $\mathbf{A}$ satisfies a Laplace equation. So do all derivatives of these components with respect to any spatial coordinate, since spatial differentiation commutes with Laplacian. Each component of $\mathbf{B}$ is a combination of such derivatives. Thus, B_x, B_y and B_z are all solutions of Laplace equation in a source-free region – if we specify their values on the boundary of the region, we ipso facto specify them everywhere inside.

This is exactly what has happened in the lower half-space. This space evidently doesn't have sources. Our choice of surface current has forced all components of $\mathbf{B}$ to be zero on the bottom of the sheet, and since the original configuration is finite, they're evidently zero at $|\mathbf{r}|=\infty$. The lower hemisphere of this together with the bottom of the sheet encompass the BC region of the lower half-space. Since $\mathbf{B}=0$ there, it must be so everywhere inside that space. Thus, we have found a solution which has the source where it is, a surface current at $z=0$ and zero $\mathbf{B}$ everywhere below. Since the overall problem is well-defined, it must

have a unique solution; since we've just found ONE solution by explicit construction, it must be THE solution.

(b) We use BIOT-SAVART's law (6.3.04) to find the magnetic field created by the infinitesimal current element of the original configuration at an arbitrary point on the surface of the superconductor i.e. at $(x,y,0)$. Evidently, $i\mathrm{d}\mathbf{l} = \mathrm{d}l\left(i_1\hat{\mathbf{x}} + i_2\hat{\mathbf{y}} + i_3\hat{\mathbf{z}}\right)$ while $\mathbf{r} = x\hat{\mathbf{x}} + y\hat{\mathbf{y}} - h\hat{\mathbf{z}}$. Using these we have

$$\mathrm{d}\mathbf{B} = \frac{\mu_0 \mathrm{d}l}{4\pi}\frac{\left(-i_2 h - i_3 y\right)\hat{\mathbf{x}} + \left(i_1 h + i_3 x\right)\hat{\mathbf{y}} + \left(i_1 y - i_2 x\right)\hat{\mathbf{z}}}{r^3} \quad . \qquad (01)$$

For the image, let's just work it out for the general case of Part (c) where its strength is α, and then set $\alpha = -1$. The image current is $i\mathrm{d}\mathbf{l} = \alpha\ \mathrm{d}l\left(i_1\hat{\mathbf{x}} + i_2\hat{\mathbf{y}} - i_3\hat{\mathbf{z}}\right)$ while the displacement vector to the surface of the superconductor is $\mathbf{r} = x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + h\hat{\mathbf{z}}$. Evaluating the cross product,

$$\mathrm{d}\mathbf{B} = \frac{\mu_0 \mathrm{d}l}{4\pi}\frac{-\alpha\left(-i_2 h - i_3 y\right)\hat{\mathbf{x}} - \alpha\left(i_1 h + i_3 x\right)\hat{\mathbf{y}} + \alpha\left(i_1 y - i_2 x\right)\hat{\mathbf{z}}}{r^3} \quad . \qquad (02)$$

Evidently, $|r|$ and hence r^3 are the same in both (02) and (03). Hence we can see that the contribution of the image is $-\alpha$ times that of the source for the x- and y-components of $\mathbf{B}$ while it is $+\alpha$ times that of the source for the z-component. This fact will be useful later.

For now, setting $\alpha = -1$ immediately shows that B_z for the image is the negative of B_z for the source at every point on the surface of the superconductor. Since we have already argued that this is the BC we need to satisfy, the given image current is indeed the correct one.

(c) This is the part which is tricky. Let's first clearly write down everything related to the actual problem, which we call Problem A. Here, we have air in $z > 0$ (region 1) and magnetic medium in $z < 0$ (region 2). There is an externally applied current in the air and no external currents in the medium. In the air we have $\mathbf{M}_1 = 0$ whereby $\mathbf{B}_1 = \mu_0\mathbf{H}_1$. This field evidently satisfies

$$\nabla\cdot\mathbf{H}_1 = 0 \qquad \nabla\times H_1 = \text{source} \quad ,$$

where "source" refers to the infinitesimal current element under study.

In the material, bound currents $\mathbf{J}_b$ and $\mathbf{K}_b$ are set up, and the material thus acquires a magnetization $\mathbf{M}_2$. By definition,

$$\mathbf{M}_2 = \chi \mathbf{H}_2 \quad \text{, whereby} \tag{03a}$$

$$\mathbf{B}_2 = \mu_0 (1+\chi)\mathbf{H}_2 \quad . \tag{03b}$$

Since there is no free current in the medium,

$$\nabla \times \mathbf{H}_2 = 0 \quad . \tag{04}$$

The divergence of $\mathbf{H}_2$ is $\nabla \cdot \mathbf{H}_2 = -\nabla \cdot \mathbf{M}_2$ into which we can substitute (03a) to get $\nabla \cdot \mathbf{H}_2 = -\chi \nabla \cdot \mathbf{H}_2$ i.e.

$$\nabla \cdot \mathbf{H}_2 = 0 \quad .$$

The volume bound current is $\mathbf{J}_b = \nabla \times \mathbf{M}$ into which we can substitute (03a) and (04) to conclude that $\mathbf{J}_b = 0$ everywhere. Note that the only fact needed to arrive at this conclusion was that there is no free current in the medium – we didn't need any of the geometry details etc for this. Hence, whenever there is a magnetic medium with no free current in it, there is no volume bound current also. Evidently, the surface bound current must be non-trivial – in our Problem A that is given by

$$\mathbf{K}_b = \mathbf{M}_2(0^-) \times \hat{\mathbf{z}} = \chi \left(\mathbf{H}_2(0^-) \times \hat{\mathbf{z}} \right) \quad , \tag{05}$$

where the 0^- in brackets indicates the value of z at which $\mathbf{M}_2$ and $\mathbf{H}_2$ need to be evaluated. The solutions for $\mathbf{H}$ and $\mathbf{B}$ in the two regions are related by the BCs (6.3.08) and (6.3.05) respectively. Moreover, we have the additional BCs that $\mathbf{H}$ and $\mathbf{B}$ are zero at $x^2+y^2+z^2=\infty$, thus providing a complete and well-defined problem for $\mathbf{H}$ and $\mathbf{B}$ everywhere.

Of course, this problem can't be solved explicitly, but before going to the next step we note that the only currents in Problem A are the source and a sheet of surface current at $z=0$. Now, enter Problem B. This problem has air everywhere, the source where it was, and the image current of unknown strength α as given in the question. We shall use a two-step strategy to prove the equivalence between Problem A and Problem B. First, we'll use the result of Part (b) to argue that Problem B is entirely equivalent to another problem, Problem C, which features the source, no material and a current sheet at $z=0$ carrying a certain current distribution whose strength is a function of α. Then we'll show that Problem C is consistent with Problem A for one specific value of α, thus demonstrating the required equivalence as well as finding the required α.

So let's get cracking with Step 1. Recall from Part (a) that in any source-free region, it is sufficient to prescribe B_x, B_y and B_z on the boundary to uniquely specify them everywhere inside. In Problems B and C, the source is common to both, so we can simply take it out and attempt to show equivalence between the magnetic field of the image and that of the surface current, in the upper half-space. With the source removed, this becomes empty space and it is sufficient to show that all components of $\mathbf{B}$ are equal each to each only on the boundary of the domain.

The boundary $|\mathbf{r}| = \infty$ is trivial since $\mathbf{B} = 0$ there in both cases. Now, let $\mathbf{S}_0(x,y)$ be the vector function describing the magnetic field on the $z = 0$ plane created by the source alone and no other current anywhere else. Of course, this is a function we can calculate explicitly – it's given by the formula (01). But we don't need the nasty details of this function, so let's stick to $\mathbf{S}_0$. As we saw from (02), in Problem B the given image alone (in the absence of the source) produces the field components $-\alpha S_{0x}$, $-\alpha S_{0y}$ and αS_{0z} on the plane.

For Problem C, we start from Part (b), using the superconducting results as a benchmark. Since we know that in this case the image is valid with $\alpha = -1$, the field components of source and image together are $2S_{0x}$, $2S_{0y}$ and 0 at $z = 0^+$. Evidently, they are zero at $z = 0^-$, so if we impose the BC (6.3.05b), we get the current distribution $\mathbf{K}_0$ in the superconductor as

$$\mu_0 \mathbf{K}_0 \times \hat{\mathbf{z}} = 2S_{0x}\hat{\mathbf{x}} + 2S_{0y}\hat{\mathbf{y}} \quad \text{, or}$$

$$\mathbf{K}_0 = \frac{2}{\mu_0}\left(-S_{0y}\hat{\mathbf{x}} + S_{0x}\hat{\mathbf{y}}\right) \quad .$$

Now removing the source, we can find the magnetic field components created by $\mathbf{K}_0$ alone at its top surface – they are S_{0x}, S_{0y} and $-S_{0z}$ respectively. Evidently, these components scale with the strength of the current, and for the proper strength they can be made equal to $-\alpha S_{0x}$, $-\alpha S_{0y}$ and αS_{0z}, which correspond exactly to Problem B. The value of current which achieves this is

$$\mathbf{K} = \frac{-2\alpha}{\mu_0}\left(-S_{0y}\hat{\mathbf{x}} + S_{0x}\hat{\mathbf{y}}\right) \quad . \qquad (06)$$

Thus, Problem B is entirely equivalent to Problem C if the current driven through the sheet is given by (06) above.

Now is Step 2 of the calculation i.e. tying up Problem C to Problem A. For this we want the currents (06) to be consistent with (05). We start from the components of $\mathbf{B}$ at $z=0^-$ i.e. below the sheet in Problem C. From (6.3.05b) the sheet alone clearly contributes αS_{0x}, αS_{0y} and αS_{0z}; adding the contribution of the source we get $(1+\alpha)S_{0x}$, $(1+\alpha)S_{0y}$ and $(1+\alpha)S_{0z}$. Forcing (05) to equal (06) implies

$$\chi \mathbf{H}_2(0^-)\times\hat{\mathbf{z}} = -\frac{2\alpha}{\mu_0}\left(-S_{0y}\hat{\mathbf{x}}+S_{0x}\hat{\mathbf{y}}\right) \quad ,$$

Now, plug in (03b) on the LHS to introduce $\mathbf{B}$ in place of $\mathbf{H}$, and then introduce the field components we just calculated – we have (reversing LHS and RHS for clarity)

$$\begin{aligned} -\frac{2\alpha}{\mu_0}\left(-S_{0y}\hat{\mathbf{x}}+S_{0x}\hat{\mathbf{y}}\right) &= \frac{\chi}{\mu_0(1+\chi)}\left(B_{2y}(0^-)\hat{\mathbf{x}}-B_{2x}(0^-)\hat{\mathbf{y}}\right) \\ &= \frac{\chi}{\mu_0(1+\chi)}\left[(1+\alpha)S_{0y}\hat{\mathbf{x}}-(1+\alpha)S_{0x}\hat{\mathbf{y}}\right] \end{aligned}$$

Evidently, the first and third expressions are consistent in form; they match up exactly if and only if

$$2\alpha = \frac{\chi}{1+\chi}(1+\alpha) \quad ,$$

i.e.

$$\alpha = \frac{\chi}{\chi+2} \quad ,$$

which is the desired answer.

(d) Having proved the image construction, now it's time to use it. As always happens, using an image is much easier than proving it (and this particular proof was especially troublesome). The image is given in terms of an infinitesimal current element of the original configuration – by linearity, we can easily extend it to every element of the original configuration i.e. to the original circuit as a whole. Let's now think of the dipole as a small loop of current. The source dipole points in the $+z$-direction, so its equivalent current loop lies in the x-y plane and carries a counter-clockwise current of some strength i when viewed from the top. The image dipole loop lies at $z=-h$ and, by the assertion we just proved,

carries a current $i\chi/(\chi+2)$ counterclockwise when viewed from the top. In other words, since the source dipole is $m\hat{\mathbf{z}}$, the image dipole is

$$\mathbf{m}_1 = \frac{\chi m}{\chi+2}\hat{\mathbf{z}} \quad .$$

Now we have to find the force on $\mathbf{m}$ due to $\mathbf{m}_1$. For this we need the magnetic field created by $\mathbf{m}_1$, and then have to find the force using the messy formula $\mathbf{F} = (\mathbf{m}\cdot\nabla)\mathbf{B}$. Fortunately, among the plethora of potential terms in this expression, most are zero in this case. The first thing is that the force has to be along the z-direction. The source and image dipoles all point along z-axis and the line joining them is also along the same axis – if there is a non-zero x- or y-force, then which way does it point ? The z-component of force is

$$F_z = m_x\frac{\partial B_z}{\partial x} + m_y\frac{\partial B_z}{\partial y} + m_z\frac{\partial B_z}{\partial z} \quad .$$

Here, m_x and m_y are clearly zero, so only the third term survives.

We use (6.3.06) to evaluate $\mathbf{B}$ at an arbitrary $z\,(>0)$ on account of the image dipole $\mathbf{m}_1$. Evidently, $\hat{\mathbf{r}} = \hat{\mathbf{z}}$ in this case and we find

$$B_z = \frac{\mu_0 m_1}{2\pi(z+h)^3} \quad .$$

To find the force, we evaluate the derivative of this at $z = h$ (which is where $\mathbf{m}$ is), multiply the result by m and then use that $m_1 = \chi m/(\chi+2)$, getting

$$\mathbf{F} = -\frac{3\mu_0\chi m^2}{32\pi(\chi+2)h^4}\hat{\mathbf{z}} \quad , \qquad (07)^*$$

which is the desired answer.

First off, we can see certain dependencies of this force. It's quadratic in the dipole strength (easily understandable since the dipole itself and the image both contribute one power) and inverse-fourth in the height above the magnetic material. The force is positive if χ is between -1 and 0 i.e. if the material is diamagnetic then the dipole is repelled away from it. The force is zero if $\chi = 0$ i.e. if there is no material at all (a trivial case). The force is negative if $\chi > 0$ i.e. paramagnet tends to attract the dipole. This is of course consistent with our

everyday experience that a bar magnet is attracted by a block of iron. When the material is a superconductor, the repulsion has the strength $3\mu_0 m^2/32\pi h^4$. The attraction also tends to the same limiting strength as χ tends to ∞; practically, χ greater than about 20 is sufficient for the force to be very close to the limit. Thus, a superconductor repels the dipole with as much force as iron attracts it.

A pathological limit appears to be $\chi = -2$. In this case, the strength of the image diverges. A material with this χ would imply $\mathbf{B} = -\mu_0\mathbf{H}$ and is certainly a weird substance. As far as I know, any material with χ more negative than -1 doesn't exist today; whether it is possible to synthesize one in future is something I am not qualified to comment about. Perhaps there is some physical principle which prevents one from approaching -2; again, this exceeds the scope of my knowledge. Hence I shall refrain from further comment about $\chi = -2$; if anyone can provide any information on this, I'll be grateful.

(e) This is of course a direct application of the previous result. The descriptions of the ornament and the door make it possible for us to use (07) with $h = d$. Since the susceptibility is large, $\chi/(\chi+2)$ is unity and the attraction between the ornament and the door is $3\mu_0 m^2 / 32\pi d^4$. Now consider the mechanical force balance on the ornament. Since the door is vertical, gravity tends to make the ornament slide down the surface. The magnetic attraction is horizontal so by itself it cannot do anything to halt the slide. But, it can employ friction in the vertical direction to balance gravity. The maximum friction force is v times the attraction (which here takes the place of the normal reaction). The lowest possible m necessary for equilibrium occurs when this maximum friction just balances the weight. Letting the mass of the ornament be q (we've used up m, M as well as μ!), we have

$$\frac{3\mu_0 m^2}{32\pi d^4} = vqg \quad ,$$

which implies that

* $$m = \left(\frac{32\pi vqgd^4}{3\mu_0} \right)^{1/2} \quad , \qquad (08)^*$$

which is the desired answer.

*

COMMENTS : Let's estimate the value of m for a typical fridge ornament. We take the mass $q = 0.05$ kg (50 grams), $v = 0.1$ and the distance $d = 2$ mm (that's about half the thickness of the magnetic disc on the back of a typical ornament). Then the formula (08) gives $m = 0.00473$ Am^2. To protect our ornament we would like a safety factor of about 10 (frictional force ten times the weight) so a more realistic m is probably 0.015 Am^2. Treating the dipole as a current loop, if the area of the loop is 1 cm^2 then the strength of the current works out to 150 A. This is quite strong, but permanent magnetic dipoles are generally strong.

A problem which sometimes occurs with ferromagnets is that under certain conditions, the material stops behaving linearly i.e. $\mathbf{M} = \chi \mathbf{H}$ ceases to hold. Two ways of going wrong are hysteresis and saturation. In hysteresis, $\mathbf{M}$ depends not only on what $\mathbf{H}$ is now but also on what $\mathbf{H}$ was at a previous time. In saturation, $\mathbf{M}$ remains a constant even though $\mathbf{H}$ keeps increasing. The effects can be minimized by suitable design of the material, and for many applications such as transformers etc. the linear approximation is good enough. For other applications such as computer memory however, hysteresis is essential, but since this book is NOT about nonlinear differential equations, we let those be.

The way I did the derivation of Part (c), the superconductor acts as a foundation on which the result for general χ stands, instead of being a special case of the more general result. The special status of superconductor arises because it imposes an outright constraint on $\mathbf{B}$ at the boundary ($\mathbf{B} = 0$) instead of specifying all BCs as relations between top and bottom, which happens for any other material. This enabled us to find the superconducting solution directly, without going into the details of the surface current etc. Then we could use this solution to prove the rest of the stuff. Perhaps there is an alternative derivation of Part (c) which does not use the superconductor as an intermediate (or cut a few corners); I don't know of one.

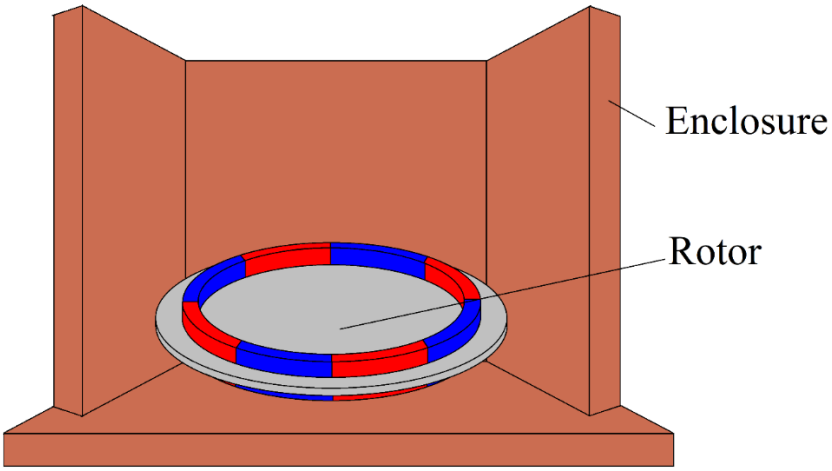
The attraction between magnets and iron of course has many other uses than fridge ornaments, for example in the lifting of heavy iron pieces. The repulsion of a magnet by a superconductor has applications in magnetic levitation. Evidently, if the dipole is mounted above a horizontal superconductor, then the repulsive force can balance its weight. It turns out that superconductor also has certain implications for stability – how that works is in the next Tute.

Example 6.3.8

Magnetic helicopter

Easier form : Part (a) only, treating it as a standalone problem.

A device for performing magnetic levitation is shown schematically in the Figure below.



It consists of a disk-shaped rotor which is fitted with one hundred permanent magnet pole pieces arranged around the circumference (only eight pole pieces shown in Figure). Gravity acts downwards, along the rotor axis. Each pole piece consists of innumerable infinitesimal magnetic dipoles which point vertically upwards and downwards in alternating pieces (these are shown in red and blue respectively). The rotor is mounted inside a five-sided enclosure made of copper – this enclosure has a floor and four walls but lacks a ceiling (the front wall has been excised in the Figure to reveal the other walls and the rotor). The rotor is then rotated at a very high speed about its axis.

The following are the dimensions of the apparatus. The rotor has external radius 0.1 m (10 cm) and the pole pieces are located in the region between radius 10 cm and 9 cm (they are mounted right on the periphery – the clearance from the periphery in the Figure is only for schematic clarity). The height of the pole pieces is 1 cm.

The internal dimensions of the enclosure are 20.6 cm on each side. The nominal operating position of the rotor is at dead centre above the enclosure floor and 3 mm above it. Each pole piece possesses a magnetic dipole moment per unit volume amounting to 10^6 SI Units.

Perform an analysis of the magnetic levitator using the sequence of steps indicated below. Note that the configurations of Parts (a) and (b) are standalone and are not immediately related to the levitator.

(a) Extend the arguments of T106 to obtain an image configuration for a situation where flux-expelling material (superconductor) occupies $x<0$, $y<0$ and $z<0$ and an ideal dipole $\mathbf{m}=m\hat{\mathbf{z}}$ is located at (x_0,y_0,z_0).

(b) A dipole of sinsuoidally time-varying strength is mounted above a semi-infinite block of conducting material. Using a trial vector potential having the spatial dependence exp j$(k_1x+k_2y+k_3z)$ where k_1, k_2 and k_3 may be real or complex, show that the magnetic field penetrates the conductor upto a distance which varies as $(\mu_0\sigma\omega)^{1/2}$.

These results will now be employed for an analysis of the levitator.

(c) Basis the result of Part (b), determine the angular velocity of the rotor for which the calculations of Part (a) will be relevant to obtain the force on it. Discuss how the evaluation of the force may be carried out.

(d) Qualitatively would you expect the rotor to be stable or unstable to translational perturbation from its nominal operating position ? Justify your answer with reference to EARNSHAW's theorem.

(e) Using a computer, determine the mass of the rotor so that the levitation force balances gravity at the equilibrium position. Determine also the response to small translational perturbations in all directions.

[Level : 6]

Our second six and once again it has a gigantic statement. As usual, go through it once and get a summary – the question is presenting some sort of design of a magnetic levitator, and our ultimate task is to find the lift force and show that the equilibrium position is stable (well, presumably – although the question doesn't say "show that it is stable," we really don't expect a prize problem to describe a levitator which DOESN'T work). As always there is a series of guided steps to take us there – having got the big picture once, let it go and focus on the individual steps.

(a) We have already seen in Part (b) of the last Tute that with $m\hat{\mathbf{z}}$ at $(0,0,z_0)$ and flux-expeller in $z<0$, the image is the dipole $-m\hat{\mathbf{z}}$ at $(0,0,-z_0)$. This time, we have a three-way flux expeller. First let's consider a simpler problem – the one where there is only one flux-expeller at $x<0$, $\mathbf{m}$ is along z-direction, and is located x_0 distance away from the expeller. The BC needed this time is that B_x must be zero at $x=0$. As always with images, we need to guess – the lucky strike is $m\hat{\mathbf{z}}$ at $(-x_0,0,0)$. In the formula (6.3.06), r from the dipole to the $x=0$ plane is obviously the same for both the source and the image. We find

$$(\mathbf{m}\cdot\hat{\mathbf{r}})\hat{\mathbf{r}}=\frac{mz}{r^2}(-x_0\hat{\mathbf{x}}+y\hat{\mathbf{y}}+z\hat{\mathbf{z}}) \quad ,$$

for the original, and

$$(\mathbf{m}\cdot\hat{\mathbf{r}})\hat{\mathbf{r}}=\frac{mz}{r^2}(x_0\hat{\mathbf{x}}+y\hat{\mathbf{y}}+z\hat{\mathbf{z}}) \quad ,$$

for the image. The x-component of their resultant is clearly zero; further since $\mathbf{m}$ has no x-component, B_x at $x=0$ for the original-image combo is indeed zero, thus proving the validity of the guess.

Now, with flux-expellers at both $x<0$ and $z<0$, we have the situation of 6.3.3 Part (b) – we'll need $+m$ (z-direction implicit) at $(-x_0,0,z_0)$ and two $-m$'s at $(\pm x_0,0,-z_0)$. These four dipoles satisfy both the BCs in a pairwise manner. Finally, we make a second extension to include the flux-expeller at $y<0$. This time we must reflect all the existing images in the y-axis. This gives a total of seven images – total eight dipoles including the original, located at $(\pm x_0,\pm y_0,\pm z_0)$. Everything in positive z must be $+m$ (one is the original) and everything in negative z must be $-m$. Evidently, this configuration will produce

pairwise cancellation of the normal component of $\mathbf{B}$ everywhere along all three planes, and hence is the desired answer.

(b) Let $z<0$ be the conductor and let there be a dipole $m\cos\omega t$ some distance above it. Inside the conductor, the PDE satisfied by any component of the vector potential is

$$\nabla^2 A = \mu_0\sigma\frac{\partial A}{\partial t} \quad , \tag{02}$$

with boundary conditions $A=0$ at infinite distance in the lower half-space, and matching conditions with a solution of inhomogeneous Laplace equation in the upper half-space which goes to zero at infinite distance. Since there is time-periodic excitation at frequency ω, we expect that A will have the form $f(x,y,z)\cos\omega t + g(x,y,z)\sin\omega t$. Plugging this into (02) and equating cos and sin ωt components each to each gives

$$\nabla^2 f = \mu_0\sigma\omega g \quad ,$$
$$\nabla^2 g = -\mu_0\sigma\omega f \quad .$$

Letting $\alpha=\mu_0\sigma\omega$, we can combine the above two equations into

$$\nabla^4 f + \alpha^2 f = 0 \quad . \tag{03}$$

This is a mess. However, instead of trying to solve it by brute force, we see what exactly the question asks for. It says to use the spatial form

$$f(x,y,z) = (\exp \mathrm{j}k_1 x)(\exp \mathrm{j}k_2 y)(\exp \mathrm{j}k_3 z) \quad , \tag{04}$$

where we understand that the actual solution is a real-valued linear combination of complex exponentials. If any of the k_i's is pure real then its $\exp(\mathrm{j}.....)$ describes oscillatory behaviour (obviously). On the other hand if say k_1 is complex with $k_1 = A + \mathrm{j}B$, then $\exp \mathrm{j}k_1 x$ is $\exp(-Bx + \mathrm{j}Ax)$, which is a decaying oscillation if B is positive. Since complex k's will come in conjugate pairs, one of them will produce a decaying exponential while the other will produce a growing exponential. On physical grounds we must throw the latter away. Plugging (04) into (03), we see that the functional form remains unchanged on taking the derivatives (so at least this IS a possible solution) and we get the relation

$$\left(k_1^2 + k_2^2 + k_3^2\right)^2 + \alpha^2 = 0 \quad \text{, or}$$

$$k_1^2 + k_2^2 + k_3^2 = \pm \mathrm{j}\alpha \quad , \tag{05}$$

which we must now interpret.

On physical grounds we expect k_1 and k_2 to be real – the x-y dependence of A comes from the source dipole whose potential can be expressed as a sum of complex exponentials using the Fourier transform. Then,

$$k_3 = \sqrt{\pm \mathrm{j}\alpha - k_1^2 - k_2^2} \quad .$$

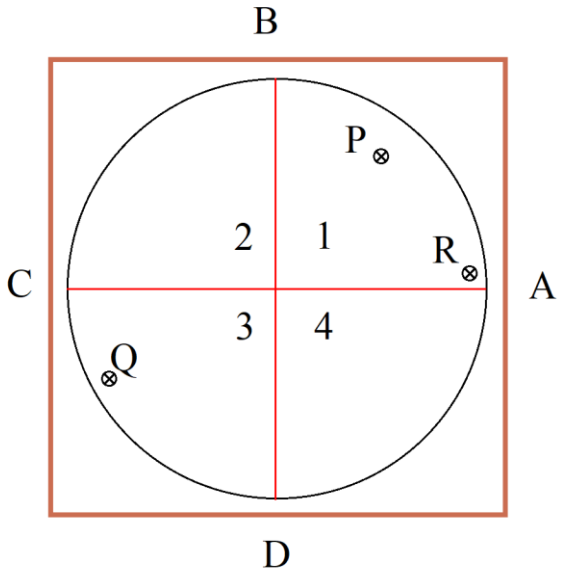
For sufficiently high σ and ω, α will be significantly greater than the typical k_1 and k_2 components present in the field of the dipole. Then, k_3 becomes $\sqrt{(\mathrm{j}\alpha)}$ (sign doesn't matter) which has an imaginary part of size $\sqrt{(\alpha/2)}$. Since $\alpha^2 = \mu_0\sigma\omega$, this indeed contributes to a decay of A in the z-direction with a constant proportional to $(\mu_0\sigma\omega)^{1/2}$, which was to be shown. The penetration depth of an electromagnetic field into a conductor is called the SKIN DEPTH in literature, and is typically obtained from a variety of derivations, none quite rigorous – this is one of them.

(c) What we obtained above is a characteristic relation between penetration depth and frequency for a time-varying dipole above a conductor; what we want is relate it to is the calculation for a dipole above a perfect flux expeller. The latter is an idealized version of the former. Evidently, the approximation of replacing one by the other will hold good if the distance from the dipole to the conductor is significantly greater than the depth to which the field penetrates it. Now for the first time we focus on the 'magnetic helicopter' which started us off on this entire task. The typical clearance between rotor and enclosure is given to be 3 mm, so we should be OK if we can restrict the penetration depth to about 0.3 mm. To be absolutely safe, let's go for another half-order of magnitude and seek $k = (0.1\ \mathrm{mm})^{-1}$ which is 10^4 SI Units. This means that $\mu_0\sigma\omega$ should be 10^8 SI Units. Now in the last Tute we saw that $\mu_0\sigma$ for copper is close to 100 SI units, so this implies that ω should be about 10^6 SI Units.

Although it is tempting to identify this ω with the rotor angular velocity, we see that this is garbage (10^6 rad/s means one crore rpm). So we have to see what exactly ω means here. Because the rotor is rotating, and the pole pieces alternate along its circumference, any particular point in space will see alternating positive and negative m. Treating this point as a dipole which is spatially fixed, we can say that it keeps flipping sign every time a new red or blue

region comes on. Now, we are given that the rotor is centipolar – there are 50 positive and 50 negative pole pieces on its circumference. So if the rotor makes N rotations per second, each point actually goes through $100N$ sign flips and $50N$ periods in one second. Thus, it changes sign with an effective frequency ω such that $\omega/2\pi = 50N$ i.e. $\omega = 100\pi N$. So for $\omega = 10^6$, we need N of about 3000, which is a rotation speed of approximately 30000 rpm. Very high, but nothing implausible – this or higher rpm is commonly attained in ultra-centrifuges, turboprop components and road vehicle engine turbochargers.

Having obtained the condition for the rotor to sit inside a perfectly flux-expelling enclosure, we must now set up the calculation for force using the results of Part (a). For this, we consider the rotor and enclosure in top view.



A, B, C and D denote the four internal walls of the enclosure, and the floor E has not been shown explicitly. The crosses labelled P, Q and R are three dipoles which will be treated as exemplars in the discussion which follows.

What we know to do is calculate the image configuration and hence the force on a dipole with infinite flux-expeller on three sides. We haven't seen or done a calculation for a dipole with two parallel flux-expelling walls in any one direction. Although such a calculation is doable (we can define an infinite set of images, ever further away, which results in successive cancellations of $\mathbf{B}$ on first one wall and then the other), it is unnecessary. For the dipole located at P, the closest elements of the enclosure are walls A and B and the floor E (the floor is equally close to all dipoles on the rotor). So as a first approximation, we can assume that

P sees walls A and B only and is unaffected by C and D. Likewise, dipole Q is closest to walls C and D and is predominantly unaffected by walls A and B. So we partion the rotor into four quadrants labelled 1 to 4 and claim that a dipole within each quadrant sees only the neighbouring walls and nothing else.

How good or bad is the approximation ? For that, we go into the nature of the force between two dipoles. The field of a dipole goes as the inverse cube of the distance; since the force features $\nabla\mathbf{B}$, it becomes inverse fourth power. So for dipole P, the distances PC and PD are at least double of PA and PB, and the thrown away terms are quite small. However, P itself is also quite an insignificant point so far as force is concerned. Recall that the rotor radius is 10 cm while the clearance from the wall is only 3 mm. The really large force terms will arise only from the points on the rotor which are extremely close to the wall. Such a point is R. To do the calculation to high accuracy, we might just as well have focussed on four narrow segments of the rotor near the four walls and considered the response from the proximal wall. Since the definition of such segments would become arbitrary, I have gone for the quadrant split factoring in the two neighbouring walls – this picks up all the really significant dipoles like R (with primary contribution RA) and a lot of small (but harmless) contributions like PA and PB. What it does not pick up (like PC and PD) is negligibly small.

Thus, to find the force on the rotor, we will partion it into four quadrants, and for each dipole in a particular quadrant, set up the seven corresponding images which will arise from the neighbouring walls. Evidently, a computer will be needed for this. This will force us to treat the continuous magnetic material of the rotor as a set of discrete point dipoles at certain specific locations. Then, we'll obtain and store the set of all images for all these dipoles. To get the force on any one dipole, we'll calculate the magnetic field contributed at its location by all the images in its quadrant; we'll then take $(\mathbf{m}\cdot\nabla)\mathbf{B}$. Since $\mathbf{m}$ is in z-direction, the non-zero component of $F_{x,y,z}$ will be $m(\partial B_{x,y,z}/\partial z)$ where the subscripts correspond. Adding the force components over all the rotor dipoles will yield the resultant force on the rotor.

(d) Although we have set up an algorithm for numerical determination of the force on the rotor, it is vital to get an analytical feel for the device – otherwise we'll have no idea of whether the numerics is returning plausible results or garbage. As I already said before starting the solution, qualitatively we would

expect the rotor to be stable simply because this fancy problem has been posed on it. That of course does not make an acceptable explanation so here goes.

To get a qualitative handle on the rotor's motion, we focus only on four dipoles located near the four walls (R in the Figure and corresponding ones in the other quadrants). The primary images will be generated across the proximal wall and across the floor, for a total of three image dipoles (we ignore the distant neighbour wall like B for R). The question to ask is, do the images attract or repel their respective sources. We'll do the calculation in terms of R, and the rest will follow by default.

Let R be positive i.e. pointing upwards. Then, the image in the floor will be located directly below it and pointing vertically downwards. With the image being the origin and R being the target point, $\left(\mathbf{m}\cdot\hat{\mathbf{r}}\right)\hat{\mathbf{r}} = -m\hat{\mathbf{z}}$ while $-\mathbf{m}$ is $+m\hat{\mathbf{z}}$; since the former term carries a 3 in front of it, the net field is along $-z$-direction. Because of the $1/r^3$ in the denominator, this becomes weaker as one moves upwards so $\partial B_z/\partial z$ is positive and the force on R is in positive z-direction. Since this image generates no B_x or B_y along the z-axis, $\partial B_{x,y}/\partial z = 0$ and the force in x- and y-directions is zero, as we would expect from symmetry anyway. So this image repels R. Now consider the image of R in the wall. This is located directly to the right of R and pointing vertically upwards. With origin at the image and R being the target, we can see that B_x exactly at R is zero. But what is B_x slightly above R ? Since $\mathbf{m}$ is along $\hat{\mathbf{z}}$, the only term in B_x arises from $\left(\mathbf{m}\cdot\hat{\mathbf{r}}\right)\hat{\mathbf{r}}$; the dot product is positive while $\hat{\mathbf{r}}$ has a negative x-component. Hence B_x at slightly above is R is negative and $\partial B_x/\partial z$ is negative. Hence R feels a force in negative x-direction, and once again it gets repelled by its image.

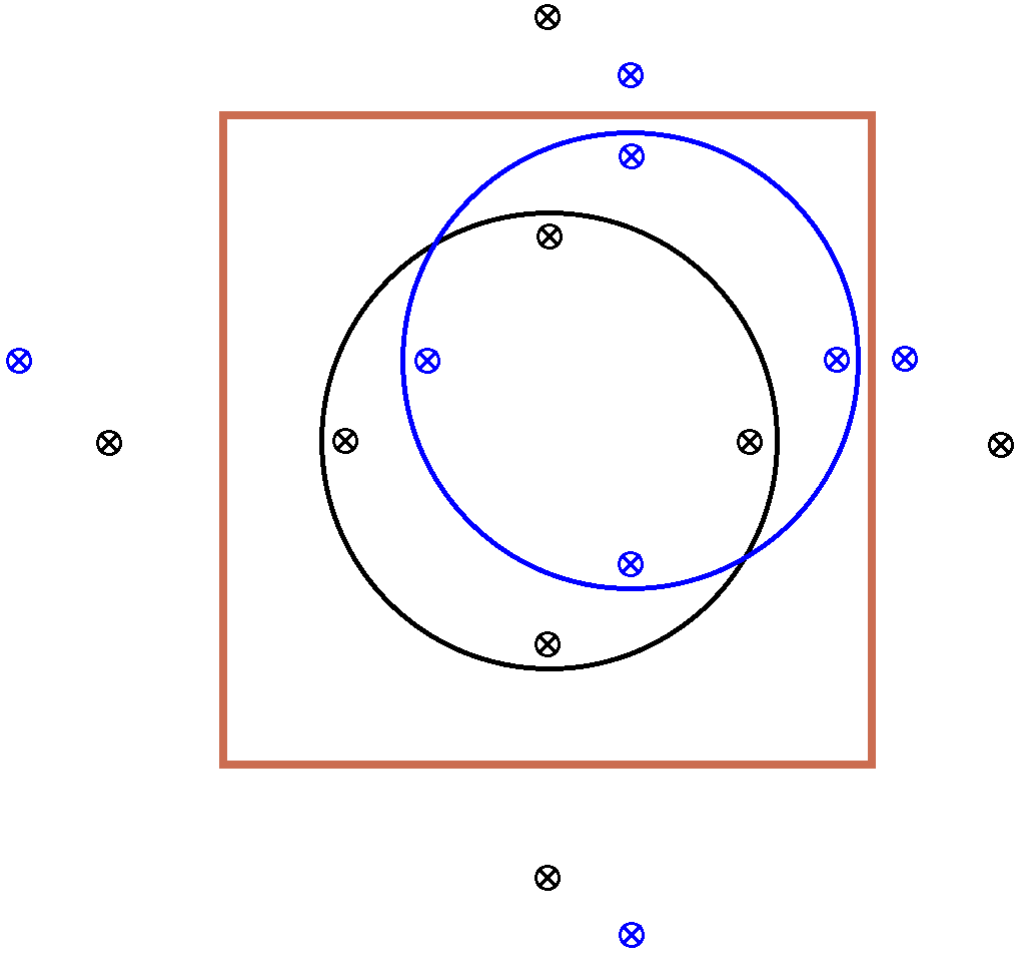
Finally we have to account for the image generated by double reflection in the wall and the floor. This image points along $-z$ and is located below and to the right of R. Now, we saw that an opposite dipole below, and a parallel dipole to the right both repel R. So, we would expect that an opposite dipole below and to the right will repel it in z-direction and attract it in x. However, this image is located considerably further from R than the two direct images – if the separation from the floor and walls are equal, this is further away by a factor of $\sqrt{2}$. Then, since force varies as inverse fourth power of separation, it will be less by a factor of 4 and we guess that this image will not overturn the effect of the

two primary images. Thus we conclude that the dipole R will be repelled from both the floor and the wall.

Once this is out, we can proceed with the stability analysis. We recall that the distance between the dipole and the image is double of that from the dipole to the wall or floor. Evidently, at equilibrium, the repulsion from the floor balances the weight of the rotor. If we move the rotor closer to the floor, then R and its image come closer to each other so the repulsion force increases, pushing the rotor upwards. If we move the rotor upwards, the repulsion weakens and gravity brings it downwards. So the rotor is stable to perturbations in z. As regards x-y motions, in equilibrium the rotor is equidistant from all four walls so the repulsive forces from all walls balance each other out. Let's give it a perturbation in the x-direction, moving it closer to the right wall. Then, R comes closer to its image and feels a stronger repulsion in the $-x$-direction; on the other hand the diametrically opposite dipole moves away from the corresponding wall and feels a weaker repulsion in $+x$-direction. Hence the resultant force on the rotor is in the $-x$-direction, i.e. opposite to the applied perturbation. Since the apparatus is symmetric w.r.t x and y, perturbations in y-direction should display identical response. This implies that the rotor is stable to perturbations in all directions. (Here we can see what would have happened if we had replaced the copper of the enclosure by iron, whose χ is positive – the enclosure would have attracted the rotor and thrown stability out the window.)

Challenge : "What your image talk boils down to is that the rotor is being repelled on all sides. But instead of this fancy induced current business, I might just as well throw away the enclosure and replace it with a stator which consists of a set of dipoles corresponding to every image of every dipole in the rotor's equilibrium position. Then also, these stator dipoles will repel the rotor dipoles, a displacement will increase the repulsion on one side at the expense of the other, the rotor will experience a restoring force, and you will conclude that it is stable. However we know that this is impossible from EARNSHAW's theorem. Therefore your argument is false."

Fortunately it's the challenge and not the argument which is false. To see this, consider the following Figure, which shows the actual rotor inside its enclosure as well as the one surrounded by the hypothetical stator.



We remove the floor and ignore the z-direction completely, focussing on the x-y motions alone. Once again we restrict ourselves to the walls which are the dipoles' immediate neighbours.

The big black circle is the equilibrium configuration of the rotor with four dipoles near the walls being shown. The black dipoles outside the enclosure are the constituents of the hypothetical stator – these locations correspond to the images of the rotor dipoles in the enclosure. Now consider the rotor in its displaced configuration, in blue (displacement exaggerated). The stator dipoles remain fixed, in black, and all stator dipoles repel all rotor dipoles. Now you can see that at the top and right sides, the mutual repulsion between rotor and stator dipoles is almost perpendicular to the rotor's displacement. At the bottom and left sides, the repulsion has a significant component along the rotor's displacement and is parallel to it in sign. Clearly, this component destabilizes the rotor. On the other hand, with the enclosure instead of the stator, the images relocate as soon as the rotor is displaced. The new images are

shown in blue outside the enclosure – clearly they repel the rotor back to its equilibrium position.

Thus, even though we used a static image to obtain the force, the magnetic helicopter is a quintessentially time-dependent configuration. The images MOVE WITH THE ROTOR, continuously positioning themselves in such a way as to make the rotor stable. To achieve this effect with a stator, we would have had to have a control software which strengthens and weakens the various dipoles accordingly as the rotor moves. Then also we would have a fundamentally time-dependent configuration, and escape EARNSHAW's theorem, which does NOT apply to time-dependent configurations. The benefit of the eddy current and image setup vis-a-vis an active control system is that here it's Uparwala who is doing the controlling, so there's no fancy setup and no risk of error. Thus, having analysed the configuration in the light of EARNSHAW's theorem, we can finally conclude that the rotor will indeed be stable to all perturbations.

(e) We set up the calculation of Part (c) on the computer. The strength of the total dipole moment of all the pole pieces will be the volume occupied by them times the $\mathbf{m}$ per unit volume of the permanent magnets (neglecting sign). This latter is given to be 10^6 SI Units (which corresponds to a neodymium permanent magnet); doing the simple calculation we get the total to be about 5 SI Units. We divide this equally among all the point dipoles we use for the numerical integration. Here I have used a total of 4500 per quadrant, more or less evenly spaced. The code is somewhat lengthy but conceptually simple. All it does is just what Part (c) says – calculate and store all the images, then calculate $\mathbf{B}$ and then $\mathbf{F}$. After checking that the code passes some test cases (zero x-y force at equilibrium, same force when displaced in x as in y etc) we now use it to report answers.

We find that the force on the rotor at a distance of 3 mm above the floor is 55 N, which supports a mass of 5.5 kg against gravity. The rotor experiences a restoring force in response to all displacements – the spring constant is 3200 N/m in the horizontal plane and just above 20000 N/m in vertical. Hence, a translational perturbation from the equilibrium causes harmonic oscillations with frequency 3.8 Hz in the horizontal plane and 9.6 Hz in vertical.

*

COMMENTS : Let's make a summary and recap of this huge calculation. Induced currents are an acceptable way of generating magnetic levitation and confinement because they aim at the caveat in EARNSHAW's theorem which says "time-independent". The thrust of the magnetic helicopter is that it uses a very high rotor polarity (we took the rotor to be centipolar) and elevated rotation speed to achieve flux exclusion inside the enclosure within a very small distance. This enables it to be operated with very small clearance from the walls and floor, resulting in the images being very close and generating a high lift force. The more we increase the speed, the smaller we can make the clearance – the latter goes as $\omega^{-1/2}$. Now since the force goes as the inverse fourth power of the clearance, the lift force and spring stiffnesses increase as ω^2. So at 100000 rpm, this rotor will support a 50+ kg weight. Here we considered pole pieces in the ring between 9 and 10 cm; more pole pieces will give rise to more lift. Further, we used permanent magnets here; if we use electromagnets on the rotor, which are excited by currents as in self-excited synchronous motors, then we can increase the lift and stiffness still further. The calculation method is very general and can easily be extended to other configurations of rotor magnets – we can decompose every such configuration into a suitable array of ideal dipoles and sum over them all.

In Part (d) we made an assumption that the images relocate instantly with changes in the position of the rotor. How valid is this assumption ? To examine this issue, we recognize that the image corresponds to the steady state of vector potential inside the conductor with the time-varying excitation. Displacements of the rotor will give rise to transients, which satisfy the homogeneous equation. Since this is 3+1 heat equation, all transients will decay – we must calculate the typical time constant to ascertain the validity of the instant-image approximation.

The governing equation (02) is

$$\frac{\partial^2 A}{\partial x^2} + \frac{\partial^2 A}{\partial y^2} + \frac{\partial^2 A}{\partial z^2} = \mu_0 \sigma \frac{\partial A}{\partial t} \quad , \tag{06}$$

(scalar A denoting any component). The homogeneous solution can have any given spatial dependence so we try an A of the form $\exp \mathrm{j}\,(k_1 x + k_2 y + k_3 z)$, hoping to use Fourier transform to generate arbitrary functions as linear combinations. But what will the typical k values be ? Consider the first configuration of Part (a)

– a dipole a height z_0 above an infinite conducting sheet. The vector potential of a dipole is given by the formula

$$\mathbf{A} = \frac{\mu_0}{4\pi}\frac{\mathbf{m}\times\mathbf{r}}{r^3} \quad .$$

Now consider what happens if one starts at $(x,y) = (0,0)$ and moves along the sheet in the x-direction. The only non-zero component is A_y – at $x = 0$ it is zero, while at $x = \infty$ (both positive and negative) it is zero again. In between it has a maximum at the location $x = z_0/\sqrt{2}$ and a minimum at $x = -z_0/\sqrt{2}$. So A_y is something with crest and trough separated by distance $\sqrt{2}z_0$; we expect that the dominant k_1's in its Fourier transform will be of order the inverse of this, which is approximately $1/z_0$. From symmetry, k_2 will be of the same size.

Now, separating variables on (06) we can easily see that a solution with spatial form $\cos k_1 x \cos k_2 y \cos k_3 z$ will decay as $e^{-\gamma t}$ where $\gamma = \mu_0\sigma(k_1^2+k_2^2+k_3^2)$. The slowest decay will come from the case $k_3 = 0$. In this case, with k_1 and k_2 being about (1 cm)$^{-1}$, the decay rate works out to $\gamma = 10^6$ s^{-1}. So all transients are killed off within a time of the order of 10^{-6} s. On the other hand, the motions of the rotor have timescale 0.1 s. So there are five orders of magnitude separating these two timescales and the approximation is for all purposes exact. Even though the logic leading to the answer was somewhat hand-waving, it is plausible enough not to be incorrect by a factor of one lakh.

What about the rotational stability ? If we tilt the rotor in any direction, then the side which goes down will feel a stronger repulsion from the floor than the side which goes up. This will amount to a torque in the direction opposite to the tilt. On the other hand, the repulsions from the walls will now act along two parallel lines instead of the same line, producing a torque parallel to the tilt. A dipole in a magnetic field also experiences a torque of $\mathbf{m}\times\mathbf{B}$; this too can be stabilizing or destabilizing according to the relative contributions from the floor and the wall. However, the rotational dynamics is not of too much concern since a rotor doing 30000 rpm has an enormous amount of gyroscopic stability – even a strongly destabilizing torque will merely cause it to precess instead of flipping over.

Finally you might wonder why I chose a rectangular box instead of something cylindrical or spherical. In 6.3.3 we'd seen image configurations for all three

geometries so wouldn't those have been more realistic ? Yes they would. BUT, those configurations don't have magnetic images, at least simple ones. A point charge outside a grounded sphere has an image point, but a magnetic dipole outside a superconducting sphere doesn't have an image dipole. So those calculations would have been far more difficult than this – we'd have had to find the induced currents explicitly. On the other hand, the principal physical conclusions i.e. stability of the rotor cannot be dependent on the geometry so we can reasonably expect that a rotor inside a cylindrical or spherical enclosure will be stable also.

*

This is the longest and hardest Unit so far and the second-most formidable in the entire book ("oh SECOND-most ?! ....."). I just hope that the reward is worth the effort. Classical electromagnetism is tailor-made for our course because MAXWELL's equations are linear PDEs – there's no question of modeling or approximation to arrive at the starting equations. And it admits some of the richest and most interesting theory and applications of differential equations. So much so in fact that although this Unit is over, we aren't done with Sir JAMES yet. All you $\mathbf{E}$ and $\mathbf{B}$ fans out there, we'll soon be back with a bang !

6.4

STEADY STATE FLUID MECHANICS

*The derivations of the basic equations often require tedious and difficult vector calculus manipulations. The **solution** of the equations however does not require this. Hence I will skip all the details of multivariate calculus and present the results straightaway.*

In continuation of the style of the last Unit, we'll see here another example where Laplace-like equations apply to a physical situation. This situation is fluid mechanics. However, unlike in electromagnetism where the governing equations are linear, the central PDE here is nonlinear as hell. This makes it intractable to solve. In most of the problems, we'll make certain heavily simplifying assumptions which will reduce the equations to linear PDEs, and easy ones at that. In one situation however, we'll be forced to confront the nonlinearities head-on. To keep things manageable, we restrict ourselves throughout to situations where the flow is steady state i.e. the velocity of the fluid at any point is independent of time.

Fluid mechanics is concerned with describing the flow of a fluid – the basic variable is the vector function $\mathbf{v}(\mathbf{r},t)$, the velocity of the fluid at every point in space and time. Auxiliary variables are the density ρ and the pressure P – both scalar functions of space and time. The **continuity equation** is

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0 \quad , \qquad \textbf{(01)*}$$

whose meaning is that the change in quantity of fluid within a volume equals the amount flowing in minus the amount flowing out (recall the similar equations for conservation of electric charge and heat energy). In cylindrical coordinates, when ρ denotes the radius, we'll use d to mean the density.

INVISCID FLOWS

We now make the assumption that the fluid has zero viscosity. It will remain in force until I take it down explicitly. The core equation is obtained by

writing NEWTON's law on every element of the fluid – the rate of change of velocity is proportional to the force. However, since the fluid is moving, the time derivative is not a simple $\mathrm{d}/\mathrm{d}t$ but a material derivative $\mathrm{d}/\mathrm{d}t + \mathbf{v}\cdot\nabla$, like the one which appeared in the convection diffusion equation of T093. The force is the negative gradient of the pressure. Putting these together we get

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} = -\frac{1}{\rho}\nabla P \quad , \tag{02}$$

which equation is due to LEONHARD EULER. This equation is valid component-by-Cartesian-component of $\mathbf{v}$; recalling that $\mathbf{v}\cdot\nabla$ means $v_x(\partial/\partial x) + v_y(\partial/\partial y) + v_z(\partial/\partial z)$, we can see that this is heavily nonlinear.

The flow is called **incompressible** if the density is independent of space and time i.e. $\rho = \text{const}$. Density fluctuations within the fluid satisfy wave equations whose solutions travel at the speed of sound, so incompressible is a good approximation if the flow speed is much less than that of sound. Since many flows do occur in such a regime, the approximation is often a good one. In this case, the continuity equation (01) reduces to $\nabla\cdot\mathbf{v} = 0$ i.e. the **velocity is divergenceless**, like a magnetic field. Moreover, in EULER's equation we can write the RHS as $-\nabla(P/\rho)$, treating the whole thing as a scalar function.

Incompressibility however does not take care of the nonlinear term. An assumption which is inherent in (02) is that the fluid is inviscid i.e. there is no viscosity. In the presence of viscosity, there are additional forces on the fluid which are not included in (02) – we will introduce viscosity later.

The **vorticity ω** is defined as the **curl of v**. With some multivariate calculus manipulations, it can be shown that (02) can be written in terms of vorticity as

$$\frac{\partial \mathbf{v}}{\partial t} + \nabla\left(\frac{1}{2}v^2\right) - \mathbf{v}\times\boldsymbol{\omega} = -\frac{1}{\rho}\nabla P \quad , \text{ or} \tag{03a}$$

$$\frac{\partial \mathbf{v}}{\partial t} + \nabla\left(\frac{1}{2}v^2\right) - \mathbf{v}\times\boldsymbol{\omega} = -\nabla\left(\frac{P}{\rho}\right) \quad , \tag{03b}$$

the second version holding true when the fluid is incompressible. Considering this latter case, the first term on the LHS vanishes in the steady state. Then, if $\boldsymbol{\omega} = 0$ we have

$$\nabla\left(\frac{P}{\rho}+\frac{1}{2}v^2\right)=0 \quad \text{, or}$$

$$P+\frac{1}{2}\rho v^2=\text{const.} \quad . \tag{04}$$

This is called BERNOULLI's principle, and holds true over the entire fluid.

When $\boldsymbol{\omega}$ is non-zero (but the flow incompressible and steady state), the relation (04) still holds but only along a streamline, which is a line tangent to the velocity field. To see this, we integrate (03b) along the streamline – the cross-product is directed perpendicular to the streamline so it vanishes, leading to the desired result. However, the constant may be different along different streamlines. We note that BERNOULLI's principle is absolutely NOT a statement of conservation of energy.

Zero vorticity is in fact one of the situations where fluid mechanics trivializes to an elementary level. A flow with $\boldsymbol{\omega}=0$ is called irrotational. Now, if the flow is **incompressible, inviscid and irrotational**, then its divergence as well as curl are zero. The second one enables us to write $\mathbf{v}=-\nabla\psi$ where ψ is the potential; the first one then implies $\nabla^2\psi=0$. This is the same setup as in electrostatics – the big difference is that for electrostatics, this is the fundamental equation, while here it came about only after the "3I" approximations. Although the first one is not too restrictive, the second and third are very much, almost cripplingly, so. This toy flow is called **potential flow** or **3I flow**.

What are the boundary conditions on 3I flow ? BCs are typically prescribed at infinite distance and at the surface of solids in contact with the fluid. The former one depends on the problem, and is usually obvious. The latter one is as follows : at the surface of a solid, the flow can only have a component parallel to the surface. Hence $v_\perp=0$. This condition is reminiscent of the BCs on electric and magnetic fields. There is however, NO constraint on the parallel component of the flow – that is free to take on whatever value comes from solving for the flow in the bulk.

VISCOUS FLOWS

Now we consider a viscous fluid. This fluid tends to resist velocity gradient or shear i.e. if adjacent layers of fluid are flowing with different speed, the viscosity tries to bring both of them to the same speed. The **viscosity** η is defined as the force per unit area per unit velocity gradient. (Its SI Unit is called Poiseuille – as typical values, the viscosity of air is $1.8\text{x}10^{-5}$, that of water is about 10^{-3} and that of honey is about 5 SI Units.) It can be shown that the viscous force acting on an infinitesimal volume element of the fluid is $\eta\nabla^2\mathbf{v}$. Adding this term to (02), and assuming the flow to be incompressible we have

$$* \qquad \frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} = -\nabla\left(\frac{P}{\rho}\right) + \frac{\eta}{\rho}\nabla^2\mathbf{v} \quad . \qquad \textbf{(05)*}$$

This is called **Navier-Stokes (NS) equation** after CLAUDE LOUIS NAVIER and Sir GEORGE GABRIEL STOKES. It is one of the most celebrated PDEs in the Literature on account of the inordinate difficulty (indeed, impossibility) of finding exact solutions.

The NS equation in its original form (05) is burdened by an additional complication coming from the pressure term. To chase it away we take the curl on the both sides. This introduces the vorticity $\boldsymbol{\omega}$ (which we recall is the curl of $\mathbf{v}$) but kills off the pressure since curl of gradient is zero. Some multivariate manipulations give the **NS equation in vorticity form** :

$$* \qquad \frac{\partial \boldsymbol{\omega}}{\partial t} + (\mathbf{v}\cdot\nabla)\boldsymbol{\omega} - (\boldsymbol{\omega}\cdot\nabla)\mathbf{v} = \frac{\eta}{\rho}\nabla^2\boldsymbol{\omega} \quad . \qquad \textbf{(06)*}$$

Of course, there is no relief from the nonlinear differential operators, but at least the pressure has gone away. If we can somehow obtain a solution (at least an approximate solution) to this equation, then we can always back-substitute $\mathbf{v}$ into (05) to determine P. We again note that (05) and (06) both hold true for incompressible flows only – compressible flows bring their own complications which we will not go into here.

The boundary conditions on viscous flow are considerably different from those on 3I flow. When a viscous fluid flows past a solid surface, the flow is by definition **stationary relative to the surface**. That means zero normal **and** zero parallel component. This BC is actually the primary difference between viscous and inviscid flows, and it is responsible for some of the most

spectacular fluid mechanical phenomena, as we'll see in the Tutes. An important property of the NS equations is that they are **invariant under Galilean transformation**. This fancy phrase means nothing but that they are the same from every reference frame which is moving at a constant velocity. Thus, if an aircraft is flying through still air, it is just as legit to perform a calculation (of lift, drag etc) from the ground frame as from a frame moving with the aircraft itself, treating the plane as stationary and the air as flowing past.

The question arises, under what circumstances can we dispense with the nonlinear term in (05) ? The other spatial derivative term in this equation is the Laplacian on the RHS – we can chuck the nonlinear term relative to the Laplacian if it is orders of magnitude smaller. If L is a typical system size and v a typical velocity, then $(\mathbf{v}\cdot\nabla)\mathbf{v}$ has the scale of v^2/L while $\nabla^2\mathbf{v}$ has the scale of v/L^2. The Laplacian dominates over the nonlinearity if $\eta v/\rho L^2 >> v^2/L$. We can rewrite this inequality as $(vL\rho/\eta) << 1$. This last quantity is so important that it has a name, **Reynolds number**. This is defined as

$$R = \frac{vL\rho}{\eta} \quad .$$

For $R<<1$, we can throw the nonlinearity out of the Navier Stokes equation and have $\eta\nabla^2\mathbf{v} = \nabla P$ in the steady state. Parallelly, we can throw out all the nonlinearities from the vorticity equation (06); it then reduces to a Laplace equation for every Cartesian component of $\boldsymbol{\omega}$.

So at low Reynolds number, we at least have some linear PDEs. Their solution is considerably difficult since only the vorticity satisfies Laplace equation. Velocity then satisfies $\nabla\cdot\mathbf{v} = 0$ and $\nabla\times\mathbf{v} = \boldsymbol{\omega}$. Low Reynolds numbers are also not physically very realistic as most everyday phenomena occur at rather high values of this parameter. For example, a cricket ball in air has R about 10000; an aircraft at cruising speed has R about ten millions. Nevertheless , low R analyses at least yield the germ of the phenomena which occur at full-blown R and can serve as starting points for expansions and approximations which throw light on this more realistic but still more difficult case.

Finally, suppose you have a solid body inside a fluid – then how do you calculate the force on the solid ? You do this by first obtaining the pressure all

over the body. Then, the force on every infinitesimal area of the body surface is $P\,\mathrm{d}A$ acting along the inward normal. Integrating all over the body gives you the force. This calculation is a very useful one since it's actually the final step in finding the lift and drag on a solid surrounded by a flowing fluid.

Example 6.4.1

Facilissimo flows

Each of the two situations below features the incompressible flow of a viscous fluid in a very simple configuration. Write BCs on the Navier Stokes equation for the velocity field $\mathbf{v}(\mathbf{r})$ and determine $\mathbf{v}$ in each case.

(a) There is an infinite plate at $z = 0$ which is held stationary. A second infinite plate is at $z = h$ and it translates in the x-direction with constant velocity v_0. Fluid fills the space between the two plates. It flows exclusively in the x-direction and the velocity is a function of z alone. The pressure is constant everywhere.

(b) There is a strip at $z = 0$ which extends from $x = 0$ to $x = L$ and infinitely in y. There is a similar strip at $z = h$. Fluid fills the intervening channel, flowing exclusively in the x-direction. A pressure of P_1 is applied at the left end of the channel and P_2 at the right end with $P_1 > P_2$. The pressure is independent of y and z and its gradient in the x-direction is a constant.

[Level : 1]

(a) A trivial problem after a long time, and we sure need the respite from difficulties. The steady state Navier Stokes equation is

$$(\mathbf{v}\cdot\nabla)\mathbf{v} = -\nabla\left(\frac{P}{\rho}\right) + \frac{\eta}{\rho}\nabla^2\mathbf{v} \quad .$$

Since the pressure is given to be constant, its gradient vanishes. The velocity is only in the x-direction so the big bad nonlinear term is $(\mathbf{v}\cdot\nabla)v_x$. However, v_x is a function of z alone so the $v_x(\partial/\partial x)$ and $v_y(\partial/\partial y)$ terms in $\mathbf{v}\cdot\nabla$ both evaluate to zero; since $v_z = 0$, the third term also follows suit and the nonlinearity is off ! So the reduced NS equation is $\nabla^2 v = 0$ where scalar v means v_x. But there's even more simplification. The Laplacian here has only the one term $\partial^2/\partial z^2$ so the equation for the flow is actually an ODE

$$\frac{\mathrm{d}^2 v}{\mathrm{d}z^2} = 0 \quad .$$

No wonder the problem is facilissimo.

Because the flow is viscous, the fluid must acquire the velocity of the two plates at the top and bottom. Hence $\underline{v(0) = 0 \text{ and } v(h) = v_0}$. The specific solution of the ODE which fits these two BCs is $\underline{v = v_0 z/h}$. Hence this is the required flow field.

(b) Very similar to Part (a) – v_x a function of z alone so off go the nonlinearities as well as the partial derivatives. The only difference is that the pressure gradient is not zero but a constant – its value must be $(P_2 – P_1)/L$. Letting $\Delta P = P_1 - P_2$ (which is positive), the pressure difference is $-\Delta P$ and the easified NS equation for this case is

$$\frac{\eta}{\rho}\frac{\mathrm{d}^2 v}{\mathrm{d}z^2} = -\frac{\Delta P}{\rho L} \quad .$$

Because the flow is viscous, it shares the velocity of the two plates hence $\underline{v(0) = v(h) = 0}$.

I will present the specific solution straightaway. We have

* $$v = \frac{\Delta P}{2\eta L} z(h - z) \quad ,$$ *

which is the desired answer.

*

COMMENTS : The first of these flows was discovered by MAURICE COUETTE and the second by JEAN POISEUILLE. Although these are extremely simple flows, we must remember that they nevertheless are exact solutions of the full Navier Stokes equation. The simplification occurred because the configurations allowed it and not because we manually threw terms out of the NS equation.

Couette flow is observed in reality if the velocity of the top plate is small. Likewise, Poiseuille flow is observed if the pressure gradient is not too large. When these limits are breached, the flows lose stability to far more complicated patterns and/or turbulence. Finding the transition point is no longer easy – for Couette flow, it was calculated by Sir GEOFFREY INGRAM TAYLOR in a 50-page paper, while the corresponding calculation for Poiseuille flow was a paper by WERNER HEISENBERG of similar length and complexity. We shall not attempt to examine or understand those or similar calculations here.

Example 6.4.2

3I Flow past a cylinder

A cylinder of infinite length and radius R is stationary relative to the laboratory, with its axis oriented along the z-axis. It is surrounded by a fluid, occupying all of space, which flows in the x-y plane. The flow is incompressible, inviscid and irrotational. Far away from the cylinder, the flow velocity is $V\hat{\mathbf{x}}$ where V is a constant.

(a) Determine the velocity everywhere.

(b) Determine the resultant force on the cylinder per unit length.

The cylinder is now made to rotate about its axis with angular velocity Ω.

(c) Again determine the resultant force.

[Level : 2]

(a) Since the flow is 3I, we write the velocity as $\mathbf{v}=-\nabla\psi$; coupled with $\nabla\cdot\mathbf{v}=0$ it gives the Laplace equation for ψ. Since the geometry is cylindrical, it is natural to write this in cylindrical polars; since the configuration extends infinitely in z, only ρ and θ are the significant variables. The boundary conditions come from the velocity. At $\rho=\infty$, $\mathbf{v}=V\hat{\mathbf{x}}$ which is $V\left(\cos\theta\hat{\boldsymbol{\rho}}-\sin\theta\hat{\boldsymbol{\theta}}\right)$. At $\rho=R_0$, $v_\rho=0$. This completes the formulation of the problem.

From the BCs we expect that the only angular harmonics present in the solution for ψ will be the $(\cos,\sin)\theta$ ones. From 6.1.4, the corresponding eigenfunctions are

$$\psi=\rho\left(A\cos\theta+B\sin\theta\right)+\frac{1}{\rho}\left(C\cos\theta+D\sin\theta\right) \quad .$$

Although the ρ terms are divergent at ∞, we cannot drop them like we could in the electromagnetic case because the velocity at ∞ is finite rather than zero. Taking the negative gradient,

$$\mathbf{v}=\left(-A\cos\theta-B\sin\theta+\frac{C}{\rho^2}\cos\theta+\frac{D}{\rho^2}\sin\theta\right)\hat{\boldsymbol{\rho}}+\left(A\sin\theta-B\cos\theta+\frac{C}{\rho^2}\sin\theta-\frac{D}{\rho^2}\cos\theta\right)\hat{\boldsymbol{\theta}} \quad . \qquad (01)$$

The BC at ∞ immediately yields $A=-V$ and $B=0$. Then the BC at $\rho=R$ yields $C=-VR^2$ and $D=0$. Plugging these into (01) we get

$$\mathbf{v}=V\left(1-\frac{R^2}{\rho^2}\right)\cos\theta\hat{\boldsymbol{\rho}}-V\left(1+\frac{R^2}{\rho^2}\right)\sin\theta\hat{\boldsymbol{\theta}} \quad , \qquad (02)^*$$

which is the desired answer.

(b) To get the force on the cylinder, we must obtain the pressure at its surface i.e. at $\rho=R$. Since the flow is 3I, BERNOULLI's principle holds and $P+dv^2/2=\text{const.}$ where d is the density. Setting $\rho=R$ in (02) and the result into BERNOULLI's principle, we get

$$P=\text{const.}-\frac{1}{2}dV^2\left(4\sin^2\theta\right) \quad .$$

Now, a constant pressure all over the surface cannot exert a net force on the cylinder so it is ignorable. A trig identity says $\sin^2\theta=(1+\cos2\theta)/2$, so the angular

variation of the pressure is $\cos 2\theta$. This function is symmetric about any diameter, so the pressure at any point on the cylinder and its diametrically opposite point are equal. The corresponding forces cancel in pairs and the resultant force on the cylinder is zero.

We note, primarily from the point of view of the next problem, that the pressure distribution over a cylinder which DOES generate a force is $(\cos, \sin)\theta$ – that has clear variation across a diameter, and the various contributions over the surface add up to give a resultant force directed from the maxima to the minima of the pressure. That in fact is the only angular harmonic which can generate pressure on a cylinder. Since the force at any point acts along the purely radial direction, its x- and y- components pick up $P\ (\cos,\ \sin)\theta$ dependences along the circumference of the cylinder. These get integrated to generate the total force. A non-zero resultant can only be produced if P goes as $(\cos,\ \sin)\,\theta$, since that gives rise to a $\cos^2\theta$ or $\sin^2\theta$ term whose average is finite along the circumference of the cylinder (same argument as in 6.3.7 for those who have read it).

(c) Since the only BC on 3I flow at the cylinder is on the radial component of $\mathbf{v}$, the rotation of the cylinder does not change the flow field in any way, and the resultant force is zero again.

*

COMMENTS : There are two reasons why potential flow (or 3I flow if we want to drum in the underlying approximations) enjoys an enduring but unjustified popularity. The first is that it is easy – in less than a page we obtained a flow field in a completely non-trivial situation ! The second is that it has a certain mathematical charm. For example, the flow field (02) is identical to the magnetic field created by inserting a cylindrical flux-expeller into a uniform magnetic field ($\mathbf{B}$ outside the cylinder also satisfies zero divergence, zero curl and zero normal component at the surface). This can lead to a lot of interesting discussion about different phenomena with the same equations, the unifying power of mathematics and the like.

The problem with this flow is that it yields absolutely nothing. A flow past a stationary cylinder with constant far-field velocity (the standard jargon for the velocity very far away from the body) is just a frame transformation on a cylinder

moving through a stationary fluid. We expect that this cylinder will experience a drag. If the cylinder is rotated while moving, it is actually supposed to generate lift – a force in y-direction, perpendicular to velocity. Neither of these forces came out of our simple little calculation. This in fact is not an accident, an unfortunate cancellation somewhere or other for a cylinder – the Kutta-Zhukovsky theorem states that an arbitrary object surrounded by 3I flow can NEVER experience a force. So potential flow is completely impotent when it comes to actually making quantitative – or even qualitative – predictions.

Then how to get drag and lift on either a cylinder or a sphere moving through fluid without and with rotation ? (We would expect that these shapes are the most basic, where analytical results should be the easiest to obtain.) From long, gruelling calculations which include viscosity and vorticity. The first such result was by GG STOKES himself – in 1854 he showed that the drag on the sphere is

$$F_D = 6\pi\eta RV \quad ,$$

where the symbols are really self-explanatory. This is one of the most famous formulae in all of fluid mechanics. In retrospect, its derivation is the easiest of the set since one can actually dispense with the nonlinear term to get this answer. The drag per unit length on the cylinder works out to be

$$F_D = \frac{4\pi\eta V}{\log\left(3.70\eta / VRd\right)} \quad ,$$

where the 3.70 is not an empirical constant but the sum of a certain series. This result comes from a mathematical tour de force where $\left(\mathbf{v}\cdot\nabla\right)\mathbf{v}$ takes centrestage. The lift on a rotating sphere evaluates to

$$F_L = 4\pi R^3 d\Omega V \quad ,$$

the simplicity of the final result belying the massive $\left(\mathbf{v}\cdot\nabla\right)\mathbf{v}$-dominated operation which produces it. Finally, the lift on a rotating cylinder is given by the expression – you'll find out for yourself in the next Tute. Cabin crew to stations please for takeoff.

Example 6.4.3

The Lift on a rotating cylinder

This problem has been rejected by European Journal of Mechanics B on technical grounds. The Reviewers cited the ad hoc nature of the approximations and the fact that they have no justification other than the retroactive one. So far, I myself have also been unable to put the assumptions on a firmer foundation. Nevertheless, I believe that this "image-like" method of starting with a trial vorticity and validating it through a self-consistency procedure has some merit, which is why I have not taken the problem down. If any of you can help me in making the basis more solid, I'll be very happy.

An infinitely long cylinder of radius R with its axis parallel to the z-axis is moving with constant speed V in the $-x$-direction through an incompressible fluid of density d and viscosity η. The cylinder is also rotating about its axis with angular velocity Ω. A frame transformation is performed so that the cylinder has no translation relative to this frame. Origin of cylindrical polar coordinates is taken at the centre of the cylinder. Letting $\mathbf{v}(\mathbf{r})$ be the velocity field in this frame, $\mathbf{v}$ can easily be seen to satisfy the Navier Stokes equation together with the boundary conditions $\mathbf{v}(\infty,\theta)=V\hat{\mathbf{x}}$ and $\mathbf{v}(R,\theta)=\Omega R\hat{\boldsymbol{\theta}}$.

An auxiliary problem is considered where the cylinder is removed and instead a given vorticity field $\boldsymbol{\omega}=(w_0/\rho)\mathrm{e}^{-k\rho}\hat{\mathbf{z}}$ is assumed to permeate all of space. Let $\mathbf{u}(\mathbf{r})$ be the solution to the auxiliary problem.

(a) Obtain the general solution for $\mathbf{u}$, keeping terms upto first harmonic in θ. Specifically include in your expansion a radial velocity which is independent of θ.

(b) Fit the $\rho=\infty$ BC and as much as possible of the $\rho=R$ BC on the actual solution $\mathbf{v}$ to the auxiliary solution $\mathbf{u}$. Obtain a condition on Ω such that the infringement of the R BC is minimized. Further impose the condition that the total vorticity contained in the flow field $\mathbf{u}$ outside $\rho=R$ be equal to $2\pi\Omega R^2$ to determine w_0 and k.

(c) Substitute **u** into the NS equation in vorticity form. Since **u** is not an exact solution, demand the best possible satisfaction of the equation to determine all remaining unknown coefficients in **u**. By stipulating that the breach of the NS equation be small everywhere, estimate the radius of the cylinder for which this calculation is valid.

It is now claimed that **u** equals **v**, the solution of the original problem.

(d) Expanding the pressure in a series of angular harmonics as $P/d = f_0(\rho) + f_1(\rho)\cos\theta + g_1(\rho)\sin\theta + ...$, substitute this and the above **v** into the NS equation to obtain ODEs for f_1 and g_1.

(e) Solve the above ODEs and determine the resultant force per unit length acting on the cylinder.

[Level : 6]

Yet another six, the statement shorter than our previous two but the task no less formidable. As usual, we start off with an outline of the calculation being asked for. The target is obviously the lift on the rotating cylinder, which needs a nonlinear analysis. Since that equation cannot be solved explicitly, the problem is essentially proposing an "image method" of generating an approximate solution. We have to first solve the "auxiliary problem" which describes the flow corresponding to a fictitious vorticity distribution (this plays the role of the fictitious charge or current configuration). We then have to show that this auxiliary solution indeed satisfies to good approximation the BCs on the original problem as well as the NS equation itself. The former is typical of the image method in electromagnetism while the latter is an extra. It is necessary in this case since $\boldsymbol{\omega}$ is just curl of $\mathbf{v}$ and not some standalone quantity like charge or current, so there is absolutely no guarantee that a fictitious vorticity and the corresponding flow will indeed satisfy the single nonlinear equation which governs them both. Having found the auxiliary solution, we must use it to calculate the pressure on the surface of the cylinder and hence the lift on the cylinder.

Outline in our heads, we narrow our focus down to the individual parts – the prospect seems much more daunting when viewed as a whole than when treated part by part. And of course there are the partial credits which we mustn't lose sight of.

(a) The problem on hand is

$$\nabla\cdot\mathbf{u}=0 \quad , \tag{01a}$$

$$\nabla\times\mathbf{u}=\frac{w_0}{\rho}\mathrm{e}^{-k\rho}\,\hat{\mathbf{z}} \quad . \tag{01b}$$

This is a linear problem so there will be a particular and a homogeneous solution, the latter corresponding to zero curl of $\mathbf{u}$. The situation here is highly reminiscent of the magnetic field equations (6.3.01c,d) in the absence of time dependence – read $\mathbf{B}$ for $\mathbf{u}$ and $\mu_0\mathbf{J}$ for $\boldsymbol{\omega}$. Although we used the differential method and nothing but the differential method in 6.3, I will cheat here and calculate the particular solution for "magnetic field" using "AMPERE's circuital rule". Evidently, the "magnetic field" or velocity must be tangential and a function of radius alone. Defining a circular circuit of radius ρ we have

$$\oint_{\text{circuit}} \mathbf{u}\cdot\mathrm{d}\mathbf{l} = \oint_{\text{area}} \boldsymbol{\omega}\cdot\mathrm{d}\mathbf{A} \quad \text{, or}$$

$$2\pi\rho u_\theta = \int_0^{2\pi} \mathrm{d}\theta' \int_0^{\rho} \rho'\,\mathrm{d}\rho' \frac{w_0}{\rho'} \mathrm{e}^{-k\rho'} \quad .$$

The easy integration yields

$$u_\theta = \frac{w_0}{k\rho}\left(1-\mathrm{e}^{-k\rho}\right) \quad .$$

We recall that this is a particular solution of (01) only.

To this we must add homogeneous solutions, which correspond to solutions of 3I flow. The potential ψ satisfies the Laplace equation and has the expansion

$$\psi = C_0 + D_0 \log\rho + \left(C_1\rho + \frac{C_2}{\rho}\right)\cos\theta + \left(D_1\rho + \frac{D_2}{\rho}\right)\sin\theta + \ldots\ldots \quad ,$$

upto the first harmonic required in the question. Taking the gradient

$$\mathbf{u}_h = \left[\frac{D_0}{\rho} + \left(C_1 - \frac{C_2}{\rho^2}\right)\cos\theta + \left(D_1 - \frac{D_2}{\rho^2}\right)\sin\theta\right]\hat{\boldsymbol{\rho}} + \left[-\left(C_1 + \frac{C_2}{\rho^2}\right)\sin\theta + \left(D_1 + \frac{D_2}{\rho^2}\right)\cos\theta\right]\hat{\boldsymbol{\theta}} \quad .$$

Combining $\mathbf{u}_h$ with $\mathbf{u}_p$ gives the general solution

$$\mathbf{u}_G = \left[\frac{D_0}{\rho} + \left(C_1 - \frac{C_2}{\rho^2}\right)\cos\theta + \left(D_1 - \frac{D_2}{\rho^2}\right)\sin\theta\right]\hat{\boldsymbol{\rho}} + \left[\frac{w_0}{k\rho}\left(1-\mathrm{e}^{-k\rho}\right) - \left(C_1 + \frac{C_2}{\rho^2}\right)\sin\theta + \left(D_1 + \frac{D_2}{\rho^2}\right)\cos\theta\right]\hat{\boldsymbol{\theta}} \quad , \qquad (02)^*$$

which is the desired answer.

The term D_0/ρ in u_ρ might look a bit suspicious – if the fluid flows radially outwards everywhere, then where does it come in from ? We might be tempted to set $D_0 = 0$ except that the question explicitly asked us to retain it. I will hold comment regarding this step until after Part (c) where its role will become apparent. Then, I will also include a brief discussion regarding its interpretation.

(b) The $\rho = \infty$ BCs on the actual solution $\mathbf{v}$ are given in the question and are the same as in the last Tute : $\mathbf{v}(\infty,\theta) = V\left(\cos\theta\hat{\boldsymbol{\rho}} - \sin\theta\hat{\boldsymbol{\theta}}\right)$. Evidently, $C_1 = V$ and $D_1 = 0$. It's $\rho = R$ which poses a problem. The actual velocity there is $\Omega R\hat{\boldsymbol{\theta}}$, which we can account for using the first term in the θ-component of $\mathbf{u}_G$; to fit it exactly however we would need D_0, $C_1 - C_2/R_0^2$ as well as $C_1 + C_2/R_0^2$ to be zero. The first would violate the question requirements while the latter two are an evident contradiction. Thus, we must go for only approximate satisfaction of this BC.

The part which we can satisfy yields

$$\frac{w_0}{kR}\left(1 - \mathrm{e}^{-kR}\right) = \Omega R \quad . \tag{03}$$

The infringing components are the terms D_0/R, whose size is yet unknown, and C_1, C_2 which are of order V. We can minimize the violation on a relative scale if the size of these terms is small compared to the part which DOES get satisfied i.e. compared to ΩR. Hence, the satisfaction of the BC gets better and better if $\Omega R >> V$, D_0/R. Qualitatively, the cylinder must rotate very fast for the BC infringement to be minimized. (Of course, fast can only be defined in relative terms – here the comparison benchmark is its translational speed V, and the mystery velocity D_0/R. This mystery will get elucidated in the next part of the solution.)

Since the R BC is approximate anyway, we let C_2 have the value zero – that makes for equal infringement on the ρ- and θ-components of $\mathbf{u}$, and, by killing some terms in (02), simplifies the subsequent steps maximally. Hence our solution now is

$$\mathbf{u} = \left[\frac{D_0}{\rho} + V\cos\theta\right]\hat{\boldsymbol{\rho}} + \left[\frac{w_0}{k\rho}\left(1 - \mathrm{e}^{-k\rho}\right) - V\sin\theta\right]\hat{\boldsymbol{\theta}} \quad . \tag{04}$$

Armed with this formula, we go into the next phase of the battle.

We are now told that the total vorticity contained in $\mathbf{u}$ outside the cylinder is $2\pi\Omega R^2$. This implies

$$\int_0^{2\pi} \mathrm{d}\theta \int_R^{\infty} \rho \mathrm{d}\rho \frac{w_0}{\rho}\mathrm{e}^{-k\rho} = 2\pi\Omega R^2 \quad .$$

Evaluating the integral,

$$\frac{w_0}{k}\mathrm{e}^{-kR} = \Omega R^2 \quad . \tag{05}$$

Dividing by R and comparing with (03) gives $\mathrm{e}^{-kR} = 1/2$ or $\underline{kR = \log 2}$. The value of w_0 then follows immediately from (05); we have $\underline{w_0 = 2\log 2\,\Omega R}$.

Now where did this total vorticity business come from ? This comes from the original studies of this phenomenon by HEINRICH MAGNUS. He modelled the cylinder as a "bound vortex" (a somewhat contrived fluid flow configuration which has infinite vorticity at the origin and zero vorticity everywhere else). To the strength of this vortex he assigned the total vorticity enclosed within the cylinder (had it been a rotating fluid), which works out to $2\pi\Omega R^2$. Arguments about diffusion of vorticity from startup of rotation to establishment of steady state flow can justify this condition rigorously. Although we are not doing bound vortex here, this condition is just as applicable – the total vorticity in the fluid equals the vorticity coming from the cylinder's rotation.

(c) The "vorticity" NS equation is (6.4.06) with the time-dependence removed :

$$(\mathbf{u}\cdot\nabla)\boldsymbol{\omega} - (\boldsymbol{\omega}\cdot\nabla)\mathbf{u} = \frac{\eta}{d}\nabla^2\boldsymbol{\omega} \quad .$$

Here, $\boldsymbol{\omega}$ is entirely in z-direction while $\mathbf{u}$ is independent of z, so $(\boldsymbol{\omega}\cdot\nabla)\mathbf{u}$ evaluates to zero. Now writing the vorticity as a scalar, we must have $(\mathbf{u}\cdot\nabla)\omega = (\eta/d)\nabla^2\omega$. The RHS is easy to evaluate; we have

$$\frac{\eta}{d}\nabla^2\omega = \frac{\eta}{d}\left(-\frac{w_0}{\rho^3} - \frac{w_0 k}{\rho^2} - \frac{w_0 k^2}{\rho}\right)\mathrm{e}^{-k\rho} \quad . \tag{06}$$

On the LHS, the only non-trivial term in $(\mathbf{u}\cdot\nabla)\omega$ (see the vector calculus review for the formula) is $u_\rho(\partial\omega/\partial\rho)$. Now, $\partial\omega/\partial\rho$ is zero harmonic in θ while u_ρ from (04) has a zero harmonic and a first harmonic component. The second of these cannot even hope to satisfy (06) – it has the wrong angular profile. The first one is not altogether hopeless, so we go with that, getting

$$(\mathbf{u}\cdot\nabla)\omega = \frac{D_0}{\rho}\left(-\frac{w_0}{\rho^2} - \frac{w_0 k}{\rho}\right)\mathrm{e}^{-k\rho} \quad .$$

In an encouraging development, this can be made to equal the first two terms of (06) if we take $D_0 = \eta/d$. So we make this choice, getting

$$* \qquad \mathbf{u} = \left[\frac{\eta}{d\rho} + V\cos\theta\right]\hat{\boldsymbol{\rho}} + \left[-\frac{w_0}{k\rho}\mathrm{e}^{-k\rho} - V\sin\theta\right]\hat{\boldsymbol{\theta}} \quad , \qquad *$$

which is our final expression for $\mathbf{u}$.

Now we have to estimate the radius of the cylinder by stipulating that the breach of NS equation be small. The primary breach occurred when we had to throw away the first harmonic terms in u_ρ in favour of the zero harmonic ones. So, this approximation will remain good so long as $\eta/d\rho$ is greater than V, i.e. within a sufficiently small radius. Outside this radius, the approximation gets worse and worse. However, the saving grace is the $\mathrm{e}^{-k\rho}$ multiplying everything – at large radius, even if our approximation is poor, the absolute size of the terms involved is negligible. So, when $\eta/d\rho$ has become of order V, $\mathrm{e}^{-k\rho}$ has to become small, i.e. if ρ_0 denote η/dV, then $k\rho_0$ must be some $O(1)$ number or larger. Now we already know what k is – it is $\log 2/R$ or approximately $1/R$, so ρ_0/R must be $O(1)$ or larger. Hence we must have $\underline{R < \eta/dV}$ for our calculation to be valid.

With our final $\mathbf{u}$ and a plausible cylinder radius on the table, now let me come back to the $1/\rho$ term in u_ρ. We just saw why it was needed – if we didn't have it, then $(\mathbf{u}\cdot\nabla)\omega$ and $\nabla^2\omega$ would not match up in functional form. So either the starting ω is completely garbage or the $1/\rho$ term is present in u_ρ. The former is highly unlikely – after all a rapidly rotating cylinder SHOULD be expected to generate a vorticity with leading term independent of angle and decreasing sharply with radius. On the other hand, the u_ρ term has a definite physical interpretation. The rotation of the cylinder gives rise to a centrifugal force which tends to push the fluid out. Indeed, a centrifugal blower is exactly this configuration – a cylinder which makes the air rotate at high speed, throwing it out radially. For the blower, air gets replenished by coming in along the axis. Likewise, for a SPHERE rotating about z-axis, the fluid gets pushed radially outwards near $\theta = 90^\circ$ and comes in from $\theta = 0$ and $\theta = 180^\circ$. Here, for a strictly infinite cylinder, the incoming flow is impossible, leading to an apparent contradiction.

However, an actually infinite cylinder is itself impossible. The best we can have is a LONG cylinder, but that will immediately allow fluid to flow in from the ends, as it does for a sphere. We can then say that our calculation deals with the

central regions of the cylinder where the inflow is at a low velocity along the z-direction, a negligible effect on its own but enough to sustain the radial outflow.

(d) Since our solution $\mathbf{u}$ matches the BC on $\mathbf{v}$ at $\rho = \infty$, matches to good approximation the BC at $\rho = R$ and satisfies the Navier Stokes equation upto a radius until the terms involved become really small, we publish an official condonation of all infringements and declare that $\mathbf{v} = \mathbf{u}$. (Of course the REAL justification of the approximations is that they will ultimately yield a plausible result – as the cricket commentator GEOFFREY BOYCOTT used to say, "no ball is a bad ball if it gets a wicket".) Then,

$$\mathbf{v} = \left[\frac{\eta}{d\rho} + V\cos\theta\right]\hat{\boldsymbol{\rho}} + \left[\frac{w_0}{k\rho}\left(1 - \mathrm{e}^{-k\rho}\right) - V\sin\theta\right]\hat{\boldsymbol{\theta}} \quad . \tag{07}$$

Now we have to use this $\mathbf{v}$ to determine P everywhere, and in particular at the surface of the cylinder.

The relation between $\mathbf{v}$ and P is the "original" NS equation (6.4.05) :

$$(\mathbf{v}\cdot\nabla)\mathbf{v} = -\nabla\left(\frac{P}{d}\right) + \frac{\eta}{d}\nabla^2\mathbf{v} \quad . \tag{08}$$

Given (07), this is a straightforward but long plug-and-chug. The first step is to expand P as given in the question : $P/d = f_0(\rho) + f_1(\rho)\cos\theta + g_1(\rho)\sin\theta + ...$ and then take its gradient. We have

$$\nabla\left(\frac{P}{d}\right) = \left(f_0' + f_1'\cos\theta + g_1'\sin\theta\right)\hat{\boldsymbol{\rho}} + \left(-\frac{f_1}{\rho}\sin\theta + \frac{g_1}{\rho}\cos\theta\right)\hat{\boldsymbol{\theta}} \quad , \tag{09}$$

where prime denotes differentiation w.r.t ρ and higher harmonics are ignored. Since the NS equation has vector derivatives, we can't afford to leave things in terms of cylindrical unit vectors but must convert to Cartesian. As in the solenoid calculation 6.3.4 Part (b), we won't convert ρ and θ in the functional dependences to x and y, but will only convert the unit vectors. The conversion formulae

$$\hat{\boldsymbol{\rho}} = \cos\theta\hat{\mathbf{x}} + \sin\theta\hat{\mathbf{y}} \qquad \hat{\boldsymbol{\theta}} = -\sin\theta\hat{\mathbf{x}} + \cos\theta\hat{\mathbf{y}} \quad ,$$

substituted into (09) yield the x- and y-components of $\nabla(P/d)$ as serieses in angular harmonics of θ. The zeroth harmonic is

$$\nabla\left(\frac{P}{d}\right)=\frac{1}{2}\left(f_1'+\frac{f_1}{\rho}\right)\hat{\mathbf{x}}+\frac{1}{2}\left(g_1'+\frac{g_1}{\rho}\right)\hat{\mathbf{y}} \quad . \tag{10}$$

These are LHSes of ODEs for f_1 and g_1, which are asked in the question; to get the corresponding RHSes we need to obtain only the zeroth angular harmonic of the expansions for the $\mathbf{v}$ terms in the NS equation.

From the review, the nonlinear term is

$$(\mathbf{v}\cdot\nabla)\mathbf{v}=\left(v_\rho\frac{\partial v_\rho}{\partial\rho}+\frac{1}{\rho}v_\theta\frac{\partial v_\rho}{\partial\theta}-\frac{v_\theta^2}{\rho}\right)\hat{\boldsymbol{\rho}}+\left(v_\rho\frac{\partial v_\theta}{\partial\rho}+\frac{1}{\rho}v_\theta\frac{\partial v_\theta}{\partial\theta}+\frac{v_\theta v_\rho}{\rho}\right)\hat{\boldsymbol{\theta}} \quad .$$

Since we want a zero harmonic when converted to Cartesian, we must keep the first harmonic terms in polar. These are

$$(\mathbf{v}\cdot\nabla)\mathbf{v}=\left[-\frac{\eta V}{d\rho^2}\cos\theta+\frac{Vw_0}{k\rho^2}\left(1-\mathrm{e}^{-k\rho}\right)\sin\theta\right]\hat{\boldsymbol{\rho}}+$$
$$\left[-\frac{Vw_0}{k\rho^2}\left(1-\mathrm{e}^{-k\rho}\right)\cos\theta+\frac{Vw_0}{\rho}\mathrm{e}^{-k\rho}\cos\theta-\frac{\eta V}{d\rho^2}\sin\theta\right]\hat{\boldsymbol{\theta}} \quad .$$

Substituting the vector conversion formulae and collecting the zero harmonic components,

$$(\mathbf{v}\cdot\nabla)\mathbf{v}=0\hat{\mathbf{x}}+\frac{Vw_0}{2\rho}\mathrm{e}^{-k\rho}\,\hat{\mathbf{y}} \quad . \tag{11}$$

The last term to be reckoned with is the vector Laplacian $(\eta/d)\nabla^2\mathbf{v}$. For this we need to convert $\mathbf{v}$ into Cartesian before taking the derivatives. Applying the conversion formulae on (07), we get

$$\mathbf{v}=\left(V+\frac{\eta}{d\rho}\cos\theta-\frac{w_0}{k\rho}\left(1-\mathrm{e}^{-k\rho}\right)\sin\theta\right)\hat{\mathbf{x}}$$
$$+\left(\frac{\eta}{d\rho}\sin\theta+\frac{w_0}{k\rho}\left(1-\mathrm{e}^{-k\rho}\right)\cos\theta\right)\hat{\mathbf{y}} \quad .$$

Evidently, the Laplacians of the x- and y-components have ZERO contribution to the θ-independent term so

$$\nabla^2\mathbf{v}=0 \quad , \tag{12}$$

at the level we are interested in.

Now substituting (10-12) into (08) and equating term by term we get

$$f_1' + \frac{f_1}{\rho} = 0 \quad , \tag{*}$$

$$g_1' + \frac{g_1}{\rho} = -\frac{Vw_0}{\rho} e^{-k\rho} \quad , \tag{*}$$

which are the ODEs asked for.

(e) What a transformation – from high-order high-dimensional nonlinear PDE to first order linear ODE, rolling back all the way past electromagnetism , heat equation , beams and BVPs , forced vibrations , free vibrations and even first order modeling . From the depths of our memory we dredge up that great concept called INTEGRATING FACTOR and obtain the solutions

$$f_1 = \frac{C_1}{\rho} \quad ,$$

$$g_1 = \frac{C_2}{\rho} + \frac{Vw_0}{k\rho} e^{-k\rho} \quad ,$$

where C_1 and C_2 are of course two arbitrary constants.

The BC that the pressure at $\rho = \infty$ be zero does not serve to determine C_1 and C_2. However, we saw in the last Tute that a $(\cos, \sin)\theta$ pressure distribution on the cylinder surface generates a force in the direction from the max to the min. Then, a non-trivial homogeneous solution would imply a force on the cylinder in the absence of rotation, which is absurd. Hence $C_1 = C_2 = 0$. Since g_1 is the $\sin\theta$ component of P/d, we have

$$P = \frac{dVw_0}{k\rho} e^{-k\rho} \sin\theta \quad . \tag{13}$$

Now set $\rho = R$ and remember the relation (05); this gives the pressure at the cylinder surface

$$P = d\Omega RV \sin\theta \quad . \tag{14}$$

The max is at the top and the min at the bottom, so the resultant force is in the $-y$-direction if Ω is positive. Although we are used to thinking of lift as "upwards", any force perpendicular to velocity is legitimate – we can just

reverse the direction of rotation to make the force go the other way. A simple integration over the cylinder circumference now gives the magnitude of lift per unit length to be

$$F_L = \pi R^2 d\Omega V \quad , \tag{*}$$

which is the desired answer.

*

COMMENTS : As a first step we confirm that our formula agrees with what is expected. From the bound vortex theory, the lift is $\mathbf{v}_0 \times \mathbf{K}$ where $\mathbf{v}_0$ is the far-field velocity and $\mathbf{K}$ the bound vortex. Our vorticity is infinite at $\rho = 0$ and is exponentially decreasing in ρ, so a bound vortex is a reasonable substitute. If Ω is positive, then w_0 is positive also, so the equivalent bound vortex is in $+z$; crossed into a $+x$-directed $\mathbf{v}_0$, it should give a $-y$ lift, which indeed we got here. It is also a known fact that a back-spin i.e. spin opposite to the direction of motion generates lift on golf and tennis balls, while a free-kick shot forward with counter-clockwise spin curls towards the left. All these are in agreement with the direction we got here. The formula itself is very similar to the result for sphere I quoted at the end of the previous Tute, so that gives more confidence in the correctness of our result.

Having performed the dull but very necessary verification (when I first did the calculation I had got the sign wrong), we now see some aspects of our result. One of the main conclusions is that it gives a physical basis to the notion of bound vortex. A delta-function vorticity at origin is contrived because vorticity is curl of velocity and not a standalone quantity. A line charge or current wire at origin is perfectly plausible because it is a definite physical object but a velocity whose curl is delta is not. The more so because the fluid doesn't even occupy the origin – where there is no fluid how can there be vorticity ? However, we see that a vorticity which goes as $e^{-k\rho}/\rho$ is indeed the starting term of an actual solution of the Navier Stokes equation. As a first approximation, this vorticity can be treated as a delta at the origin.

Our calculation also shows the Alpha Bravo Charlie of lift – VORTICITY, VISCOSITY and NONLINEARITY. Vorticity was of course essential – take that away and you are left with 3I flow, which does not have lift. Viscosity was needed to enforce

the BC at the cylinder surface. Without that BC, the fluid around the cylinder would not rotate even if the cylinder did, and the flow would not acquire the necessary vorticity. The effect of viscosity is to set up a BOUNDARY LAYER around the moving body (in this case the cylinder and more generally a wing or similar object), within which the flow pattern is substantially different from 3I flow. Outside this layer, the flow is close to potential. But even though 3I flow permeates the bulk of space, all the lift gets generated within the layer.

The presence of the layer is not too apparent from the flow field (07). But it becomes clear if we examine the $\sin\theta$ component of pressure, (13). This is the actual lift generator – it is practically zero all the way from infinity to even a couple of radii away from the cylinder; it then rises sharply as one moves further inwards to hit the value (14) on the surface. The thickness of the layer is of order R, which we saw in Part (c) should be of order η/dV. An explicit calculation for the thickness of the boundary layer has been performed by PAUL BLASIUS in a Cartesian geometry – he has obtained the thickness of the layer to be $(\eta/dV)^{1/2}$, very similar to what we have here.

Challenge : "Your star formula for lift does not contain η. Presumably it would hold true even if η were zero. Then how you can say that viscosity is essential for lift ?" We already saw where viscosity plays its role – in enforcing the rotational boundary condition at the cylinder surface. It just happens that the final formula is independent of viscosity, at least to largest order. As a simplified but somewhat similar example, let me propose the following. Suppose you have a damped harmonic oscillator which you are driving at some constant frequency ω well off the resonance. Then you will claim that the long-time response is going to be sinusoidal oscillations whose amplitude is gain $G(\omega)$ times the excitation amplitude. For low γ and off-resonance ω, G is independent of γ at least to largest order, and a leading approximation of the response formula will also not feature γ. This might lead you to the conclusion that the damping is unnecessary – you would have observed the same response even if γ were zero. But clearly that is not the case. For γ EXACTLY zero, your observation would depend on initial condition and agree with the formula for one single isolated IC. But for ANY non-zero γ, however small, the homogeneous solutions would get killed off, and the response formula, independent of γ, would become valid.

The salient roles of viscosity and vorticity immediately negate the popular but incorrect explanation of lift based on BERNOULLI's principle. This explanation states that the airflow is slower on the bottom side of a lifting object and faster on the top; then since $P + dv^2/2 = \text{const.}$ the pressure is higher on the bottom than the top and the body rises up. However, BERNOULLI's principle holds only for inviscid and irrotational flow, while the lift comes from its exact converse.

Finally, who contributed the non-trivial term to the $\sin\theta$ pressure ? The nonlinearity. The ONLY non-zero term on the RHS of the ODE for g_1 came from $(\mathbf{v}\cdot\nabla)\mathbf{v}$ – it did not come from $\nabla^2\mathbf{v}$. This is why a linear analysis including viscosity and vorticity still can't produce lift. Indeed, the drag on the sphere due to STOKES was from a linear analysis – that calculation also yields the velocity field around the moving sphere. Another linear calculation produces the velocity field around a sphere rotating inside a fluid at rest (it also gives a drag torque). We would expect that adding the two flow fields would yield the flow and hence the lift on a moving-plus-rotating sphere. However, following that approach the lift comes out to zero. As I mentioned in the comments to the last Tute, the calculation which does give the lift has the nonlinearity at its core.

Like the three other calculations of that Tute, this is a low Reynolds number estimation i.e. I have assumed that VRd/η is small. I used this fact twice, once when permitting the infringement of the BC at the cylinder surface and again when dropping the $V\cos\theta$ term in u_ρ in preference to the $\eta/d\rho$ term. Evidently, if the Reynolds number is raised, the first thing we have to do is tack on additional $(\cos,\ \sin)\theta$ to the trial vorticity function so that we can match more terms in $(\mathbf{u}\cdot\nabla)\boldsymbol{\omega}$. The drag on the cylinder came out to zero in this case because I allowed an infringement of the BC at $\rho = R$. Incorporating more terms in the trial vorticity function will very likely allow an exact (or at least more accurate) satisfaction of that BC and a non-zero result for the drag. I hope that this "image method" can also be applied to obtain nonlinear solutions for lift and drag in other configurations.

So how does this relate to the more familiar lift on a wing ? There is in fact a mathematical trick called Zhukovsky transformation (beyond our scope) which can transform a flow past a cylinder into a flow past a wing (or airfoil as it is called in Literature). This transformation also shows that if the wing is inclined

at an angle to the free-stream velocity (called the angle of attack) then the velocity in the boundary layer acquires a significant swirling component, like in this case. Hence, our calculation is indeed strongly related to that most spectacular fluid mechanical phenomenon which can keep a 300 ton aircraft aloft.

Of course, we have to remember that the Reynolds number used in our calculation was low, while that for an aircraft is extremely high. But it sometimes happens that a low Reynolds number analysis does show the same behaviour as a higher Reynolds number calculation (if that is at all possible). That is not true for Stokes drag since the low-speed theoretical result says drag proportional to V while the experimental observation is drag proportional to V^2. But it does happen to hold true for the lift calculations, both for the sphere and cylinder. To see this, we look at the experimentally obtained formula for the lift on a wing moving through static air at speed V. It is $F_L = (1/2)C_L dAV^2$ where C_L is the coefficient of lift (a number dependent on size and shape of the wing), d the density of air and A the surface area of the wing. Our final formula $F_L = \pi R^2 d\Omega V$ gives lift per unit length – multiplying by the cylinder length would introduce the surface area. This formula also features the density. Finally, it has the product ΩV. The Ω arose because we generated swirling flow around the cylinder by making it rotate. A wing produces the swirl by virtue of geometry alone – the swirl speed is proportional to V so our Ω gets replaced by V, producing a resultant dependence of V^2.

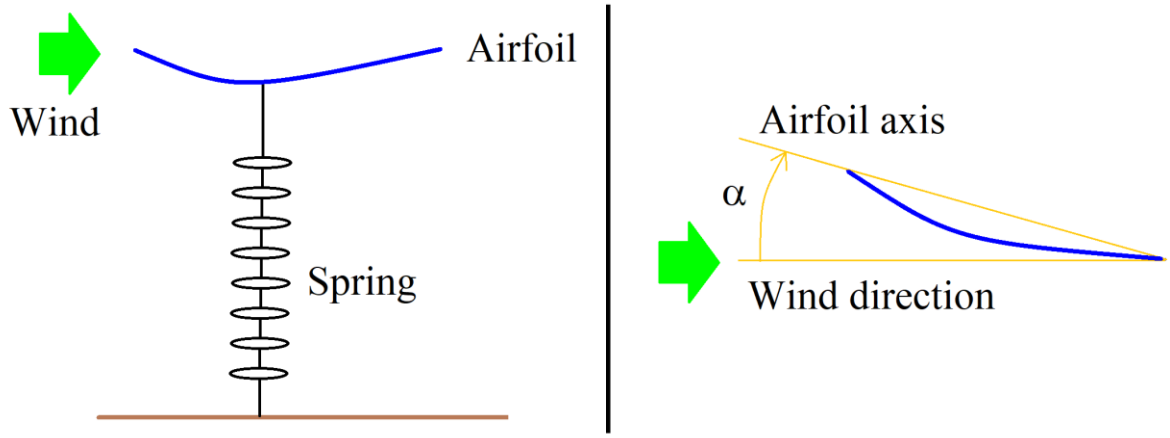
*

Two maximally rated problems in a space of double that number – this was really not my goal when I set out to write DELTA. But it just happens that both maglev and aircraft are among the greatest wonders of the modern world. This book would (at least in my mind) remain incomplete unless I could show you how they work. So here they were, and I sincerely hope that you liked them. Now I promise I will cut back on the throttle. The next Tute is an application of what we saw here but it's far simpler – it uses lift to only write a linear ODE.

Example 6.4.4

Aeroelastic flutter

A bridge undergoing aeroelastic flutter can be modelled as shown in the left panel of the Figure below.



The bridge consists of an airfoil of mass m connected to ground via a vertical spring (k) and damper (C). A wind blows from the left at constant velocity V. It is known that when the axis of the airfoil makes an angle α with the far-field velocity of the incoming flow (the angle is shown in the right panel, measured clockwise from the flow direction to the airfoil axis), the airfoil experiences an upward force equal to $-K\alpha V^2$ where K is a positive constant.

(a) A disturbance sets the bridge into motion. Assuming that the vertical velocity of the airfoil is much less than V, and that the steady state airflow profiles are valid at all times, write an ODE for the displacement z of the airfoil as a function of time.

(b) Describe the nature of solutions of the above ODE and interpret your results physically.

[Level : 2]

(a) We've come so long from springs and dampers that we've almost forgotten how to write the equation of motion of a mass-spring-damper system. Hopefully without resorting to furtive glances at 3.3 we can muster

$$m\ddot{z} = -k\left(z - l\right) - C\dot{z} - mg + F_L \quad , \qquad (01)$$

where z goes vertically upwards measured from ground, l is the spring's natural length and F_L is the lift force, the reason why this problem is here instead of 300 pages back.

What is the origin of lift on the airfoil ? It's not that its motion causes it to pitch up and down like an aircraft. Rather, when it is moving, if you go into a frame which moves with it, then the far-field wind velocity is no longer horizontal but makes an angle to the airfoil axis. This makes α non-zero and produces a lift.

So let $\mathbf{v} = v\hat{\mathbf{z}} = \dot{z}\hat{\mathbf{z}}$ be the vertical velocity of the airfoil at any instant while $V\hat{\mathbf{x}}$ is the far-field velocity of the wind. Since we are told that the steady state flow fields are valid throughout, we perform the transformation into a reference frame which has instantaneous velocity $\mathbf{v}$. In this frame, the far-field velocity of the wind is $V\hat{\mathbf{x}} - v\hat{\mathbf{z}}$ while the axis of the airfoil is of course along x-axis. Comparing with the Figure, we can see that this implies α is NEGATIVE. The size of α is the usual $\arctan(v/V)$, which for small v reduces to v/V. Now, it is given that positive α means negative lift, hence negative α implies positive or upward lift. In place of V^2 of the lift formula, we should now substitute the square of the new far-field velocity, $V^2 + v^2$. For small v/V however α already throws in a factor of v/V, so we can ignore completely the v's in the V^2 term. Taking all this into account, the lift term works out to $F_L = KV^2(v/V)$ which is KVv.

For the final step, we substitute this into (01), getting

$$m\ddot{z} = -k\left(z - l\right) - C\dot{z} - mg + KV\dot{z} \quad .$$

Dividing by m and letting $\beta = K/m$ we have

$$\ddot{z} + \left(\gamma - \beta V\right)\dot{z} + \omega^2 z = \omega^2 l - g \quad , \qquad (02)^*$$

which is the ODE asked for.

(b) The coefficient of the first derivative term in (02) is positive when $V = 0$ on account of the damping. But as V increases, this coefficient decreases and ultimately turns negative. When the coefficient is positive, we get the usual exponentially decaying solutions. When it turns negative however, we get

exponentially growing solution even in the presence of damping. The crossover occurs for $V = \gamma/\beta$. This implies that, if the wind speed is sufficiently high, the bridge becomes completely unstable. The smallest perturbation sets it into oscillations which grow with exponentially increasing amplitude. At sufficiently large amplitude, there are two possibilities. In the bright scenario, the lift becomes weaker than the prediction of the formula and the spring and damper can then match the excitation, preventing further growth of the amplitude. In the dark scenario the bridge breaks up into pieces.

*

COMMENTS : Although this model of bridge is grossly simplified, it does show how a negative damping term – something quite elusive except in artificial electronic circuits – can enter the equation of motion of a structure subjected to high winds. The spectacular failure of Tacoma Narrows which I had mentioned in 3.3.10 was something similar to this. Although that was much more complicated , with a torsional mode of the bridge coupling to a translational mode, the basic phenomenon was this – a gale force wind gave rise to negative damping via lift generation.

You might also wonder at the sign of the lift – aircraft wings go UP when they pitch up, while this went the other way; surely that's not physical. But aircraft wings do as they do only because they are designed that way. When a wing pitches up into an air stream coming from the left, it develops around it a clockwise swirling flow which pushes it upward. (If you've read the last Tute, you can understand this statement immediately, else just treat it as a fact.) The cross-sectional shape of the wing is engineered very carefully so that this may happen. A bridge on the other hand is not supposed to be aerodynamic at all – it should be insensitive to airflow around it. But the shape of bridge is determined primarily by other considerations, and it might very well happen that a positive pitch-up in a right-moving wind causes a counter-clockwise swirling motion and a downward force. Another everyday example of flutter is when a wind blows past a Venetian blind – beyond a certain wind speed the slats of the blind often flop around violently. Each slat can be modelled as an airfoil on a spring (which restores the slat to equilibrium in absence of wind); it loses stability to flutter when the wind speed becomes too high.

Finally we can also see that the only way to stabilize a flutter-prone bridge is to add damping. This is easier said than done since the bulk of the damping has to come from the intrinsic properties of the material and structure of the bridge rather than an external dashpot – the $6\pi\eta RV$ drag on a typical dashpot is not nearly of the order of magnitude required to control a rogue bridge. Increasing the stiffness is far easier – use titanium, composites or a hundred other fancy materials and you will get it. But increasing stiffness will have zero effect when the troublesome term is on velocity rather than position.

* * * * *

This page intentionally left blank

7

THE WAVE EQUATION

This Chapter presents a discussion of the wave equation. Separation of variables as well as D'ALAMBERT's formula are presented for the homogeneous equation. Inhomogeneous problems and the vibrations of beams are also considered. The Chapter concludes with applications of the theory.

7.1

INTRODUCTION TO WAVE EQUATION

The wave equation is a dynamic equation like the heat equation – time is one of the independent variables. At the same time, it is second order in all independent variables, like Laplace equation. In one space and one time dimension, it reads as

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} \quad , \tag{01}$$

where u as usual is the generic dependent variable and c, a constant, is the **wave propagation velocity**. (Why c represents the wave velocity, that we'll see in due course of time; why c instead of v, that's because c is the speed of the most famous wave of all, and that scintillating example has made the generic notation almost a universal convention.) Using prime for space and dot for time, we can also write this as $u'' = \left(1/c^2\right)\ddot{u}$.

At first glance, the wave equation looks a lot like Laplace equation if you rename t to y – the factor of $1/c^2$ and can be absorbed by a simple axis rescaling. The sign however is the reverse – Laplace would have had a negative on the RHS while this has positive. Since sign is the most important component of a differential equation, it is no surprise that the solutions of Laplace and wave equations have absolutely nothing in common.

What about the IC/BC situation ? As you would expect, we need two specifications of u (or u') at two values of x and all values of t, as well as two specifications of u (or $\dot{u}$) at two values of t and all values of x. Practically, what happens is that the BC indeed features u or u' at two values of x and all t, while the IC features u as well as $\dot{u}$ at the same value of t (usually $t = 0$). Why this is so, we can best understand if we think of the archetypal example of a system which satisfies the wave equation. This is a stretched string which runs between two fixed points – the dependent variable is the displacement of the string perpendicular to its length. (The derivation of this fact, we'll do later when we come to the applications of the wave equation – for now we

treat this as a given and focus on the equation as primarily a mathematical construct.) Evidently, the displacement at the two ends at all time is zero by definition, while to start the system into vibrations, we either pluck it or strike it. These two correspond to the initial position and velocity as a function of position, and motivate the nature of the IC.

Separation of variables on the equation leads to trigonometric functions in both x and t, the details of which we'll see in the Tutes. There is another method however of solving the 1+1 wave equation, which was invented by JEAN D'ALAMBERT. What he did was, he considered (01) defined on the infinite real line, x belongs to $[-\infty, \infty]$. He imposed plausibility BCs that u remain finite at $x = \pm\infty$ and general ICs $u(x,0) = f(x)$ and $\dot{u}(x,0) = g(x)$ for well-behaved functions f and g. He then showed that the solution of (01) which satisfies these BCs and ICs is

$$u(x,t) = \frac{1}{2}\left[f(x+ct) + f(x-ct) \right] + \frac{1}{2c} \int_{x-ct}^{x+ct} g(s)\,\mathrm{d}s \quad . \qquad \textbf{(02)*}$$

This is called **D'ALAMBERT's formula**. As usual, the rather elaborate derivation will come in Tutes. Although the formula is for infinite spatial domain, it can be applied to finite domains also by performing a suitable extension of the IC function.

In higher spatial dimensions, the $\partial^2/\partial x^2$ term gets replaced by ∇^2. The case of 3-dimensional wave with spherical symmetry is easy to handle – the substitution $u(r,t) = (1/r)\, w(r,t)$ reduces the problem to a 1+1 wave equation for w. Complicated formulae for more general situations in two and three dimensions also exist, but it turns out that we won't need those for any of the problems here.

Example 7.1.1

Simple wave equation problem

The transverse displacement $u(x,t)$ of the stretched string x belongs to [0, 1] satisfies the 1+1 wave equation

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{4}\frac{\partial^2 u}{\partial t^2} \quad ,$$

with BCs $u(0,t) = u(1,t) = 0$ and ICs $u(x,0) = f(x)$, $\dot{u}(x,0) = 0$, where $f(x)$ is the function which is unity in the interval [2/5, 3/5] and zero everywhere else.

(a) Separate variables and obtain the eigenfunction expansion of $u(x,t)$.

(b) Fit the IC to obtain the specific solution $u(x,t)$.

Now consider the same problem with the two ICs reversed i.e. $u(x,0) = 0$ and $\dot{u}(x,0) = f(x)$.

(c) Again determine $u(x,t)$.

[Level : 2]

I could have made this Read-Only and given an easier IC but I think you have had enough of PDE practice by now to take this in your stride. Trying this by yourself is also an excellent practice for handling an unseen PDE in the exam hall.

(a) By now we are masters of the art of separating variables. We break up $u(x,t) = X(x)T(t)$, plug into the wave equation and divide by XT. Then invoke the quintessential sep var argument (memory recall : if a function of x alone equals a function of t alone then both must be equal to a constant). This gives

$$\frac{X''}{X} = \frac{1}{4}\frac{\ddot{T}}{T} = \lambda \quad .$$

As usual we have the three cases $\lambda>0$, $\lambda<0$ and $\lambda=0$. The first case produces exponential functions in both space and time, the second produces trigonometric functions in both variables, and the third produces constant and linear functions.

The fail-safe way to proceed from here is to consider all three cases and employ the BCs in each case. A double-zero BC on X is something we've seen many times before – instinctively (hopefully) we know that the trigs case will yield eigenfunctions and all other cases will throw up the trivial solution. (Jargon : a double zero BC for a string is often called "clamped at both ends".)

So letting $\lambda = -k^2$, we have the ODEs

$$X'' + k^2 X = 0 \quad ,$$
$$\ddot{T} + 4k^2 T = 0 \quad ,$$

with the respective solutions

$$X = (\cos,\ \sin)kx \quad ,$$
$$T = (\cos,\ \sin)2kt \quad .$$

The BCs quickly yield $A = 0$ and $k = n\pi$, so the eigenfunction expansion for u is

$$u(x,t) = \sum_n \sin n\pi x \left(C_n \cos 2n\pi t + D_n \sin 2n\pi t\right) \quad , \qquad (01)^*$$

which is the desired answer. The oscillation frequencies $2n\pi$ are often called the EIGENFREQUENCIES of the system. Note that this solution is temporally periodic with time period 1 – substituting $t+1$ for t into (01) gives us back every spatial eigenfunction in its original form, and hence reproduces the solution identically.

Solutions of the wave equation in finite-sized domains are almost always periodic in time with the lowest eigenfrequency.

(b) So far it's been a routine algebra hack, like on any other equation; now is the step when things become special to this equation. There are two families of arbitrary constants, C_n and D_n; to determine both, we'll need the two ICs position and velocity. Fitting the position first, we set $t=0$ into (01); this yields

$$u(x,0)=\sum_n C_n \sin n\pi x = f(x) \quad ,$$

where $f(x)$ is the step-like function given in the question. This of course is a Fourier series problem for C_n.

The velocity is the $\partial/\partial t$ of (01), which is

$$\dot{u}(x,t)=\sum_n \sin n\pi x\left(-2n\pi C_n \sin 2n\pi t + 2n\pi D_n \cos 2n\pi t\right) \quad .$$

Setting $t=0$ here and fitting the velocity IC gives

$$\dot{u}(x,0)=\sum_n 2n\pi D_n \sin n\pi x = 0 \quad ,$$

implying that all the D's are zero.

Finally, we have to evaluate the Fourier coefficients C_n. The function this time is not something we've encountered before so we use the coefficient formula (5.1.3.05), accounting for the antisymmetry of f and the sines :

$$C_n = \frac{2}{L}\int_0^L f(x)\sin\frac{n\pi x}{L}\,\mathrm{d}x \quad .$$

Of course $L=1$ in this case. Since $f(x)=1$ in $[2/5,\ 3/5]$ and zero everywhere else, the integral evaluates to

$$C_n = \frac{2}{n\pi}\left(\cos\frac{2n\pi}{5}-\cos\frac{3n\pi}{5}\right) \quad .$$

Setting this into (01), we have the specific solution

* $$u(x,t)=\sum_n \frac{2}{n\pi}\left(\cos\frac{2n\pi}{5}-\cos\frac{3n\pi}{5}\right)\sin n\pi x\cos 2n\pi t \quad , \qquad (02)^*$$

which is the desired answer.

(c) The eigenfunction expansion remains unchanged from Part (a). After that, the position IC $u(x,0)=0$ gives $C_n=0$ for all n. The velocity IC yields

$$\dot{u}(x,0)=\sum_n 2n\pi D_n \sin n\pi x = f(x) \quad ,$$

which implies that $2n\pi D_n$ [rather than D_n itself as was the case in Part (b)] is the n^{th} Fourier coefficient of $f(x)$. So we have

$$D_n=\frac{1}{n^2\pi^2}\left(\cos\frac{2n\pi}{5}-\cos\frac{3n\pi}{5}\right) \quad ,$$

leading to the specific solution

$$u(x,t)=\sum_n \frac{1}{n^2\pi^2}\left(\cos\frac{2n\pi}{5}-\cos\frac{3n\pi}{5}\right)\sin n\pi x \sin 2n\pi t \quad , \qquad (03)^*$$

which is the desired answer.

*

COMMENTS : In Part (a) there is a different argument which we can use to eliminate the cases $\lambda>0$ and $\lambda=0$. We say, drawing on physical experience, that the wave equation always has solutions which are periodic in time. This constraint eliminates the undesirable λ's at the outset. Note however that these two cases might not always produce the trivial solution mathematically. For example, if the BCs were that $\partial u/\partial x=0$ at the two ends, then the case $\lambda=0$ will produce eigenfunctions. It might be possible to specially construct BCs such that the exponential solutions are also present. However, from a physical perspective, such solutions are meaningless. Just as the harmonic oscillator describes the vibrations of a mass on a spring, the wave equation describes spatiotemporal vibrations of a continuous medium. Exponential decay (we won't even consider growth) is not a behaviour which such a system shows. The constant and linear solutions are trivial physically even if they aren't so mathematically. A constant is just a constant displacement. The linear term is translation with a constant velocity. Both these motions are force-free and don't excite vibrations of the system, hence they are of little interest.

Other than being a basic wave equation demonstrator and a typical final exam question, this problem also shows the differences in the vibration profiles of a plucked and a struck string. The position IC of Part (b) is a good model of a guitar

string, where the central portion is plucked with the plectrum (or with fingers) and the string then allowed to vibrate. We can see that the n^{th} harmonic has an amplitude of $1/n$ for large n. The velocity IC of Part (c) is a model of a piano string, where the central portion is struck with a hammer which imparts an instantaneous initial velocity. We can see that the n^{th} harmonic has an asymptotic amplitude of $1/n^2$. The timbre of a musical instrument is indeed determined by the relative amplitudes of the higher harmonics present in the tone – two different instruments playing the same note vibrate at the same fundamental frequency and differ only in the higher frequencies.

So this was guitar and piano – what about violin ? That is a completely different story as the bow introduces a nonlinear excitation term into the wave equation. Unlike these two, that is not a low-hanging fruit – to gather it in our basket we'll have to climb all the way up to the top of the tree.

Example 7.1.2

D'ALAMBERT's formula

Consider the 1+1 wave equation

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} \quad ,$$

defined in x belongs to $[-\infty, \infty]$ with BCs u remains finite at $|x|=\infty$ and ICs $u(x,0)=f(x)$ and $\dot{u}(x,0)=g(x)$.

(a) Define new variables $\xi = x+ct$ and $\eta = x-ct$ and express the wave equation in terms of ξ and η.

(b) Basis the above result, derive D'ALAMBERT's formula (7.102).

(c) Solve Part (b) of the previous problem using this formula.

(d) Considering Part (c) of the previous problem, determine the motion of the midpoint of the string using this formula.

[Read Only]

(a) Like 5.4.7, here's another variable transformation problem, and a popular one at that. As usual, the first thing is to see which way we should do the substitution. The original equation has derivatives w.r.t x and t; for the new equation we want derivatives w.r.t ξ and η. So we should express $\partial/\partial x$ and $\partial/\partial t$ (and the second derivatives) in terms of $\partial/\partial\xi$ and $\partial/\partial\eta$ and then substitute these expressions in place of the corresponding terms in the original equation. The actual derivative conversion is via Chain Rule (yes, you really need complete familiarity with this one). We have

$$\begin{aligned}\frac{\partial u}{\partial x} &= \frac{\partial u}{\partial \xi}\frac{\partial \xi}{\partial x} + \frac{\partial u}{\partial \eta}\frac{\partial \eta}{\partial x} \\ &= \frac{\partial u}{\partial \xi} + \frac{\partial u}{\partial \eta}\end{aligned} \quad .$$

Taking another derivative w.r.t x readily yields

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial \xi^2} + 2\frac{\partial^2 u}{\partial \xi \partial \eta} + \frac{\partial^2 u}{\partial \eta^2} \quad .$$

On the t side we have

$$\begin{aligned}\frac{\partial u}{\partial t} &= \frac{\partial u}{\partial \xi}\frac{\partial \xi}{\partial t} + \frac{\partial u}{\partial \eta}\frac{\partial \eta}{\partial t} \\ &= c\left(\frac{\partial u}{\partial \xi} - \frac{\partial u}{\partial \eta}\right)\end{aligned} \quad ,$$

the second derivative is

$$\frac{\partial^2 u}{\partial t^2} = c^2\left(\frac{\partial^2 u}{\partial \xi^2} - 2\frac{\partial^2 u}{\partial \xi \partial \eta} + \frac{\partial^2 u}{\partial \eta^2}\right) \quad ,$$

and all terms of the original equation have been converted to the new variables.

Now we actually plug these into the original PDE :

$$\begin{aligned} PDE \rightarrow & \frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} \\ \rightarrow & \frac{\partial^2 u}{\partial \xi^2} + 2\frac{\partial^2 u}{\partial \xi \partial \eta} + \frac{\partial^2 u}{\partial \eta^2} = \frac{1}{c^2}c^2\left(\frac{\partial^2 u}{\partial \xi^2} - 2\frac{\partial^2 u}{\partial \xi \partial \eta} + \frac{\partial^2 u}{\partial \eta^2}\right) \end{aligned} \quad ,$$

and after obvious cancellations,

$$\frac{\partial^2 u}{\partial\xi\partial\eta}=0 \quad , \tag{01}$$

which is the desired answer.

The above process is the general way of effecting a substitution in a PDE – it can take an arbitrary PDE from the existing variables to a set of given new variables. In this particular case, there is a shortcut to (01) which we can use if we know the answer already (which we do thanks to Monsieur D'ALAMBERT). We can factor the original wave equation as

$$\frac{\partial^2 u}{\partial x^2}-\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}=0\rightarrow\left(\frac{\partial}{\partial x}+\frac{1}{c}\frac{\partial}{\partial t}\right)\left(\frac{\partial}{\partial x}-\frac{1}{c}\frac{\partial}{\partial t}\right)u=0 \quad ,$$

and then work the Chain Rule in the reverse direction.

Inverting the definitions of ξ and η to get $x=(\xi+\eta)/2$ and $y=(\xi-\eta)/2c$, we use Chain Rule to write

$$\frac{\partial}{\partial\xi}=\frac{1}{2}\frac{\partial}{\partial x}+\frac{1}{2c}\frac{\partial}{\partial t} \qquad \frac{\partial}{\partial\eta}=\frac{1}{2}\frac{\partial}{\partial x}-\frac{1}{2c}\frac{\partial}{\partial t} \quad ,$$

whereby it follows at once that the wave equation is $\partial^2 u/\partial\xi\partial\eta=0.$

(b) What we have on the table is (01) i.e. $u_{\xi\eta}=0$ where ξ and η are two independent variables. Evidently, any arbitrary function of either variable satisfies the equation by default, since the partial derivative w.r.t the other variable identically vanishes. By linearity, the sum of two such functions is also a solution. So the most general form of the solution is $u_G(\xi,\eta)=\varphi(\xi)+\psi(\eta)$ where φ and ψ are arbitrary functions. If we write these in terms of x and t, we have the general solution as $u_G(x,t)=\varphi(x+ct)+\psi(x-ct).$

Before proceeding with the IC fit, let's spend a moment to see that this identifies c as the speed of wave propagation. An arbitrary function of $x-ct$ is a function which moves to the right with speed c. Just as in 5.4.7, if we consider an observer who moves right at speed c, what we call x, he will call $x-ct$, and to him the function ψ will appear a constant function. Likewise, a second observer who moves left at speed c will define her own position to be $x+ct$ and the function φ will for her be a constant. So the general solution (and statements about general solution are very robust since they cannot be affected by BCs and ICs) supports

only two kinds of solutions – waves which travel to the left and to the right, both with speed c.

Now fit the ICs. The position is $u(x,0) = f(x)$; setting $t = 0$ into u_G gives $\varphi(x) + \psi(x) = f(x)$. The velocity is $\dot{u}(x,0) = g(x)$. We already got $\partial u/\partial t$ while performing the variable transformation; setting u_G into that expression and then $t = 0$ gives

$$c\left(\frac{\mathrm{d}\varphi}{\mathrm{d}\xi}(x) - \frac{\mathrm{d}\psi}{\mathrm{d}\eta}(x)\right) = g(x) \quad .$$

Here and henceforth it's very important to keep track of who is a function of what. The dependent variable u is a multivariate function; it can be one of x and t or one of ξ and η. The functions φ and ψ however are univariate functions of their respective arguments. Originally defined w.r.t ξ and η, they can of course be defined on any other argument also. Likewise, f and g are also univariate functions, originally both of x.

So we have got two relations connecting φ and ψ to f and g. Now we will no longer think of them as functions of x,ξ,η etc but as functions of some abstract dummy variable p. In this abstract space we'll obtain a pair of functional equations for φ and ψ and substitute the appropriate function arguments for p only at the end. In terms of p we rewrite the IC relations as

$$\varphi(p) + \psi(p) = f(p) \quad , \tag{02a}$$

$$\varphi_p(p) - \psi_p(p) = \frac{g(p)}{c} \quad , \tag{02b}$$

where the subscript p means $\mathrm{d}/\mathrm{d}p$. Integrating (02b) gives

$$\varphi(p) - \psi(p) = \frac{G(p)}{c} \quad ,$$

where G is an antiderivative of g. This can easily be combined with (02a) to get

$$\varphi = \frac{1}{2}\left(f + \frac{G}{c}\right) \qquad \psi = \frac{1}{2}\left(f - \frac{G}{c}\right) \quad , \tag{03}$$

where everything is function of p.

Now, what is $G(p)$? It is just $\int g(p)\mathrm{d}p$. But a more pedantic way of writing it is to say $G(p)=\int_O^p g(s)\mathrm{d}s$ where O is some initial point (it may or may not be zero, and its value will eventually be irrelevant) and s is a dummy variable (in fact a dummy's dummy since p is already a dummy). We have seen this pedantic notation twice before – in 1.1 while writing the integrating factors formula and in 3.1 while writing the variation of parameters formula. In both those cases, it was overfuss, but now it is essential. So we now write (03) as

$$\varphi(p)=\frac{1}{2}f(p)+\frac{1}{2c}\int_O^p g(s)\mathrm{d}s \quad , \tag{04a}$$

$$\psi(p)=\frac{1}{2}f(p)-\frac{1}{2c}\int_O^p g(s)\mathrm{d}s \quad . \tag{04b}$$

These finally are two functional equations for φ and ψ in terms of f and g.

Now come back from abstract space to earth. The solution of the wave equation is $\varphi(x+ct)+\psi(x-ct)$, while the IC fit has given us φ and ψ in terms of f and g. Setting $p=x+ct$ into (04a) and $p=x-ct$ into (04b) we have

$$\varphi\left(x+ct\right)=\frac{1}{2}f\left(x+ct\right)+\frac{1}{2c}\int_O^{x+ct} g(s)\mathrm{d}s \quad ,$$

$$\psi\left(x-ct\right)=\frac{1}{2}f\left(x-ct\right)-\frac{1}{2c}\int_O^{x-ct} g(s)\mathrm{d}s \quad .$$

Then add the two together :

$$\begin{aligned} u(x,t)&=\varphi\left(x+ct\right)+\psi\left(x-ct\right)\\ &=\frac{1}{2}\left[f\left(x+ct\right)+f\left(x-ct\right)\right]+\frac{1}{2c}\int_O^{x+ct} g(s)\mathrm{d}s-\frac{1}{2c}\int_O^{x-ct} g(s)\mathrm{d}s \end{aligned} \quad .$$

We're almost there; the only fly in the ointment is the pesky O attaching to the integrals. So we flip the sign and swap the integration limits on the second integral. This turns it into a positive integral of g from $x-ct$ to O. Now this adds to the first term, which is an integral of g from O to $x+ct$. Their resultant is a single integral of g from $x-ct$ to $x+ct$: O goes out of the picture. So we have

$$u(x,t)=\frac{1}{2}\left[f(x+ct)+f(x-ct)\right]+\frac{1}{2c}\int_{x-ct}^{x+ct} g(s)\mathrm{d}s \quad ,$$

which finally is D'ALAMBERT's formula (7.1.02).

This is by no means an easy derivation. On exams, it can appear as one of the harder questions if the prof has an inclination towards derivations. Otherwise, questions testing the application of the formula are far more common. So you should certainly remember the formula even if you decide to let go of the derivation. Two such application questions are the parts coming up now.

(c) The question here is 7.1.1 Part (b) which is wave equation with $c=2$ defined on the string $[0,\ 1]$ clamped at both ends. Initial displacement is 1 in the region $[2/5,\ 3/5]$ and zero everywhere while initial velocity is zero everywhere.

The primary step in applying D'ALAMBERT's formula is to note that the formula is defined only on the infinite string and NOT on any finite domains ! Before using the formula, you MUST perform an extension of the problem from the finite to the infinite domain. The nature of the extension must come from the eigenfunction expansion – since all eigenfunctions are sines, they are antisymmetric and periodic, and hence the extension must also be the same. This is of course similar to the extensions we used for solving the first few heat equation problems (5.1.2 to 5.1.6). There is a big difference however, in that the extension was optional for the heat equation (we can also do them using eigenfunction orthogonality) but is obligatory here.

So the extended position IC $f(x)$ consists of a positive step between 2/5 and 3/5 and a negative step between $-3/5$ and $-2/5$, then another negative step between 7/5 and 8/5, another positive between $-8/5$ and $-7/5$ etc. The velocity IC is zero, extended or otherwise. D'ALAMBERT's formula now says $u=\left[f(x+2t)+f(x-2t)\right]/2$. At $t=0$, both these components sit on top of each other and the extended IC is the sum of the left- and right-moving waves. As t starts increasing from zero, these two components drift off in their respective directions. Although the problem is extended, the actual wave motion is visible only in the fundamental cell $[0,\ 1]$, where the string really exists. The extension to the other cells is a purely mathematical construct which will enable us to visualize the motion in the domain of interest.

The step IC at the centre of the string breaks up into left and right-moving components of half the height. Things are fine until they reach the boundaries of the string i.e. $x = 0$ and $x = 1$. Although in the infinite problem, these steps continue moving beyond the cell boundaries, the negative steps from the adjacent cells ($[-1, 0]$ and $[1, 2]$) now enter into the fundamental cell. These are the reflected waves. They approach each other and meet in the centre, producing a big negative step (size equal to original). Then this step again breaks up as the two components move off towards the boundaries. The next entrants into the fundamental cell are the positive steps from the next-to-nearest neighbour cells $[-2, -1]$ and $[2, 3]$. These again approach each other and meet in the centre, producing a big step just like the IC. From this point onwards of course the cycle goes on for ever.

(d) This time, the initial position $f(x) = 0$ while the initial velocity $g(x)$ is the step function. As before, we extend antisymmetrically and periodically, so that the extended $g(x)$ consists of the positive step inside the fundamental cell together with the alternating negative and positive steps in the adjacent cells and the next and the one after.

Now what has been asked for is the motion of the midpoint (which we call P) of the string i.e. $u(1/2,t)$. From D'ALAMBERT's formula,

$$u(1/2,t) = \frac{1}{4}\int_{1/2-2t}^{1/2+2t} g(s)\,\mathrm{d}s \quad .$$

This is an integral of the alternating steps over an interval whose centre is at $x = 1/2$ and whose width increases in time as $4t$. At $t = 0$, the integral is of course zero. As time starts to evolve, the integration interval covers more and more of the fundamental step. Since the step is a constant function, the value of the integral i.e. $u(1/2,t)$ also increases linearly and the point P moves up with a constant velocity. The displacement saturates when the interval covers the entire fundamental step, which happens at $t = 1/20$ (since the step extends for a distance 1/10 on either side of the centre, and $c = 2$). The value of the maximum is 1/4 times the area under the fundamental step, which is 1/20, implying that the speed of P during the ascent is 1.

For a while starting from this point, although the integration interval gets wider and wider, the integral does not pick up any contribution because g is zero

outside the fundamental step. So P remains stationary during this time. The next contributions occur when the limits touch the boundaries of the negative steps in the adjacent cells. This happens at a time of 4/10 unit after the attainment of the first maximum. Now, the contribution is negative, so $u(1/2,t)$ starts to decrease; the rate of decrease is constant because the steps are constant functions. This implies that P moves downwards with constant velocity. The decrease is complete when the integration interval has swelled to accommodate the entirety of both the negative steps in the two fundamental cells. At this time, the value of the displacement is 1/4 times the total area under one positive and two negative steps, which is $-1/20$. Again, the speed of P during the descent is 1.

From this point onwards, the motion is predictable. After a stationary period at the minimum, P again moves upwards towards its maximum, then a halt, then downwards towards the minimum and so forth, all motions taking place with constant vertical speed 1.

*

COMMENTS : Since we just solved the same problem using two methods, let us compare between the two. Sep var is the workhorse of PDE – a machine which takes any PDE, BCs and ICs, and spits out answers (at least in some form). D'ALAMBERT's formula holds only for the wave equation and is perhaps more difficult to apply, but it can produce explicit answers rather than serieses.

In fact, the result of Part (c) here can also be got from the separable solution (T11202) without too much hassle. That solution is

$$u(x,t)=\sum_n \frac{2}{n\pi}\left(\cos\frac{2n\pi}{5}-\cos\frac{3n\pi}{5}\right)\sin n\pi x\cos 2n\pi t \quad .$$

Now use the $\sin C\cos D$ identity from the trig review :

$$u(x,t)=\sum_n \frac{2}{n\pi}\left(\cos\frac{2n\pi}{5}-\cos\frac{3n\pi}{5}\right)\frac{1}{2}\left(\sin n\pi(x+2t)+\sin n\pi(x-2t)\right) \quad .$$

We already know that the sum over n of the messy coefficient times $\sin n\pi x$ is the IC function $f(x)$. It immediately follows that the sum of the above series is $[f(x+2t)+f(x-2t)]/2$, which is what we got from D'ALAMBERT's formula.

The result of Part (d) here however is not so easy to get from sep var. The separable solution for $u(1/2,t)$ is [obtained by setting $x=1/2$ in (T11203) and simplifying the $\sin(n\pi/2)$'s]

$$u(1/2,t)=\sum_{\text{odd } n}\frac{(-1)^{n+1}}{n^2\pi^2}\left(\cos\frac{2n\pi}{5}-\cos\frac{3n\pi}{5}\right)\sin 2n\pi t \quad .$$

We've already seen in 4.3.3 that Fourier expansion is a one-way operation. So, it is not at all an obvious conclusion that the sum of the above series is the simple up-and-down function we described in Part (d) here.

In conclusion, is sep var always worse than D'ALAMBERT ? No. For one, before using D'ALAMBERT's formula, we had to do an extension. The nature of the extension came from sep var. For another, the separable solutions quickly told us that the harmonic content of the guitar string is $1/n$ while that of the piano string is $1/n^2$. The explicit solutions we found here do not yield this information directly. The bottom line is that both are useful techniques in their own right, and you should have both up your sleeve.

Example 7.1.3

Vibrations of a drumhead

The vertical displacement $z(x,y,t)$ of the drumhead x belongs to [0, L] with $L > 1$ and y belongs to [0, 1] satisfies the 2+1 wave equation

$$\frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial y^2} = \frac{\partial^2 z}{\partial t^2} \quad .$$

The BCs are that $z = 0$ along all four walls of the drum.

(a) Determine the temporal oscillation frequencies of the eigenfunctions (or modes) having the two lowest frequencies.

(b) If the second mode is to have a frequency 3/2 times that of the first or fundamental mode, determine L. Also determine the frequencies of the three lowest modes.

The drummer generates sounds by striking the drumhead with a hammer at some point $(x,y) = (a,b)$. This can be modelled as an IC $z(0) = 0$ and $\dot{z}(0) = \delta(x-a)\delta(y-b)$. It is known that the third mode of the drum is dissonant to the first and second modes.

(c) Where should the drumhead be struck so that the third mode is not produced at all ? Justify your answer.

[Level : 3]

(a) No surprise in how we start this problem off – we separate variables. Writing $z = X(x)Y(y)T(t)$, plugging into the 2+1 wave equation and dividing by XYT we have

$$\frac{X_{xx}}{X} + \frac{Y_{yy}}{Y} = \frac{\ddot{T}}{T} \quad ,$$

then proceeding similarly as in 6.1.3 and 6.2.1,

$$X_{xx} = \lambda_1 X \qquad Y_{yy} = \lambda_2 Y \qquad \ddot{T} = (\lambda_1 + \lambda_2)T \quad .$$

The BCs are also old hat : $X = \sin(n_1 \pi x / L)$ with $\lambda_1 = -(n_1^2 \pi^2 / L^2)$ and $Y = \sin(n_2 \pi y)$ with $\lambda_2 = -n_2^2 \pi^2$. This then tells us

$$\ddot{T} + \left(\frac{n_1^2 \pi^2}{L^2} + n_2^2 \pi^2 \right) T = 0 \quad ,$$

which has the solutions

$$T = (\cos,\ \sin)\omega_{n_1,n_2} t \quad \text{, where} \qquad \text{(01a)}$$

$$\omega_{n_1,n_2} = \pi \left(\frac{n_1^2}{L^2} + n_2^2 \right)^{1/2} \quad . \qquad \text{(01b)}$$

I went quickly upto here because I have covered nearly identical calculations in 6.1.3 and 6.2.1 in far greater detail – if you have confusion about any of the steps so far, go back to those Tutes at once and then fill out the details for this one.

Now we want the two modes which have the lowest oscillation frequency. Equation (01b]) gives the frequency formula, and we evidently have to find the two lowest ω_{n_1,n_2}. Since ω is a sum of squares, the thing is quite easy – the lower the two n's, the lower the frequency. The only thing to be careful of is that neither n can be zero – that corresponds to the trivial solution and not an eigenfunction ! The lowest frequency is of course with $n_1 = n_2 = 1$; that gives

* $$\omega_{1,1} = \pi \left(\frac{1}{L^2} + 1 \right)^{1/2} \quad . \qquad *$$

The next one can be either 1,2 or 2,1. Now, the question tells us that $L > 1$. Setting this into (01b) shows that 2,1 will have a lower frequency than 1,2. So the second lowest frequency is with $n_1 = 2$ and $n_2 = 1$, giving

* $$\omega_{2,1} = \pi\left(\frac{4}{L^2}+1\right)^{1/2} \quad .$$ *

The above two modes make up the desired answer.

Note that at this stage, without knowing the value of L, it is impossible to say anything about which mode is the next-highest. If L is close to 1, it is likely to be 1,2 (for a square drum 1,2 and 2,1 have the same frequency) while if L is very large (say 100), then it will be 3,1. So, basis what we have so far, only these two modes can be identified uniquely.

(b) We now have a condition on the first two modes : the frequency of the second is 3/2 times that of the first. This means

$$\left(\frac{4}{L^2}+1\right)^{1/2} = \frac{3}{2}\left(\frac{1}{L^2}+1\right)^{1/2} \quad .$$

Squaring yields $1/L^2 = 5/7$ so $\underline{L = \sqrt{(7/5)}}$. This gives the values of the first two frequencies as $\underline{\omega_{1,1} = \pi\sqrt{(12/7)}}$ and $\underline{\omega_{2,1} = \pi\sqrt{(27/7)}}$. Clearly, they satisfy the 3:2 relationship required in the question.

We suspect that the third mode is going to be 1,2 and not 3,1. Indeed, 1,2 has a frequency of $\pi\sqrt{(33/7)}$ while 3,1 has $\pi\sqrt{(52/7)}$. So the third highest frequency is $\underline{\omega_{1,2} = \pi\sqrt{(33/7)}}$.

(c) This requires IC fit so we write out the eigenfunction expansion explicitly. We have

$$z(x,y,t) = \sum_{n_1,n_2} \left(\sin\frac{n_1\pi x}{L}\right)(\sin n_2\pi y)\left(C_{n_1,n_2}\cos\omega_{n_1,n_2}t + D_{n_1,n_2}\sin\omega_{n_1,n_2}t\right) \quad ,$$

where L looks cleaner than the $\sqrt{(5/7)}$ which it denotes. The IC has $z(0) = 0$ so all the C's go away. The time derivative is

$$\dot{z}(x,y,t) = \sum_{n_1,n_2} \omega_{n_1,n_2} D_{n_1,n_2}\left(\sin\frac{n_1\pi x}{L}\right)(\sin n_2\pi y)\cos\omega_{n_1,n_2}t \; .$$

Setting $t = 0$ here gives the double Fourier series for the IC :

$$\sum_{n_1,n_2} \omega_{n_1,n_2} D_{n_1,n_2}\left(\sin\frac{n_1\pi x}{L}\right)(\sin n_2\pi y) = \delta(x-a)\delta(y-b) \quad .$$

Again we have already seen how to handle this in 6.1.3. The coefficient ωD (I'm letting go of the subscripts for clarity) is the double integral of the IC function and the eigenfunctions over the rectangle , multiplied by the squares of the normalization constants. These constants are $\sqrt{(2/L)}$ for x and $\sqrt{2}$ for y. This means

$$\omega_{n_1,n_2} D_{n_1,n_2} = \frac{4}{L}\int_0^L \mathrm{d}x \int_0^1 \mathrm{d}y \; \delta(x-a)\delta(y-b)\sin\frac{n_1\pi x}{L}\sin n_2\pi y \quad .$$

An integral of function times delta picks out the function value at the location of the delta; this tells us that

$$D_{n_1,n_2} = \frac{1}{\omega_{n_1,n_2}}\sin\frac{n_1\pi a}{L}\sin n_2\pi b \quad . \qquad (02)$$

Now, what we want to do is kill the mode 1,2; for this we should choose a and b such that $D_{1,2}$ evaluates to zero. So either $\sin(\pi a/L)$ must be zero or $\sin 2\pi b$ must be zero. The first condition implies $a = 0$ or $a = L$ i.e. the drumhead must be struck at its boundary, which is a trivial and meaningless condition. The second condition implies $b = 1/2$, which is plausible. Hence, to eliminate the third mode, the player must strike the drumhead at <u>any point whose ordinate is 1/2</u>.

*

COMMENTS : In case you are scoffing at rectangular drum, the analysis of the circular drum features Bessel functions while that of elliptical drum brings out Mathieu equations. What happens then is that the basic physical concepts behind the drum get drowned in a sea of mathematical complexity. On the other hand, this simple shape makes for easy maths but does not throw away any fundamental physics of drums, as I will now show.

The first thing is that a lot of the drum's sound comes from the fundamental and low modes. The Fourier coefficient (02) has a $1/\omega$ which is approximately $1/n_1$ and $1/n_2$. This weak dependence happened because delta is an intractable function whose all Fourier coefficients are of order unity. In reality, the initial velocity will not be actually a delta but some better-behaved function of

position, whose coefficients will have $1/n$ or $1/n^2$ decay rates (n meaning both n_1 and n_2). This will arrange for quite a rapid decay of the higher harmonics.

Next, we saw that changing the shape of the rectangle changes the frequency of the higher modes relative to that of the fundamental mode. The frequency ratio 3:2 ensures that the second mode is consonant with the first – if the fundamental mode is the note C then 3/2 frequency corresponds to G and the two together are part of the tonic chord on C. If the drum had been square then the second mode would have had a frequency of 1.58 times the first, which makes it C and A^b, a totally different chord. In real drums also, elaborate shaping is done to ensure that the second and (as many as possible) higher harmonics are consonant with the fundamental, so that the drum produces a rich, on-topic tone.

What Part (c) shows is that the location of striking the drumhead makes a big difference to the tone quality. The 1,2 mode in this example is indeed dissonant – if the first two are C and G, then this is A. However, hitting the drumhead in a particular spot can kill it (practically, reduce its amplitude greatly). In a like manner, we can perform calculations for the amplitudes of various higher harmonics as functions of the striking location. For example, if the drumhead is hit at $x = L/2$, that will kill the 2,1 mode. Hitting at the centre will kill everything upto 3,1 (excluding the fundamental). Indeed, the variation of tone quality with the location of striking is well known to timpani players – blows near the centre are said to sound ‘thin’, off-centre strikes are ‘rich’ while too close to the periphery the sound becomes ‘tinny’. Modern compositions by GEORGE GERSHWIN and BELA BARTOK sometimes explicitly include such effects in the score.

Here we can also see that the eigenfrequencies of the drum are not all integer multiples of the lowest one, hence the solution starting from an arbitrary IC will not be periodic with the lowest frequency. Indeed, the various frequencies are related in irrational ratios to each other, so the solution will not be periodic at all but quasiperiodic (see 2.3.5).

A related but orders of magnitude more difficult question is, can one hear the shape of a drum ? That is, suppose we have a drumhead of arbitrary shape in the plane, constrained to zero displacement all along the periphery and having a given wave speed c. Then, is it possible for two different drumheads to have

all their eigenfrequencies equal, each to each ? The one-dimensional equivalent of this problem – can one hear the length of a string – is trivial. A string of length L with double zero BC has eigenfrequencies $n\pi c/L$ so different L means different eigenfrequencies, simple as that. The two-dimensional question was posed explicitly by MARK KAC in 1966. Mathematicians battled the issue until 1992 when CAROLYN GORDON, DAVID WEBB and WOLPERT SCOTT showed with an explicit example that the answer is NO. They constructed two drumheads of equal area, equal perimeter but different shape whose all eigenfrequencies were identical. Needless to say, I cannot understand or reproduce a word of their paper beyond the abstract.

Example 7.1.4

Heat or wave

The function $u(x,t)$ satisfies the PDE

$$\frac{\partial^2 u}{\partial x^2} = \alpha \frac{\partial^2 u}{\partial t^2} + (1-\alpha)\frac{\partial u}{\partial t} \quad ,$$

where α is a real number between 0 and 1. The domain of definition is x belongs to $[0, L]$ with BCs $u(0,t) = u(L,t) = 0$. The problem is made well-defined by specifying some ICs whose exact nature is not of interest.

Drawing reference to 2.1.6 where appropriate , discuss the similarities and differences between the nature of the solutions in the cases α is very close to, and α is exactly equal to, 1 and 0.

[Level : 2]

EXTRA :

Show that in the limit $\alpha \rightarrow 1$ but α not exactly equal to 1, information about disturbances can propagate instantaneously instead of being constrained by the maximum permissible speed $c = 1$.

First we do the mechanical part – separate variables assuming an arbitrary value of α. Writing $u = XT$, plugging and dividing by XT we have

$$X'' = \lambda X \quad ,$$
$$\alpha \ddot{T} + (1-\alpha)\dot{T} = \lambda T \quad .$$

Double zero BC on x so $\lambda = -(n^2\pi^2/L^2)$, $X = \sin(n\pi x/L)$. Now plug this λ into the T equation :

$$\ddot{T} + \frac{1-\alpha}{\alpha}\dot{T} + \frac{n^2\pi^2}{\alpha L^2}T = 0 \quad . \tag{01}$$

This is a damped harmonic oscillator. Depending on the values of α and n, the damping can be less than, greater than, or equal to the critical, and the solutions will be the corresponding decaying oscillations, monotonic decay and critical decay. This completes the routine part of the process, now we go on to examine the limiting cases.

The case $\alpha = 1$ is evidently the wave equation. For α just less than 1, we have an $O(1)$ second derivative term and a very small first derivative term on the RHS. This is a damping term, and can model say the internal friction in a vibrating string. Considering (01), when $1-\alpha$ is very small, we expect that the equation is underdamped for all n. Hence the solutions are

$$T = \exp\left(\frac{\alpha-1}{2\alpha}t\right)(\cos,\ \sin)\omega_{dn}t \quad \text{, where} \tag{02a}$$

$$\omega_{dn} = \frac{\left[\frac{4n^2\pi^2}{\alpha L^2} - \left(\frac{1-\alpha}{\alpha}\right)^2\right]^{1/2}}{2} \quad , \tag{02b}$$

i.e. ω_{dn} is the n^{th} damped frequency. Evidently, if we set $\alpha = 1$ in the above, the exponential decay rate tends to zero and ω_{dn} tends to the n^{th} eigenfrequency of the wave equation (with $c = 1$), so the limit appears to be a very smooth one.

However , just as in 2.1.6, the exponential decay ensures that the difference between $\alpha \to 1$ and $\alpha = 1$ is very significant . Since every mode in the damped equation carries the temporal decay factor , the solutions of this equation always die down, whatever the IC, unlike in the undamped equation where the solutions keep oscillating forever.

We now consider the opposite limit $\alpha \to 0$. In this case the characteristic roots of (01) are

$$r_{+,-} = \frac{1}{2}\left(\frac{\alpha - 1}{\alpha} \pm \sqrt{\frac{(\alpha-1)^2}{\alpha^2} - \frac{4n^2\pi^2}{\alpha L^2}} \right) \quad .$$

As α gets smaller and smaller, the first term inside the surd dominates the second until a larger and larger value of n. For these n, both roots are real and represent exponential decay. Evidently, the root r_- with negative sign is huge negative – it's approximately $-1/\alpha$. The root r_+ with positive sign is more interesting since the $-1/\alpha$ outside the radical cancels the $1/\alpha$ coming from the leading term inside.

Let's evaluate this root by Taylor expanding the surd. We have

$$\left(\frac{1 - 2\alpha + \alpha^2}{\alpha^2} - \frac{4n^2\pi^2}{\alpha L^2} \right)^{1/2} = \frac{1}{\alpha}\left[1 - \left(2 + \frac{4n^2\pi^2}{L^2} \right)\alpha + ... \right]^{1/2} \quad ,$$

where we drop higher order terms in α. This then gives

$$\begin{aligned} r_+ &= \frac{1}{2}\left[\frac{\alpha - 1}{\alpha} + \frac{1 - \alpha - 2n^2\pi^2\alpha / L^2}{\alpha} \right] \quad . \\ &= -\frac{n^2\pi^2}{L^2} \end{aligned}$$

This is evidently the decay constant associated with the n^{th} spatial eigenfunction of the 1+1 heat equation. As α becomes smaller, more and more harmonics obey this decay equation to greater accuracy and the solution tends to that of the heat equation.

Just as in 2.1.6 Part (c), there is the issue that the equation with small α accommodates two ICs whereas the heat equation takes just one. As happened in that problem, a large initial velocity will cause increasing disparity between the two solutions. The limiting case is where the two equations are excited with a velocity IC alone – the second order equation will show some finite solution which decays to zero while the heat equation will show the trivial solution. Nevertheless, since the solutions of the equations with both zero and small α go to zero in time, that too with the same decay constants for at least the first few modes, the match between the finite and zero α equations is pretty decent. *****

COMMENTS : We now take on the extra part about the absence of a MPS for signal propagation in the damped wave equation. We have already seen how c as a propagation speed arises in the undamped equation – either from D'ALAMBERT's formula or from a trig identity on the eigenfunction expansion. So now let's see how the slightest difference in α can allow for instantaneous signal propagation.

Consider a simple IC – an initial delta function displacement in the middle of the interval and zero initial velocity. Ignoring questions of convergence we expand the delta in a Fourier series as

$$\delta\left(x-\frac{L}{2}\right)=\frac{2}{L}\sum_{n}\sin\frac{n\pi}{2}\sin\frac{n\pi x}{L} \quad . \tag{03}$$

For the regular wave equation, $\sin(n\pi x/L)$ would tack on $\cos\omega_n t$, where $\omega_n = n\pi/L$, to generate the delta waves which move left and right which speed 1. The important thing here is, EVERY component sine wave moves with speed 1.

For the damped wave equation however, $\sin(n\pi x/L)$ would tack on $\cos\omega_{dn} t$, where ω_{dn} is slightly different from ω_n. Using the trig identity, this means that each sine wave would travel at a slightly different speed. This might appear harmless – after all nothing is travelling at infinite speed. But the speed difference is enough to put paid to an MPS. We see this first qualitatively, then quantitatively, using the delta IC as example. Qualitatively, we know that the component sine waves of the delta add up to infinity at the 'centre of the waves' and zero everywhere else. When all waves move together, they always add up in the SAME PHASE, so they continue to give identically zero except at the centre, which moves with speed 1. When the component waves move at different speeds, their relative phases at a remote point start drifting immediately – now, it is no longer guaranteed that the waves in the new phases will still add up to zero. In fact they don't, as the quantitative analysis will now show.

Using (03), the solution of the proper wave equation with the delta IC is

$$u_1(x,t)=\frac{2}{L}\sum_{n}\sin\frac{n\pi}{2}\sin\frac{n\pi x}{L}\cos\frac{n\pi t}{L} \quad , \tag{04}$$

where we use u_1 and u_2 to denote the solutions of the undamped and damped wave equations. For the damped equation, we let $\gamma = 1-\alpha$ so that γ can be a small parameter. The solution of this equation is

$$u_2(x,t) = \frac{2}{L}\sum_n \sin\frac{n\pi}{2}\sin\frac{n\pi x}{L} e^{-\gamma t/2}\left(\cos\omega_{dn}t + \frac{\gamma}{2\omega_{dn}}\sin\omega_{dn}t\right) \quad , \qquad (05)$$

where ω_{dn} is as in (02b), and we have approximated γ/α as γ since α is almost 1.

Now we simplify the above solution using smallness of γ. Terms of order γ arise here from various sources – there is a γ in ω_{dn}, there is another γ in the exponential, and one more in the coefficient of the sin term. For MPS analysis, we are interested in extremely small times, so we Taylor expand the exponential as $e^{-\gamma t/2} = 1-\gamma t/2$. The actual source of the MPS violation is the γ in the ω_{dn} terms, so that γ is the key player in the whole game, the opposite of an ignorable quantity. It is sufficient however to consider $O(1)$ terms with the shifted phases – the $O(\gamma)$ and smaller-sized phase-shifted terms are negligible compared to this leading term. So we can drop the exponential and the sin term in (05) to write

$$u_2(x,t) = \frac{2}{L}\sum_n \sin\frac{n\pi}{2}\sin\frac{n\pi x}{L}\cos\omega_{dn}t \quad . \qquad (06)$$

Taylor expanding (02b) gives

$$\omega_{dn} = \frac{n\pi}{L}\left(1-\frac{\gamma^2L^2}{2n^2\pi^2}\right) \quad ,$$

and substituting this into (06) gives

$$u_2(x,t) = \frac{2}{L}\sum_n \sin\frac{n\pi}{2}\sin\frac{n\pi x}{L}\cos\left(\frac{n\pi}{L}\left[1-\frac{\gamma^2L^2}{2n^2\pi^2}\right]\right)t \quad . \qquad (07)$$

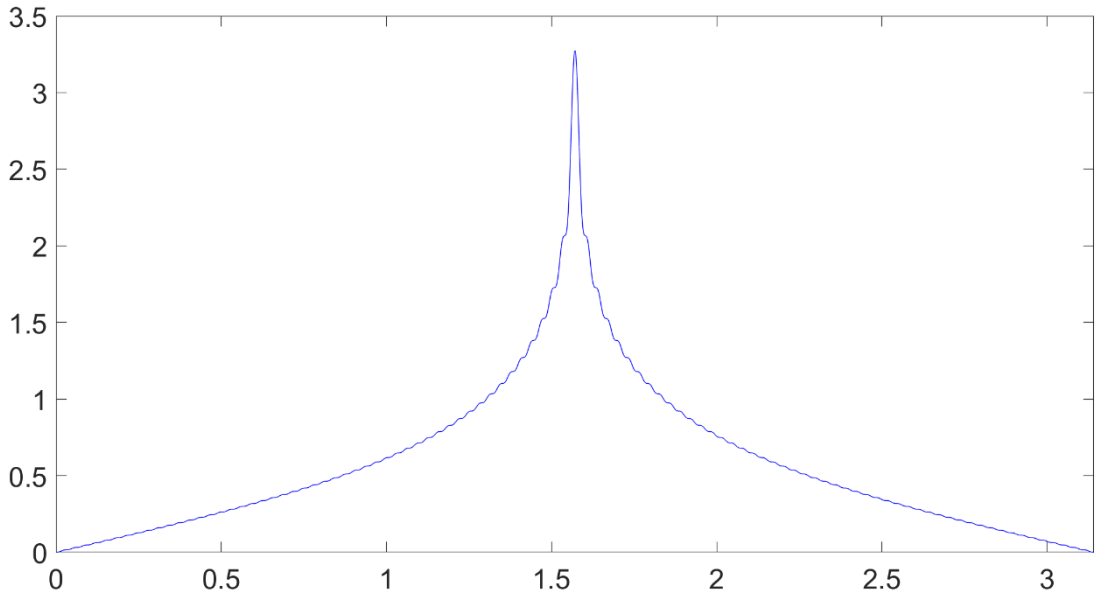
Now we compare between (04) and (07). The natural thing to compute is the difference $u_2(x,t)-u_1(x,t)$ at some extremely small time $t = \varepsilon$. Using the trig identity $\cos C - \cos D = 2\sin([C+D]/2)\sin([D-C]/2)$, that works out to

$$u_2(x,\varepsilon) - u_1(x,\varepsilon)$$
$$= \frac{4}{L}\sum_n \sin\frac{n\pi}{2}\sin\frac{n\pi x}{L}\sin\left(\frac{n\pi}{L}\left[1-\frac{\gamma^2L^2}{4n^2\pi^2}\right]\varepsilon\right)\sin\frac{\gamma^2L^2}{4n^2\pi^2}\varepsilon \quad .$$

We assume that ε is so small that $\sin(\ldots..\varepsilon)$ in the above can be treated as just $(\ldots..\varepsilon)$. Then, dropping γ^4 in comparison with γ^2, the above simplifies to

$$u_2 - u_1 = \frac{1}{\pi}\gamma^2\varepsilon^2 \sum_n \frac{1}{n} \sin\frac{n\pi}{2} \sin\frac{n\pi x}{L} \quad .$$

This is $\gamma^2\varepsilon^2$ times a Fourier series, which we sum using a computer, taking $L = \pi$. The result is below.



Clearly, this Fourier series is non-zero everywhere in the interval $[0, L]$. This implies that the difference $u_2 - u_1$ is not zero anywhere. Now, we know what $u_1(x,\varepsilon)$ is – from D'ALAMBERT's formula it is the sum of two deltas shifted by distance ε, so it is identically zero for all x outside $[L/2-\varepsilon, L/2+\varepsilon]$. Thus, at time ε, $u_2(x,t)$ is of order $\gamma^2\varepsilon^2$ and not zero everywhere inside $[0, L]$. Although the size of u_2 is infinitesimal, the concept of an MPS has been completely trashed – any immediate response, howsoever small, means that at least some information about the initial disturbance has spread instantaneously all over the interval.

Then is the damped wave equation a garbage ? Fortunately, no. A first order damping term is in any case usually a stand-in for more complicated processes such as the viscoelastic action of a string. Although we expect the string to have a strictly finite information propagation speed, and this model equation does have instant responses, from a practical point of view their size is very small (the leading term goes as γ^2). The big response will occur only when the bulk of the

disturbance reaches the point in question, and this will still happen at the wave speed. So a leading order correction to the motion of a realistic guitar or piano string arising from the damping will be given accurately by an equation of this form.

The damped wave equation will NOT work when we are dealing with signal transmission, and are interested in the time of arrival of the signals. In such a case we'll have to work with the actual wave equation, and then maybe throw in an attenuation factor onto the resulting solution. It is a well-known fact that the heat equation allows instantaneous information propagation – this is one of the drawbacks of the equation as a model of heat conduction or chemical diffusion. For example, if you set off the infinite rod with a delta in the middle, the temperature everywhere becomes finite in an instant. What this analysis shows is that the infinite MPS enters the picture the moment we introduce the first derivative term – the size of the acausal responses is proportional to the square of the size of this term. The thing to be careful of is that a heat equation with a wave-like term thrown in does NOT represent heat conduction with a finite wave speed. Indeed, this phenomenon is modelled using complicated nonlinear PDEs with integro-differential terms.

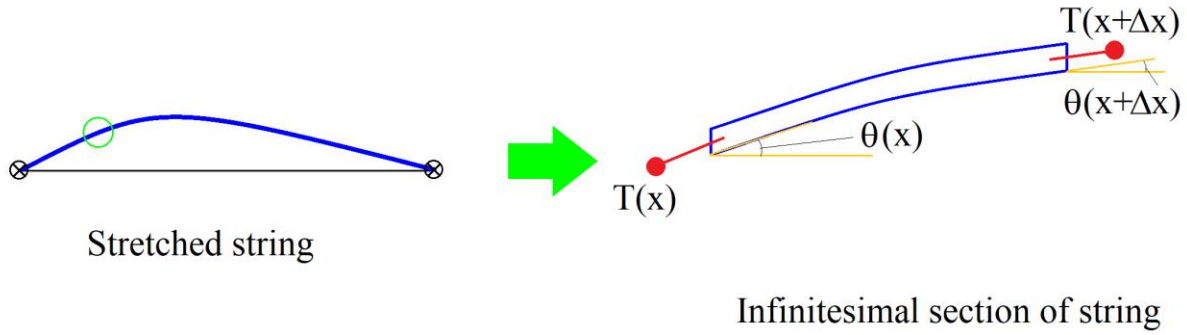
7.2

SCENARIOS INVOLVING WAVE EQUATION

Here we see some basic physical situations where the wave equation crops up. These show us the ubiquitousness of this equation and will help us to model physical systems using this equation. The actual modeling, we'll do in a later Unit (N°. 27).

STRETCHED STRING

The archetypal wave equation situation is a string which is clamped at the two ends and stretched to tautness. Let x be the position along the string and y be the perpendicular displacement of the string. It is of course a function of both space and time. Let $\theta(x)$ be the angle made by the tangent to the string at any point, $T(x)$ be the tension at any point and the constant ρ_1 its mass per unit length. Then, we consider an infinitesimal section of the string, as shown in the Figure, and do a force balance on it.



We have

$$\Delta m \frac{\partial^2 y}{\partial t^2} = T(x+\Delta x)\sin\theta(x+\Delta x) - T(x)\sin\theta(x) \quad ,$$

where Δm is the mass of the section under study. Expanding the RHS,

$$\Delta m \frac{\partial^2 y}{\partial t^2} = \frac{\Delta T}{\Delta x}\Delta x \sin\theta(x) + T(x)\frac{\Delta(\sin\theta(x))}{\Delta x}\Delta x \quad . \qquad (01)$$

The tensions also have an x-component – writing the force balance in that direction gives

$$\Delta m\frac{\partial^2 x}{\partial t^2} = T\left(x+\Delta x\right)\cos\theta\left(x+\Delta x\right) - T(x)\cos\theta(x)$$
$$= \frac{\Delta T}{\Delta x}\Delta x\cos\theta(x) + T(x)\frac{\Delta\left(\cos\theta(x)\right)}{\Delta x}\Delta x \quad .$$

By definition of a string, there is no motion in the x-direction hence

$$\frac{\Delta T}{\Delta x}\Delta x\cos\theta(x) + T(x)\frac{\Delta\left(\cos\theta(x)\right)}{\Delta x}\Delta x = 0 \quad . \qquad (02)$$

This does not look very promising until we make the **small displacement approximation** – we say that the slope of the string at every point is very small compared with unity. Then, $\sin\theta$ becomes $\Delta y/\Delta x$ while $\cos\theta$ becomes 1 plus negligible terms. Then, cancelling the common factor of Δx in (01) and replacing the Δ's by d's,

$$\rho_1\frac{\partial^2 y}{\partial t^2} = y\frac{\partial T}{\partial x} + T\frac{\partial^2 y}{\partial x^2} \quad .$$

On the other hand, the second term in (02) becomes $\partial/\partial x$ of unity, which is zero, implying that $\partial T/\partial x = 0$ i.e. ***T* is a constant**. Then, the first term goes out of the above RHS and we have

$$\frac{\partial^2 y}{\partial x^2} = \frac{\rho_1}{T}\frac{\partial^2 y}{\partial t^2} \quad ,$$

which is what we want to see. Note that the **speed of wave is $c = \sqrt{(T/\rho_1)}$**. The tension T is determined by how taut the string is made – in pianos for example, it is varied by tightening or loosening screws at either end. Such screw action is exactly what piano tuners do to bring the piano in tune.

In some problems, the string may be subjected to an external force. This force takes the form of a line load (force per unit length) $q(x,t)$ applied at every point on the string. In such a case, a term $q\Delta x$ gets added to the RHS of (01); carrying through the following steps yields

$$\frac{\partial^2 y}{\partial t^2} - \frac{T}{\rho_1}\frac{\partial^2 y}{\partial x^2} = \frac{1}{\rho_1}q\left(x,t\right) \quad ,$$

which is an inhomogeneous wave equation.

ELASTIC BAR

Consider a bar of uniform cross-section which can elongate or contract by small amounts along its length. Let x be the positional coordinate along the bar and u be the amount of extension (signed) as a function of space and time. We assume that the extension is small enough so that x can be measured relative to the undeformed bar. The bar has density ρ (usual 3-dimensional density), cross-sectional area A and YOUNG's modulus E. The **stress** σ is the force per unit cross-sectional area of the bar, while the **strain** ε is defined as $\partial u/\partial x$. The material property relation is $\sigma/\varepsilon = E$.

Now we focus on an infinitesimal element of the bar, located at position x and having a length Δx. The mass of the element is $\Delta m = \rho A \Delta x$ while the force is $A\left[\sigma\left(x+\Delta x\right)-\sigma(x)\right]$, which is $A(\Delta\sigma/\Delta x)\Delta x$. Writing NEWTON's law, cancelling the common factor of $A\Delta x$ and replacing Δ's by d's, we have

$$\rho \frac{\partial^2 u}{\partial t^2} = \frac{\partial \sigma}{\partial x} \quad .$$

Now apply stress/strain = YOUNG's modulus; this gives

$$\rho \frac{\partial^2 u}{\partial t^2} = E \frac{\partial \varepsilon}{\partial x} \quad .$$

Finally, ε is defined to be $\partial u/\partial x$ so

$$\frac{\partial^2 u}{\partial x^2} = \frac{\rho}{E} \frac{\partial^2 u}{\partial t^2} \quad .$$

Note that the **speed of wave is $c = \sqrt{(E/\rho)}$**.

As in the string, there might be an external line load $q(x,t)$ applied on the bar. This term adds on to NEWTON's law at the starting step of the derivation – continuing through the steps we find the inhomogeneous wave equation

$$\frac{\partial^2 u}{\partial t^2} - \frac{E}{\rho} \frac{\partial^2 u}{\partial x^2} = \frac{1}{\rho A} q\left(x,t\right) \quad .$$

There are two types of common and physically plausible BCs on the bar. One option is to **clamp an end to the wall**, which sets $u = 0$ at the end. The other option is to have the end **perfectly free**. Mathematically, this means that the end has no stress i.e. $\sigma = 0$ whereby $\varepsilon = 0$ and hence $\partial u/\partial x = 0$ at the end.

LIGHT AND SOUND

As we have already seen in L22, electromagnetic fields in free space satisfy the wave equation. These waves are light waves. The density fluctuations in a fluid also satisfy the wave equation. These waves are sound waves. We will do the derivation of this equation in a Tute. More often than not, light and sound waves propagate freely through space instead of being confined within a cavity. In such a case, a wave of the form $\cos(kx-\omega t)$ or $\sin(kx-\omega t)$ [or with y, z, r etc in place of x] is said to have **wavenumber** k and **wavelength** $2\pi/k$.

VIBRATION OF BEAMS

A beam (as opposed to a string) satisfies an equation similar to the wave equation but with reversed sign and higher order derivatives w.r.t x. Since we did not go into derivation even for the static case (N^{o}. 13), we will not do the derivation here also but present the equation straightaway :

$$EI\frac{\partial^4 y}{\partial x^4} = -\rho_1\frac{\partial^2 y}{\partial t^2} + q(x,t) \quad ,$$

where EI is the flexural rigidity and ρ_1 the mass per unit length of the beam. The BCs on this equation are the same as those for static beams – rigid support, movable support, pin, free end etc.

Example 7.2.1

Sound wave

Derive the equation of motion for a sound wave in a fluid in one spatial dimension.

[Read Only]

We consider a fluid which has to largest order zero flow velocity, constant density and constant pressure everywhere. The variable which satisfies the wave equation is a small fluctuation in density which is added on to this constant density. This fluctuation in turn sets up small pressure fluctuations and small velocities inside the fluid.

The equations involved are the continuity equation (L2301)

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho v) = 0 \quad , \tag{01}$$

and EULER's equation (L2302)

$$\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial x} = -\frac{1}{\rho}\nabla P \quad , \tag{02}$$

both of them modified suitably to account for the single spatial dimension involved. Finally, we need a third equation. This is a material property connecting fluctuations in P about the mean value P_0 to fluctuations in ρ about the mean value ρ_0. It is called the adiabatic bulk modulus K and is defined as

$$K = \rho_0 \frac{\Delta P}{\Delta \rho} \quad . \tag{03}$$

These are the basic three equations from which we will get the wave equation.

To start off we set $\rho = \rho_0 + \Delta\rho$ into (01). That gives

$$\frac{\partial}{\partial t}(\rho_0 + \Delta\rho) + \frac{\partial}{\partial x}(\rho_0 v + \Delta\rho v) = 0 \quad .$$

In this expression, ρ_0 is a constant so it has no spatial or temporal derivatives. The term $\Delta\rho v$ is negligible since it is the product of two small terms, and can be dropped. Implementing these we have

$$\frac{\partial}{\partial t}\Delta\rho + \rho_0 \frac{\partial v}{\partial x} = 0 \quad . \tag{04}$$

In a like manner, we develop (02). We have

$$(\rho_0 + \Delta\rho)\left(\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial x}\right) = -\frac{\partial}{\partial x}(P_0 + \Delta P) \quad .$$

Again, ρ_0 and P_0 are constants while terms of second and higher order in $\Delta\rho$ and v are negligible. Then, we substitute (03), writing $(K/\rho_0)\Delta\rho$ in place of ΔP. Then what we get is

$$\rho_0 \frac{\partial v}{\partial t} = -\frac{K}{\rho_0} \frac{\partial}{\partial x} \Delta\rho \quad , \qquad (05)$$

and the equation is almost ready.

For the final step we take a $\partial/\partial t$ of (04). That produces a mixed derivative term $\partial^2 v / \partial x \partial t$, which we write as $\partial/\partial x$ of $\partial v/\partial t$ and then substitute in (05). This gives

$$\frac{\partial^2 \Delta\rho}{\partial x^2} = \frac{\rho_0}{K} \frac{\partial^2 \Delta\rho}{\partial t^2} \quad ,$$

which is the wave equation. Note that the wave velocity is $c = \sqrt{(K/\rho_0)}$. At this step, if we had taken $\partial/\partial x$ instead of $\partial/\partial t$ on (04), then we would have got an identical equation for v.

As often happens, in higher dimensions the equation for $\Delta\rho$ comes out to be the same with $\partial^2\Delta\rho/\partial x^2$ getting replaced by $\nabla^2\Delta\rho$. Likewise, every Cartesian component of the velocity vector $\mathbf{v}$ satisfies the many+1 wave equation with the same wave speed $\sqrt{(K/\rho_0)}$. For air at 20 °C and sea level, the speed of sound works out to about 1235 km/hr. It's an interesting fact that the speed of sound was something which Sir ISAAC NEWTON had calculated WRONGLY – he had assumed sound transmission in air to be isothermal while actually it is adiabatic. This error caused the answer to come out 15 percent short of true, and was corrected by LAPLACE. Actually, NEWTON's mistakes didn't stop there. Having found the wrong answer, he included in the Principia extensive details of a fake experiment which purportedly brought out the truth behind his calculation. The founder of the scientific practice was thus also the originator of scientific malpractice.

For air at typical aircraft cruising altitudes however, the sonic speed is 1060 km/hr, very close to NEWTON's false value. This speed is typically called Mach 1, and the ratio of the aircraft speed to the sound speed is called Mach number. Aircraft cruising speeds are often specified in terms of Mach number; unless explicitly stated otherwise, 1060 km/hr is the relevant conversion factor.

Example 7.2.2

Baby and big wave equations

An elastic bar is discretized as a line (chain) of N identical masses m connected to each other by identical springs of constant k, as in T035. One says that the chain becomes the bar in the limit that the number of masses tends to infinity.

(a) As N increases, how should m and k be scaled so as to get a physically plausible result ? Justify your answer.

(b) Using suitable numerical work, present a comparison between the solutions of the discrete system for various N and the known solution of the wave equation for the continuous bar.

[Level : 2]

(a) To write the equations of motion for the masses, we copy shamelessly from T035. Part (a) there has what we would now call double zero BCs – here the BC is not specified explicitly (because it doesn't matter – surely the way we have to scale the system cannot depend on what's happening at the two ends). So let's go with this most basic BC. Let N denote the number of significant masses and let two extra masses #0 and #N+1 correspond to the end points, as in T062. Let $\mathbf{u}$ denote the vector of displacements of the masses from their equilibrium positions (it was $\mathbf{x}$ in T035, but here x is the position along the bar). Then we have, taking $N=4$ for display purposes,

$$m\frac{\mathrm{d}^2}{\mathrm{d}t^2}\begin{bmatrix}u_1\\u_2\\u_3\\u_4\end{bmatrix}=k\begin{bmatrix}-2&1&0&0\\1&-2&1&0\\0&1&-2&1\\0&0&1&-2\end{bmatrix}\begin{bmatrix}u_1\\u_2\\u_3\\u_4\end{bmatrix}\quad\text{, or}$$

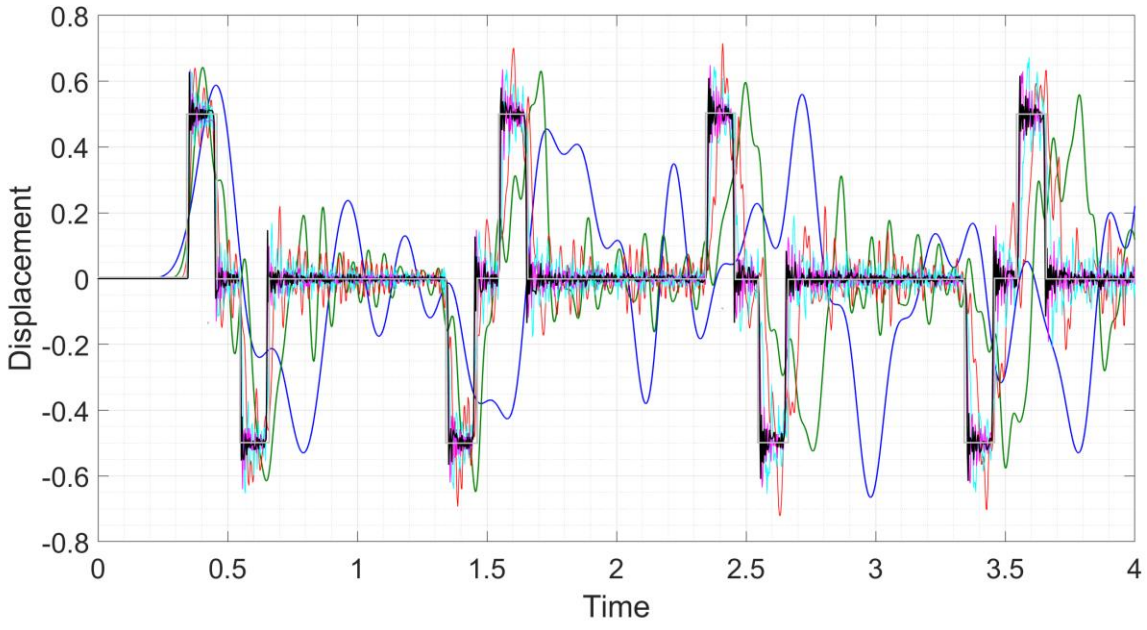
$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\begin{bmatrix}u_1\\u_2\\u_3\\u_4\end{bmatrix}=\frac{k}{m}\begin{bmatrix}-2&1&0&0\\1&-2&1&0\\0&1&-2&1\\0&0&1&-2\end{bmatrix}\begin{bmatrix}u_1\\u_2\\u_3\\u_4\end{bmatrix}\quad. \qquad (01)$$

The scaling on m is quite easy. Physical plausibility suggests that the total mass of the bar must be a constant; if we break it up into more and more point masses, then each mass must become smaller. This implies that m varies as $1/N$.

The scaling on k is more difficult since the spring constant has no such obvious physical interpretation. We get it by looking at the limiting behaviour. In the limit, (01) has to reduce to $\ddot{u}=c^2u''$ where c^2 is a finite number. Evidently, the matrix on the RHS has to become a constant times the discrete second-derivative matrix $\mathbf{Q}$ of T062 (note that $\mathbf{Q}$ of that Tute features the same BCs as here). Indeed, the entries of this matrix are exactly those of $\mathbf{Q}$; what is needed is the factor of $1/h^2$, which is N^2, which sits outside. This factor must come from k/m; since the $1/m$ contributes an N, the k must contribute another N. Hence k scales as N.

(b) This question appears rather vague – indeed, it is intentionally open-ended, being more a project type of question than an exam problem. What it expects is that you will spend some time thinking about how to set up a scheme which can demonstrate the desired comparison, and then play around with the computer to arrive at a set of plots which can do it effectively. As such, this question does not have a unique solution (like certain BVPs), and I am by no means claiming that my solution below is the best one.

What I have done is take $c = 1$, take the bar to occupy the interval x belongs to $[0, 1]$, clamp both ends to $u = 0$ and then have a position-only IC which is unity in the interval $[0.45, 0.55]$ and zero everywhere else. In the discrete versions, I have assigned unit displacement to all masses which lie between 0.45 and 0.55. Then, the quantity I have plotted is $u(0.9,t)$ as a function of time. I found that this was clearer than showing profiles of u vs. x at different t or u vs. t at different x – both of these tended to become messy unless I included a huge number of pictures. In the below Figure I show $u(0.9,t)$ vs. t for N taking the values 20, 60, 200, 600, 2000 and 6000. The successive time traces are in blue, green, red, cyan, magenta and black respectively. In particular, I have highlighted the first two and the last time traces ($N = 20,\ 60$ and 6000) by putting them in thicker lines.



Finally, I have overlaid on the plot a hand-drawing of the analytical solution for the bar in a silver line. This line is slightly shifted from true in some places so that

it does not overlap and blot out the other lines. (The plot benefits a lot from zooming in, so I am putting up the Figure file online at www.shayak.in/Shayakpapers/DELTA/T117_01.png.)

We know from D'ALAMBERT's formula that the solution for the elastic bar features u at first sitting stationary, then jumping up to 0.5 when the wave arrives, then coming back down, then jumping down to -0.5 almost immediately as the reflected wave arrives, again coming back to zero, and so on. What we can see is that the blue line (20 point masses) is nothing like that at all while the black line (6000 masses) shows a very similar profile (with GIBBS' phenomenon due to convergence issues with the Fourier series at a jump). Indeed, the blue and green lines are not even periodic – they are quasiperiodic as we would expect from the results of T035 Part (b). As the number of masses increases, the plot becomes more and more regular, with the non-periodicity being minimal in the black trace.

Another interesting feature of the plot is the time when the mass first starts to move. In the continuous limit, it should show no motion at all until the wave arrives. In the discrete versions, that won't happen – the n^{th} mode does not have a frequency proportional to n so the different component waves will move at different speeds, leading to small but instantaneous responses as in T115. Indeed, the jump up from zero to finite displacement becomes sharper and sharper as the number of masses increases.

*

COMMENTS : Like T082, this is another example where the solutions of the discretized system converge to the known continuous solution as the grid gets finer. Evidently, it demonstrates again how a PDE is the limit of a system of ODEs. The most surprising feature is the number of discrete masses which was required to get a good solution. For baby heat equations, $N = 4$ was already reasonable, while $N = 50$ was indistinguishable from the PDE solution. For baby wave on the other hand, $N = 60$ is lousy, the hundreds (red, cyan) are bad, and even 6000 leaves something to be desired – after all, the fluctuations are quite large (GIBBS' phenomenon only accounts for a single overshoot right at the discontinuity, not the others). Why this difference ?

To go into this issue, we look at the eigenvalues and vectors (normal modes) of $\mathbf{Q}$. We had already calculated these in T062. The vectors are

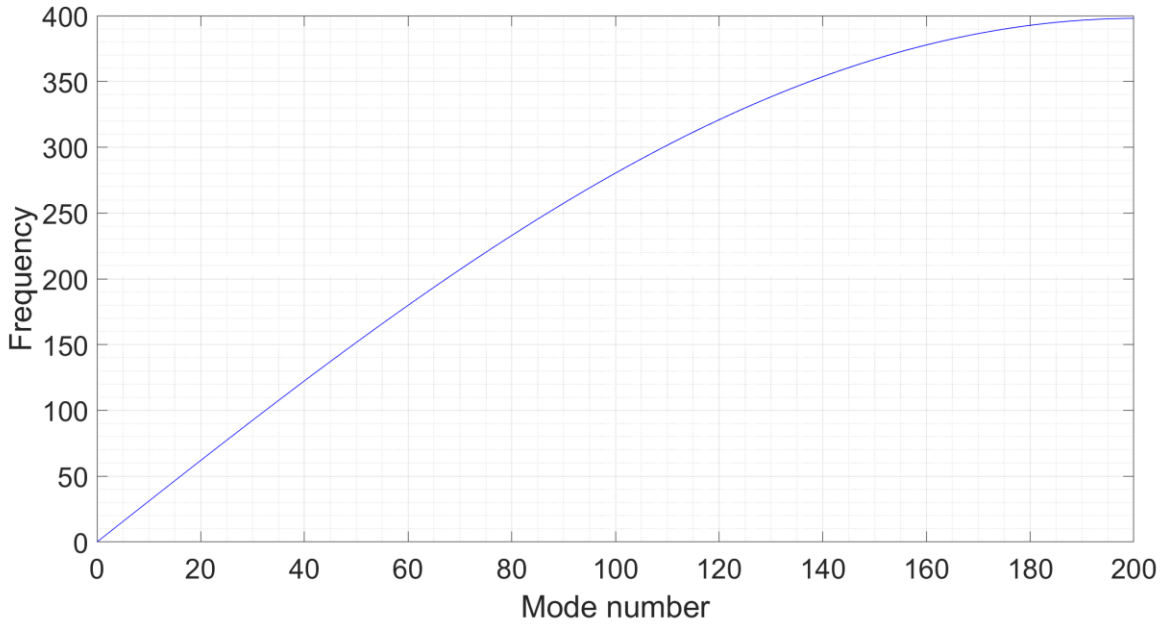
$$\mathbf{v}_n = \left[\sin n\pi h;\quad \sin 2n\pi h;\quad .....;\quad \sin Nn\pi h\right] \quad ,$$

where n is an integer between 1 and N. These are the discrete forms of their continuous counterparts and pose no problem. The eigenvalues on the other hand are

$$\lambda_n = \frac{2}{h^2}\left(\cos n\pi h - 1\right) \quad ,$$

which reduces to $-n^2\pi^2$ in the limit $nh << 1$. The eigenfrequencies are the square roots of their negatives – in the continuous case ($h \to 0$) they turn into $n\pi$ just as they should. But in the discrete case how much is the difference ?

To see this, we prepare a plot of the n^{th} eigenfrequency vs n for $N = 200$.



The limiting case would of course show a straight line plot. Here, things are good upto a while but then the curve really starts to flatten out. If these higher modes are present in the motion in a significant amount, then we will get markedly different wave speeds and a messy quasiperiodic time trace instead of a clean, periodic one. That's exactly what is happening here – the IC has a Fourier coefficient which goes as $1/n$ so a lot of modes are present, and the high order modes move totally out-of-sync with the lower ones. Only when we go to tens

of thousands of N do all the significant modes lie within the linear region and produce an acceptable plot.

In the heat equation, the eigenvalues of $\mathbf{Q}$ are of course the same. The big difference however is in their interpretation – there they are exponential decay constants. The higher the mode number, the greater the decay rate. So if you take an IC with a high number of modes (which we DID do in T082 – sawthooth wave also has $1/n$ coefficients), the higher ones just get killed off at a rate different from what the continuous limit would suggest, but they nevertheless die out very quickly. Then the lower ones evolve as they would in the continuous case. At very small times, there will indeed be significant differences between the discrete and continuous solutions on account of the approximation, but then the offending modes will disappear below machine precision. Here, the eigenvalues are oscillation frequencies so nothing gets killed off. If you have intractable modes to start with, then they remain on board throughout and produce an inaccurate solution. So you have to go for an N which does not contain these modes in the first place.

This difference in discretization accuracy between heat and wave equations is something we must keep in mind when solving PDEs on a computer. Indeed, stuff like this is what makes numerical solution of PDE an entirely non-trivial task. I won't go into more details here, but if you are interested, please consult any book on finite element methods.

Example 7.2.3

Vibrating cantilever

The deflection $y(x,t)$ of a cantilever beam satisfies the PDE

$$EI\frac{\partial^4 y}{\partial x^4} = -\rho_1 \frac{\partial^2 y}{\partial t^2} \quad ,$$

defined in x belongs to $[0, L]$ with BCs $y(0) = y'(0) = y''(L) = y'''(L) = 0$.

(a) Obtain a transcendental equation for the eigenfrequencies of the cantilever (this equation cannot be solved in closed form).

(b) When the eigenfrequency is sufficiently high, the above equation can be solved approximately. Use this fact to determine the eigenfrequencies of the high order modes of the cantilever.

(c) Would you expect the spatial eigenfunctions of the cantilever to be orthogonal ? Justify your answer.

[Level : 3]

(a) How do we start an unfamiliar PDE problem ? We separate variables. Writing $y = X(x)T(t)$, plugging, dividing by XT and invoking the sep var argument gives

$$\frac{EI}{\rho_1}\frac{X^{(4)}}{X} = -\frac{\ddot{T}}{T} = \lambda \quad .$$

For oscillatory solutions, $\lambda = \omega^2$ (the T equation has a minus sign so the eigenvalues are positive) where ω is the oscillation eigenfrequency which is to be determined. The cases $\lambda < 0$ and $\lambda = 0$ give rise to unphysical temporal behaviour and can be rejected. Then, the X equation becomes

$$X^{(4)} - \frac{\omega^2 \rho_1}{EI} X = 0 \quad .$$

We have already seen how to solve this fourth order ODE in T026 part (a); letting $\beta = (\rho_1/EI)^{1/4}\omega^{1/2}$, the general solution is

$$X_G = C_1 \cos\beta x + C_2 \sin\beta x + C_3 \cosh\beta x + C_4 \sinh\beta x \quad .$$

The BCs are evidently $X(0) = X'(0) = X''(L) = X'''(L) = 0$. To prepare for the BC fit we compute the three derivatives of X_G. They are

$$X_G' = -\beta C_1 \sin\beta x + \beta C_2 \cos\beta x + \beta C_3 \sinh\beta x + \beta C_4 \cosh\beta x \quad ,$$
$$X_G'' = -\beta^2 C_1 \cos\beta x - \beta^2 C_2 \sin\beta x + \beta^2 C_3 \cosh\beta x + \beta^2 C_4 \sinh\beta x \quad ,$$
$$X_G''' = \beta^3 C_1 \sin\beta x - \beta^3 C_2 \cos\beta x + \beta^3 C_3 \sinh\beta x + \beta^3 C_4 \cosh\beta x \quad .$$

Setting $x = 0$ into X_G and X_G' gives $C_1 = -C_3$ and $C_2 = -C_4$. Setting $x = L$ into X_G'' and X_G''' and then using these relations gives

$$\left(-\beta^2 \cos\beta L - \beta^2 \cosh\beta L\right)C_1 + \left(-\beta^2 \sin\beta L - \beta^2 \sinh\beta L\right)C_2 = 0 \quad ,$$
$$\left(\beta^3 \sin\beta L - \beta^3 \sinh\beta L\right)C_1 + \left(-\beta^3 \cos\beta L - \beta^3 \cosh\beta L\right)C_2 = 0 \quad .$$

Cancelling the common factors of β^2 and β^3 from the first and second equation above, we get a matrix equation of the type $\mathbf{MC} = 0$ where $\mathbf{C} = [C_1;\ C_2]$. Either $\mathbf{C} = 0$ or $\mathbf{M}$ is non-invertible. The first case means $C_1 = C_2 = C_3 = C_4 = 0$ and represents the trivial solution for the cantilever. Hence the relevant condition is $\mathbf{M}$ is non-invertible or $\det \mathbf{M} = 0$. This yields

$$\det\begin{bmatrix} -\cos\beta L-\cosh\beta L & -\sin\beta L-\sinh\beta L \\ \sin\beta L-\sinh\beta L & -\cos\beta L-\cosh\beta L \end{bmatrix}=0 \quad ,$$

or, expanding out the determinant and simplifying the obvious trig,

$$\cosh^2\beta L-\sinh^2\beta L+2\cos\beta L\cosh\beta L+1=0 \quad .$$

Now, the $\cosh^2$–$\sinh^2$ can be simplified further; either we know the relevant hyperbolic identity or express the hyperbolics in terms of exponentials to get $\cosh^2\beta L-\sinh^2\beta L=1$. This brings the condition down to

$$\cos\beta L\cosh\beta L=-1 \quad .$$

Finally since $\beta=(\rho_1/EI)^{1/4}\omega^{1/2}$, we get

* $$\cos\left(\left[\frac{\rho_1}{EI}\right]^{1/4}L\omega^{1/2}\right)\cosh\left(\left[\frac{\rho_1}{EI}\right]^{1/4}L\omega^{1/2}\right)=-1 \quad , \qquad (01)^*$$

which is the transcendental equation asked for.

(b) The question says that the eigenfrequency should be high enough for the transcendental equation (01) to be solvable. Evidently, this means that the thing inside the cos and cosh should far exceed unity i.e. $(\rho_1/EI)^{1/2}L^2\omega >> 1$. When this holds true, the cosh is gigantic – it is exponential of something much greater than 1 and is like infinity. Hence, to make $\cos\beta L\cosh\beta L$ equal -1, the cos must be practically zero i.e. βL must equal $m\pi/2$ where m is an odd number (technically, βL should be a hair's breadth away from this value, in such a direction as to make the cos negative). Letting $m=2n+1$ for natural number n and expanding this out, we get the eigenfrequency

$$\omega=\frac{(2n+1)^2\pi^2}{4L^2}\left(\frac{EI}{\rho_1}\right)^{1/2} \quad .$$

Finally, since the frequencies involved are large, n is large also. We can drop the '1' from $2n+1$ and write

* $$\omega=\frac{n^2\pi^2}{L^2}\left(\frac{EI}{\rho_1}\right)^{1/2} \quad , \qquad *$$

which is the desired answer.

This is quite different from the case of the string – there the n^{th} eigenfrequency goes as n but here it goes as n^2. Some features are common though between the

two situations – the shorter, lighter and stiffer the object (stiffness is tension for a string and flexural rigidity for a beam), the higher its oscillation frequencies. Note that the higher eigenfrequencies here are not integer multiples of the lowest one, so the vibrations of a beam released from an arbitrary IC will in general NOT be periodic in time.

(c) The spatial eigenfunctions don't look at all easy. For one we have to solve the matrix equation for non-trivial C_1 and C_2; then, everything depends on some transcendental equation whose solution we don't know exactly. For orthonormality, an approximate eigenfrequency might just not be good enough – a brute force evaluation of the integral will quite possibly leave residual terms of the size of the errors allowed. However, we have already seen once how to find orthogonality in a non-trivial case in T083 – look at the equation and not the functions.

Just as in there, we discretize the problem using matrix method and see whether the matrix involved is symmetric. The discretized equation will read as

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{y} = -\frac{EI}{\rho_1}\mathbf{S}\mathbf{y} \quad ,$$

where $\mathbf{y}$ is the vector of displacements and $\mathbf{S}$ is the fourth derivative matrix with the appropriate BCs. Constructing this matrix is also an old hat – we did just that in T067. What we found there was (using $N=6$ for display)

$$\mathbf{S} = \frac{1}{h^4}\begin{bmatrix} 6 & -4 & 1 & 0 & 0 & 0 \\ -4 & 6 & -4 & 1 & 0 & 0 \\ 1 & -4 & 6 & -4 & 1 & 0 \\ 0 & 1 & -4 & 6 & -4 & 1 \\ 0 & 0 & 1 & -4 & 5 & -2 \\ 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix} .$$

Clearly, this matrix is symmetric so its eigenvectors are orthogonal. Since the eigenfunctions of the cantilever are the limit of these vectors as $h \to 0$, we conclude that they are orthogonal also.

7.3

ELEMENTARY QUANTUM MECHANICS

In this Unit we will refer to the fundamental imaginary entity as i *rather than* j *out of deference to the physicists who conceived and developed this field.*

Quantum mechanics says that a particle is not an object with a definite position. Rather, it can be represented as a wave-function which is a complex-valued function of space and time. This function is denoted by ψ. ψ is fundamentally complex in the sense that its real and imaginary parts have no separate existence or significance (unlike say the complex quantities in the induction motor space phasor model). We will follow the customary notation (of this field) to denote complex quantities without boldface, bars or any special symbols. The evolution of ψ is governed by a PDE due to ERWIN SCHROEDINGER. In 1+1 dimensions, the Schroedinger equation for $\psi(x,t)$ for a particle of mass m is

* $$-\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2}+V(x)\psi=\mathrm{i}\hbar\frac{\partial\psi}{\partial t} \quad , \qquad \textbf{(01)*}$$

where $\hbar$ (pronounced "h-bar") equals $1.055\text{x}10^{-34}$ SI Units and is a fundamental constant of Nature while $V(x)$ is the potential which is externally applied (say through an electric field). As usual, in higher dimensions, the $\partial^2/\partial x^2$ gets replaced by Laplacian and $V(x)$ becomes $V(\mathbf{r})$.

Although Schroedinger equation might look like heat equation rather than wave equation with a second derivative on space and first derivative on time, the i on the $\partial/\partial t$ means that its solutions are closer to those of wave equation. The physical interpretation of the wave-function is that the probability of finding the particle between point x and $x+\mathrm{d}x$ is $|\psi|^2\mathrm{d}x$. This immediately imposes a constraint on ψ. Since the total probability of finding the particle somewhere or other in space is unity, we have $\int_{-\infty}^{\infty}|\psi|^2\,\mathrm{d}x=1$. If the particle is confined within a particular region of space, say the interval [0, L] then the integration must run over that interval. This requirement is generally called

normalization of the wave-function. This physical interpretation of ψ means that it is **continuous everywhere**, which usually acts as BCs on quantum mechanical problems.

If we separate variables on (01), we get

$$\frac{1}{X}\left(-\frac{\hbar^2}{2m}X''+V(x)X\right)=\mathrm{i}\hbar\frac{\dot{T}}{T}=\lambda \quad .$$

Now, it so happens that the separation constant λ is actually the **energy** of the particle, E. Thus, the only permissible energies of the particle are actually the eigenvalues E_n of the BVP for X; the corresponding time-evolutions are

$$T_n(t)=\mathrm{e}^{-\mathrm{i}(E_n)t/\hbar} \quad .$$

This is of course why this subject is called quantum mechanics – the energy is quantized and can't take any desired value. To find the allowed energies, we focus only on the spatial part of the equation. Thus, the energies are the eigenvalues of the **time-independent Schroedinger equation**

$$\left(-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2}+V(x)\right)\psi=E\psi \quad , \qquad \textbf{(02)*}$$

together with the suitable BCs. It is an amateurish error however to label (02) as the actual Schroedinger equation – that is a description of the spatiotemporal dynamics of the wave-function and not just a prescription for finding energies.

The most counter-intuitive (at least for me) part of quantum mechanics comes now. With any classical system (such as one governed by heat, Laplace and wave equations) we are given ICs and know exactly what our variable is going to look like at any point in space and time. With quantum mechanics however, that is not the case. At any time, the particle can be in any arbitrary linear combination of the allowed eigenfunctions (called eigenstates here), and we don't know which one it is. If we try to measure, we end up tampering with the wave-function. Physically observable quantities such as position and momentum, which are numerical values in classical theory, are **operators** here. Thus, position is represented by the operator x (i.e. the position operator acting on a function is x times that function), momentum is represented by $-\mathrm{i}\hbar\left(\partial/\partial x\right)$ and so on. If we attempt a measurement of any of these quantities

then the particle instantly goes into an eigenstate of the associated operator and starts evolving afresh after the measurement is over.

For instance, the eigenfunction of the position operator is a delta function $\delta(x-x')$. The eigenfunction of the momentum operator is a plane wave e^{ikx} for any real k. So suppose we have a particle in a potential and we try to measure momentum. Then, before the measurement we don't know what the wave-function is; the measurement itself yields a certain value (k), but it also converts the wave-function to a plane wave having that particular k.

In that case, can the measurement of any observable yield any value whatever and we have no idea of what it's going to be ? No, it's not that hopeless either. The average value (or **expectation value**) of an operator O is defined as

$$\langle O \rangle = \int_{-\infty}^{\infty} \psi^* O \psi \, dx \quad .$$

This expectation value is over ensemble and not over time. This means that if you have say one lakh identically prepared systems and you measure O on each of them, then the average over all those systems will be $\langle O \rangle$. It does not mean that if you measure O on the same system one lakh times in succession then you will get that average value – the first measurement itself will tamper with the wave-function.

In connection with expectation values, we have the famous **uncertainty principle** (due to WERNER HEISENBERG). For this, we need the concept of standard deviation of an observable. This is defined as $\Delta O = \langle O^2 \rangle - \langle O \rangle^2$. Then, the uncertainty principle states that if Δx be the standard deviation of the position measured over a large number of systems and if Δp be the standard deviation of the momentum, then $\Delta x \Delta p \geq \hbar / 2$. The name comes from the fact that standard deviation represents the uncertainty in measurement – the greater the deviation, the less precise the measurement. As it turns out, position and momentum are not the only quantities which satisfy this principle – there are other pairs also which we will not go into. Fortunately, the small size of $\hbar$ (10^{-34} SI Units) means that the uncertainty product for a classical object is tiny, and the principle consequently moot – if we can measure both position and momentum of say a train upto 10^{-4} SI Unit

precision (let alone the 10^{-17} which Herr WERNER permits us), then we are more than happy.

As in N^{o}. 14, the treatment in this Unit will be somewhat skin deep. Two factors are responsible for this. The part of quantum mechanics which deals explicitly with solving PDEs runs into special functions very quickly. There is another part of quantum mechanics which is concerned with things like two-state systems – there the equations are simpler but the physical context is too involved for this book. Nevertheless, quantum mechanics features a huge amount of differential equations and linear algebra, so even if this particular Unit is brief, the things you have learnt in this course as a whole will certainly be of help in studying this subject.

Example 7.3.1

Easy energy eigs

(a) A particle of mass m is confined in the one-dimensional box x belongs to $[0, L]$ by a potential well which is zero inside the interval but infinity outside, preventing the particle from exiting the region. Thus $\psi = 0$ everywhere outside the interval, and by continuity at the ends of the interval. Determine the allowed energies.

(b) A particle of mass m is confined inside the two-dimensional box x belongs to $[0, L_1]$ and y belongs to $[0, L_2]$ by a potential well which is zero inside the region and infinity outside. Determine the allowed energies.

[Level : 1]

(a) Since $V=0$ inside the well, we have

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2}=E\psi \quad ,$$

$$\psi"+\frac{2mE}{\hbar^2}\psi=0 \quad ,$$

with the double zero BCs $\psi(0)=\psi(L)=0$. Note how the continuity of ψ brings about the BCs on the problem. By now you have seen this BVP so many times it probably makes you puke; letting λ be $2mE/\hbar^2$, the eigenvalues are $\lambda=n^2\pi^2/L^2$. This immediately gives

$$* \qquad E=\frac{\hbar^2 n^2\pi^2}{2mL^2} \quad , \qquad *$$

which is the desired answer. The corresponding eigenfunctions are $\sin(n\pi x/L)$ although nobody has asked for them.

(b) Equation :

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}\right)=E\psi \quad .$$

Sep var :

$$\underset{=\lambda_1}{\frac{X_{xx}}{X}}+\underset{=\lambda_2}{\frac{Y_{yy}}{Y}}=-\frac{2mE}{\hbar^2} \quad .$$

BC fit :

$$X=\sin\frac{n_1\pi x}{L_1} \qquad \lambda_1=-\frac{n_1^2\pi^2}{L_1^2} \quad ,$$

$$Y=\sin\frac{n_2\pi y}{L_2} \qquad \lambda_2=-\frac{n_2^2\pi^2}{L_2^2} \quad .$$

Solution :

$$* \qquad E=\frac{\hbar^2\pi^2}{2m}\left(\frac{n_1^2}{L_1^2}+\frac{n_2^2}{L_2^2}\right) \quad . \qquad *$$

We are done.

*

COMMENTS : This problem was, for us, a triviality. That was ONLY because we are by now differential equations experts and can bat away a simple problem like this, whatever the context in which it might appear. By contrast, the 1D box is a very standard problem in an introductory quantum mechanics course – GRIFFITHS for example spends more than two pages on it (if we include the proof, left as an exercise to the reader, of the fact that the cases $\lambda < 0$ and $\lambda = 0$ produce the trivial solution). The 2D box is also very standard fare on the mid-semester or final examinations of such a course, and is not necessarily considered as marks on a platter by either teacher or learner. So these examples show you why it is a good thing to master PDEs in general before going on to hardcore physics courses – you'll find the maths a breeze.

Example 7.3.2

Quantum harmonic oscillator

A quantum particle of mass m satisfies the 1+1 Schroedinger equation with the potential energy given by $V(x)=(1/2)m\omega^2x^2$ where ω is a constant.

(a) Write a BVP whose eigenvalues are the allowed energies.

(b) It is given that the energies are $E_n=\left(n-1/2\right)\hbar\omega$. Basis this result, what important conclusion can be drawn regarding the eigenfunction expansion ?

A particular solution of the 1+1 Schroedinger equation with the above potential is

$$\psi\left(x,t\right)=\left(\frac{m\omega}{\pi\hbar}\right)^{1/4}\exp\left(-\frac{m\omega}{2\hbar}\left(x-A\cos\omega t\right)^2\right)\exp\mathrm{i}\theta(t) \quad , \qquad (*)$$

where A is a parameter and $\theta(t)$ is some function.

(c) Using the probability distribution function, explain the physical significance of the solution presented above.

[Level : 2]

(a) The given potential energy corresponds to the harmonic oscillator because the potential energy of a classical spring-mass system is $(1/2)m\omega^2x^2$. To get the BVP for the energies we plug V into the time-independent Schroedinger equation (L2602) :

$$* \qquad \frac{\hbar^2}{2m}\frac{\mathrm{d}^2\psi}{\mathrm{d}x^2} - \frac{1}{2}m\omega^2x^2\psi + E\psi = 0 \quad . \qquad (01)*$$

This is of course the ODE. To complete the BVP we need the BCs. Evidently, the potential energy is infinity at $x = \pm\infty$ so the wave-function must be zero there. Hence $\underline{\psi(\infty) = \psi(-\infty) = 0}$. The ODE together with the BCs make up the BVP asked for.

(b) The eigenfunction expansion of 1+1 Schroedinger equation is

$$\psi(x,t) = \sum_n C_n X_n(x)\exp\mathrm{i}\left(-\frac{E_n t}{\hbar}\right) \quad .$$

We are given the energy eigenvalues but not the eigenfunctions. So we plug in the values while keeping the functions as unknown X_n's, getting

$$\psi(x,t) = C_1X_1(x)\exp\mathrm{i}\left(-\frac{\omega t}{2}\right) + C_2X_2(x)\exp\mathrm{i}\left(-\frac{3\omega t}{2}\right) + C_3X_3(x)\exp\mathrm{i}\left(-\frac{5\omega t}{2}\right) + \ldots\ldots \quad .$$

The complex exponentials here are periodic – the first one has fundamental period $4\pi/\omega$, the next one has fundamental period $4\pi/3\omega$ and so on. But now, just as for the wave equation in a finite domain, the second and higher eigenfrequencies are exact integer multiples of the first one, so the eigenfunction expansion as a whole is <u>temporally periodic with period $4\pi/\omega$</u>.

(c) We first calculate the probability distribution $|\psi|^2$. The mod squared of $\mathrm{e}^{\mathrm{i}\theta(t)}$ is just unity so that factor is ignorable. The rest is real, so mod squared means just plain squared, and we have

$$|\psi|^2 = \left(\frac{m\omega}{\pi\hbar}\right)^{1/2}\exp\left(-\frac{m\omega}{\hbar}(x - A\cos\omega t)^2\right) \quad .$$

This is a Gaussian function whose centre is at $x = A\cos\omega t$ and whose width is constant. Thus the probability distribution function remains unchanged in shape

while its centre i.e. the position of its maximum oscillates with amplitude A and frequency ω. This is a physically significant solution since oscillation with constant amplitude and frequency ω is exactly what a classical harmonic oscillator does. Hence this wave-function shows behaviour which is very similar to that of the equivalent classical system.

*

COMMENTS : This problem is a clear demonstrator of the mathematical difficulties inherent in quantum mechanics. A hundred Tutes back, we solved the classical harmonic oscillator in one line. Now, the quantum version threw up a BVP whose eigenvalues we had to take as a given (the solutions of (01) are of course special functions). There IS a trick, thanks to PAUL DIRAC, of getting to the eigenvalues without solving (01) but that's a lot of fancy quantum mechanics and little or no differential equations.

Here we can see some similarities between a classical system and its quantum equivalent. The eigenfunction expansion of the quantum oscillator is periodic, and the special solution of Part (c) shows behaviour which is extremely similar to that of the classical particle. We shall see a little more of the connection between quantum and classical mechanics in the next Tute.

Finally, how can we generate the special solution (*) and similar results ? We can do it by juggling with the special function solutions of (01). The function $\theta(t)$ turns out to be a big mess. The parameter A denotes the centre of the Gaussian at $t = 0$ i.e. it is the equivalent of $x(0)$ for the classical particle. For the Gaussian to oscillate in this special manner with no change in shape, its width must be a certain specific value (the value we used in the quoted solution). Gaussian ICs of different widths don't show such perfect oscillations but even then, the centre oscillates regularly and the width also oscillates in time.

Example 7.3.3

Quantum free particle

The treatment here is based on the textbook "Quantum Mechanics" by RICHARD ROBINETT. This problem is algebraically very heavy.

Consider the 1+1 Schroedinger equation defined in x belongs to $[-\infty,\infty]$ with $V=0$ everywhere.

(a) Obtain the eigenfunction expansion for $\psi(x,t)$.

Now consider a wave-function whose initial value is

$$\psi(x,0)=\frac{1}{\pi^{1/4}(\hbar\alpha)^{1/2}}\exp\left(-\frac{x^2}{2\hbar^2\alpha^2}+\mathrm{i}\frac{p_0 x}{\hbar}\right) \quad ,$$

where α and p_0 are constants. This represents a particle which moves to the right with initial momentum p_0.

(b) Drawing reference to T085, determine $\psi(x,t)$ and $|\psi(x,t)|^2$. Hint : All the integrals involved are doable, and maybe looked up in a table.

(c) Determine $\langle x \rangle$ and Δx. Same hint as Part (b) applies.

(d) Compare the results of this problem with the solutions for similar ICs on the heat equation and wave equation.

[Level : 3]

Why does $V = 0$ make a harder Tute than $V = (1/2)m\omega^2 x^2$? Because we'll actually have to solve Schroedinger equation here instead of being given the solutions on a platter.

(a) The equation is

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2} = \mathrm{i}\hbar\frac{\partial\psi}{\partial t} \quad .$$

Separating variables as usual,

$$\frac{X''}{X} = -\mathrm{i}\frac{2m}{\hbar}\frac{\dot{T}}{T} = \lambda \quad .$$

Since the T equation has an imaginary coefficient, the time evolution will always be oscillatory irrespective of whether λ is positive or negative. So we look to the X equation to narrow down our choices.

If λ is positive then we have the spatial solutions e^{kx} and e^{-kx} (where $k = \sqrt{\lambda}$ as always). One of these blows up in the negative region and the other one in the positive region. Since the wave-function cannot be infinite, these are useless. If λ is negative then we have the usual $\cos kx$ and $\sin kx$ solutions, which are more tractable, and can be used to make up really tractable functions (i.e. functions which vanish at infinity) using the Fourier transform (which we saw in T085). Here however, since everything is complex, we write the two solutions as $\mathrm{e}^{\mathrm{i}kx}$ and $\mathrm{e}^{-\mathrm{i}kx}$. The case $\lambda = 0$ features constant and linear solutions which are both inadmissible. Thus, λ is negative with the spatial solutions $\mathrm{e}^{\mathrm{i}kx}$ and $\mathrm{e}^{-\mathrm{i}kx}$.

The corresponding temporal equation is

$$\dot{T} + \mathrm{i}\frac{\hbar k^2}{2m}T = 0 \quad ,$$

with solution

$$T = \exp\left(-\mathrm{i}\frac{\hbar k^2}{2m}t\right) \quad .$$

Putting all this together we have

$$\psi(x,t) = \sum_k \left(A_k\,\mathrm{e}^{\mathrm{i}kx} + B_k\,\mathrm{e}^{-\mathrm{i}kx}\right)\exp\left(-\mathrm{i}\frac{\hbar^2 k^2}{2m}t\right) \quad .$$

Finally, since every k is allowed, the sum over k becomes an integral. By letting this integral run from $-\infty$ to ∞ we can ignore the $\mathrm{e}^{-\mathrm{i}kx}$ terms and combine A_k and B_k into a single complex-valued function $f(k)$. We thus have

$$\psi(x,t) = \int_{-\infty}^{\infty} \mathrm{d}k\, f(k) \exp(\mathrm{i}kx) \exp\left(-\mathrm{i}\frac{\hbar k^2}{2m}t\right) \quad , \qquad (01)^*$$

which is the eigenfunction expansion asked for.

(b) Substituting $t = 0$ into (01) and fitting it to the IC given in the question we have

$$\int_{-\infty}^{\infty} f(k)\mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}k = \frac{1}{\pi^{1/4}(\hbar\alpha)^{1/2}} \exp\left(-\frac{x^2}{2\hbar^2\alpha^2} + \mathrm{i}\frac{p_0 x}{\hbar}\right) \quad .$$

Letting $g(x)$ denote this IC function, the Fourier transform inversion theorem which we saw in T085 tells us that

$$f(k) = \frac{1}{2\pi}\int_{-\infty}^{\infty} g(x)\mathrm{e}^{-\mathrm{i}kx}\,\mathrm{d}x \quad . \qquad (02)$$

Now the integral fun starts.

Nasty integral #1 : (02) evaluates to

$$f(k) = \frac{\alpha^{1/2}}{\pi^{1/4}} \exp\left(\frac{-\alpha^2\hbar^2(k-k_0)^2}{2}\right) \quad ,$$

where $k_0 = p_0/\hbar$. Plugging this into (01) gives nasty integral #2; that evaluates to

$$\psi(x,t) = \frac{1}{\pi^{1/4}(\hbar\alpha)^{1/2}\left(1+\mathrm{i}\frac{t}{m\hbar\alpha}\right)^{1/2}} \times$$

$$\exp\left(-\frac{\left(x-\frac{p_0}{m}t\right)^2}{2\hbar^2\alpha^2\left(1+\mathrm{i}\frac{t}{m\hbar\alpha}\right)}\right) \exp\mathrm{i}\left(\frac{p_0}{\hbar}x - \frac{p_0^2}{2m\hbar}t\right) \quad , \qquad (03)^*$$

which is the desired expression for ψ.

Now we need the absolute value of this (remember, that's the actual probability distribution of finding the particle at position x at time t). The mod squared of an $\exp \mathrm{i}(\ldots)$ factor is easy – it's just 1. The rest is more algebra, which leads to

$$* \qquad |\psi|^2 = \frac{1}{\pi^{1/2}\hbar\alpha\left(1+\dfrac{t^2}{m^2\hbar^2\alpha^4}\right)^{1/2}} \exp\left(-\frac{\left(x-\dfrac{p_0}{m}t\right)^2}{\hbar^2\alpha^2\left(1+\dfrac{t^2}{m^2\hbar^2\alpha^4}\right)}\right) , \qquad (04)*$$

the second quantity required by the question.

(c) The expectation value of x is easy. Looking at (04) we can see that the probability distribution is a Gaussian function centred at $x = (p_0/m)t$. Because the Gaussian is completely symmetric, the expectation value of x must be at its centre. Hence $\underline{<x> = (p_0/m)t}$.

To calculate the standard deviation, we need the expectation value of x^2. This is given by

$$\left\langle x^2 \right\rangle = \int_{-\infty}^{\infty} \psi^* x^2 \psi \, \mathrm{d}x \quad ,$$

with ψ as in (03). Nasty integral #3 and the attendant manipulations finally yield

$$* \qquad \Delta x = \frac{1}{\sqrt{2}}\left(\alpha^2 + \frac{\hbar^2}{m^2\alpha^2}t^2\right)^{1/2} \quad , \qquad (05)*$$

which is the desired answer.

(d) Finally we have something which requires concepts rather than algebra. What the solution (03) basically says is that the Gaussian wave-function keeps moving to the right with constant velocity p_0/m, while its width continues to increase as per the formula (05). At large time, the width is proportional to the time. If we consider the 1+1 wave equation with IC as a Gaussian which moves to the right, then the disturbance will continue moving rightwards with the same speed. However, D'ALAMBERT's formula says that it will retain its shape completely unlike (03), which spreads out. Heat equation on the other hand cannot support a moving Gaussian as an IC – we can only give it a static Gaussian. Then, the Gaussian spreads out in time as we saw in T085 (the initial delta

function turns into Gaussian as soon as it starts to evolve, and continues that way). The width there is perfectly linear in time.

Thus, the Schroedinger equation, at least in this case, combines solution features of the heat and wave equation. It has the rightward motion from the wave equation but also has the spreading out of the width from the heat equation. This hybrid is due to the presence of the $\partial/\partial t$ term with an imaginary coefficient – the i changes the behaviour significantly from the heat equation case but cannot perfectly simulate a $\partial^2/\partial t^2$ either.

*

COMMENTS : This is the most algebraically tedious problem in the entire book. Difficulty of mathematical manipulation is in fact one of the hallmarks of quantum mechanics. In the classical subjects, physically motivated tricks often substitute for a mathematical tour de force – one can make some approximations, guess a solution and so on. But in quantum mechanics, that option hardly ever exists. We already saw that the harmonic oscillator wasn't really doable at our level. The simplest spherical problem, the hydrogen atom, is solved in terms of associated Legendre and associated Laguerre polynomials. As GRIFFITHS says in the preface to his excellent introductory textbook, "we encounter Legendre, Hermite and Laguerre polynomials, spherical harmonics, Bessel, Neumann and Hankel functions, Airy functions and even the Riemann zeta function". Yes, that's an introductory quantum book. By contrast, in electromagnetism, we can go all the way upto maglev without ever having to see a special function.

Coming back to the results we got here, the part about the centre of the wave-function moving at speed p_0/m carries over from classical mechanics. By contrast, the part about its spreading out in time has no classical counterpart. This is similar to what happened for the oscillator in the last Tute – the centre followed classical dynamics while the width has no classical equivalent. So how does quantum turn into classical ? In one line, quantum mechanics turns into classical mechanics in the limit $\hbar \to 0$. Sometimes, the truth of this premise is evident. Here, the width (05) has $\hbar$ multiplying the temporally increasing part. In the special solution of Part (c) of the last Tute, the argument of the exponential has $\hbar$ in the denominator – set it to zero and the Gaussian becomes

infinitely narrow, i.e. a delta function. For more complicated systems, the limiting process is not at all transparent. For instance, one big question is, Schroedinger equation is linear but classical systems are often nonlinear. So if a classical system is chaotic, what happens to the corresponding wave-function ?

The study of these and other questions is fascinating, and has produced a wealth of excellent literature. As for us however, we won't be concerned any more about the details of the limit. Instead, we'll TAKE that limit. Having seen our necessary share of how microscopic particles work, we'll now make a triumphant return to the certain, deterministic world where position is what you measure on a scale, velocity is its derivative, momentum is velocity times mass, energy is momentum squared upon twice mass etc. And yes, we'll turn that i back into a j.

7.4

MODELING WITH THE WAVE EQUATION

We shall proceed to the examples directly.

Example 7.4.1

Stability transition

A rope (string) of length L and linear mass density ρ_1 hangs vertically under its own weight from a fixed point P. The bottom end of the rope is constrained using a massless slider to move in the vertical direction alone. A motor is attached to point P to make it rotate with angular velocity ω about a vertical axis. In an experiment, the rope is started from rest and ω increased at an extremely slow rate.

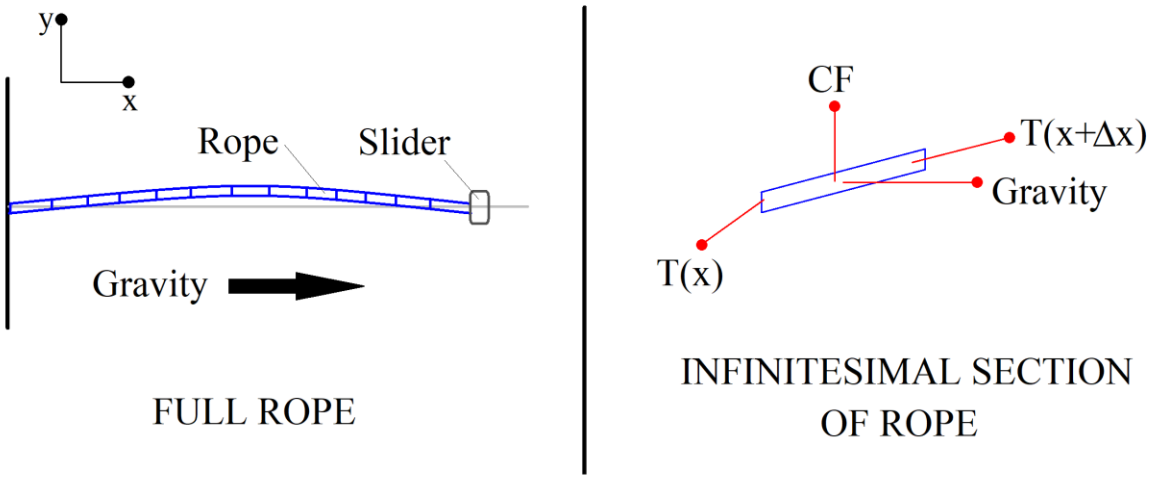
(a) Considering a small transverse displacement of the rope from the trivial state (hanging vertically downwards), write a PDE problem for the displacement as a function of space and time. Hint : The relevant forces are gravity and the centrifugal force.

(b) Instead of hanging vertically, suppose the rope were set up horizontally, stretched to a constant tension τ and then rotated about a horizontal axis. Determine the value of ω at which the trivial solution would first become unstable.

(c) Discretize the equation obtained in Part (a) using the matrix method. Considering the parameter values $L = 0.5$ m and $\rho_1 = 0.1$ kg/m, determine the value of ω at which the vertically hanging rope loses stability.

[Level : 4]

(a) For this part we need to understand the details behind the derivation of the wave equation for a string. Let x be the vertically downward direction, and y be the radial direction (so that our target function can be $y(x,t)$ and not some non-standard thing). Then, the rope is shown in the Figure below (analysing from a rotating frame).



Following L25 we do a force balance on the infinitesimal element of the rope between x and $x+\Delta x$. As usual, let $\theta(x)$ denote the angle between the tangent to the rope and the horizontal. We do the x-direction force balance first. The rope has no motion in this direction, so we have

$$\begin{aligned}(\rho_1\Delta x)\ddot{x} &= T(x+\Delta x)\cos\theta(x+\Delta x) - T(x)\cos\theta(x) + (\rho_1\Delta x)g \\ &= 0\end{aligned} \quad .$$

Now we invoke the fact that the transverse displacement y is small. This means that $\cos\theta = 1$ and $\sin\theta = \theta = \Delta y/\Delta x$. Plugging that into the above RHS, Taylor expanding $T(x+\Delta x)$, replacing the Δ's by d's and cancelling the common factor of Δx we get

$$\frac{\mathrm{d}T}{\mathrm{d}x} + \rho_1 g = 0 \quad .$$

This "differential equation" for $T(x)$ is below the level of 1.1.1; the associated BC must be that the tension at the free end is zero i.e. $T(L) = 0$. Solving this system gives

$$T = \rho_1 g(L-x) \quad , \qquad (01)$$

and the x-direction force balance has yielded all it can.

Now we do the force balance in y. The centrifugal force is $(\rho_1 \Delta x)\omega^2 y$ and it pushes the rope outwards (this is the effect which makes the vertical state go unstable). So we have

$$(\rho_1 \Delta x)\ddot{y} = T(x+\Delta x)\theta(x+\Delta x) - T(x)\theta(x) + (\rho_1 \Delta x)\omega^2 y \quad .$$

Expanding, replacing Δ by d, replacing θ by y' and factoring out Δx we have

$$\ddot{y} = \frac{T}{\rho_1} y'' + \frac{T'}{\rho_1} y' + \omega^2 y \quad . \tag{02}$$

It is a ready check that if $T = \text{const.}$ and $\omega = 0$ then we recover the known equation from the standard case.

Finally, plugging (01) into (02) gives us

$$g(L-x)\frac{\partial^2 y}{\partial x^2} - g\frac{\partial y}{\partial x} + \omega^2 y = \frac{\partial^2 y}{\partial t^2} \quad , \tag{03}*$$

which is the PDE asked for. The BCs are simple. Letting $x = 0$ be the point P, the displacement is obviously zero there, so $\underline{y(0) = 0}$. At the other end $x = L$, the slider again constrains the radial displacement to zero so $\underline{y(L) = 0}$.

(b) Equation (03) does not look at all friendly, which is why we have to do it using matrix method. Before that, the simple case considered in this part gives us insight regarding what's going on here.

With the rope horizontal and stretched to a constant tension τ (T from this point onwards will become the separable function of time), the equation in the presence of rotation is

$$\frac{\tau}{\rho_1} y'' + \omega^2 y = \ddot{y} \quad . \tag{04}$$

Before blindly separating variables, we assess what we need out of here. This is a stability calculation like the one in 5.4.3 – $y(x,t)$ is not an actual shape of the rope but only a small perturbation upon the trivial solution, which is the real shape function. The trivial state is stable if the perturbation remains bounded in time (in an undamped mechanical system, oscillations are the best we can get, no decay). This perturbation can have arbitrary shape and can thus be expressed as an arbitrary linear combination of the spatial eigenfunctions of the problem,

which are $\sin(n\pi x/L)$. Each eigenfunction will have its own temporal evolution; even if one of them grows in time then the trivial solution becomes unstable.

So we now substitute $(\sin kx)T(t)$ into (04) where $k = n\pi/L$ for any n. Cancelling the common factor of $\sin kx$ yields

$$\ddot{T} = \left(-\frac{k^2\tau}{\rho_1} + \omega^2 \right) T \quad .$$

So long as the expression inside the brackets is negative, the solutions will be oscillatory; if it turns negative, then the exponentially growing solution will enter the picture. At sufficiently low ω, the thing will be negative for all k; as ω increases, the first one to turn positive will be the one with lowest k. This corresponds to $n = 1$, so that will be the first mode to cause the destabilization of the trivial solution.

Given this, the critical value of ω comes out immediately : we have

$$\omega = \left(\frac{\pi^2\tau}{\rho_1 L^2} \right)^{1/2} \quad , \qquad *$$

which is the desired answer.

(c) For matrix method, we follow the examples of 5.2.2 and 7.2.2. We discretize space into N significant and two boundary lattice sites x_i and at each site define a y_i which evolves continuously in time. The step size h becomes $L/(N+1)$. As usual, $\mathbf{y} = [y_1;\ y_2;\ \ldots..;\ y_N]$ denotes the variable vector. The second spatial derivative with the double zero BCs is just the matrix $\mathbf{Q}$ from 7.2.2. The first derivative with a zero on one side is the matrix $\mathbf{P}$ from 4.2.4. The multiplier x is the matrix $\mathbf{X}$ whose diagonal elements are $x_1, x_2, \ldots.., x_N$ (again 4.2.4). Finally, a multiplier of a constant (like L, g etc.) is the constant times $\mathbf{I}$, the identity matrix of size N. Using these, (03) becomes

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{y} = g(L\mathbf{I} - \mathbf{X})\mathbf{Q}\mathbf{y} - g\mathbf{P}\mathbf{y} + \omega^2\mathbf{I}\mathbf{y} \quad ,$$

where, using our typical display size of $N = 4$,

$$\mathbf{Q}=\frac{1}{h^2}\begin{bmatrix}-2 & 1 & 0 & 0\\ 1 & -2 & 1 & 0\\ 0 & 1 & -2 & 1\\ 0 & 0 & 1 & -2\end{bmatrix},\quad \mathbf{P}=\frac{1}{h}\begin{bmatrix}1 & 0 & 0 & 0\\ -1 & 1 & 0 & 0\\ 0 & -1 & 1 & 0\\ 0 & 0 & -1 & 1\end{bmatrix},$$

$$\mathbf{X}=\begin{bmatrix}h & 0 & 0 & 0\\ 0 & 2h & 0 & 0\\ 0 & 0 & 3h & 0\\ 0 & 0 & 0 & 4h\end{bmatrix},\quad \mathbf{I}=\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix}.$$

Thus we can write the matrix formulation of the problem as

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{y}=\mathbf{M}\mathbf{y}\quad \text{, where} \tag{05a}$$

$$\mathbf{M}=g\left(L\mathbf{I}-\mathbf{X}\right)\mathbf{Q}-g\mathbf{P}+\omega^2\mathbf{I}\quad . \tag{05b}$$

Now we have to interpret how to use this equation.

In the constant tension case, at very low speed, all the modes were oscillatory. We suspect that the same will hold true here – after all a hanging rope can't jump out of the vertical the moment the rotation starts. So, in the oscillatory regime, we seek a solution of (05a) of the form $\mathbf{y}=\mathbf{v}_n\cos\Omega_n t$. Evidently, $\mathbf{v}_n$ is an eigenvector of $\mathbf{M}$ and $-\Omega_n{}^2$ the corresponding eigenvalue. For all eigenvalues of $\mathbf{M}$ to be negatives of frequency squares, they must all be real and negative. As the rotation speed ω is increased, the smallest of the eigenfrequencies will become smaller and smaller and ultimately become zero. Thus, at this point the largest eigenvalue of $\mathbf{M}$ will become zero while all others will still remain negative. Hence the criterion for the transition must be that the largest eigenvalue of $\mathbf{M}$ just becomes zero while all other eigenvalues are real and negative.

What about complex eigenvalues of $\mathbf{M}$? On physical grounds we do not expect that to happen before the first stability transition. A complex eigenvalue would mean a complex frequency, which is trouble since one of $\mathrm{e}^{\mathrm{j}\Omega t}$ and $\mathrm{e}^{\mathrm{j}\Omega^* t}$ with complex Ω represents exponential growth. So, if the qualitative behaviour of the rope under constant tension is to carry over to this situation (and that is a VERY plausible assumption), we would expect the transition to occur in the manner I

described above. A complex eigenvalue of $\mathbf{M}$ before the first transition would likely mean an incorrect formulation of the problem.

How finely should we discretize ? While 7.2.2 might suggest thousands and thousands of lattice sites, we don't need that here. This is because the multitudinous lattice sites were needed to take care of the high modes whereas Part (b) clearly showed that the instability comes from the lowest mode, which is numerically the most tractable.

Everything on the table, we finally substitute the given parameter values into (06b) and do the eigenworks. We take $g = 9.8$ rather than 10 since for a computer it doesn't matter. Note that ρ_1 does not appear in the equation at all ! True to prediction, all eigenvalues of $\mathbf{M}$ are negative real when $\omega = 0$. The numerical value of the largest one (including the sign!) also converges very quickly with increasing N. For $N = 10$, it is -173, for $N = 20$ it's -170, for $N = 50$, it's -167 and for $N = 500$, -166. The critical rotation speed at which it becomes zero turns out to be 12.9 rad/s or 123 rpm.

*

COMMENTS : The rotating string is a non-trivial problem, as evident from the fact that a certain popular differential equations textbook (and now I'll have to be euphemistic) presents a treatment which is lacking clarity in some places. The ONLY thing our work tells us is the speed at which the vertically hanging state loses stability. It does NOT say anything about what the shape becomes beyond this point. The possible equilibrium shapes of the rope come from static force balances normal and tangential to the string – similar to what we were doing in Part (a) but without the small y restriction. These equations are nonlinear, and their solutions will yield equilibrium geometries of the rope. Then, the stability of each one can be calculated by assuming a perturbation and writing its dynamics. Here we bypassed the first part by considering a state (trivial solution) which is an obvious solution to the equilibrium equations, and then did the second part.

The purpose of the slider at the end of the rope is to force the BC there. The slider ensures that $y = 0$ but it can freely rise up to accommodate a higher mode shape (this shape will be curved and must necessarily hang down by a shorter

length than the straight state). Using a pin instead of a slider i.e. constraining the end to remain at $x = L$ even in higher modes would result in forces coming from the pin and change the problem altogether. But what would happen for a real rope with a free end ? Then, the model of the rope as a string would fall through completely. We would have to consider the lengthwise stresses and strains along the rope and impose the BC that the hanging end is stress-free. An elastic bar model would also not suffice because that can't accommodate perpendicular deflections. So that would require a lot more elasticity theory than we would like to go into here.

Together with 5.4.3, this problem shows you the two most common ways of performing a stability analysis. In 5.4.3 we had the nonlinear equation given to us – we found the fp, considered perturbations off it, and then assumed the perturbation to be small and linearized . Here , we don't have the nonlinear equation – indeed , the equation of motion of a string in an arbitrarily curved state is not at all easy to find. Rather , we used physical argument to obtain a fixed point of whatever the equation may be (of course there's no hanky panky involved here – hanging straight down is a perfectly genuine fixed point behaviour which the rope shows). Then we built up linearized equations of motion for the perturbations , assuming the smallness at the outset . This is sometimes called ad hoc linearization. The former method is more general but this latter has tremendous utility especially in mechanical systems where nonlinear equations are often extremely hard to derive on account of the geometry and the constraints.

The principle of obtaining stability was of course the same in both cases – visualize an expansion of the potential perturbation in terms of the system's eigenfunctions and stipulate that every single eigenfunction must decay or oscillate in time. PDE stability calculations are often monumentally difficult, two such examples being HEISENBERG's and TAYLOR's fluid flow calculations mentioned in 6.4.1. The backbone of such calculations however is always this.

Example 7.4.2

Vibrations of an aircraft wing

Consider a beam of length L, flexural rigidity EI and mass per unit length ρ_1 which is free at the end $x = 0$ and is attached to a point mass m at the end $x = L$.

(a) Write down the PDE and BCs satisfied by the beam.

(b) Trying a temporal behaviour of the form $\cos\omega t$, obtain a transcendental equation for the eigenfrequencies.

Such a beam can model the wing of an aircraft if one takes the point mass to equal half that of the fuselage. For a particular aircraft, the parameter values are given to be $L = 30$, $EI = 10^{10}$, $\rho_1 = 2500$ and $m = 10^5$ SI Units.

(c) Determine the lowest eigenfrequency of vibration of the wing.

[Level : 3]

EXTRA :

(d) Show that the spatial eigenfunctions $X_1(x)$, $X_2(x)$, ….. of the beam are not orthogonal in the usual sense.

(e) Obtain the relation satisfied by the eigenfunctions of the beam.

(f) In the eigenfunction expansion $f(x) = C_1X_1(x) + C_2X_2(x) + ...$ use the relation obtained above to determine the coefficient C_k.

(a) The equation is the usual beam equation from 7.2.3

$$EI\frac{\partial^4 y}{\partial x^4} = -\rho_1 \frac{\partial^2 y}{\partial t^2} \quad . \tag{*}$$

The BCs at $x = 0$ are also very standard, we have $\underline{y''(0) = y'''(0) = 0}$. Only the end $x = L$ poses complications due to the mass being attached there. Evidently, the mass cannot constrain the displacement or slope of the beam at $x = L$. Hence the BCs at that end cannot be coming on y and y'. The second derivative y'' is the torque. Now if the mass is exerting a torque on the beam, then the beam will also exert a torque on the mass. But a finite torque on a point mass will cause it to rotate with infinite angular acceleration, which is ridiculous. Hence, $\underline{y''(L) = 0}$. The last BC is on the shear force. Since the mass is attached to the beam, the mass and the beam end move together, or $y_m(t) = y_b(L,t)$ where subscripts m and b refer to mass and beam respectively.

The force which m feels is obviously $m\ddot{y}_m(t) = m\ddot{y}_b(L,t)$. This force has to be coming from the beam, so by Sir ISAAC's Third, the mass exerts the equal and opposite force on the beam. From beam theory, that force must be equal to $-EIy'''$. Hence we have the fourth BC

$$EIy'''(L,t) = m\ddot{y}(L,t) \quad , \tag{*}$$

where y again refers to the beam and the negative sign takes the "opposite" in NEWTON's Law into account. The PDE with these four BCs make the problem asked for.

(b) Trying $y(x,t) = X(x)\cos\omega t$ and plugging into (01), we readily end up with $X_{(4)} - (\omega^2\rho_1/EI)X = 0$, just as we did in 7.2.3. The first two BCs are $X''(0) = X'''(0) = 0$. The third one is $X'''(L) = 0$. The last one finally says

$$EIX'''(L)\cos\omega t = -mX(L)\omega^2\cos\omega t \quad ,$$

or, $X'''(L) + (m\omega^2/EI)X(L) = 0$.

Now try the general solution as in T118,

$$X_G = C_1\cos\beta x + C_2\sin\beta x + C_3\cosh\beta x + C_4\sinh\beta x \quad ,$$

where $\beta = (\rho_1/EI)^{1/4}\omega^{1/2}$. Getting the double and triple derivatives from that Tute, and setting $x = 0$ gives $C_3 = C_1$ and $C_4 = C_2$. Going over to $x = L$, the double prime BC gives

$$-\beta^2 C_1 \cos\beta L - \beta^2 C_2 \sin\beta L + \beta^2 C_1 \cosh\beta L + \beta^2 C_2 \sinh\beta L = 0 \quad .$$

Finally, the triple prime BC gives a messy equation which I won't waste space by writing now because it will become apparent almost immediately.

If we take these last two equations (one unwritten) for C_1 and C_2, we get the usual $\mathbf{MC} = 0$, and for nontrivial solution we impose the condition that $\det \mathbf{M} = 0$. This means

$$\det\begin{bmatrix} -\cos\beta L + \cosh\beta L & -\sin\beta L + \sinh\beta L \\ m_{21} & m_{22} \end{bmatrix} = 0 \quad \text{, where}$$

$$m_{21} = \beta^3\left(\sin\beta L + \sinh\beta L\right) + \frac{m\omega^2}{EI}\left(\cos\beta L + \cosh\beta L\right) \quad ,$$

$$m_{22} = \beta^3\left(-\cos\beta L + \cosh\beta L\right) + \frac{m\omega^2}{EI}\left(\sin\beta L + \sinh\beta L\right) \quad .$$

The det is not as bad as it looks – although it has a plethora of terms, there's a lot of cut-cut-cancel-cancel. That followed by the identities $\sin^2 + \cos^2 = 1$ and $\cosh^2 - \sinh^2 = 1$ gives

$$* \qquad \beta^3 - \beta^3 \cos\beta L \cosh\beta L + \frac{m\omega^2}{EI}\sin\beta L \cosh\beta L = 0 \quad , \qquad *$$

which is the transcendental equation asked for.

(c) We plug the parameter values into a code which varies ω, evaluates the equation and hunts for a root. The lowest value it returns is $\omega = 25$ rad/s which is about 4 Hz.

*

COMMENTS : First some physical aspects of the problem. The parameter values correspond approximately to a large aircraft. Then, this result is about a factor of 2 off from the actual answer, which is in the range of 2-2.5 Hz. The error is the result of a largish number of simplifications in the modeling. Firstly, a constant value of EI for the wing is heuristic at best, since the wings of an aircraft are broad near the base and narrow near the tip. Moreover, the calculation for the second moment of area is also very approximate, since the cross-sectional shape is non-trivial. For the entire EI part I just replaced the actual wing by a rectangular parallelopiped of averaged dimensions. Secondly, we neglected the

effect of the jet – that is like a point mass hanging from the beam about one-third of the way down. Thirdly, the masses of the wings and fuselage are approximate also, being clubbed together from various sources – the only data which I found to good accuracy is the total mass of the aircraft, a trivial general-knowledge parameter. And then there are all the struts, stiffeners and hundred other contraptions inside the wing which all go into changing its effective EI.

Nevertheless, this problem shows how a quasi-heuristic and simple calculation can often yield results which agree with reality at least upto order of magnitude, and sometimes a little better. Given an actual wing, finding its eigenfrequencies theoretically will be either impossible or nightmarish – certainly nothing at our level. But with a little physically motivated cutting of corners, we did prise out an answer which is very reasonable.

This rather low fundamental frequency of vibration becomes clear whenever the aircraft hits an air pocket. In mild turbulence, the type when the Air India captain switches on the seatbelt signs but passengers loiter about in the aisles, the primary motion of the plane is a small vertical oscillation at about a couple of cycles per second. The reason for this is that the turbulent air excites the wings with a large number of frequencies, but the wing shows a resonance and hence a pronounced response at its natural frequencies. The smallest of these also happens to be the one most easily perceptible.

Now we'll do the extra. Those questions examine a mathematical subtlety inherent in this problem. I didn't put it in the bulk because we won't need the trick anywhere else, and the stuff I'm going to do now is quite difficult. So read on only if you are attracted to linear algebra.

The problem as it stands is reasonably easy – we did the force balance to get the BC, separated variables on the PDE and found the lowest eigenfrequency. We could easily have found more frequencies and their corresponding spatial eigenfunctions. But supposed I had included an additional assignment – given an IC displacement of the beam, determine the subsequent motion. Then, all hell would have broken loose because the spatial eigenfunctions do not satisfy the orthogonality relation $\int_0^L X_m(x)X_n(x)\,\mathrm{d}x = 0$ which we have by now taken almost for granted.

(d) To see why this is the case, we discretize the problem using matrix method as usual. The beam shape becomes a discrete vector $\mathbf{y} = [y_1; y_2; \ldots..; y_N]$ with all the y_i's being time-dependent. The ODE system satisfied by this vector is

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{y} = -\frac{EI}{\rho_1}\mathbf{S}\mathbf{y} \quad ,$$

where $\mathbf{S}$ is the fourth-derivative matrix which implements the suitable BCs. To construct $\mathbf{S}$, we start from the raw fourth derivative matrix $\mathbf{S}_+$ just as we did in 4.2.4 (which you will probably want to review before proceeding here). That matrix is ($N = 4$ for display)

$$\mathbf{S}_+ = \frac{1}{h^4}\begin{bmatrix} 1 & -4 & 6 & 4 & 1 & 0 & 0 & 0 \\ 0 & 1 & -4 & 6 & -4 & 1 & 0 & 0 \\ 0 & 0 & 1 & -4 & 6 & -4 & 1 & 0 \\ 0 & 0 & 0 & 1 & -4 & 6 & -4 & 1 \end{bmatrix} \quad ,$$

where h is the step size. We recall that $\mathbf{y}^{(4)} = \mathbf{S}_+\mathbf{y}_+$ where $\mathbf{y}_+$ consists of four boundary points over and above the significant ones. Implementation of the BCs kills off the extra terms in $\mathbf{y}_+$. At $x = 0$ there are the no-force, no-torque BCs, which in 4.2.4 were present at $x = L$. These will cause the same modifications as we saw there but in the top left corner instead of bottom right. Thus, the first two columns of $\mathbf{S}_+$ can be lopped off with the modification

$$\mathbf{S}_+ = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 \\ -2 & 5 & -4 & 1 & 0 & 0 \\ 1 & -4 & 6 & -4 & 1 & 0 \\ 0 & 1 & -4 & 6 & 4 & 1 \end{bmatrix} \quad .$$

Now come to the bottom right corner. The first BC is easy – $y''(L) = 0$. We take a 2nd derivative of y at site $\#N$ to get $(y_{N+1} - 2y_N + y_{N-1})/h^2 = 0$ or, $y_{N+1} = 2y_N - y_{N-1}$. The last BC is the troublesome (or at least non-standard) one – $y'''(L,t) - (m/EI)\ddot{y}(L,t) = 0$. Taking a right side third derivative at site $\#N$ we have

$$y'''(L) \to \frac{y_{N+2} - 3y_{N+1} + 3y_N - y_{N-1}}{h^3} \quad .$$

Since the vector on the LHS of the discretized equation has only N entries and no additional ones, the last y to have a well-defined double dot is y_N. Then, the BC gets discretized as

$$y_{N+2} - 3y_{N+1} + 3y_N - y_{N-1} - \frac{h^3 m}{EI}\ddot{y}_N = 0 \quad .$$

Substituting the known value of y_{N-1} into this equation gives us y_{N+1} and y_{N+2} in terms of the significant y's as

$$y_{N+1} = 2y_N - y_{N-1} \quad ,$$

$$y_{N+2} = 3y_N - 2y_{N-1} + \frac{h^3 m}{EI}\ddot{y}_N \quad .$$

If we set $m = 0$ i.e. no point mass and hence no shear force, we recover the relations we had found in 4.2.4; since that problem represents exactly that BC, this is an important consistency check.

The extra double dot term on the above RHS now messes things up. In its absence, the discretized PDE would have been $\ddot{\mathbf{y}} = -(EI/\rho_1)\mathbf{S}\mathbf{y}$ with

$$\mathbf{S} = \frac{1}{h^4}\begin{bmatrix} 1 & -2 & 1 & 0 \\ -2 & 5 & -4 & 1 \\ 1 & -4 & 5 & -2 \\ 0 & 1 & -2 & 1 \end{bmatrix} \quad .$$

However, as things stand, the equation is

$$\ddot{\mathbf{y}} = -\frac{EI}{\rho_1}\mathbf{S}\mathbf{y} - \frac{m}{h\rho_1}\left[0; \quad 0; \quad .....; \quad 0; \quad \ddot{y}_N\right] \quad ,$$

which we can write in a matrix form as

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1+\frac{m}{h\rho_1} \end{bmatrix} \frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{y} = -\frac{EI}{\rho_1}\mathbf{S}\mathbf{y} \quad . \qquad (01)$$

Letting $\mathbf{M}$ be the matrix on the left, we can see that $\mathbf{M}$ is invertible since the inverse is a diagonal matrix whose elements are the reciprocals of those of $\mathbf{M}$. Then, multiplying by $\mathbf{M}^{-1}$ on the both sides we have

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{y} = -\frac{EI}{\rho_1}\mathbf{M}^{-1}\mathbf{S}\mathbf{y} \quad . \tag{02}$$

Getting Matlab to multiply the two matrices on the right shows that their product is NOT symmetric. Hence its eigenvectors are not expected to be orthogonal. Since the eigenfunctions are the continuous limit of these eigenvectors, we don't expect them to be orthogonal also.

(e) Although the matrix in (02) is not symmetric, it turns out that all is not lost. Looking for an oscillatory state, we plug in $\mathbf{y} = \mathbf{X}\cos\omega t$ into (01) [$\mathbf{X}$ is the discrete equivalent of the $X(x)$ eigenfunction]. A little rearrangement of terms gives

$$\frac{EI}{\rho_1}\mathbf{S}\mathbf{X} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\dfrac{m\omega^2}{h\rho_1} \end{bmatrix}\mathbf{X} = \omega^2\mathbf{I}\mathbf{X} \quad ,$$

where $\mathbf{I}$ is the identity matrix.

This is an eigenvalue problem having the special structure $\mathbf{B}\mathbf{v} = \lambda\mathbf{v}$ where $\mathbf{B}$ looks like

$$\mathbf{B} = \mathbf{A} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \alpha\lambda \end{bmatrix} \quad ,$$

with $\mathbf{A}$ symmetric. In our problem, $\lambda = \omega^2$, and $\alpha = -m/\rho_1 h$. The eigenvalue equation $\mathbf{B}\mathbf{v} = \lambda\mathbf{v}$ means

$$\left(\mathbf{A} + \begin{bmatrix} & & \\ & & \\ & & \alpha\lambda \end{bmatrix}\right)\mathbf{v} = \lambda\mathbf{v} \quad \text{, or}$$

$$\mathbf{A}\mathbf{v} + \alpha\lambda\left[0....0; \quad v_N\right] = \lambda\mathbf{v} \quad , \tag{03}$$

where v_N is the N^{th} element of $\mathbf{v}$.

To obtain the relation between different eigenvectors of $\mathbf{B}$, we carry out the equivalent of the first few steps of the proof of the spectral theorem which I had presented all the way back in 2.3. Let λ_1 and λ_2 be two distinct eigenvalues of $\mathbf{B}$, and $\mathbf{v}_1$ and $\mathbf{v}_2$ the corresponding eigenvectors. Then, consider the dot product $\mathbf{v}_1 \cdot \mathbf{A}\mathbf{v}_2$. Using (03) it is

$$\begin{aligned} \mathbf{v}_1 \cdot \mathbf{A}\mathbf{v}_2 &= \mathbf{v}_1 \cdot \left(\lambda_2 \mathbf{v}_2 - \alpha \lambda_2 \left[0 \ldots . 0; \quad v_N \right] \right) \\ &= \lambda_2 \left(\mathbf{v}_1 \cdot \mathbf{v}_2 - \alpha v_{1N} v_{2N} \right) \end{aligned} \quad . \tag{04}$$

On the other hand, we can write $\mathbf{v}_1 \cdot \mathbf{A}\mathbf{v}_2$ as $\mathbf{A}\mathbf{v}_1 \cdot \mathbf{v}_2$ due to the transposition trick followed by $\mathbf{A}$ is symmetric. Using (03) this is

$$\mathbf{A}\mathbf{v}_1 \cdot \mathbf{v}_2 = \lambda_1 \left(\mathbf{v}_1 \cdot \mathbf{v}_2 - \alpha v_{1N} v_{2N} \right) \quad . \tag{05}$$

Now force equality between the RHSes of (04) and (05); since λ_1 does not equal λ_2 by definition, we must have

$$\mathbf{v}_1 \cdot \mathbf{v}_2 - \alpha v_{1N} v_{2N} = 0 \quad . \tag{06}$$

This is the condition which will give the relation satisfied by the beam's eigenfunctions.

Now we come back to the beam to see what the condition (06) means. The eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ become two discretized eigenfunctions $\mathbf{X}_m$ and $\mathbf{X}_n$ of the beam. Our α contains a factor of $1/h$; we pull it out and transfer it to the dot product term getting

$$h\left(\mathbf{X}_m \cdot \mathbf{X}_n \right) + \frac{m}{\rho_1} X_{m,N} X_{n,N} = 0 \quad .$$

Now h times the dot product of $\mathbf{X}_m$ and $\mathbf{X}_n$ is nothing but $\int_0^L X_m X_n \mathrm{d}x$, since h becomes $\mathrm{d}x$ in the limit. At the same time, the last elements of the vectors $\mathbf{X}_m$ and $\mathbf{X}_n$ are $X_m(L)$ and $X_n(L)$. Then, we have

$$* \qquad \int_0^L X_m(x) X_n(x) \, \mathrm{d}x + \frac{m}{\rho_1} X_m(L) X_n(L) = 0 \quad , \tag{07}*$$

which is the relation asked for.

(f) We can easily use the above relation to expand an IC in terms of the eigenfunctions. Given the relation

$$f(x) = C_1 X_1(x) + C_2 X_2(x) + ... + C_k X_k(x) + ... \quad ,$$

for unknown constants C, then, to find C we consider the following quantity :

$$Q = \int_0^L f(x) X_k(x) \mathrm{d}x + \frac{m}{\rho_1} f(L) X_k(L) \quad .$$

Plugging in the expansion of f, this evaluates to

$$Q =$$

$$C_1 \int_0^L X_1 X_k \mathrm{d}x + C_2 \int_0^L X_2 X_k \mathrm{d}x + ... + C_k \int_0^L X_k X_k \mathrm{d}x + ... \quad .$$

$$+ \frac{m}{\rho_1} \left[C_1 X_1(L) X_k(L) + C_2 X_2(L) X_k(L) + ... + C_k X_k(L) X_k(L) + ... \right]$$

Now, from (07), all but one of the terms here are zero; the one which isn't gives

$$* \qquad C_k = \frac{\int_0^L f(x) X_k(x) \, \mathrm{d}x + \frac{m}{\rho_1} f(L) X_k(L)}{\int_0^L X_k(x) X_k(x) \, \mathrm{d}x + \frac{m}{\rho_1} X_k(L) X_k(L)} \quad , \qquad *$$

which is the desired answer. This formula enables us to fit an IC to the beam with a mass at the end.

This problem is a very simple example of a "non-self-adjoint" problem i.e. one where the equivalent matrix is not symmetric. In this case, we were able to get a close-to-orthogonality relation because of the special nature of the deviation from symmetry. However, in completely asymmetric problems, no such option exists. The only thing we can do is prove (or take for granted on physical grounds) the linear independence of the various eigenfunctions and perform a Gram-Schmidt orthonormalization process on the first few of them (as many as we want to take). Then, we fit the ICs using orthonormality on the transformed functions, and subsequently invert the transformation to get things back in terms of the original eigenfunctions. This hair-raising calculation is beyond the scope of this book.

Example 7.4.3

Catenary cable

The overhead electric line (OHE) used to supply power to electric locomotives has a linear density of 0.7 kg/m. Each section of OHE is mounted between two masts spaced 70 m apart and has a maximum sag of 0.1 m at the centre. This can be modelled as a string which satisfies the wave equation and is constrained to zero displacement at the two ends.

(a) Assuming the shape of the OHE to be a parabola, determine the tension and the wave velocity.

A locomotive enters the section moving with constant velocity and its pantograph exerts a constant upward force on the OHE. This force can be treated as a line load which equals 700 N/m in a region of width 0.1 m centred at the pantograph, and zero everywhere else. The OHE is completely quiescent prior to entry of the locomotive into the section.

(b) Expanding the forcing function in terms of the system eigenfunctions, determine the displacement of the catenary from its equilibrium as a function of space and time, upto the time when the loco exits the section. You may ignore terms which are of the order of the ratio of the length of the pantograph (0.1 m) to that of the section.

(c) Determine the motion of the OHE after the loco exits the section.

(d) Plot a graph of the oscillatory vertical force exerted by the OHE on the mast at either of the boundary points, if the speed of the locomotive is 130 km/hr.

[Level : 3]

(a) We are given the shape of the OHE (parabola), the span and the maximum sag. If the origin is at the centre point of the cable, then the coordinates of the end are (0.1,35) all in metres. This immediately gives the equation of the parabola to be $y = ax^2 = (0.1/35^2)\,x^2$.

Now we do the force balance as in 7.4.1, assuming small transverse displacements (an excellent approximation in this case). The differences from that Tute are that there is no force here in the x-direction , and in the y-direction there is gravity instead of the centrifugal force. In the x-direction, $\mathrm{d}T/\mathrm{d}x = 0$ and T is a constant. In the y-direction we have

$$T'\,y' + Ty'' = \rho_1 g \quad .$$

Since T is a constant, $T' = 0$; since $y = ax^2$, $y'' = 2a$. This is consistent, so a constant tension does indeed satisfy the equilibrium equation. We then have $T = \rho_1 g/2a$, which evaluates to 42 kN. The wave speed is $c = (T/\rho_1)^{1/2}$ which is 245 m/s. This is considerably faster than the train speed – it wouldn't do to have a "supersonic" train !

(b) This is the crux of the problem, where we have to find the solution of the wave equation with an inhomogeneous term. Before that, we shift axes so that $x = 0$ is one end of the OHE section and $x = L$ (which equals 70 m) is the other end (these of course are the BCs we are most comfortable with). We define $y(x,t)$ to be the displacement of the OHE from its equilibrium. This displacement evidently satisfies the inhomogeneous 1+1 wave equation – from 7.2 we have

$$\frac{\partial^2 y}{\partial t^2} - c^2 \frac{\partial^2 y}{\partial x^2} = \frac{1}{\rho_1} q(x,t) \quad , \qquad (01)$$

where $q(x,t)$ is the step of width 0.1 and strength 700 which moves with the train velocity.

The question tells us that we should proceed by expanding the forcing function in the systemic eigenfunctions. These of course are $\sin(n\pi x/L)$ but how does the force follow ? To crack that bit, we recognize that the forcing function – a function which moves to the right without changing its shape – is like the solution of a wave equation where the wave velocity is v rather than c. (Just to be clear, the forcing function is not actually a solution of a wave equation – it only BEHAVES like one.) In a slight fudge which is suggested by the question, we now assume $t = 0$ to be when the entire pantograph is just inside the section,

ignoring the $0.1/v$ seconds it takes for the panto to enter. The initial forcing function is then the Fourier expansion of the step-like function of height 700 defined in the interval $[0, 0.1]$. This expansion is

$$q(x,0) = q_0 \sum_n \frac{2}{n\pi}\left(1-\cos\frac{n\pi l}{L}\right)\sin\frac{n\pi x}{L} \quad , \tag{02}$$

where, since symbols are much cleaner than numbers, q_0 denotes 700 and l denotes 0.1.

Now, for a wave equation, we already know how to deal with the "IC" – simply tack on a $\cos(n\pi vt/L)$ to each spatial eigenfunction. This creates the half-sized function which moves to the right and another half-sized function which moves to the left. This latter goes out of the interval of interest as soon as the time starts to evolve (it takes time l/v to do so). The former is the moving force which is of interest – it has the right shape but is of half size. To counteract this, we must add on a factor of 2 to the expansion (02), and then tack on the temporal evolutions. Doing this gives

$$q(x,t) = q_0 \sum_n \frac{4}{n\pi}\left(1-\cos\frac{n\pi l}{L}\right)\sin\frac{n\pi x}{L}\cos\frac{n\pi vt}{L} \quad . \tag{03}$$

This expansion indeed generates a forcing function which is a right-moving step with speed v in the time interval from 0 to L/v. After that, the negative steps from the adjacent cells (this expansion implies antisymmetric periodic extension as in 7.1.2) enter the domain but those are spurious – at $t = L/v$ the loco has left the section so $q = 0$ thereafter.

Now we substitute (03) into (01). By inspection, we should try a particular solution which is of the form

$$y_p(x,t) = \sum_n C_n \sin\frac{n\pi x}{L}\cos\frac{n\pi vt}{L} \quad .$$

An easy calculation then gives

$$C_n = \frac{4q_0L^2}{\rho_1 n^3\pi^3}\left(1-\cos\frac{n\pi l}{L}\right)\frac{1}{c^2-v^2} \quad . \tag{04}$$

The last task is to fit the ICs, which are that the OHE is quiescent at $t = 0$ i.e. initial position and velocity are both zero. Evidently, the y_p we found above satisfies

the latter condition – to satisfy the former one, we need a homogeneous solution which has as IC

$$y_h(x,0) = \sum_n -C_n \sin\frac{n\pi x}{L} \quad ,$$
$$\dot{y}_h(x,0) = 0 \quad .$$

The time evolution tacks on the usual cos and sin components; to account for the zero initial velocity we must have the cosines only. This gives

$$y_h(x,t) = \sum_n -C_n \sin\frac{n\pi x}{L}\cos\frac{n\pi ct}{L} \quad .$$

Finally we put these together to assemble the specific solution : that is

$$y(x,t) = \sum_n C_n \sin\frac{n\pi x}{L}\left(\cos\frac{n\pi vt}{L} - \cos\frac{n\pi ct}{L}\right) \quad , \qquad (05)^*$$

with the C_n's as defined in (04) and is the desired answer.

(c) After the loco exits the section i.e. at $t = t_0 = L/v$, the OHE satisfies the homogeneous wave equation. The ICs for this part have to be the position and velocity we get by substituting $t = t_0$ into (05) – this is the same as the matching process we used in 4.4 and also in 5.4.5. We can write this homogeneous solution as a Fourier series

$$y(x,t) = \sum_n \sin\frac{n\pi x}{L}\left(D_n \cos\frac{n\pi c(t-t_0)}{L} + E_n \sin\frac{n\pi c(t-t_0)}{L}\right) \quad , \qquad (06)$$

which effectively resets time to zero at $t = t_0$.

Now, at $t = t_0$, the position and velocity as obtained from (05) are

$$y(x,t_0) = \sum_n C_n \sin\frac{n\pi x}{L}\left(\cos n\pi - \cos\frac{n\pi c}{v}\right) \quad , \text{ and}$$
$$\dot{y}(x,t_0) = \sum_n C_n \sin\frac{n\pi x}{L}\left(\frac{n\pi c}{L}\cos\frac{n\pi c}{v}\right) \quad .$$

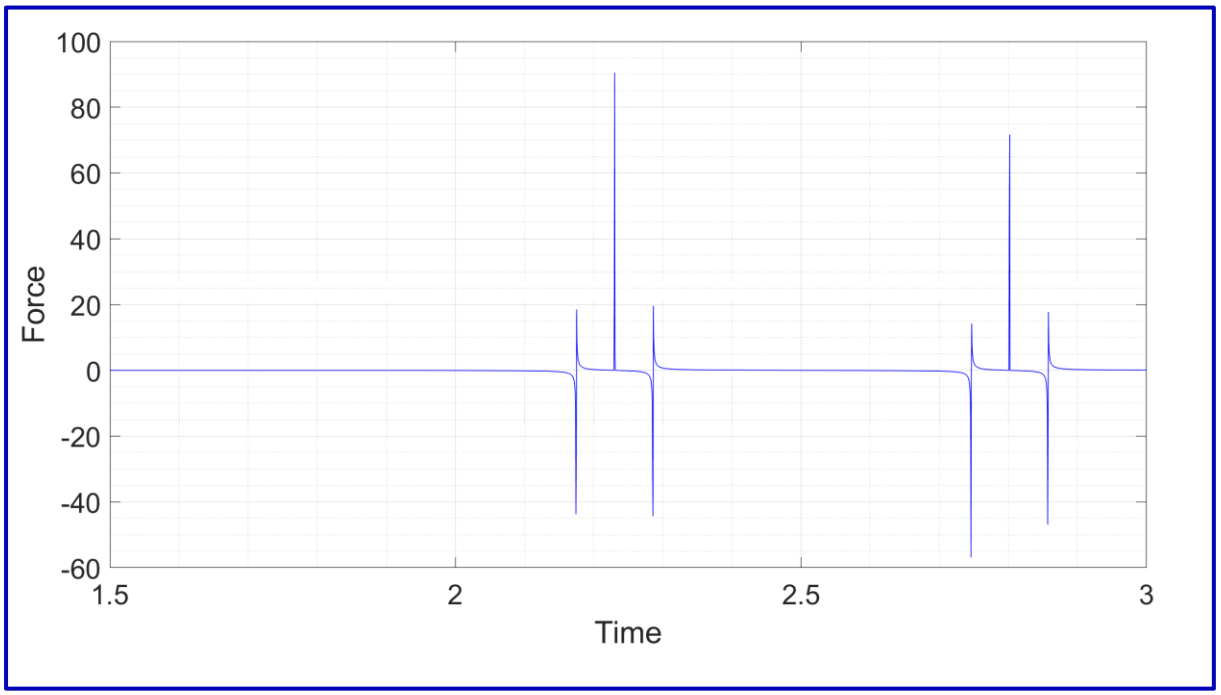
In elementary Fourier series questions of the course exams, they always cut marks unless you simplify $\cos n\pi$ to $(-1)^n$; by now, such things are mere trifles. Equating these to the ICs on (06) yields

$$D_n = C_n\left(\cos n\pi - \cos\frac{n\pi c}{v}\right) \quad ,$$

* $$E_n = C_n \cos \frac{n\pi c}{v} \quad .$$ *

Plugging these two formulae into (06) yield the desired answer.

(d) Before running the plotter, we must figure out what to plot. The desired quantity is the oscillatory component of vertical force exerted by the OHE on the mast at either end. We take the end $x = L$ here. Now, the vertical force is just the y-component of tension, which is Ty' evaluated at $x = L$. Since y is the deviation from equilibrium, this formula automatically produces the forces generated over and above the constant equilibrium force. Further, since the oscillatory component is asked for, we don't need to factor the equilibrium into the picture. The logic worked out, we now present a plot of the force using the solution formulae (05) and (06). Because the factor of $1 - \cos(n\pi l / L)$ in (03) can generate a slow convergence of the series, we throw in ten thousand harmonics. The result is below.



We can see that the oscillations are initiated after the loco passes the section. The shape of the oscillation is peculiar with sudden spikes superposed on a nearly quiescent background.

*

COMMENTS : An oscillatory force of size almost 100 N with very high frequency components – this is a non-trivial loading which the OHE imposes upon the mast.

This joints holding OHE to mast become weaker in time due to fatigue and need periodic maintenance to prevent snapping of OHE. These OHE oscillations and the resulting stresses are the reason why there is a default MPS of 100 km/hr imposed on any train with double headed electric loco. If a second panto hits the OHE which has already been set into oscillation by the first one, then the two oscillations can add constructively and generate a very large vibration amplitude. With a slow enough train speed, the damping in the OHE (something else we neglected here) can mitigate the effect of the first one before the second arrives. But at higher speed that is not possible. Moreover, the interference between the two pantos increases with increasing speed, which can be seen from a tedious calculation. These factors lead to this restriction being imposed.

A consequence of this cap can be seen in the mechanism of electricity supply to the coaches in a Rajdhani rake – the power for driving the AC, lights, fan etc is supplied either from the loco (HOG capable WAP7, only for Howrah and Mumbai Rajdhani) or from diesel generator cars at the two ends (all other Rajdhanis). The first option involves complicated modifications to the loco and the power cars, while the second one is evidently polluting and wasteful. An apparently obvious solution would have been to replace the diesel generator car by a power car which draws electricity from the OHE. However, this is why that cannot be done.

In foreign railways also, the OHE poses constraints on trains with multiple pantos. Speeds of less than 200 km/hr are usually achieved with loco-hauled trains; single high-powered locos are always preferred to double heading. For higher speeds still, EMU type trainsets prove more efficient; there the OHE requires special modifications to handle multiple pantographs one after another. Even so, the number of pantos per train is generally restricted to two, and these are placed eight or more coaches apart.

The parabola shape of the OHE happens to be an exact solution of the linearized equilibrium equations. The full nonlinear equilibrium equations are also exactly solvable in this case; the solution is a cosh function. For small transverse displacement, cosh indeed looks like a parabola. Another way we could have found the particular solution in Part (b) is to see that a RHS function $q(x,t) = f(x-vt)$ automatically generates a particular solution which is $F(x-vt)/(c^2-v^2)$, where F is the second anti-derivative of f. This would not have saved any steps because to fit the ICs we would have had to find the homogeneous solutions anyway, and that would have forced an eigenfunction expansion of the particular solution.

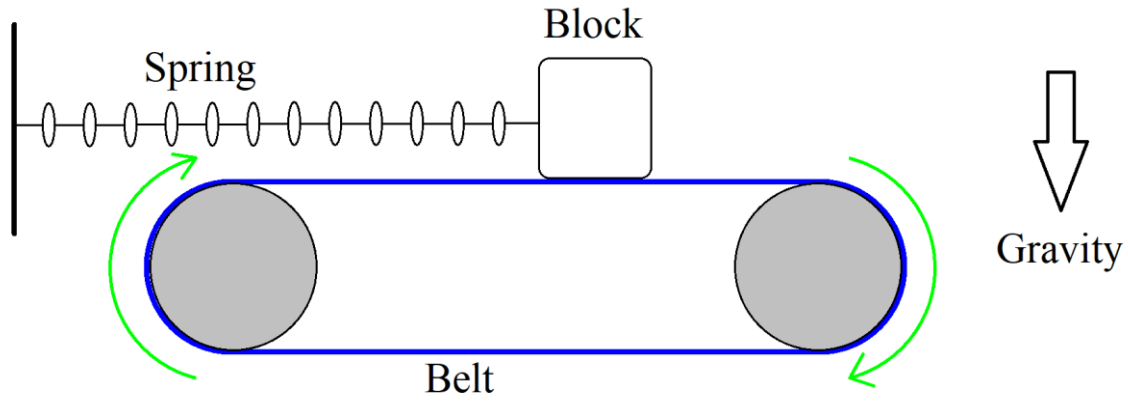
In summary, there are two technical take-home messages of this Tute. The first is that an inhomogeneous problem can sometimes be solved easily by expanding the forcing function in a series of the eigenfunctions of the homogeneous system. This happens because the response of the system to an excitation in the shape of an eigenfunction is also in the shape of an eigenfunction. We'll need this trick later, in another problem. The second is, this calculation gives you a reminder about the method of matched boundary conditions, and shows you its application in a relatively simple wave equation setting. The next Tute again uses this method, but this time I'll notch up the difficulty to max. So put on your thinking caps and get ready.

Example 7.4.4

Stick-slip friction and the violin

Read the passage given below and answer the questions that follow.

Stick-slip friction refers to the commonly observed phenomenon that the coefficient of kinetic friction between two bodies in contact with each other is less that that of static friction. This can lead to interesting kinds of oscillatory phenomena, as the following example demonstrates. Consider the apparatus shown in the Figure below, consisting of a block of mass m attached to a spring of constant k and mounted on a conveyor belt moving with speed v_0. Stick-slip friction acts between the mass and the belt. The coefficient of static friction is μ_s while that of kinetic friction is μ_k with $\mu_k < \mu_s$.

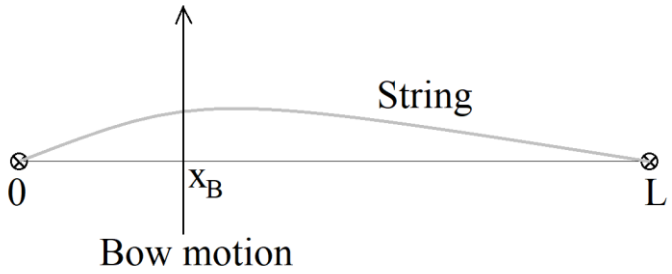


The analysis of the motion can be simplified if one makes the assumption that $\mu_k = 0$. In this case, the qualitative nature of the motion is apparent. One starts from a time when the mass is moving with the belt with the spring force being balanced by the static friction force. Evidently, this motion state will continue until the extension of the spring becomes equal to $\mu_s mg/k$. Beyond this point, the mass will start to slip. Since $\mu_k = 0$ by the assumptions, the dynamics will now be that of the spring and mass alone. This motion will continue until the mass acquires at some point in its

trajectory the velocity v_0 of the conveyor belt. At this time, it will become at rest relative to the belt and static friction will come into effect, causing the cycle to start over again.

The result of this alternating static and kinetic friction is the generation of an oscillation at fixed amplitude and frequency. This is called a limit cycle. It is noteworthy however that all initial conditions might not lead to this limit cycle. For instance, if the mass is started off with such a low oscillation amplitude as to never reach the velocity v_0, then it will continue oscillating like a spring-mass system alone.

The vibration of a violin string is also governed by stick-slip friction. Consider the string as shown in the Figure below.



It is tuned to a constant tension T and clamped at the two ends $x = 0$ and $x = L$. The displacement $y(x,t)$ of the string satisfies the damped 1+1 wave equation

$$\frac{\partial^2 y}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 y}{\partial t^2} + \gamma\frac{\partial y}{\partial t} \quad ,$$

with the damping constant γ being very small. The bow is drawn across the string at the point $x = x_B$ with constant velocity v_0. It applies a concentrated force (delta function line load) F at x_B, perpendicular to the plane of the page. The friction between the bow and the string is stick-slip with the static value being μ_s and the kinetic value again being assumed to be zero.

To obtain the motion of the string, one starts tracking the state of the system at the point where the string just transitions from a stick phase to a slip phase of motion. From this point onwards, its displacement satisfies the homogeneous wave equation. It enters the stick phase when the velocity of the string at the bow point

becomes equal to that of the bow. During this phase, the point $x = x_B$ moves upwards with constant velocity v_0, and the two parts to its left and right can be assumed to satisfy the wave equation separately. The stick phase ceases when the normal component of tension at $x = x_B$ exceeds the friction force, and the next cycle of motion begins at this time.

(a) Let the eigenfunction expansion of the displacement of the string during the slip phase be

$$\sum_n \sin\frac{n\pi x}{L} e^{-\gamma t/2}\left(C_n \cos\frac{n\pi ct}{L} + D_n \sin\frac{n\pi ct}{L}\right) \quad ,$$

with t being measured from the start of slip. Quantitatively develop the argument given in the passage to obtain a relation between the C_n's and D_n's in one cycle and the next.

(b) Implement the above procedure on a computer, using the following parameter values (all in SI Units) : $L = 0.325$, $T = 47$, fundamental frequency $\omega_0 = 2\pi(294)$, damping constant 1 percent of ω_0, $c = 193$, $F = 0.2$, $\mu_s = 0.1$ and v_0 varying between 0.06 and 0.1. Discretize the entire string into 300 lattice sites including boundaries and consider the first 60 spatial harmonics in your eigenfunction expansion. Take $x_B = \#100$. Start the routine in the slip state and use the following special initial condition : $y(x,0)$ consisting of two straight lines between the clamps and the bowing point, with the displacement at x_B having the value $1/16500$, and $\dot{y}(x,0) = 0$. Present suitable plots to illustrate the motion of the string. Examine the effect of varying the IC functions.

(c) Considering $v_0 = 0.08$ and the special IC mentioned above, determine the fundamental frequency of oscillation of the string.

[Level : 6]

Here's the Big One for the wave equation. General gist : the motion of the violin string is similar to that of the spring block on the belt, where the stick-slip friction is responsible for driving limit cycle oscillations. So as a warm-up let's do that problem quantitatively (this is not obligatory for the actual problem, but that is very difficult and I think this is a good preparation).

Letting x be the displacement of the mass from the equilibrium point of the spring, the stick→slip transition occurs and the slip phase starts when the inward force of the spring balances the restoring force of friction i.e. at $x = x_0 = \mu mg/k = \mu g/\omega^2$ (where the subscript s on μ is redundant since we are assuming the kinetic friction coefficient to be zero, and ω is of course the natural frequency of the system). At this time, it is moving with the belt so its velocity is v_0. Letting $t = 0$ be the start of slip, the solution during slip is evidently

$$x = x_0 \cos \omega t + \frac{v_0}{\omega} \sin \omega t \quad ,$$

which we can rewrite as

$$x = \left(x_0^2 + \frac{v_0^2}{\omega^2} \right)^{1/2} \cos\left(\omega t - \varphi\right) \quad , \qquad (01)$$

where $\varphi = \arctan\left(v_0/\omega x_0\right)$.

The slip → stick transition occurs when the velocity of the mass becomes equal to v_0 i.e. the block becomes stationary relative to the belt. So we first calculate the velocity. Differentiating (01) we get

$$\dot{x} = -\omega \left(x_0^2 + \frac{v_0^2}{\omega^2} \right)^{1/2} \sin\left(\omega t - \varphi\right) \quad .$$

Let the time when this equals v_0 be t_1. Then, t_1 satisfies

$$\sin\left(\omega t_1 - \varphi\right) = -\frac{v_0}{\left(\omega^2 x_0^2 + v_0^2\right)^{1/2}} \quad .$$

Plugging this into (01) and writing the cos in terms of the sin gives

$$x(t_1) = \pm \left(x_0^2 + \frac{v_0^2}{\omega^2} \right)^{1/2} \left(1 - \frac{v_0^2}{\left(\omega^2 x_0^2 + v_0^2\right)^2} \right)^{1/2} \quad . \qquad (02)$$

The positive root is an unphysical solution since the velocity and position cannot both be positive on the return path of the mass from a positive displacement. The negative root is the actual solution for $x(t_1)$. At this point, the block sticks to the belt.

We have now described the motion of the block completely. When slip starts, it evolves as per (01) from the point x_0 to the point x_1 which is given by (02). It then sticks to the belt and gets carried by it out to the point x_0 at uniform velocity v_0. Then, it starts slipping again and the cycle begins afresh. The trajectory $x(t)$ of the periodic state thus looks like a portion of a sinusoidal curve joined to a straight line.

Further quantitative analysis of the condition (02) and the resulting amplitude and period of motion under various limiting cases is no longer of interest. Rather, what we note is that the analysis requires a matching of the slip and stick solutions i.e. the method of matched boundary conditions. The technique will be the same for the violin string. The bow is the equivalent of the belt while the string serves the purpose of the spring-mass. Of course, being a PDE, the complications are orders of magnitude greater – the previous example and also 5.4.5 act as good review material.

(a) We start from the homogeneous damped wave equation, independent of context. We have already seen how to solve this in 7.1.4 – with double zero BCs at $x = 0$ and $x = l$ (the more usual L is the length of the violin string here, while this analysis is general) the eigenfunction expansion is

$$y(x,t) = \sum_n \sin\frac{n\pi x}{l} e^{-\gamma t/2}\left(A_n \cos\omega_{dn} t + B_n \sin\omega_{dn} t\right) \quad ,$$

where the n^{th} damped frequency is

$$\omega_{dn} = \frac{\sqrt{\dfrac{4n^2\pi^2c^2}{l^2} - \gamma^2}}{2} \quad .$$

Because the damping is light, we will let ω_{dn} be equal to $n\pi c/l$, ignoring the correction – that might be important for analysis of signal propagation but is not essential to understand the violin's motion. Then, pulling the decaying exponential outside the summation, we have

$$y(x,t) = \mathrm{e}^{-\gamma t/2} \sum_n \sin \frac{n\pi x}{l} \left(A_n \cos \frac{n\pi ct}{l} + B_n \sin \frac{n\pi ct}{l} \right) \quad . \tag{03}$$

Indeed, the question itself tells us this form of the solution – it does not contain γ inside the frequencies.

In fact, we'll go further with the light damping approximation. We'll say that the velocity profile is

$$\dot{y}(x,t) = \mathrm{e}^{-\gamma t/2} \sum_n \sin \frac{n\pi x}{l} \left(-\frac{n\pi c}{l} A_n \sin \frac{n\pi ct}{l} + \frac{n\pi c}{l} B_n \cos \frac{n\pi ct}{l} \right) \quad . \tag{04}$$

This expression looks like something written by a calculus ignoramus, but the neglected terms are of order $\gamma/(n\pi c/l)$, which is given to be 1 percent for the lowest frequency and is still smaller for the higher harmonics. The reason we can be so cavalier about the damping is because it is not an essential phenomenon in the dynamics. As we saw for the mass-spring-belt example, the damping was not required to drive the mass into the limit cycle – the stick-slip friction automatically kills off the homogeneous motions. I am not taking $\gamma = 0$ here so that spontaneous oscillations of the violin string, which occur when the bow is not coupling to it, die off in time. Then, we can easily tell these motions apart from the limit cycle and also prove the role of the friction in providing the energy for driving the oscillations.

Given this, we now have the Fourier formulae – if at $t = 0$ the initial position be $f(x)$ and the initial velocity be $g(x)$ then

$$A_n = \frac{2}{l} \int_0^l f(x) \sin \frac{n\pi x}{l} \mathrm{d}x \quad , \tag{05a}$$

$$B_n = \frac{2}{n\pi c} \int_0^l g(x) \sin \frac{n\pi x}{l} \mathrm{d}x \quad . \tag{05b}$$

The basic tools required for our analysis are now on the ground.

Now we come to the violin string itself. Suppose that a phase of slip starts at some time $t = t_0$. From this time onwards, the string as a whole satisfies the damped wave equation. We can write the displacement as

$$y^{(1)} = \sum_n \sin \frac{n\pi x}{L} \mathrm{e}^{-\gamma(t-t_0)/2} \left(C_n \cos \frac{n\pi c(t-t_0)}{L} + D_n \sin \frac{n\pi c(t-t_0)}{L} \right) , \tag{06}$$

where the superscript (1) refers to the slip phase and the C_n's are D_n's are all unknown. (The equivalent IC functions whose Fourier coefficients are C_n and D_n are of course unknown also – that's how matched boundary conditions works, things become known only at the last step.)

Just as for the block, slip continues until a time t_1 when the velocity of the string at the bow point x_B becomes identically equal to v_0. Thus, the condition for transition from slip to stick is

$$\dot{y}^{(1)}\left(x_B, t_1\right) = v_0 \quad . \tag{07}$$

Of course, we won't be able to solve this condition analytically as we did for the block – that's why Part (b) says to encode everything as a computer routine.

The dynamics in the stick phase is more interesting. Here, the bow drags the point $x = x_B$ at constant speed v_0. If $y^{(1)}(x_B, t_1) = y_1$ (a quantity which comes out when checking for the slip→stick transition), then the subsequent motion of this point is $y^{(2)}(x_B, t) = y_1 + v_0(t - t_1)$, where the superscript (2) denotes the stick phase. But what about the motion of the rest of the string ?

What we do is, we split up the string into two parts – one running from 0 to x_B and one from x_B to L. We let L_1 and L_2 ($=L-L_1$) denote the lengths of these two segments (of course L_1 and x_B are numerically equal, but their connotations are different – L_1 is the LENGTH of the sub-string while x_B is the COORDINATE of the bow). Because one boundary point of each sub-string is moving, they will both satisfy an inhomogeneous wave equation rather than the homogeneous one, and the general solution will be a particular or steady state solution on top of the homogeneous motions. The steady state solution for each sub-string is not difficult to guess – since the total displacement of the moving end is extremely small and almost perpendicular to the string itself, this motion is a ROTATION about its respective clamp at constant angular velocity.

For the left sub-string, this rotation is trivial to write – it is

$$y^{(2)}_{ssL} = \frac{\left(y_1 + v_0(t - t_1)\right)x}{L_1} \quad ,$$

(subscript L denotes left) and its derivative is

$$\dot{y}_{ssL}^{(2)} = \frac{v_0 x}{L_1} \quad .$$

For the right sub-string, the steady state displacement is the straight line connecting the point $(x_B,\ y_1+v_0(t-t_1))$ to the point $(L,0)$. By either deriving, remembering or looking up (I myself used the last option) the formula for straight line between two points, we arrive at

$$y_{ssR}^{(2)} = \left(y_1 + v_0(t-t_1)\right)\left(1 - \frac{x-x_B}{L_2}\right) \quad .$$

Its derivative is then an easy matter :

$$\dot{y}_{ssR}^{(2)} = v_0\left(1 - \frac{x-x_B}{L_2}\right) \quad ,$$

and the steady state solutions of the two sub-strings are out.

If we subtract off the particular solutions, then the remaining motions exhibited by the sub-strings are just solutions of the damped wave equation with double zero BCs. Hence they satisfy the relations (03-5) with the length being L_1 and L_2 as appropriate, and the initial time being t_1 instead of zero. So we can write

$$y_L^{(2)}(x,t) = y_{ssL}^{(2)} + \mathrm{e}^{-\gamma(t-t_1)/2}\times \sum_n \sin\frac{n\pi x}{L_1}\left(E_n \cos\frac{n\pi c(t-t_1)}{L_1} + F_n \sin\frac{n\pi c(t-t_1)}{L_1}\right) \quad . \qquad (08)$$

The situation for the right sub-string is almost the same except for one subtlety (where I tripped up initially) – that sub-string is located between $x = x_B$ and $x = L$ while the eigenfunctions would have been pure sines only if it had been in the interval $[0,\ L_2]$. Hence, for the actual eigenfunctions, we must shift the argument of the sines by x_B. Then,

$$y_R^{(2)}(x,t) = y_{ssR}^{(2)} + \mathrm{e}^{-\gamma(t-t_1)/2}\times \sum_n \sin\frac{n\pi(x-x_B)}{L_2}\left(G_n \cos\frac{n\pi c(t-t_1)}{L_2} + H_n \sin\frac{n\pi c(t-t_1)}{L_2}\right) \qquad (09)$$

.

Finally we have to evaluate the coefficients E_n to H_n (these also can't be undetermined – only the original C_n's and D_n's are, and everything else has to

come out in their terms until we close the chain). These come from the continuity of y and $\partial y/\partial t$ at the transition from slip to stick (the continuity at the bow point itself, we built in explicitly while defining the transition, while continuity everywhere else will have to come now). For the left sub-string, the initial position and velocity are $y_L^{(1)}(x,t_1)$ and $\dot{y}_L^{(1)}(x,t_1)$, while for the right sub-string they are $y_R^{(1)}(x,t_1)$ and $\dot{y}_R^{(1)}(x,t_1)$, where by the L and R subscripts we now understand the portions of the original string in the two intervals $[0,\ x_B]$ and $[x_B,\ L]$.

Since there is a steady state solution involved here, the thing which gets expanded in the eigenfunctions is of course the IC minus the steady state solutions evaluated at $t = t_1$ (see 5.3.1 etc for easier examples). This then gives

$$E_n = \frac{2}{L_1}\int_0^{L_1}\left(y_L^{(1)}(x,t_1) - y_{ssL}^{(2)}(x,t_1)\right)\sin\frac{n\pi x}{L_1}\,\mathrm{d}x \quad , \tag{10a}$$

$$F_n = \frac{2}{n\pi c}\int_0^{L_1}\left(\dot{y}_L^{(1)}(x,t_1) - \dot{y}_{ssL}^{(2)}(x,t_1)\right)\sin\frac{n\pi x}{L_1}\,\mathrm{d}x \quad , \tag{10b}$$

for the left sub-string and

$$G_n = \frac{2}{L_2}\int_0^{L_2}\left(y_R^{(1)}(x,t_1) - y_{ssR}^{(2)}(x,t_1)\right)\sin\frac{n\pi x}{L_2}\,\mathrm{d}x \quad , \tag{11a}$$

$$H_n = \frac{2}{n\pi c}\int_0^{L_2}\left(\dot{y}_R^{(1)}(x,t_1) - \dot{y}_{ssR}^{(2)}(x,t_1)\right)\sin\frac{n\pi x}{L_2}\,\mathrm{d}x \quad , \tag{11b}$$

for the right. Here I have dispensed with the $x - x_B$ on the sine terms by shifting the limits of integration from 0 to L_2 instead of x_B to L.

So this is the stick solution – now when does it go back to slip ? The block did it when the spring pulled it so hard that the friction could not keep up with it. Here, the force which will oppose the friction is the tension. Dragging the bow point will create a kink in the wire – the greater the displacement of the bow point, the more pronounced the kink. A kink means a discontinuity in the derivative y' at $x = x_B$. The normal component of tension is of course $T\sin\theta$ which is Ty' – a jump in the derivative will give rise to a macroscopic force at the bow point. The bow point will stick so long as this force remains less than the friction, which is μF where F is the bow force. Thus, the condition for stick going to slip is

$$T\left(y_L^{(2)}{}'(x_B,t_2)-y_R^{(2)}{}'(x_B,t_2)\right)=\mu F \quad . \tag{12}$$

This of course defines the time of transition as t_2.

Finally end in sight – what we want are the coefficients C_n and D_n at the time t_2 which is the start of the next phase of slip. As usual, the transition will be constrained by continuity of position and velocity at $t=t_2$. We already have the stick solution (08-11) at t_2; now we have to choose the coefficients C_n and D_n such that plugging them into (06) with t_0 replaced by t_2 produces a new slip solution which obeys the required ICs at $t=t_2$. Evidently,

$$C_n=\frac{2}{L}\int_0^L y(x,t_2)\sin\frac{n\pi x}{L}\,\mathrm{d}x \quad , \tag{13a}$$

$$D_n=\frac{2}{n\pi c}\int_0^L \dot{y}(x,t_2)\sin\frac{n\pi x}{L}\,\mathrm{d}x \quad , \tag{13b}$$

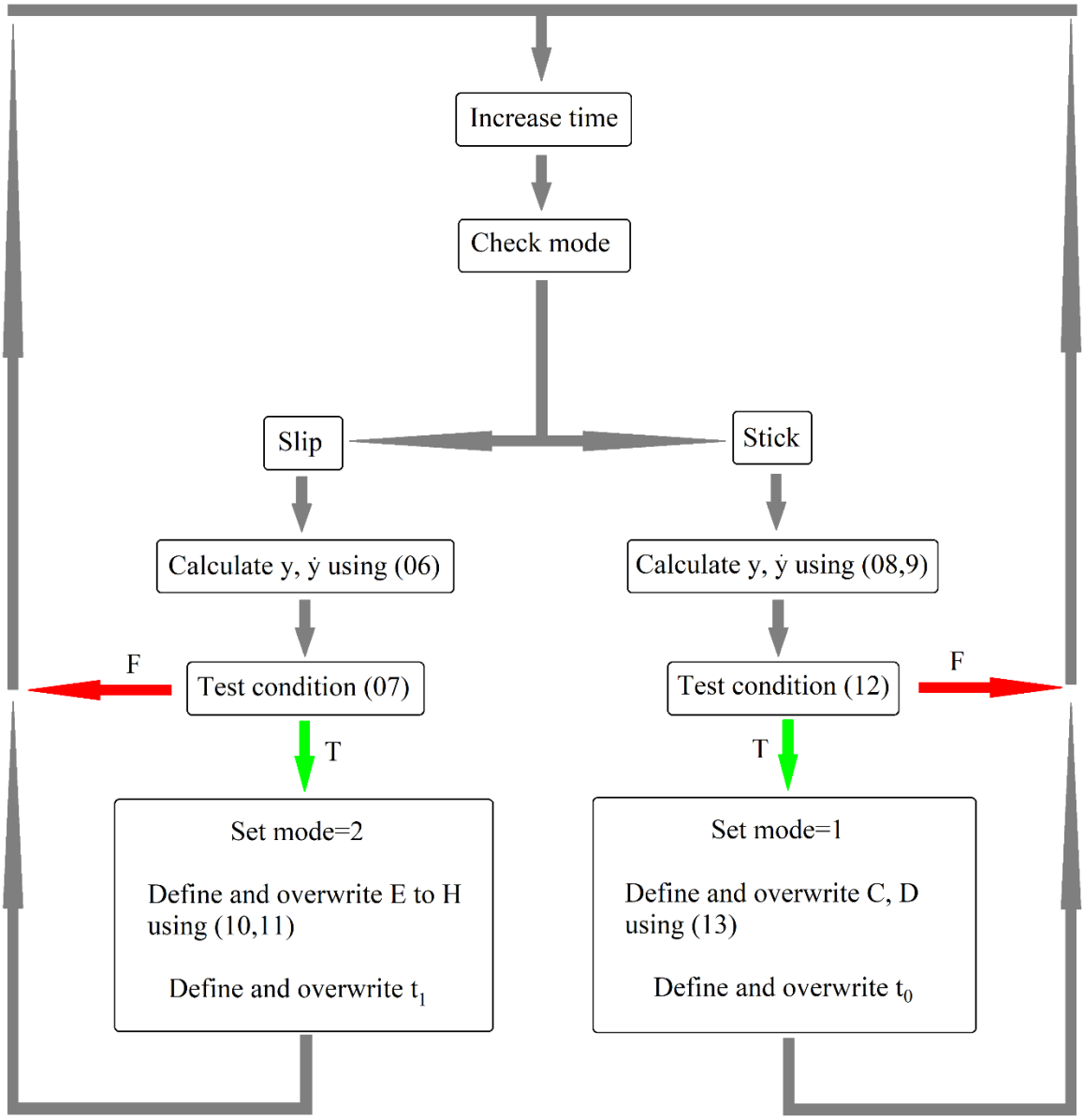
and we have obtained the required relation between the C_n's and D_n's at the start of one cycle and that of the next.

(b) Now we have to implement the above process on a computer. Since the end product is a recursion between the C's and D's from one cycle to the next, we'll have to start the thing off with some plausible initial values. These are given in the question so we don't need to worry any further about this part, and can focus on implementing the logic itself.

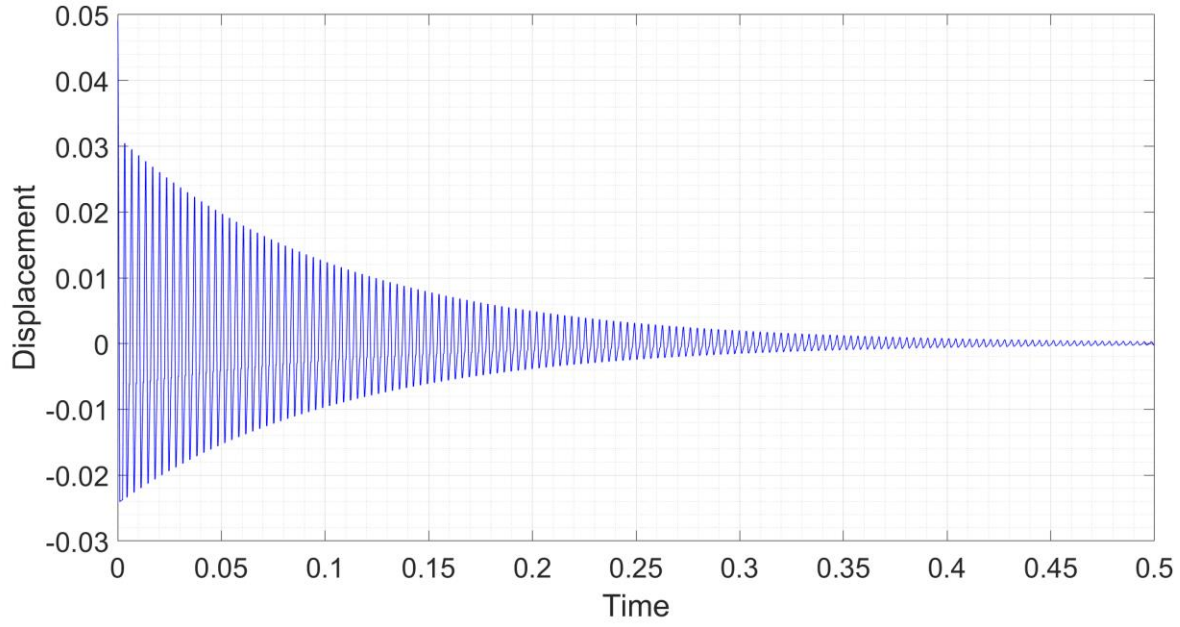
The code is somewhat non-trivial (especially for a CSE amateur like me, and I kept making mistakes) – reproducing it in full is too long but I will give a reasonably detailed outline of what the program does. The initial steps are easy – discretize the space and define the eigenfunctions of the whole string and the two sub-strings. As required, we partion space into 300 points including boundaries, and initialize the first 60 spatial eigenfunctions. We then initialize the arrays for C_n to H_n, all of size 60 (i.e. the array C equals $[C_1;\ C_2;\ \ldots..;\ C_{60}]$ etc). We also initialize the arrays y and y_d (which is $\partial y/\partial t$) both of size 300 (same size as the x-array). Finally we create a variable called mode whose value 1 denotes slip and 2 denotes stick.

To run the program we loop over the variable t which starts from zero and increases with each iteration. Since the lowest natural frequency of the string is about 300 Hz, I have gone for a time incrementation of 10^{-5} s at each step. The initial C_n's are as given in the question while D_n to H_n are trivial. The initial mode is 1 (slip) while $t = t_0 = 0$. At every t we calculate the elements of the array y and y_d (which are of the same size as x), use these to check for conditions and continue the solution.

The central block of the program implements the following algorithm, which I present as a flow chart.



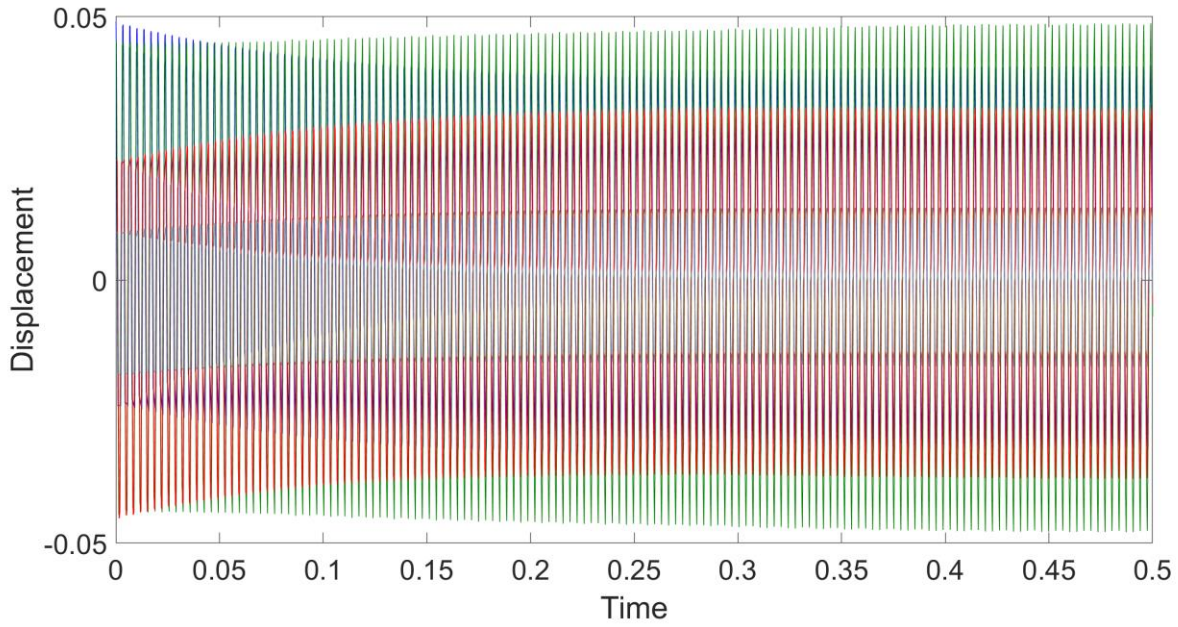
We now present the output of the program. The parameters are all as specified in the question. The first plot considers $x_B = \#100$ as required in the question. The bow speed is $v_0 = 0.06$. The output is the curve of $y(x,t)$ with x corresponding to lattice site #81. Time is measured in seconds and displacement in mm (conventions I will use in all plots).



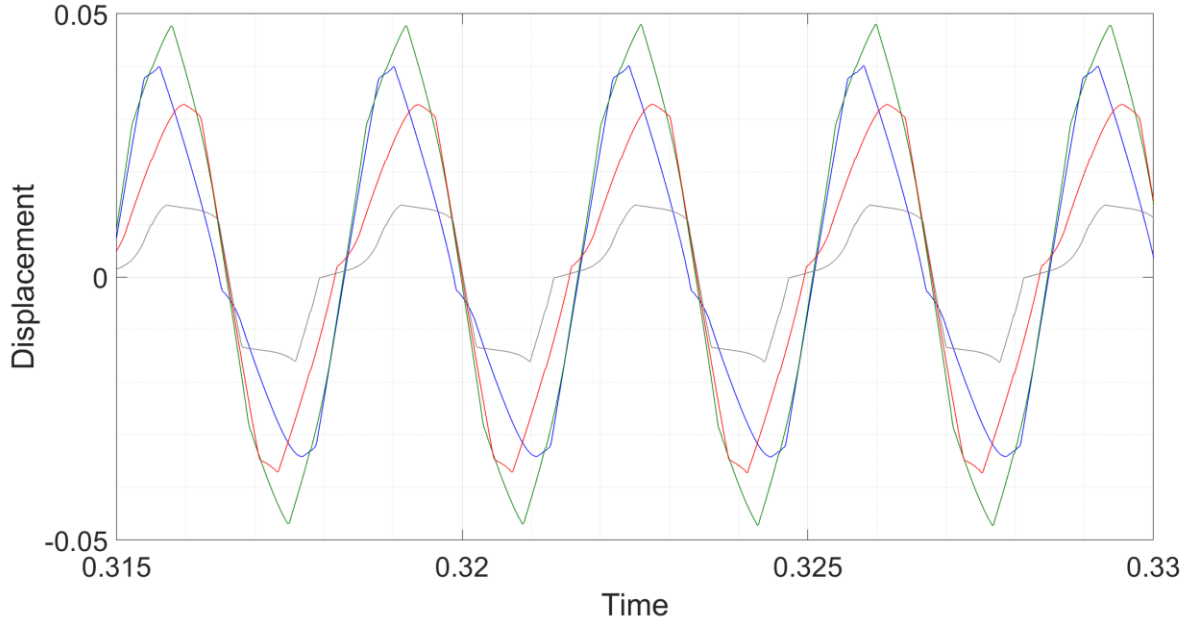
We can see that the oscillations are damped. These are generated by the IC and not by the bow, showing that the bow has not coupled to the string. (The absence of γ in the equations would have resulted in an undamped oscillation here, making it difficult to recognize the bow-driven limit cycle when it occurs.) Raising the bow speed to 0.08 produces the limit cycle. The cycle persists in the range of bow speed from about 0.07 to 0.09 – it again disappears when $v_0 = 0.1$. The amplitude of the limit cycle increases initially upto a maximum at about $v_0 = 0.085$; it then drops and the cycle eventually vanishes. If the bow point is relocated to #50, then also we get a similar behaviour, with the limit cycle motion appearing at about $v_0 = 0.06$, increasing in amplitude and then disappearing at 0.09. Varying the IC for a little bit (about 5 percent!) around the prescribed one causes the limit cycle to remain as it is. More significant variation however causes it to disappear.

We now select the operating point $x_B = \#100$, $v_0 = 0.08$ and the special IC of the question so that we can characterize the limit cycle motion in detail. This time we simultaneously plot $y(\#81,t)$, $y(\#150,t)$, $y(\#225,t)$ and $y(\#270,t)$ as

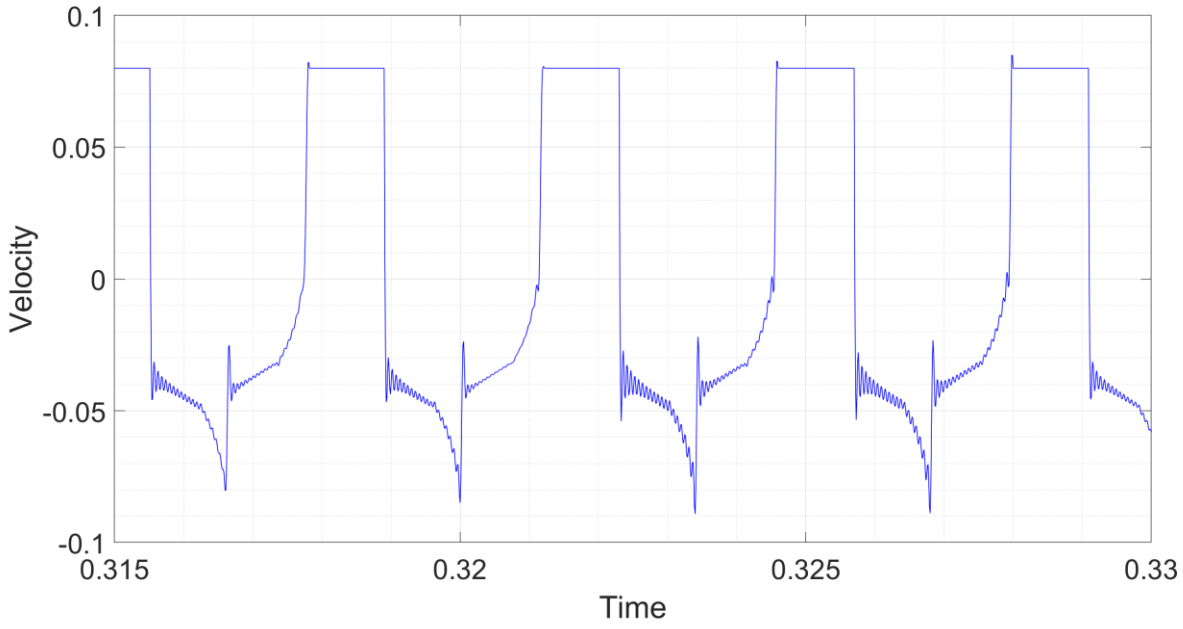
functions of time. The successive traces are in blue, green, red and grey. Clearly, they all oscillate together, with undiminished amplitude and identical fundamental frequency.



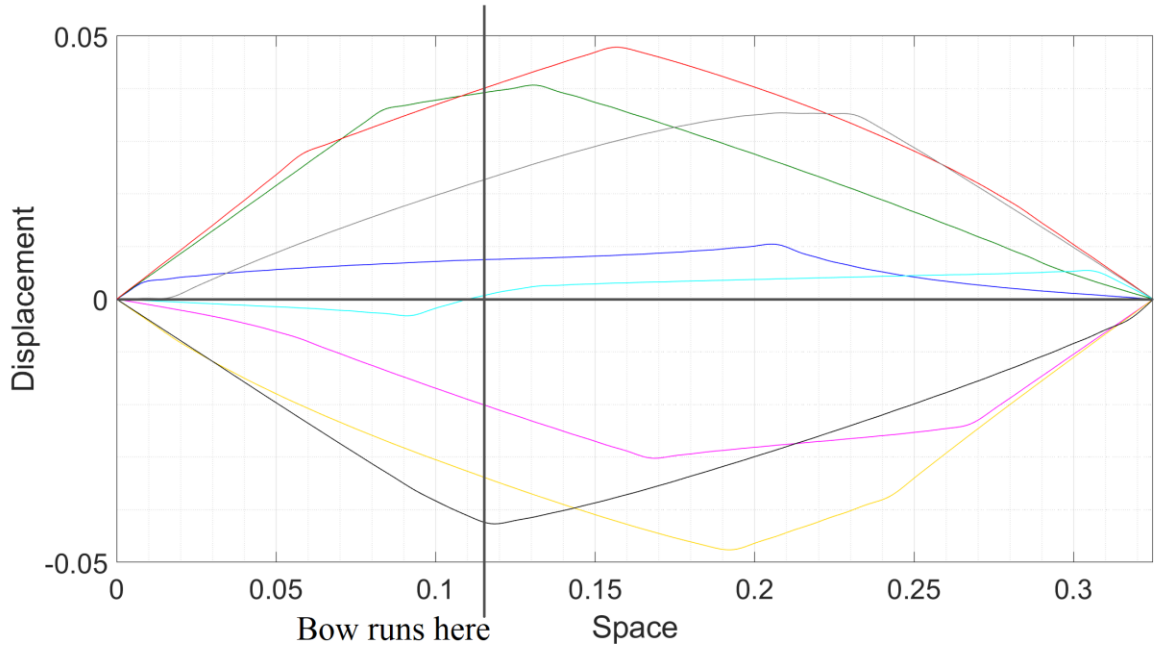
Zooming in on the plot shows that the shapes of the oscillations at different points are quite different. At sites #81 and #150, the shape is close to the triangle wave while further away from the bow and closer to the end, it becomes like a cross between a square wave and a sawtooth wave.



We now plot the velocity of the bow point as a function of time, showing only the zoomed in plot. Clearly, there are periods of stick (velocity equals 0.08 which is the bow speed) alternating with periods of slip (velocity going up and down freely, but not crossing the bow speed).



Finally we show the function $y(x)$ at several time points over one period. The successive time traces are in the standard livery of blue, green, red, grey, cyan, magenta, yellow and black.

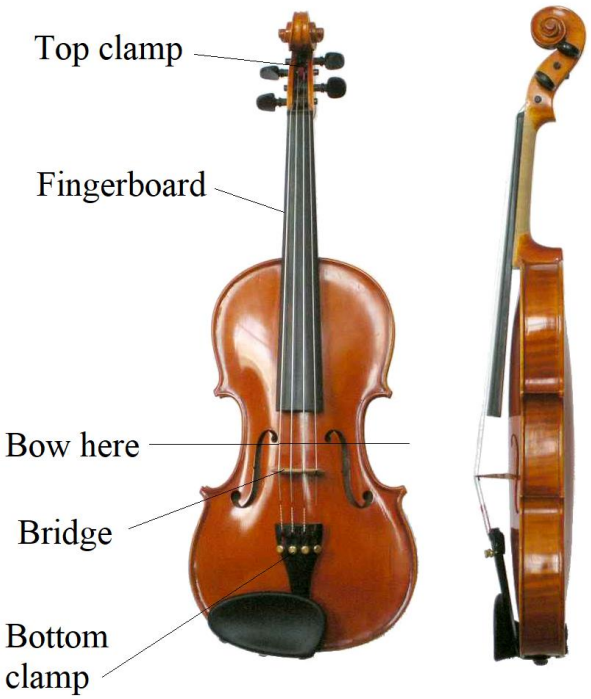


These plots are ample demonstration of the unique nature of PDE limit cycle oscillations in the violin string.

(c) We measure the time taken for 20 oscillations and find the fundamental frequency to be 294 Hz. This is equal to the fundamental natural frequency of the string to within parameter precision.

*

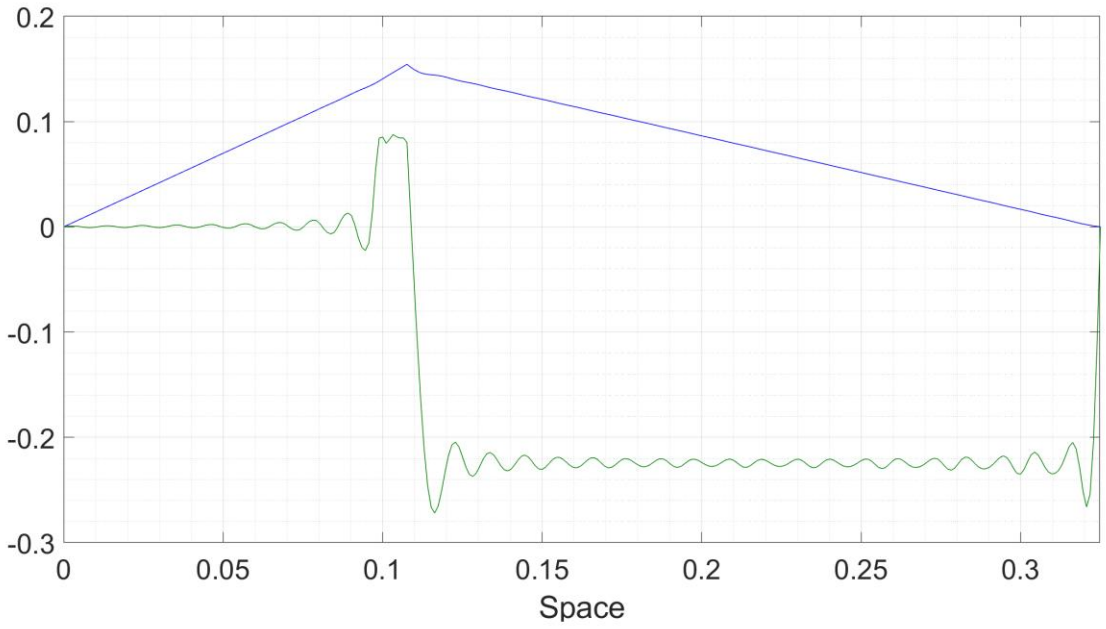
COMMENTS : This is probably the hardest a 1+1 PDE can get. In this case, the low dimensionality is not a modeling assumption but a genuine feature of the system. An actual violin is shown in the Figure below (at the end of all the calculation, which is probably somewhat funny), the image being taken from Wikipedia.



The way the violin is played is as follows. The string is clamped at the top and bottom. The top part runs over the fingerboard – to change the frequency of the note, the player uses her fingers to clamp the wire to any desired position on this board. Near the bottom, the string passes over a bridge which transmits the vibrations to the violin body – the dominant sound output of the violin therefore has the harmonics of the bridge's motion. The string is bowed usually at a point

between the bridge and the fingerboard – here I took it as 1/3 of the distance between the lower and upper clamps (which I had arbitrarily assigned the values 0 and L respectively). The time trace of $x = \#81$ is supposed to represent the oscillations of the bridge (again the position is approximate but indicative). In the model, we took the bridge to be massless, which is an approximation – a heavy bridge CAN be incorporated into the computer routine if one wants but the complexity increases greatly and in a dull way, reducing (at least in my opinion) the clarity of the calculation.

For those states which don't go the limit cycle, some of them end up at the damped oscillation we saw earlier, where the bow doesn't couple. Some more however end up at a state which the computer shows as nearly a steady state. The time trace of any point is almost a constant and the corresponding displacement (blue) and velocity profiles (green) as functions of x look as follows.

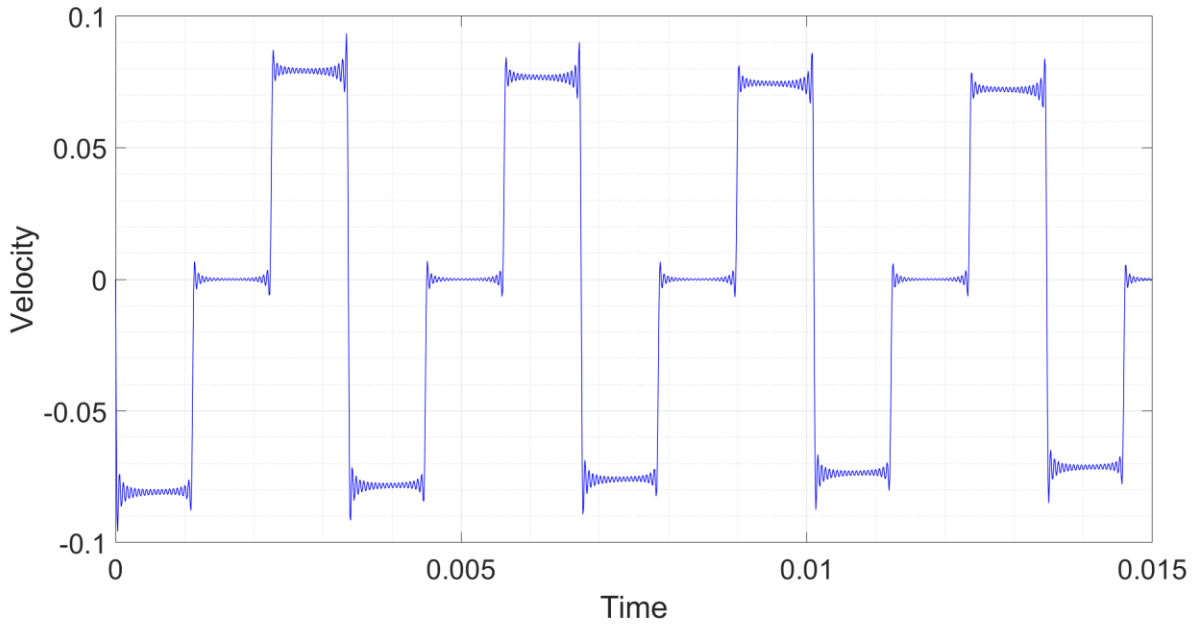


It appears that the computer finds in this case a virtually instantaneous transition from slip to stick and vice versa. According to the computer, the moment the string is released from this IC, the velocity of x_B becomes v_0, while the moment the bow point is dragged up a little bit, the tension difference becomes equal to μF. At present I am not sure of whether this is a genuine quasi-fixed point of the system or whether it is an artefact of the program I have used here (Fourier series convergence issues etc). This state generally manifests if the

bow force is too high, and branches out of the limit cycle which appears initially but fails to stabilize. Since it IS true that excess bow force does not generate the desirable sounds from the instrument, we can probably conjecture without risk of error that even if the fixed point is spurious, the system does converge to a state which is different from the limit cycle and is musically useless.

The small range of parameter space where the limit cycle is stable shows why it is so difficult to play the violin properly – the margin of error in bow force, bow speed etc is quite low. The IC of the question is generated so as to produce a normal component of tension at x_B equal to μF, so that it very nearly does recreate a state where stick transitions to slip. When the string starts to be played, we do expect that the bow will take it to such a configuration before it slips and the dynamics start off in full force.

The parameter values I used are typical ones for the second string of the violin, which is tuned to the note D above middle C. It is quite surprising that the frequency of the limit cycle oscillation should be almost exactly equal to the natural frequency of the string. After all, the full system is forced and the forcing frequency might well have been different from the natural one. A possible reason for this coincidence is that in the limit cycle oscillation, the intervention of the bow plays a marginal role. The bow pushes the string up at constant velocity in a part of the oscillation cycle where it would have been going up anyway.



The above plot bolsters this conjecture. It shows $\dot{y}(x_B,t)$ as a function of time – very similar to the thing I showed a while back except that this time the action of the bow ceases at the initial time of the plot. This of course is a solution of the homogeneous damped wave equation, and the fundamental frequency is going to be that of the lowest mode of the string. The shape of the oscillation is remarkably similar to the previous case, which had the driving function – the primary effect of the sticking was to replace the oscillations near the top by a straight line at the bow velocity. This also ensured that the amplitude of the oscillations stayed constant, because the bow dragged the string to the same displacement in every cycle – here we can see a small decrease of amplitude at this timescale itself and of course the amplitude goes to zero in a macroscopic time interval.

Herein we also see a basic difference between the sound of the violin and of the piano (T112, one of the easiest wave equation problems – listening to a violin-piano double sonata who would guess that one of the instruments is trivial to analyse and the other so complicated). For the piano, the sound is generated by the homogeneous motion of the string – since some damping is always present, the sound invariably dies out some time after the note is struck. It is desirable to have the damping as low as possible so that when required, chords may be held "for as long as the sonority of the instrument permits" (HANS VON BULOW). When the key is lifted, the note stops playing because of a complicated mechanism which rushes up and hits the wire, ceasing its motion instantly. By contrast, the violin is an excited system – the sound is generated by the external forcing of the bow and can be sustained as long as the player wants. The wind instruments are also excited externally by the player's breath. In these instruments, notes can be held indefinitely long, and damping is made quite high so that the note stops sounding quickly after the excitation is removed.

The complicated dynamics of the violin string makes it a system worth further study. And yet, some of the best violins were manufactured by ANTONIO STRADIVARI (1644-1737) and GIUSEPPE GUARNERI (1698-1744), both of whom lived in an age when "partial", "differential" and "equation" were three separate words and separate concepts. We saw this for the Japanese paper pot (T090) and again we see it here – it sometimes doesn't require fancy equations to create a technological marvel.

Yet, that's not a reason to avoid doing the problem academically. For example, it is contended that the Italian masters' violins sound better than modern instruments – although the majority of tests do not show any statistical evidence of this, some trained listeners can indeed tell the difference under blindfolded conditions. If academic knowledge about the violin rises to a sufficiently high level, then it should be possible to factory-make an instrument which has the same sound quality as a Stradivarius or Guarnerius or any other violin of choice. It should also be possible to use cheaper and more environmentally friendly materials (for example, synthetic compounds instead of wood in the body or horsehair in the bow) in the construction of violins and other musical instruments. While individual geniuses such as STRADIVARI and GUARNERI can operate on the basis of intuition alone, a systematic design or development process does indeed require modeling, equations, solutions and simulations. And that's what makes calculations like this one worthwhile.

*

After this mammoth I would really have liked to present an easy problem. However it so happens that the only way to preserve topical continuity of the following few problems is to have a Level 4 in the next slot. A change of topic and a respite from difficulties will come after that.

Example 7.4.5

Resonance box

In many musical instruments, a sound box or resonance box is used to amplify the sound output. This problem presents an analysis of this amplification.

(a) The 3+1 wave equation in spherical coordinates with spherical symmetry (no dependence of the waves on θ and φ) is

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right)=\frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} \quad .$$

Letting $w(r,t)=(1/r)u(r,t)$, show that w satisfies the 1+1 wave equation in r and t.

From the above result it follows that if a point monopole source i.e. a spherically symmetric delta function excitation with the temporal dependence of $\cos\omega t$ is placed at the origin, the waves radiated by it have the form $(q/r)\cos(kr-\omega t)$ where q is the strength of the source and $\omega/k=c$. This result will be useful in what follows.

Consider a situation where the air is subjected to a z-direction vibration alone. Then, letting u be the z-displacement of the air, $u(\mathbf{r},t)$ can be taken to satisfy the inhomogeneous 3+1 wave equation

$$\rho\frac{\partial^2 u}{\partial t^2}-E\nabla^2 u=f(\mathbf{r},t) \quad ,$$

where ρ is the density of the air, E the suitable elastic modulus and $f(\mathbf{r},t)$ the force per unit volume applied externally. For an oscillating point source of strength q, $f(\mathbf{r},t)$ has the form $q\,\delta^{(3)}(\mathbf{r})\cos\omega t$. The boundary conditions on u within a confined volume are that the displacement at a wall is zero, while the air at a hole opened to the atmosphere is stress-free.

(b) A flute is a column of air of length L oriented parallel to the direction of vibration and opened to the atmosphere at both ends. The vibrations may be assumed independent of all coordinates

perpendicular to the flute's axis. Determine its fundamental frequency.

The resonance box is a cube of side length L with a small hole at the centre of the top face. When a source with time dependence $\cos\omega t$ is placed inside the box it creates a particular solution corresponding to vibrations of the air inside the box. In the presence of even an infinitesimal damping, the homogeneous motions decay in time and only these steady state vibrations are significant. These may be obtained by expanding the spatial dependence of the forcing function in the eigenfunctions of the box.

(c) Suppose that an oscillating point source of strength F_0 is mounted at the centre of the resonance box. The frequency ω of the source is extremely close to the lowest eigenfrequency of the box, if one ignores the hole. Determine the steady state vibrations of the air inside the box.

The effect of the hole on the region outside the box can now be taken into account by making the assumption that the air in the hole is stress-free. Then, considering the hole to be absent, one first calculates the stresses on the hypothetically inserted section of wall. One follows up by saying that the solution for the box with the hole equals the solution for a box with the hole patched up and an external stress added, which is the negative of that experienced by the fictitious patch of wall. This excitation causes sound to be radiated outside the box. Far away from the box, it can be treated as a point source whose total strength is the integral of the stress over the hole.

(d) The point source considered in Part (c) is first mounted in open air and then at the centre of the resonance box of side L which has a small hole of area A at the centre of one face. Determine the ratio of the sound amplitudes in the presence and the absence of the box at a point far away from the instrument.

[Level : 4]

With a two-page problem statement and x, y, z and t all on stage, this problem is rather intimidating. However, once we get the hang of the concepts, the manipulations will turn out to be quite simple. So, as usual, we proceed step by step.

(a) This is a variable transformation question and has nothing to do with the main point of the problem (like the initial parts of the Takahashi problem 4.4.4, for those who've read that). So we get rid of this before even thinking about how to do any of the rest. Plugging $w(r,t) = (1/r)\,u(r,t)$ into the wave equation, we get on the LHS

$$\begin{aligned} LHS \rightarrow & \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right) \\ &= \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\left[-\frac{1}{r^2}w + \frac{1}{r}\frac{\partial w}{\partial r}\right]\right) \\ &= \frac{1}{r^2}\left(-\frac{\partial w}{\partial r} + \frac{\partial w}{\partial r} + r\frac{\partial^2 f}{\partial r^2}\right) \\ &= \frac{1}{r}\frac{\partial^2 f}{\partial r^2} \end{aligned}$$

On the RHS, things are very simple – the $1/r$ is unaffected by the differentiation w.r.t t so RHS → $(1/c^2)\,\partial^2 u/\partial t^2 = (1/c^2 r)\,\partial^2 w/\partial t^2$. If we drop the common factor of 1/r from the both sides, we evidently get the 1+1 wave equation for $w(r,t)$. A very easy manoeuvre and some partial credits in our kitty.

(b) This too is nothing fancy – vibrations independent of the perpendicular direction is just a way of saying that the wave equation for the flute is 1+1. The two ends are open to the atmosphere so they are stress-free. From 7.2, this means that $\partial u/\partial x = 0$ at the two ends. The spatial eigenfunctions for this setup we've already seen in 5.1.4 – they are $\cos(n\pi x/L)$. In wave equation , each spatial eigenfunction gets multiplied by $(\cos, \sin)n\pi ct/L$. Hence the eigenfrequency of the n_{th} mode is $n\pi c/L$. The fundamental or lowest frequency is for $n = 1$ and evaluates to $\underline{\pi c/L}$.

This trivial exercise at this point might appear like a distraction, a detour in the logical development of the main problem. However, what it shows is that the stress-free BCs at an open end are extremely plausible – for a flute of length 0.6 m the lowest frequency is 283 Hz, which agrees very well with the known fact

that a flute of close to that length should have a fundamental frequency corresponding to C# above middle C. That a free air column should have this BC at the end is not at all an obvious premise – it's not an elastic bar hanging in air but air in air. So an experimental verification of the BC is a sensible exercise, seeing as we're going to apply it to the resonance box.

(c) Finally, we have the inhomogeneous 3+1 problem rather than a warm-up. The equation is

$$\ddot{u} - c^2\nabla^2 u = \frac{1}{\rho} F_0 \delta\left(x - \frac{L}{2}\right)\delta\left(y - \frac{L}{2}\right)\delta\left(z - \frac{L}{2}\right)\cos\omega t \quad , \tag{01}$$

where we have divided LHS and RHS of the given equation by ρ and have accounted for the fact that the oscillatory point source is at the centre of the box. The BCs are $u = 0$ on all the walls of the box.

To crack this, we begin by generalizing the delta function to some arbitrary function with oscillatory time dependence. So we let the equation be

$$\ddot{u} - c^2\nabla^2 u = f(\mathbf{r})\cos\omega t \quad .$$

Gut says, try particular solution $u_p = R(\mathbf{r})\cos\omega t$. This goes through, giving

$$\left(c^2\nabla^2 + \omega^2\right)R(\mathbf{r}) = -f(\mathbf{r}) \quad , \tag{02}$$

together with the BCs $R(\mathbf{r}) = 0$ on the walls. This doesn't look at all helpful, so the question tells us what to do – we have to break up f into the eigenmodes of the box (this is a 3D extension of what we did in 7.4.3).

So let f now be one of these modes, corresponding to a frequency Ω. We suspect that R will also be the same mode, times some proportionality constant. Now, if R is an eigenmode of the box with eigenfrequency Ω, then $R(\mathbf{r})\cos\Omega t$ satisfies the wave equation $c^2\nabla^2\left(R\cos\Omega t\right) - \partial^2\left(R\cos\Omega t\right)/\partial t^2 = 0$, wherefrom it follows at once that $c^2\nabla^2 R = -\Omega^2 R$. Plugging this into (02) we get

$$R(\mathbf{r}) = \frac{f(\mathbf{r})}{\Omega^2 - \omega^2} \quad . \tag{03}$$

What a similarity between this formula and the response of a single driven harmonic oscillator, which we considered all the way back in 3.1.5 Part (a)! Again, in the case of $\omega = \Omega$ we should be seeing the linearly growing solution. Deriving that might make a good final exam problem but it's not really of

relevance to us because of the inevitable presence of damping, and also of some detunation away from the resonance.

Now, we can expand the triple delta of (01) into the eigenmodes and use the response formula (03) for each of them. At this stage of the course you should be able to write down the box's eigenfunctions by instinct – they are $\sin(n_1\pi x/L)\sin(n_2\pi x/L)\sin(n_3\pi x/L)$ for natural numbers n_1, n_2 and n_3. If you have a confusion about this step, please again go back to 7.1.3 and its references, write out the sep var steps, fit BCs and arrive here.

The corresponding expansion of the delta is also the natural extension of 7.1.3 : we must have

$$\delta\left(x-\frac{L}{2}\right)\delta\left(y-\frac{L}{2}\right)\delta\left(z-\frac{L}{2}\right)$$
$$=\sum_{n_1,n_2,n_3} D_{n_1,n_2,n_3}\sin\frac{n_1\pi x}{L}\sin\frac{n_2\pi y}{L}\sin\frac{n_3\pi z}{L}\,.$$

The eigenfunctions are orthonormal, the normalization constant for each one in each direction is $\sqrt{(2/L)}$ and the corresponding Fourier coefficient is

$$D_{n_1,n_2,n_3}=\frac{8}{L^3}\int_0^L dx\int_0^L dy\int_0^L dz\; g(x,y,z)\sin\frac{n_1\pi x}{L}\sin\frac{n_2\pi x}{L}\sin\frac{n_3\pi x}{L}\quad,$$

where $g(x,y,z)$ denotes the delta. Although this expansion is for a different purpose – this is for solving an inhomogeneous system while the usual stuff is for the IC fit – don't get confused; the process is the exact same. Since g is a triple delta, the above integral is easy to evaluate – it picks out the value of the sines at the position of the delta, which is $(L/2,L/2,L/2)$. Thus we have

$$D_{n_1,n_2,n_3}=\frac{8}{L^3}\sin\frac{n_1\pi}{2}\sin\frac{n_2\pi}{2}\sin\frac{n_3\pi}{2}\quad. \tag{04}$$

Now of course we can calculate the natural frequency Ω_{n_1,n_2,n_3} for each mode and use this decomposition together with (03) to get the answer. However, we are told that the excitation frequency is very close to the lowest eigenfrequency of the box, which corresponds to the mode 1,1,1. Then, the $\Omega^2-\omega^2$ in the denominator of (03) will result in a tremendous amplification of this mode relative to the others and the dominant component of the vibration, for all

practical purposes, will come from the first mode alone. Letting ω_0 be the eigenfrequency of this mode, (04) gives $D_{1,1,1} = 8/L^3$ and then (03) gives the particular solution

$$u_p(x,y,z,t) = \frac{8F_0}{\rho L^3\left(\omega_0^2 - \omega^2\right)} \sin\frac{\pi x}{L} \sin\frac{\pi y}{L} \sin\frac{\pi z}{L} \cos\omega t \quad , \qquad (05)$$

which is the desired answer. It is a ready check that the quantity on the RHS has the dimensions of displacement as indeed it should.

(d) The calculation above would have been an end in itself if we'd been sitting inside the box and listening to the amplified sounds. However, that's not where we are – we are sitting in the auditorium several metres away from the instrument and listening only to what's coming out of the box. For this purpose of course, we need the hole – a closed box would radiate zero. The smallness of the hole ensures that the dynamics inside the box undergoes negligible change – the bulk of the box is still enclosed so (05) will hold to very good accuracy. But the hole will radiate some of the sound from the box to the outside.

It's now that we invoke the fact that an air chamber exposed to atmosphere at a point experiences a stress-free BC there. This of course is an assumption, just as it is for the flute – since it holds so well there, it should work here also. If the box were solid, then the portion of wall where the hole is would have been experiencing some stress – a force would have been required to constrain the displacement to zero there. But in the real problem, there's no stress at the hole. So that problem is like no hole plus an external stress which is the reverse of that experienced on the missing hole. This external excitation gives rise to motions in the region outside the box (the hole-less box contributing zero), and hence produces the sounds that we hear. At considerable distance from the instrument (i.e. distance much greater than instrument dimension), the details of this source are immaterial – all that matters is the total strength. (Recall that the same thing happens in electrostatics – a complicated charge distribution from close up appears like a point charge from far away, as in 6.3.2.) Thus, we can replace the stress over the hole by a delta function force of strength equal to the integral of the stress over the area of the hole.

Since the hole is small, we can treat the stress distribution across it to be constant, equal to its value at the centre. Since the hole is given to be at the

centre of the face $z = L$, the coordinates of the hole centre are $(L/2, L/2, L)$. The strain is the $\partial/\partial z$ of (05) evaluated there; the stress is E times that, which is

$$\sigma = -\frac{8F_0 E}{\pi\rho L^2\left(\omega_0^2 - \omega^2\right)}\cos\omega t \quad .$$

Finally, the total force is the area A times the stress, and the equivalent excitation for the outside is its negative. That evaluates to

$$F = \frac{8F_0 EA}{\pi\rho L^2\left(\omega_0^2 - \omega^2\right)}\cos\omega t \quad . \qquad (06)$$

Now we are at the stage where we can present results. In the absence of the reso box, the sound would have been produced by the point source of strength F_0. In the presence of the box, the equivalent point source has total strength given by (06). Since the wave equation is linear, the amplitude of the sound is obviously proportional to the strength of the source. Hence, the gain achieved by the resonance box is $\underline{8EA/\pi\rho L^2(\omega_0{}^2 - \omega^2)}$.

*

COMMENTS : The resonance box is extremely frequency-selective – it is useless unless the excitation frequency equals the resonance frequency, almost upto an identity. For example, if we take a box of side length 0.5 m, the fundamental frequency evaluates to 3700 rad/s. Let the box have a circular hole of diameter 5 cm. Then substituting typical values for air into the gain formula, we get an amplification factor of 5 if the detunation of the excitation is about 0.1 rad/s off the resonance. So how can the concept have practical utility ?

It happens because of two factors – (i) the box typically utilizes higher modes and not the fundamental mode and (ii) the excitation in a musical instrument invariably has some frequency spread. To amplify the first factor, the higher up the mode scale one goes, the closer their frequencies get. For example, the second lowest mode of the box, which is 1,1,2, has a frequency 40 percent higher than the fundamental. On the other hand, the mode 11,10,10 has a frequency only 3 percent higher than 10,10,10. So if the box is designed such that the source frequency is in the higher mode region, then a randomly chosen

frequency will end up close to a resonance anyway. As regards the second factor, the excitation in a musical instrument is never a pure sine wave but a complicated one (as we saw for the violin). This means that the excitation contains a range of frequencies centred at the basic frequency. So even if the basic frequency is a little detuned from a resonance, there will be some component which will be right on the resonance and gets amplified. Then of course, it is up to the designer of the box to ensure that the resonances of the box correspond to C, D^b, D, E^b etc, the notes of the musical scale. This is achieved by changing the shape and size of the box and the hole.

The big wooden box behind the strings of violins and guitars is a resonance box. The violin has a hole of complicated shape, like an integral sign. The guitar usually has a circular hole. Indeed, one of the factors differentiating a good instrument from a bad one is the design of the box. In a well-designed instrument like STRADIVARI's violin, the box is of just the right size and shape so that the vibrations of the strings make its very walls resound with harmony, melodious or clanging depending on who is playing. In a mediocre violin, the resonances are located away from the pitches of musical notes, resulting in a weak sound of low quality.

We note however that there are many more resonance mechanisms involved in the violin and guitar – two of them are the resonances of the plates making up the box and the Helmholtz resonance of the box (resonance with a vibration mode where air flows in and out of the box via the hole, different from the internal resonances of the box). Violins are usually designed such that the Helmholtz resonance coincides with a musical note in its lower register – the plate resonances are distributed all over the range. The internal resonances we saw here generally lie in the higher register and are useful for amplifying the higher harmonics in the waveform. These harmonics are not insignificant however as they contribute to the difference between a violin and say a piano playing the same fundamental note, as we saw in 7.1.1.

The piano doesn't use a resonance box but a resonance board – the strings are mounted on a plate whose natural frequencies correspond to the notes being played. This is similar to the resonance from the plates of the box in violin and guitar. In wind instruments, the air column itself is the resonance box – the excitation is generated by the player's breath (sometimes assisted by a reed, as

in oboe and clarinet) and contains a range of frequencies. Then, the desired ones are picked up and amplified by the column and radiated out to the audience. This is very similar in principle to the mechanism we discussed here.

Thus we see that resonance in PDE is an extremely important phenomenon in music. The mechanism is quite similar to that of ODEs as the formula (03) shows. But as often happens, a single variable in an ODE problem gets replaced by a whole function in the equivalent PDE – thus, the thing that shows the resonance is not the position of one mass on one spring but a special shape of displacement of the air. This equivalence between numbers in ODEs and functions in PDEs helps us to effectively understand PDE phenomena.

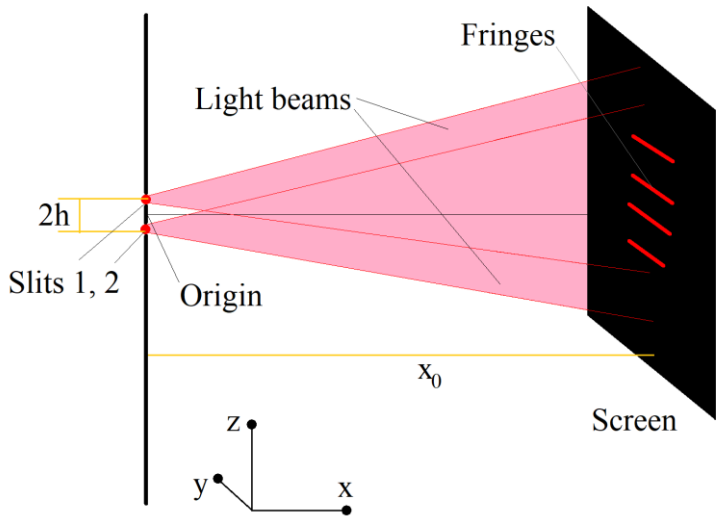
*

Now for a change of track – having had enough of sound waves, we'll go on to light waves. And, the good news is that the opening problem there is really really easy.

Example 7.4.6

Light interference

In YOUNG's double slit experiment, two beams of light radiated from two slits impinge on a screen kept some distance away from it. The setup is shown in the Figure, which also defines the axis conventions and dimensions.



Each slit can be assumed to be radiating a wave whose equation is $u(r,t) = (1/r)\cos(kr-\omega t)$ where r is the radius in spherical polar coordinates centred at the slit. The two waves interfere and generate a pattern on the screen. There is a bright fringe at a point where the amplitudes add completely in phase and double, and a dark fringe where the amplitudes add completely out of phase and cancel.

(a) Will the centre of the screen i.e. the point on the screen with $(y,z) = (0,0)$ be bright or dark ? Justify your answer.

(b) How far along the z-axis must one move from the centre to encounter the first point with the opposite shade ?

[Level : 1]

Thank Uparwala for Level 1.

(a) At the centre point of the screen, both waves have travelled equal distance to get there. Hence, r for both of them is the same and the amplitudes add, producing a bright fringe.

(b) It is reasonable to assume that the size of the fringe pattern will be much smaller than the separation x_0 between slits and screen. Hence, at the location of the first dark fringe, the amplitude factor of $1/r$ will be the same for both waves. Then, we'll get a dark fringe at a point on the screen where the two waves are opposite in phase. At this point, $(kr_1 - \omega t) - (kr_2 - \omega t) = \pi$ where r_1 and r_2 are the distances to the point from the two slits respectively. Of course, the phase difference can be $\pi + 2n\pi$ for any n, but the first dark fringe will occur for the smallest phase difference.

This is the condition which will determine the location of the dark fringe. If y and z be its coordinates on the screen, then

$$r_1 = \left(x_0^2 + y^2 + (z-h)^2 \right)^{1/2} \quad ,$$

$$r_2 = \left(x_0^2 + y^2 + (z+h)^2 \right)^{1/2} \quad .$$

Then, Taylor expanding the surds and setting $r_1 - r_2 = \pi/k$ yields

$$x_0 \left(1 + \frac{y^2 + (z-h)^2}{2x_0^2} - 1 - \frac{y^2 + (z+h)^2}{2x_0^2} \right) = \frac{\pi}{k} \quad .$$

Simplifying this expression, we have $-2zh = \pi x_0 / k$ or, $z = -\pi x_0 / 2kh$ as the location of the first dark fringe. By symmetry of the problem, there will also be a dark fringe at $z = +\pi x_0 / 2kh$. Hence, the distance we have to move from the centre to get to a dark fringe is $\pi x_0 / 2kh$.

*

COMMENTS : THOMAS YOUNG's (same fellow as the elastic modulus) double slit experiment (YDSE) is today almost a triviality – the experiment is one of the first in a basic physics laboratory. But when YDSE was first performed in 1801, neither the process nor the outcome was anything of a joke. The experiment

established for once and for all that light was a wave, settling the age-old and heated dispute between the followers of ISAAC NEWTON (particles or “corpuscles”) and CHRISTIAN HUYGENS (wave). Shooting little balls (Newtonian corpuscles) at a double slit apparatus will obviously not give a fringe pattern but will result in all balls accumulating at two definite locations corresponding to the two slits. So YDSE was an irrefutable and incontrovertible demonstration of the wave nature of light. The details of the original YDSE were also not straightforward because for a concrete result one needs the two waves from the two slits to be perfectly in phase with each other. Two random candles at two points never generate a fringe pattern because their phases relative to each other vary continuously, washing out the pattern and producing a time-averaged addition of intensities rather than amplitudes.

To get around this problem, what YOUNG did was to orchestrate the interference between the two parts of the same beam of sunlight which he split using a mirror and routed onto the screen. Today, the thing is far simpler because we use a laser beam, which, by definition of laser comes entirely in the same phase.

So where did the wave equation actually enter the picture here ? It did when I quoted the form of the light waves $(1/r)\cos(kr-\omega t)$. You can easily check that if $\omega/k = c$ then this satisfies the 3D wave equation in spherical coordinates $\nabla^2 u = \left(1/c^2\right)\ddot{u}$ (we already did a more general calculation in the last Tute but a four can’t be prereq for a single). Its interpretation as a wave which is moving radially outwards with diminishing amplitude is quite transparent, which is why we learn about YDSE much earlier than we learn about the wave equation itself.

A century after the original YDSE, MAX PLANCK confused the issue of wave vs. particle by declaring light to be made up of quantum particles called photons. So how can they show interference ? They can because of the uncertainty in position of a quantum particle. If you shoot photon after single photon at a double slit apparatus, then one will end up here, another will end up there, according to some quantum mechanical probability distribution. In the experiments, photons are usually identified by a flash or scintillation on the screen at the location where it strikes. You send a few photons and the scintillation pattern will appear random. You send sufficiently large number of

photons and there will be maximum hits at the locations of the classical bright fringes and minimum hits at the location of the classical dark fringes. As the number of photons tends to infinity, the fraction of photons received in the classical bright fringe area tends to unity, and the fraction of photons in the classical dark area to zero.

Example 7.4.7

Localizer – aircraft landing system

A point monopole source of waves of strength q oscillating with frequency ω generates waves which have the form $(q/r)\cos(kr-\omega t)$ where $\omega = k/c$ and r is the radius in spherical coordinates centred at the source. A dipole source consists of two monopole sources $+q\cos\omega t$ and $-q\cos\omega t$ located at the points $(0,0,h)$ and $(0,0,-h)$ respectively. The separation $2h$ is significantly smaller than the wavelength of the emitted waves.

(a) Determine the time-averaged intensity of the waves at a point located far away from the sources. Express your answer in terms of spherical coordinates centred at the origin.

A vehicle moves about on the plane $z = \text{const.}$ far away from the sources. It carries on board a sensor which measures the intensity of the waves and guides the vehicle to the point of maximum intensity.

(b) Where does the vehicle end up asymptotically ? Justify your answer.

[Level : 2]

(a) If you read 7.4.5, you of course know the wave field of a point monopole source already – I had to repeat it here because of our prerequisite rules. To find the amplitude of the waves at a remote point as a result of the dipole source, we must add the waves coming from the two monopole sources. Thus, we must calculate the sum

$$u(\mathbf{r},t) = q\left(\frac{1}{r_1}\cos(kr_1 - \omega t) - \frac{1}{r_2}\cos(kr_2 - \omega t)\right) \quad , \qquad (01)$$

where

$$r_{1,2} = \left(x^2 + y^2 + (z \mp h)^2\right)^{1/2} \quad ,$$

are the distances of the observation point from the positive and negative sources respectively. Letting $r = (x^2 + y^2 + z^2)^{1/2}$, we do some quick Taylor series expansions and find

$$r_{1,2} = r\left(1 \mp \frac{zh}{r^2}\right) \quad ,$$

$$\frac{1}{r_{1,2}} = \frac{1}{r}\left(1 \pm \frac{zh}{r^2}\right) \quad ,$$

correct to first order in h/r which is a small parameter.

Now if we expand the sum (01) in a series of h/r, we'll get two kinds of terms – one where we treat the $1/r$ as different for the two sources and the cosines as equal for both and the other where we treat the $1/r$ as equal and the cosines as different (we expect that terms where both are different will be smaller in size than these two). Terms of the first kind go as $1/r_1 - 1/r_2$ which is of size zh/r^3. The h is just a constant; since $z = r\cos\theta$, the dependence is $(\cos\theta)/r^2$. Indeed, this term is like the static potential of an electric dipole (two equal and opposite sources and $1/r$ contribution from each), so the $1/r^2$ dependence is no surprise.

On the other hand, the term of the second kind is

$$u(\mathbf{r},t) = \frac{q}{r}\left[\cos(kr_1 - \omega t) - \cos(kr_2 - \omega t)\right] \quad .$$

The trig review tells us

$$u(\mathbf{r},t)=\frac{2q}{r}\sin\left(\frac{k(r_1+r_2)-2\omega t}{2}\right)\sin\left(\frac{k(r_2-r_1)}{2}\right) \quad .$$

Using the Taylor expansion formulae, the first sin is just $\sin(kr-\omega t)$; the second one is $\sin(kzh/r)$ which is kzh/r. Then, using $z=r\cos\theta$ this gives

$$u(\mathbf{r},t)=\frac{2qhk}{r}\cos\theta\sin(kr-\omega t) \quad . \qquad (02)$$

This is a $1/r$ term, and at large distances is infinitely larger than the $1/r^2$ term we saw previously. Hence at those distances, which are asked for in the question, (02) is the formula for u.

Finally, we need the time-averaged intensity and not the amplitude. At every point, u has a cos or sin time dependence; squaring will give $\cos^2\omega t$ or $\sin^2\omega t$ whose long time average is 1/2. Hence the averaged intensity is

$$* \qquad |u|^2_{\text{avg}}=\frac{2q^2h^2k^2}{r^2}\cos^2\theta \quad , \qquad *$$

which is the desired answer.

(b) The vehicle sits on the "x-y" plane in far field, hunting for the maximum intensity. Evidently, this max occurs at $\theta=0$ i.e. on the z-axis. Hence the vehicle ends up asymptotically on the z-axis.

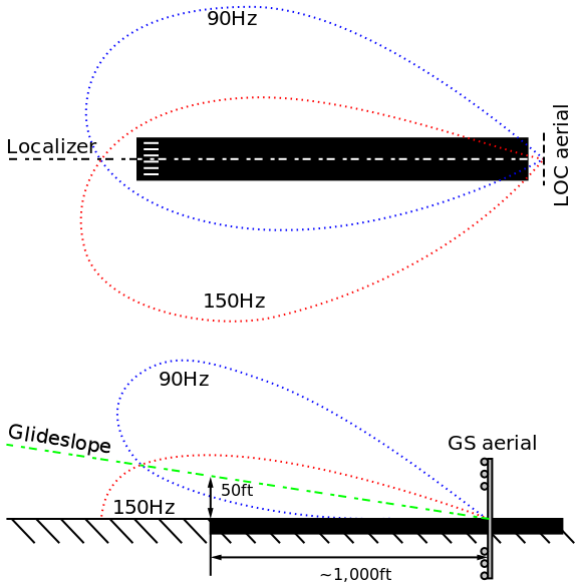
*

COMMENTS : The electromagnetic fields radiated by an oscillating dipole go as $1/r$. This is an extremely long-ranged dependence – the potential of a static dipole goes as $1/r^2$ and the field as $1/r^3$. This property plays a critical role in what is probably one of the most spectacular applications of dipole radiation in science and technology. That is of course the problem topic, the localizer in the instrument landing system of an aircraft.

The localizer is what ensures that an aircraft approaching an airport from high in the sky invariably ends up at the touchdown point on the runway in one smooth descent – no last-minute climbs, drops or swerves, let alone go-arounds. It is a dipole source located just beside the runway at the touchdown point. It projects into the sky a beam of light waves with the maximum intensity being along a line parallel to the runway in the horizontal plane and having a (usually) 3 percent vertical slope. This line is called GLIDESLOPE. The beam is strong

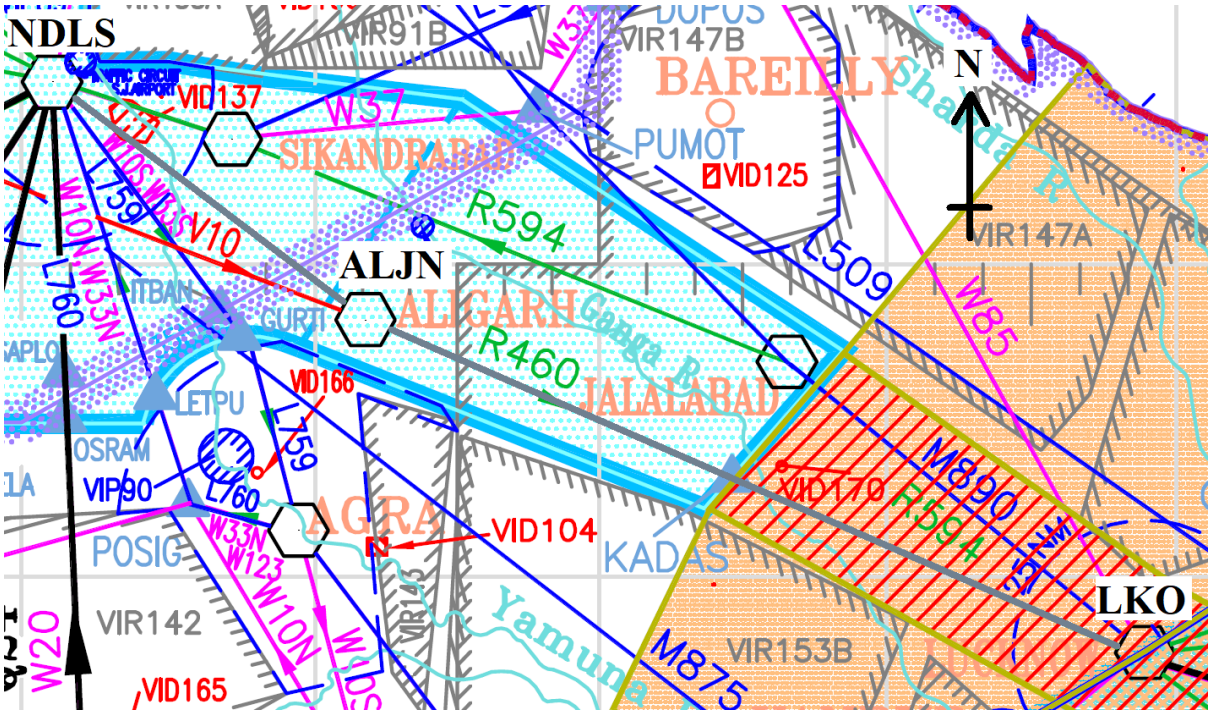
enough to be detected many kilometres away. A sensor just like the one in Part (b) measures the intensity and directs the pilot to the region of maximum intensity. When the aircraft is a few kilometres away from the airport, the radar controllers take it to a point near the glideslope where it can catch the waves from the localizer – thereafter the sensor continually forces the pilot to align with the peak and hence with the slope. This ensures that the aircraft comes smoothly down the slope.

Well, almost. There IS a little bit of simplification in the presentation I gave above – the real thing has two such dipole waves, each at a different frequency, and the glideslope is characterized by equal intensity of both frequencies. The physics behind the thing of course remains the same – the $1/r$ dependence and $\cos^2\theta$ distribution of intensity of the dipole waves which results in a sharp peak along the slope itself, projecting for an enormous distance. A schematic of this is shown in the below Figure, taken from Wikipedia.



The balloon-like intensity curve is in fact the hallmark of the dipole radiation profile.

The monopole and dipole waves also have a tremendous application in aircraft navigation at their cruising altitude. They are what ensure that planes don't deviate from their planned courses, don't get lost up at 35000 feet. In overland sections (country airspaces), dedicated corridoors called ATS routes are defined along which planes are intended, and constrained, to fly. For example, here is a section of Indian airspace, in the Uttar Pradesh region.



The pic (a pilot's map) has a lot of junk which you should ignore. Focus on the cities of Nayi Dilli (NDLS), Aligarh (ALJN) and Lucknow (LKO). The ATS route R460 is shown as a thick slate grey line connecting these cities. All flights in this region must mandatorily travel along R460 – the captain cannot take any arbitrary path through the air (R460 continues to Kolkata via Varanasi, Gaya, Dhanbad and is the standard route for all Dilli-Kolkata flights). This of course raises the question, how do we define R460 – it's not that there are concrete highways or railway tracks in the sky.

The way we define the route (and keep planes on it) is through a device called VOR (very-high-frequency omnidirectional range-finder). Dilli, Aligarh and Lucknow all have VORs, shown in the map as hexagons. VOR actually consists of a monopolar and a dipolar source which radiate electromagnetic waves far out into the air (several hundred kilometres). As we saw here, a monopole radiates waves at equal intensity in all directions. A dipole on the other hand radiates a strongly directional beam with a sharp intensity peak along the dipole axis (the balloon shape). In VOR, both the monopole and dipole radiate at the same very high frequency (about 100 MHz). The monopole wave is then modulated by a sinusoidal carrier wave at 30 Hz. The dipole on the other hand is mounted in the horizontal plane, modulated sinusoidally at 30 Hz and rotated about a vertical

axis at 30 cycles per second, so that the intensity balloon is oriented due North at one instant, due East at the next, then due South, then West and so on. Wherever an aircraft may be, it receives the monopole waves continuously and the dipole waves once every 1/30 of a second, when the balloon hits it. Now, the phases of modulation and rotation are adjusted such that at any location θ degrees clockwise from North measured from the VOR, the phase difference between the modulated monopolar and dipolar waves is θ. These waves are picked up by instruments on board the aircraft, which measure the phase difference and thus immediately give the angle of the aircraft's current position relative to the VOR.

An ATS route now gets defined as the straight line connecting two VORs. For example, the first section of R460 is NDLS-ALJN. On this line, the angle from NDLS is 127° (remember, measured clockwise from North – for some reason the aviation people don't measure anticlockwise from East) while the angle from ALJN is 307°. While flying this section, the aircraft instruments are tuned to pick up the signals from both NDLS and ALJN VOR; if the angles are exactly 127° and 307° then it's on R460; otherwise it needs to correct course. After passing Aligarh, a Down flight will let go of NDLS and catch onto LKO VOR instead. This time, the two angles need to be 114° and 294° ex-ALJN and ex-LKO respectively. In this way, a brilliant combination of monopole and dipole waves keep aircraft on their designated courses in the sky at all times. ATS routes break down when crossing oceans – in this case the pilots rely on internal navigation systems (gyroscopes etc) to stay on course until they can catch onto signals from VOR mounted on the shore. Nowadays, VOR and internal navigation are being replaced by GPS in a lot of sections, especially in the United States; for this form of navigation, none of this discussion is relevant. Nonetheless, VOR and ATS routes remain very much in existence in many parts of the world and VOR is expected to serve as backup to GPS in all foreseeable future.

As we saw on a few occasions before (like 1.2.2, 3.3.9 and even 4.4.4), a technological marvel is often not a virtuoso exhibition of back -breaking mathematics – it is likelier to be a simple concept applied in a novel and ingenious way . Mathematical tours de force are however more likely than simple but profound applications to guarantee glory to their creators . The localizer was invented by an American named ANDREW ALFORD. An Internet

search for him reveals far more results about a certain businessman. In the same way, search results for ISAO TAKAHASHI generally begin with a Taekwondo instructor. By comparison, scientists who work in mathematically ferocious subjects which are incomprehensible to all but specialists often enjoy celebrity status among the common people. This is very unfortunate but it's true.

Coming back to the maths, our treatment of an oscillating electric or magnetic dipole as two monopole sources of waves is undoubtedly a little crude. For one, what is u ? Potential ? Any one component of electric field ? One of magnetic field ? Which one ? Also, MAXWELL's equations feature point charges and point currents but not point "wave sources". Applicationally, why does VOR have megahertz waves modulated down to 30 Hz – why not simply radiate waves at 30 Hz ? All this begs for a more realistic calculation of the radiation fields of an electromagnetic source. So here it comes.

Example 7.4.8

Dipole radiation and magnetic levitation

A spherical shell of radius R centred at the origin carries a surface current $\mathbf{K} = K_0 \sin\theta \cos\omega t \hat{\boldsymbol{\varphi}}$ where ω is completely arbitrary.

(a) Starting from MAXWELL's equations, write a PDE for $\mathbf{E}(\mathbf{r},t)$. Separate variables, substitute the appropriate temporal and angular dependences and verify that the radial equation is solved by the two functions

$$f(kr) = \frac{\cos kr}{kr} - \frac{\sin kr}{k^2 r^2}, \qquad g(kr) = \frac{\sin kr}{kr} + \frac{\cos kr}{k^2 r^2} \quad ,$$

with $k = \omega / c$.

The BCs on the PDE are the usual conditions on **E** and **B** at the sphere surface and the constraint that outside the sphere, there must be outgoing waves only. After imposing these BCs, limit is taken that $R \to 0$ and $K_0 \to \infty$ keeping the dipole moment $m = 4\pi K_0 R^3 / 3$ a constant (reference may be drawn to exampld 6.3.5).

(b) Determine $\mathbf{E}(\mathbf{r},t)$ everywhere outside the dipole.

(c) Determine $\mathbf{B}(\mathbf{r},t)$ everywhere outside the dipole and verify that the result is consistent with 6.3.5 in the limit $\omega \to 0$.

The dipole is now mounted at the point $(0,0,h)$ and the region $z < 0$ filled with an electrical conductor of conductivity σ.

(d) Arguing as in 6.3.8 Part (b), show that the electric field penetrates the conductor to a depth which varies inversely as ω.

(e) Treating the conductor as a perfect flux-expeller and using an image construction, determine the time-averaged force acting on the dipole when h is large.

[Level : 6]

When I'd written the original draft of this book, this problem wasn't there – I merged two Level 1's in an earlier Section to create the extra slot. I find it somewhat surprising that the climactic problem of this entire second Part of the book should come almost as an afterthought. In the earlier version, the resonance box problem (now example 7.4.5) occupied position 7 and acted as a pseudo-climax – for all its 3+1 spatiotemporal dynamics, it somehow lacked enough weight to act as a capstone PDE example.

As usual, we start with an outline. Not much to say here, since it's a short problem statement going the whole way – must be the actual calculations which are hard. But basically, there's a dipole sitting above a metal sheet – task is to show that the metal acts like a flux-expeller and generates a lift on the dipole. This is very similar to the configuration of example 6.3.8 – the only difference is that there the dipole oscillated slowly making the slow electrodynamic approximation was valid, while here the dipole oscillates rapidly and the complete dynamics becomes relevant. Presumably, there will be some important quantitative differences from 6.3.8 in the final answer, which is why we're embarking on this difficult exercise in the first place.

(a) In 6.3, I'd pulled off a manoeuvre which had required "beating a retreat". That manoeuvre had led to a wave equation and we'd come out of it by assuming the frequency to be slow. This time, we are charged with exploring the territory from where we'd backed out then. The only source of charge or current present in the system is oscillating shell, which can be described as $\mathbf{J}(\mathbf{r},t)=K_0\delta(r-R)\sin\theta\cos\omega t\hat{\boldsymbol{\varphi}}$. MAXWELL's equations then read as

$$\nabla\cdot\mathbf{E}=0 \quad , \tag{01a}$$

$$\nabla\times\mathbf{E}=-\frac{\partial}{\partial t}\mathbf{B} \quad , \tag{01b}$$

$$\nabla\cdot\mathbf{B}=0 \quad , \tag{01c}$$

$$\nabla\times\mathbf{B}=\mu_0\varepsilon_0\frac{\partial}{\partial t}\mathbf{E}+\mu_0K_0\delta(r-R)\sin\theta\cos\omega t\hat{\boldsymbol{\varphi}} \quad . \tag{01d}$$

Now do as in 6.3; curl the second one, kick out the divergence and plug the fourth one. That gives

$$\nabla^2\mathbf{E}-\frac{1}{c^2}\frac{\partial^2\mathbf{E}}{\partial t^2}=\mu_0\omega K_0\delta(r-R)\sin\theta\sin\omega t\hat{\boldsymbol{\varphi}} \quad , \tag{02}$$

where we recall that $\mu_0\varepsilon_0 = 1/c^2$. So, $\mathbf{E}$ satisfies a homogeneous 3+1 wave equation everywhere except at the surface of the shell – as in 6.3.2 and 6.3.4-5 , we expect that the shell will provide BCs which match the internal and external solutions . Since the source term is entirely in $\hat{\boldsymbol{\varphi}}$ direction , we expect that the net $\mathbf{E}$ too will be in that direction, although the equation (02) itself will hold only for Cartesian components of $\mathbf{E}$ (again recall 6.3.4-5).

The homogeneous 3+1 wave equation in spherical coordinates takes the cake as the nastiest PDE in the book :

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial E}{\partial r}\right)+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial E}{\partial\theta}\right)+\frac{1}{r^2\sin^2\theta}\frac{\partial^2 E}{\partial\varphi^2}-\frac{1}{c^2}\frac{\partial^2 E}{\partial t^2}=0 \quad ,$$

where scalar E denotes any Cartesian component. As usual, open out the derivatives (see 6.1.5), break up $E = R(r)\,\Theta(\theta)\,\Phi(\varphi)\,T(t)$ and divide by $R\Theta\Phi T$. That gives

$$\frac{1}{R}\left(R_{rr}+\frac{2}{r}R_r\right)+\frac{1}{r^2}\frac{1}{\Theta}\left(\Theta_{\theta\theta}+\frac{\cos\theta}{\sin\theta}\Theta_\theta\right)+\frac{1}{r^2\sin^2\theta}\frac{1}{\Phi}\Phi_{\varphi\varphi}-\frac{1}{c^2}\frac{1}{T}\ddot{T}=0 \quad .$$

The T term here is an obvious constant – moreover, since the excitation is periodic with frequency ω, we expect that T also will be $(\cos,\ \sin)\,\omega t$ and the constant evaluates to $-\omega^2$. Then,

$$\frac{1}{R}\left(R_{rr}+\frac{2}{r}R_r\right)+\frac{1}{r^2}\frac{1}{\Theta}\left(\Theta_{\theta\theta}+\frac{\cos\theta}{\sin\theta}\Theta_\theta\right)+\frac{1}{r^2\sin^2\theta}\frac{1}{\Phi}\Phi_{\varphi\varphi}+\frac{\omega^2}{c^2}=0 \quad .$$

Next, multiply by $r^2\sin^2\theta$ and the Φ thing breaks off into

$$\frac{\mathrm{d}^2\Phi}{\mathrm{d}\varphi^2}-\lambda_1\Phi=0 \quad . \qquad (03)$$

Then divide what remains by $\sin^2\theta$ to get

$$\frac{1}{R}\left(r^2R_{rr}+2rR_r\right)+\frac{1}{\Theta}\left(\Theta_{\theta\theta}+\frac{\cos\theta}{\sin\theta}\Theta_\theta\right)+\frac{\lambda_1}{\sin^2\theta}+\frac{\omega^2r^2}{c^2}=0 \quad .$$

The Θ thing now severs ties with the rest :

$$\frac{\mathrm{d}^2\Theta}{\mathrm{d}\theta^2}+\frac{\cos\theta}{\sin\theta}\frac{\mathrm{d}\Theta}{\mathrm{d}\theta}+\left(\frac{\lambda_1}{\sin^2\theta}-\lambda_2\right)\Theta=0 \quad , \qquad (04)$$

leaving behind an ODE for R :

$$r^2\frac{\mathrm{d}^2R}{\mathrm{d}r^2}+2r\frac{\mathrm{d}R}{\mathrm{d}r}+\left(k^2r^2+\lambda_2\right)R=0 \quad , \tag{05}$$

where $k=\omega/c$ and the equation has separated fully. Note that (03) and (04) are the same as what we got in example 6.1.5 – only (05) is different , and less friendly . (It is NOT a foregone or obvious conclusion that the extra t dependence in the wave equation vis-a-vis Laplace equation would change only the R equation and not the angular ones – because of this we couldn't take any shortcut to (03-5) without risking error.)

Now, plug in what we already know. $\sin\theta(\cos,\sin)\varphi$ are solutions of the angular equations (03) and (04); since (02) features just this type of source term, we expect that E_x and E_y will also have the same angular dependences. (Every component of $\mathbf{B}$ also satisfies a wave equation, but the calculation is bit simpler if we go for $\mathbf{E}$ first.) From 6.3.5 Part (b), the constants λ_1 and λ_2 evaluate to -1 and -2 respectively. Then, plugging this into (05) gives

$$r^2\frac{\mathrm{d}^2R}{\mathrm{d}r^2}+2r\frac{\mathrm{d}R}{\mathrm{d}r}+\left(k^2r^2-2\right)R=0 \quad , \tag{06}$$

which is the radial equation.

For the next step, we are given the solutions and merely asked to verify them. I'll do it explicitly for only one of them, which is $R=f(kr)$. We have

$$f(kr)=\frac{\cos kr}{r}-\frac{\sin kr}{k^2r^2} \quad ,$$

$$\frac{\mathrm{d}f}{\mathrm{d}r}=-\frac{\sin kr}{r}-\frac{2\cos kr}{kr^2}+\frac{2\sin kr}{k^2r^3} \quad ,$$

$$\frac{\mathrm{d}^2f}{\mathrm{d}r^2}=-\frac{k\cos kr}{r}+\frac{3\sin kr}{r^2}+\frac{6\cos kr}{kr^3}-\frac{6\sin kr}{k^2r^4} \quad .$$

Plugging these into (06) results in a zero, showing that $f(kr)$ indeed satisfies the ODE. A similar and equally boring calculation shows that $g(kr)$ also satisfies the ODE, and the check is complete.

(b) The BCs on $\mathbf{E}$ and $\mathbf{B}$ at the shell are that $\mathbf{E}$ is continuous (the shell is not a CHARGE sheet), the normal component of $\mathbf{B}$ is also continuous and the tangential component of $\mathbf{B}$ has a jump proportional to the surface current. Now, $\mathbf{E}$ has two components, a $\cos\omega t$ and a $\sin\omega t$, both of which have the spatial dependences $(f,g)\,kr\sin\theta\,(\cos,\sin)\,\varphi$. Since the current is $\cos\omega t$, the jump in $\mathbf{B}$

comes only from its $\cos\omega t$ component; since $\mathbf{E}$ depends on $\partial\mathbf{B}/\partial t$, the jump will come from the $\sin\omega t$ component of $\mathbf{E}$. So let's focus on this component first.

The natural thing to ask is, are f and g both physically plausible solutions inside and outside the shell. Outside, they both are, since both go as $1/r$. To see inside, let's Taylor expand the trigs. For f we have

$$\begin{aligned} f(kr) &= \frac{\cos kr}{kr} - \frac{\sin kr}{k^2r^2} \\ &= \frac{1}{kr}\left(1 - \frac{1}{2}k^2r^2 + \ldots..\right) - \frac{1}{k^2r^2}\left(kr - \frac{1}{6}k^3r^3 + \ldots..\right) \quad , \\ &= -\frac{1}{3}kr \end{aligned}$$

so it is tractable. For g on the other hand, the $(\cos kr)/k^2r^2$ is an obvious troublemaker at $r=0$. Now we account for the fact that the φ-dependences of E_x and E_y are such as to generate the resultant $\mathbf{E}$ in a circumferential direction. Then, we can write

$$\mathbf{E}_1 = A_1 f(kr)\sin\theta\sin\omega t\hat{\boldsymbol{\varphi}} \quad , \tag{07a}$$

$$\mathbf{E}_2 = \left(A_2 f(kr) + B_2 g(kr)\right)\sin\theta\sin\omega t\hat{\boldsymbol{\varphi}} \quad , \tag{07b}$$

inside and outside the shell respectively. This is a partial $\mathbf{E}$ since it's the $\sin\omega t$ component only, nevertheless, I'll just write it as $\mathbf{E}$ instead of $\mathbf{E}^{\sin}$ or something, which looks too cumbersome. Please keep this in mind.

Continuity of $\mathbf{E}$ at the shell $r=R$ (the variable R now onwards denotes the shell radius and not the separable function of r) gives the equation

$$A_1 f(kR) = A_2 f(kR) + B_2 g(kR) \quad , \tag{08}$$

and we have one relation connecting A_1, A_2 and B_2.

Now we account for the jump in $\mathbf{B}$. Its component tangent to the shell is of course B_θ; from (01b) and (07) we have inside the shell

$$-\frac{\partial B_{\theta 1}}{\partial t} = -A_1\left(\frac{f(kr)}{r} + kf'(kr)\right)\sin\theta\sin\omega t \quad ,$$

where f' means the derivative of f w.r.t kr and we have of course looked at the multivariate calculus review to get that spherical curl. This implies that

$$B_{\theta 1} = -\frac{A_1}{\omega}\left(\frac{f(kr)}{r} + kf'(kr)\right)\sin\theta\cos\omega t \quad .$$

Likewise, outside the shell,

$$B_{\theta 2} = \left[-\frac{A_2}{\omega}\left(\frac{f(kr)}{r} + kf'(kr)\right) - \frac{B_2}{\omega}\left(\frac{g(kr)}{r} + kg'(kr)\right)\right]\sin\theta\cos\omega t \quad . \qquad (09)$$

The BC implies

$$\left[B_{\theta 2} - B_{\theta 1}\right]_{r=R} = \mu_0 K_0 \sin\theta\cos\omega t \quad .$$

The current is evidently consistent in functional form with the internal and external B_θ's we just found. Imposing the match yields

$$\begin{aligned} &A_1\left(\frac{f(kR)}{R} + kf'(kR)\right) - A_2\left(\frac{f(kR)}{R} + kf'(kR)\right) \\ &-B_2\left(\frac{g(kR)}{R} + kg'(kR)\right) = \mu_0 K_0 \omega \end{aligned} \quad , \qquad (10)$$

and we have a second relation between the three unknown constants.

If you thought continuity of the radial component B_r would yield the third, you're mistaken. To calculate B_r we use (01b) and (07) for a second time to get

$$B_{r1} = \frac{2A_1 f(kr)}{\omega r}\cos\theta\cos\omega t \quad , \qquad (11a)$$

$$B_{r2} = \left(\frac{2A_2 f(kr)}{\omega r} + \frac{2B_2 g(kr)}{\omega r}\right)\cos\theta\cos\omega t \quad . \qquad (11b)$$

Equating the two at $r = R$ leads to the condition (08) again – things would have become TOO easy if the final answers came tumbling out already. Instead, right now we're at three unknowns and two equations.

To make progress, we bring the $\cos\omega t$ component of $\mathbf{E}$ into focus. Like (07), this must be having the form

$$\mathbf{E}_1 = C_1 f(kr)\sin\theta\cos\omega t\hat{\boldsymbol{\varphi}} \quad , \qquad (12a)$$

$$\mathbf{E}_2 = \left(C_2 f(kr) + D_2 g(kr)\right)\sin\theta\cos\omega t\hat{\boldsymbol{\varphi}} \quad , \qquad (12b)$$

inside and outside. $\mathbf{E}$ is continuous at the shell so

$$C_1 f(kR) = C_2 f(kR) + D_2 g(kR) \quad .$$

The $\cos\omega t$ component of $\mathbf{E}$ generates a $\sin\omega t$ component of $\mathbf{B}$; since the current does not have this component, B_θ is continuous also. The BC is the same as (10)

with zero in place of $\mu_0 K_0 \omega$ (there might be some overall minus signs involved, but, since the equation is homogeneous, it doesn't matter). So we write

$$C_1\left(\frac{f(kR)}{R}+kf'(kR)\right)-C_2\left(\frac{f(kR)}{R}+kf'(kR)\right)$$
$$-D_2\left(\frac{g(kR)}{R}+kg'(kR)\right)=0 \quad ,$$

and we've doubled up to six unknowns, four equations; no apparent improvement from (08,10).

However, there is another BC up our sleeve – the outgoing wave condition which we've not used so far. Outgoing waves look like $\cos(kr-\omega t)$ and $\sin(kr-\omega t)$, no trigs of $kr+\omega t$ allowed. Outside the dipole, the largest order terms in both (07b) and (12b) have a radial dependence of $1/r$. At this order, (07b) yields $A_2 \cos kr \sin\omega t + B_2 \sin kr \sin\omega t$, while (12b) yields $C_2 \cos kr \cos\omega t + D_2 \sin kr \cos\omega t$. To bunch them all up into trigs of $kr-\omega t$ we must have $D_2 = -A_2$ and $C_2 = B_2$. Bingo ! We now have a full set of equations ! Before writing them out all together, let me define $F(kR)$ as $f(kR)/R + kf'(kR)$ and $G(kR)$ as $g(kR)/R + kg'(kR)$, so that the equations don't spill into two lines for no valid reason. Then we have

$$f(kR)A_1 - f(kR)A_2 - g(kR)B_2 = 0 \quad , \qquad (13a)$$
$$F(kR)A_1 - F(kR)A_2 - G(kR)B_2 = \mu_0 K_0 \omega \quad , \qquad (13b)$$
$$g(kR)A_2 - f(kR)B_2 + f(kR)C_1 = 0 \quad , \qquad (13c)$$
$$G(kR)A_2 - F(kR)B_2 + F(kR)C_1 = 0 \quad , \qquad (13d)$$
$$C_2 = B_2 \quad , \qquad (13e)$$
$$D_2 = -A_2 \quad . \qquad (13f)$$

We can solve this system to determine the unknown constants and hence $\mathbf{E}$ everywhere.

To simplify the solution, we now use the fact that the dipole has R vanishingly small and K_0 infinitely large, so that $K_0 R_0^3 = 3m/4\pi$ where m is the dipole moment (note that the given dipole is in the $\hat{\mathbf{z}}$ direction). At really small kR, the asymptotic forms $f(kR) = -kR/3$ and $g(kR) = 1/k^2R^2$ are relevant, and (13a-d) become

$$-\frac{k}{3}A_1+\frac{k}{3}A_2-\frac{1}{k^2R^3}B_2=0 \quad ,$$
$$-\frac{2k}{3}A_1+\frac{2k}{3}A_2+\frac{1}{k^2R^3}B_2=\frac{3\mu_0\omega m}{4\pi R^3} \quad ,$$
$$\frac{1}{k^2R^2}A_2+\frac{kR}{3}B_2-\frac{kR}{3}C_1=0 \quad ,$$
$$\frac{1}{k^2R^3}A_2-\frac{2k}{3}B_2+\frac{2k}{3}C_1=0 \quad .$$

We can of course solve this system outright but some insight saves a little tedium. Since R will be shrunk to zero while k remains as is, the above equations must be consistent in powers of R. This implies that B_2 and C_1 are of order unity while A_1 and A_2 are both of order $1/R^3$. Then, for dimensional consistency, A_1 and A_2 must both be of the form k^2/R^3. Finally, we can set $k=1$, replace the RHS of the second equation by 1, solve the resulting system (which now has only numerical coefficients) on Matlab and plug the algebraic parameters back into the results. What we get is

$$A_1=-\frac{3\mu_0 m\omega}{4\pi kR^3} \qquad B_2=C_1=C_2=\frac{\mu_0 m\omega k^2}{4\pi} \qquad A_2=D_2=0 \quad . \qquad (15)$$

Given this, (07b) and (12b) immediately yield $\mathbf{E}$ outside the dipole as

* $$\mathbf{E}(\mathbf{r},t)=\frac{\mu_0 m\omega k^2}{4\pi}\left(\frac{\cos(kr-\omega t)}{kr}-\frac{\sin(kr-\omega t)}{k^2r^2}\right)\sin\theta\hat{\boldsymbol{\varphi}} \quad , \qquad (16)*$$

which is the desired answer.

(c) We've got $\mathbf{E}$, and now want $\mathbf{B}$ outside the dipole. Most of the work is already done but there's a couple of tedious manipulations needed to complete the solution. Equations (11b) and (09) give the radial and tangential components of $\mathbf{B}$ corresponding to the $\sin\omega t$ component of $\mathbf{E}$. For the $\cos\omega t$ component of $\mathbf{E}$, we can again get the $\mathbf{B}$'s by using (01b) on (12b); doing so we find

$$B_{r2}=\left(-\frac{2C_2 f(kr)}{\omega r}-\frac{2D_2 g(kr)}{\omega r}\right)\cos\theta\sin\omega t \quad ,$$

while

$$B_{\theta 2}=-\left[C_2\left(\frac{f(kr)}{r}+kf'(kr)\right)+D_2\left(\frac{g(kr)}{r}+kg'(kr)\right)\right]\sin\theta\cos\omega t \quad .$$

Finally, using (15) [fortunately $D_2 = 0$ so some terms go away] and expanding the derivatives of f and g we get

$$\mathbf{B}(\mathbf{r},t) = \frac{\mu_0 mk}{2\pi}\cos\theta\left[\frac{\sin(kr-\omega t)}{r^2} + \frac{\cos(kr-\omega t)}{kr^3}\right]\hat{\mathbf{r}}$$

$$+\frac{\mu_0 mk^2}{4\pi}\sin\theta\left[\begin{array}{l} -\frac{1}{r}\cos(kr-\omega t) + \frac{1}{kr^2}\sin(kr-\omega t) \\ +\frac{1}{k^2 r^3}\cos(kr-\omega t)\end{array}\right]\hat{\boldsymbol{\theta}} \quad , \qquad (17)^*$$

which is the desired answer.

Now is the consistency check with 6.3.5 – whether the above result reduces to the static case when $\omega = 0$. First check the current itself – since we defined the current to be $\cos\omega t$, it becomes a constant when we set $\omega = 0$ (if we'd defined the current as $\sin\omega t$, then the current as well as the fields would have become zero in the limit, which would then have become ill-defined). Since $k = \omega/c$, this of course means $k = 0$. Looking at (17), in the r-component of $\mathbf{B}$, the second term survives in this limit since the cos becomes unity, while in the θ-component, it's the third term which survives. Collecting the survivors, we get

$$\mathbf{B} = \frac{\mu_0 m}{4\pi}\left[\frac{2\cos\theta\hat{\mathbf{r}} + \sin\theta\hat{\boldsymbol{\theta}}}{r^3}\right] \quad ,$$

which agrees exactly with what we found in 6.3.5, Likewise, the electric field (16) becomes identically zero when $k = 0$. This is consistent with our knowledge that a static magnetic dipole produces no electric field, and the check passes.

(d) In the space $z < 0$ we don't have sources but a conductor of conductivity σ. Since $\mathbf{J}=\sigma\mathbf{E}$ we have in (01d)

$$\nabla\times\mathbf{B} = \mu_0\sigma\mathbf{E} + \frac{1}{c^2}\frac{\partial}{\partial t}\mathbf{E} \quad .$$

Now taking the curl of (01b) and using the above leads to the governing equation

$$\nabla^2\mathbf{E} = \frac{1}{c^2}\frac{\partial^2\mathbf{E}}{\partial t^2} + \mu_0\sigma\frac{\partial\mathbf{E}}{\partial t} \quad ,$$

which is a damped wave equation for $\mathbf{E}$. As in 6.3.8 Part (b), we try an electric field (any component) of the form $p(\mathbf{r})\cos\omega t + q(\mathbf{r})\sin\omega t$. This leads to

$$\nabla^2 p = \mu_0 \sigma \omega q - k^2 p \quad ,$$
$$\nabla^2 q = -\mu_0 \sigma \omega p - k^2 q \quad ,$$

where $k = \omega / c$ as before. Using the first equation to express q in terms of p and substituting into the second, we get a single fourth-order PDE for p :

$$\nabla^4 p - 2k^2 \nabla^2 p + k^4 p + \mu_0^2 \sigma^2 \omega^2 p = 0 \quad . \tag{18}$$

The spatial form we try for p is $\exp \mathrm{j}(k_1 x + k_2 y + k_3 z)$, with the understanding that the actual solution is a real-valued linear combination of complex exponentials. We recall that a real k_i denotes undiminished propagation of the waves while complex k_i implies propagation with decay. Setting this trial p into (18) gives a quadratic equation for $X = k_1^2 + k_2^2 + k_3^2$:

$$X^2 - 2k^2 X + k^4 + \mu_0^2 \sigma^2 \omega^2 = 0 \quad ,$$

which has the two roots

$$X = \frac{2k^2 \pm \sqrt{-4\mu_0^2 \sigma^2 \omega^2}}{2} = k^2 \pm \mathrm{j}\mu_0 \sigma \omega \quad .$$

On physical grounds we expect k_1 and k_2 to be real and thus have

$$k_3 = \sqrt{k^2 - k_1^2 - k_2^2 - \mathrm{j}\mu_0 \sigma \omega} \quad .$$

Thus, k_3 has the form $\sqrt{(A + \mathrm{j}B)}$ where A is variable but B is fixed; hence it is necessarily complex. For maximum penetration depth we want its imaginary part to be as small as possible. Now, the square root of $A + \mathrm{j}B$ is (we can do it like we found the roots in T026, or you can just look it up)

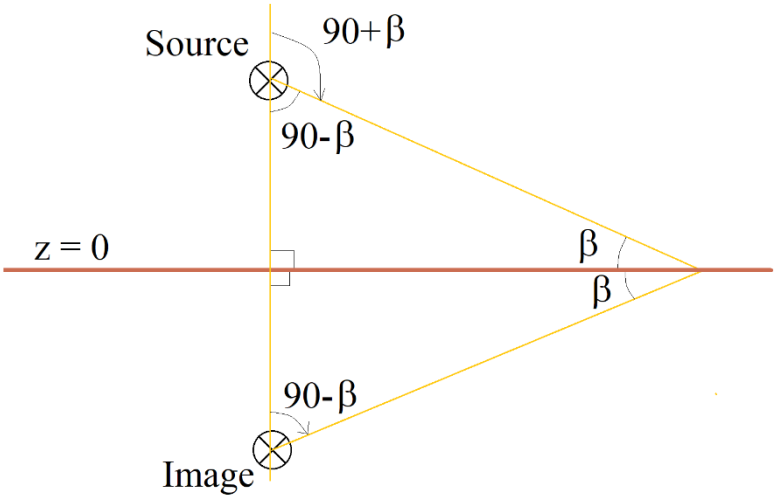
$$\sqrt{A + \mathrm{j}B} = \pm \left[\left(\frac{A + \sqrt{A^2 + B^2}}{2} \right)^{1/2} + \mathrm{j} \left(\frac{\sqrt{A^2 + B^2} - A}{2} \right)^{1/2} \right] \quad .$$

The minimum of the imaginary part will occur for the largest possible positive A, which occurs if $k_1 = k_2 = 0$; in this case $A = k$. Now if $B >> A$ then this minimum value goes as $(\mu_0 \sigma \omega)^{1/2}$ while if $B << A$ then it goes as $\mu_0 \sigma \omega / 2\sqrt{k}$. Either way, the greater the value of ω, the larger the imaginary part and the smaller the penetration depth, which was to be shown.

We note that the presence of the conductor is essential to achieve the screening of the field. In the case $\sigma = 0$, X would be purely real, and an $\mathbf{E}$ having

$k_1^2 + k_2^2 + k_3^2 = k^2$ would propagate with no diminution, as we would expect. This is an easy but important check on our results.

(e) Since the conductor can now be treated as a perfect flux-expeller, we must have the BC on the electric field $\mathbf{E} = 0$ at $z = 0$. We expect that this will be satisfied by an image configuration which replaces the flux-expeller with a dipole at $(0,0,-h)$; the thing to see is whether the image should be the same as the source or its negative. Now, the electric field outside the dipole is given by (16). Considering the source and image and an arbitrary point $(x,y,0)$ on the plane, at any instant of time, the r and t parts of $\mathbf{E}$ are the same for both the dipoles. As for the $\sin\theta$ component, the Figure below shows that θ for the source is $90^o+\beta$ while θ for the image is $90^o-\beta$ for some angle β between 0 and 90^o.



Since $\sin(\pi/2+\beta) = \sin(\pi/2-\beta)$ for $0 \le \beta \le \pi/2$, this factor is also identical for both dipoles. Thus, the image must be the negative of the source i.e. if the source dipole is $m\cos\omega t\hat{\mathbf{z}}$, the image is $-m\cos\omega t\hat{\mathbf{z}}$. This is the same as what we got in T106 Part (d) for $\chi = -1$ (perfect flux-expeller), and is consistent – in a different regime of operation, the quantitative features will of course be different but the SIGNS also flipping is too far-fetched.

To find the force on the source, we start off as in 6.3.7 Part (d). The formula $\mathbf{F} = (m\cdot\nabla)\mathbf{B}$ reduces to $F = m(\partial B_z/\partial z)$. B_z at the source point comes from both source and image but only the latter contributes a force. Thus, we must find B_z at the point $(0,0,h)$ on account of the image $-m\cos\omega t$ at $(0,0,-h)$.

At this point, we stop copying 6.3.7 since the magnetic field there was completely different from what it is here.

Evidently, for the image point, θ=0, $\hat{\mathbf{z}} = \hat{\mathbf{r}}$ and $B_z = B_r$. Thus, we should use the formula (17), after replacing m by $-m$. This gives

$$B_r = -\frac{\mu_0 mk}{2\pi}\left(\frac{\sin kr\cos\omega t - \cos kr\sin\omega t}{r^2}\right) \quad .$$

This incorporates only the leading term in $1/r$; since the question asks for the large h behaviour, this is sufficient. Evidently, the $\partial/\partial z$ of this B_r (or B_z) is the same as its ∂/∂r. Replacing r by z and retaining only the leading order terms in $1/z$ we have,

$$\frac{\partial B_z}{\partial z} = -\frac{\mu_0 mk^2}{2\pi}\frac{\cos kz\cos\omega t + \sin kz\sin\omega t}{z^2} \quad , \qquad (19)$$

and the long calculation is almost done.

Last step – for the force we want $m(\partial B_z/\partial z)$ averaged over time – since m is also $\cos\omega t$, the average with the $\cos\omega t$ component of (19) evaluates to 1/2 while that with the $\sin\omega t$ component of (19) vanishes. Substituting $z = 2h$ we get

* $$F = -\frac{\mu_0 m^2 k^2\cos 2kh}{16\pi h^2} \quad , \qquad (20)*$$

which is the desired answer.

*

COMMENTS : The force (20) is positive (i.e. a lift force) for suitable values of h, when $\cos 2kh$ is negative. Since it's an image-generated force, it is appropriate for a maglev application (recall how we discussed in 6.3.8 that images make good maglevs since they move with the sources). The primary point of the calculation is that the distance dependence of the force is $1/h^2$. In the calculation of 6.3.8, we were working in the slow electrodynamic limit i.e. we were essentially dealing with a static dipole, whose field is $1/h^3$ and force is its gradient which is $1/h^4$. Indeed, in 6.3.7 Part (d) we found that the force on the static dipole above a superconducting sheet is $3\mu_0 m^2/32\pi h^4$; in the slow limit, the force on an oscillating dipole above a metal sheet works out to half of that. Here, by exploiting the long-ranged dependence of the fields in the high-frequency limit, we have got a force with a more tractable dependence on h.

On another track, we note that the electric and magnetic fields (16,19) are EXACT solutions of MAXWELL's equations for the oscillating point magnetic dipole – there are no approximations involved anywhere. Equation (01) is the governing equation and (02) a transformed version. For the macroscopic sphere, (07) and (12) are its exact solutions and the BC fitting equations (13) are also exact. To proceed to (16), we only shrank the dipole to a point and did nothing else. Also noteworthy is that we got this answer directly by solving the wave equation, instead of using the retarded potentials as an intermediate (forget this sentence if you don't know what they are – obviously you didn't need them).

You might also ask how did I know that $f(kr)$ and $g(kr)$ are solutions of the radial equation . Like the angular distributions (see comments to 6.1.5), those too came from special functions theory. But the monopole and dipole terms have very simple structures on an absolute scale and, as we saw more than once, are capable of yielding a lot of useful information . So I will strongly recommend that you know these angular distributions and the corresponding radial functions for both Laplace and wave equations. For monopole, the radial solutions of the wave equation are $(\sin kr)/kr$ and $(\cos kr)/kr$ respectively. For the "electric dipole" $\cos\theta$ dependence , the separation constant λ_2 is again -2 so the radial functions are the f and g we used here.

At this stage we can also answer why the waves from a VOR (see the previous Tute) are megahertz modulated down to hertz instead of being radiated at hertz. If we made the current oscillate at hertz then k would have been very small (about 10^{-8}) and the magnetic field would have been dominated by the $1/r^3$ term of a static dipole. The electric field would have been zero for all practical purposes. There is no way these fields could have been radiated at Lucknow and picked up by a plane at Aligarh. The high frequency (large k) is essential to ensure that the waves propagate out to a huge distance; thereafter we modulate it down so that the aircraft can pick up the phase difference between two 30 Hz waves (relatively easy) instead of two megahertz waves (almost impossible).

Example 7.4.9

Train jerks

A train can be modelled as a loco of mass 123 tons plus 20 coaches of mass 49 tons each, all connected by couplers of spring constant $k = 2.5 \text{x} 10^8 \text{ N/m}$. The couplers also exert a nonlinear damping force, which can be treated as Coulomb friction between every pair of coaches with a finite coefficient μ_k of kinetic friction and zero coefficient μ_s of static friction. Let x_i be the displacement of the i_{th} mass from its equilibrium position when all couplers are unstretched, in a reference frame which is instantaneously moving with the train. Then, the damping force on each mass is given by

$$F_D = -C\left[\text{sgn}\left(\dot{x}_i - \dot{x}_{i-1}\right) + \text{sgn}\left(\dot{x}_i - \dot{x}_{i+1}\right)\right] \quad ,$$

where sgn denotes signum function and the numerical value of C is 15000 N. The train is coasting through a caution order at a constant speed when it approaches a termination board. After clearing the board, the loco pilot applies the throttle suddenly, resulting in an immediate rise of the tractive effort of the loco to its rated value of 220 kN.

(a) Drawing reference to 2.3.5 as required, write a system of coupled ODEs for $\mathbf{x}(t)$, describing the motion of the train after the sudden application of throttle. Also write the ICs.

The method of matched boundary conditions can be used to generate the solution for $\mathbf{x}(t)$. Here, one recognizes that the nonlinear friction force between any pair of coaches is piecewise constant – it is full-strength attractive if the coaches are separating away from each other and full-strength repulsive if they are approaching towards each other. Over small intervals of time, when the relative velocity between no two coaches changes sign, the forcing vector as a whole is a constant. During this interval, the equation of motion is a linear inhomogeneous system which can be solved exactly. This solution remains valid until there is a

change in the forcing vector, at which time one uses the current position and velocity vectors as initial conditions for the new linear system in the next interval.

(b) Implement the above qualitative description as an algorithm to determine $\mathbf{x}(t)$ with the help of a computer. Hint : The eigenvectors of the relevant matrix are not orthogonal.

(c) The ride quality in a coach is determined primarily by the jerk, which is defined as the second time-derivative of its velocity. Plot the velocities of the 2^{nd}, 8^{th}, 14^{th} and 18^{th} coaches as functions of time and qualitatively compare the ride quality.

[Level : 3]

(a) The setup is similar to the baby wave equation with two free ends – the differences are that the first mass here is heavier than the others and that there is nonlinear damping present here. We let x_1 be the displacement of the loco (from equilibrium, as the question says) and x_2, x_3, ….., x_{21} the displacements of the coaches. For a typical coach we have

$$m\ddot{x}_i = k\left(x_{i-1} - 2x_i + x_{i+1}\right) - C\left[\operatorname{sgn}\left(\dot{x}_i - \dot{x}_{i-1}\right) + \operatorname{sgn}\left(\dot{x}_i - \dot{x}_{i+1}\right)\right] \quad ,$$

where m is the coach mass. The exceptions are the last coach and loco. For the former we have $m\ddot{x}_{21} = k\left(x_{20} - x_{21}\right) - C\operatorname{sgn}\left(\dot{x}_{21} - \dot{x}_{20}\right)$ while for the loco, $M\ddot{x}_1 = k\left(-x_1 + x_2\right) - C\operatorname{sgn}\left(\dot{x}_1 - \dot{x}_2\right)$, where M is the loco mass. Further, since the loco is applying traction, there will also be an external forcing term present in the equation, which will be non-zero for the loco alone.

We can combine these facts into a matrix equation, which we display for a three-coach toy train rather than the real McCoy :

$$\begin{bmatrix} M & 0 & 0 & 0 \\ 0 & m & 0 & 0 \\ 0 & 0 & m & 0 \\ 0 & 0 & 0 & m \end{bmatrix} \begin{bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \\ \ddot{x}_3 \\ \ddot{x}_4 \end{bmatrix}$$

$$= k \begin{bmatrix} -1 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} - C \begin{bmatrix} \operatorname{sgn}\left(\dot{x}_1 - \dot{x}_2\right) \\ \operatorname{sgn}\left(\dot{x}_2 - \dot{x}_3\right) \\ \operatorname{sgn}\left(\dot{x}_3 - \dot{x}_4\right) \\ 0 \end{bmatrix} - C \begin{bmatrix} 0 \\ \operatorname{sgn}\left(\dot{x}_2 - \dot{x}_1\right) \\ \operatorname{sgn}\left(\dot{x}_3 - \dot{x}_2\right) \\ \operatorname{sgn}\left(\dot{x}_4 - \dot{x}_3\right) \end{bmatrix} \quad .$$

$$+ \begin{bmatrix} F; & 0; & 0; & 0 \end{bmatrix}$$

Letting $\mathbf{M}$ be the mass matrix, $\mathbf{Q}$ the "discrete second derivative" matrix attaching to k, $\mathbf{f}_1$ the external forcing vector and $\mathbf{f}_2$ the sum of the two damping vectors, we can write this as

$$* \qquad \frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{x} - k\mathbf{M}^{-1}\mathbf{Q}\mathbf{x} = \mathbf{M}^{-1}\mathbf{f}_1 - C\mathbf{M}^{-1}\mathbf{f}_2(\dot{\mathbf{x}}) \quad , \qquad (01)^*$$

which is the ODE system asked for. We have explicitly indicated the dependence of $\mathbf{f}_2$ on the velocity vector to highlight that it is not a constant vector, as is usually the case.

For the IC, we note coasting through a caution order at constant speed means that in the steady state all springs are transmitting no force. Since by the problem assumptions they do not have any static friction, this implies that they must all be at their natural lengths. Since the springs are damped, they'll arrive at this steady state within few seconds or so of the caution's beginning. Hence, when the termination board arrives, the springs shall all be quiescent and at equilibrium. The IC is thus $\underline{\mathbf{x}(0) = \mathrm{d}\mathbf{x}/\mathrm{d}t\,(0) = 0}$.

(b) We rewrite (01) as

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathbf{x} - k\mathbf{A}\mathbf{x} = \mathbf{g}_1 + \mathbf{g}_2 \quad , \qquad (02)$$
$$= \mathbf{g}$$

thus defining $\mathbf{A} = \mathbf{M}^{-1}\mathbf{Q}$, $\mathbf{g}_1 = \mathbf{M}^{-1}\mathbf{f}_1$, $\mathbf{g}_2 = -C\mathbf{M}^{-1}\mathbf{f}_2$ and $\mathbf{g} = \mathbf{g}_1 + \mathbf{g}_2$. The vector $\mathbf{g}_1$ is the loco traction which is a constant, while $\mathbf{g}_2$ is the damping which is dependent on the velocity vector in a nonlinear manner. Nevertheless, as the question tells us, $\mathbf{g}_2$ is constant over intervals where the relative velocity between no two coaches changes in sign. In such intervals, we can solve the system exactly, and then use boundary matching to proceed from one interval to the next.

Preliminary to solving the system, we get Matlab to calculate the eigenvalues and vectors of $\mathbf{A}$. Matlab is allergic to inverting matrices so we feed in the fact that the inverse of the diagonal matrix $\mathbf{M}$ is simply another diagonal matrix whose entries are the reciprocals of those of $\mathbf{M}$. We see that all eigenvalues of $\mathbf{A}$ are negative except for one which is zero. The corresponding eigenvector is [1; 1; …..; 1] just as it was in 2.3.5 – the reason for its appearance here is the same also. Hereafter, we shall denote by $\mathbf{v}_0$ the constant eigenvector, and by $\mathbf{v}_1$ to $\mathbf{v}_{20}$ the eigenvectors of the non-zero eigenvalues λ_1 to λ_{20}.

Equation (02) in an interval of constant $\mathbf{g}_2$ is a linear inhomogeneous ODE system so we start from the homogeneous solutions. Setting $\mathbf{g} = 0$, the resultant equation has the fundamental solutions $\mathbf{v}_0$, $\mathbf{v}_0 t$ and $\mathbf{v}_n \cos \omega_n t$, $\mathbf{v}_n \sin \omega_n t$ for n going from 1 to 20, with $\omega_n = \sqrt{-k\lambda_n}$. Now we have to find particular solution of (02). For this, we expand the inhomogeneity $\mathbf{g}$ in the eigenvectors i.e. we write

$$\mathbf{g} = a_0\mathbf{v}_0 + a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \ldots.. + a_{20}\mathbf{v}_{20} \quad . \qquad (03)$$

As the question warns us, the $\mathbf{v}$'s are not orthonormal ($\mathbf{A}$ is not symmetric), so to find the a's we cannot take a dot product but must do $\mathbf{a} = \mathbf{V}\backslash\mathbf{g}$. Although the $\mathbf{v}$'s are not orthonormal, they ARE linearly independent so this expansion is unique. (Read the rest of this paragraph ONLY if you had studied the extra parts of 7.4.2. You will notice that the equation structure $\mathbf{M}\ddot{\mathbf{x}} = k\mathbf{Q}\mathbf{x}$ is somewhat similar to (7.4.2.01) – there also we had a corner-heavy mass matrix attaching to the time derivative and screwing orthogonality up. But we were able to prise out a modified orthogonality condition there, so the $\mathbf{v}$'s here should also satisfy some kind of relation similar to (7.4.2.06) – surely we can exploit that instead of doing the crude $\mathbf{a} = \mathbf{V}\backslash\mathbf{g}$? But a golden rule of problem solving is to never use higher grade ammunition than required. This is not a PDE system but an array of 21 masses . The computer is on board anyway to find the eigenvalues and vectors , and later to actually generate the solution . So why not just make it solve $\mathbf{V}\backslash\mathbf{g}$ instead of hand -calculating the modified orthogonality relation.)

Since the excitation is constant in time, we guess the particular solution $\mathbf{x}_p$ to be a constant also. Expanding this in the eigenvectors as $\mathbf{x}_p = b_0\mathbf{v}_0 + b_1\mathbf{v}_1 + \ldots . b_{20}\mathbf{v}_{20}$, we plug into (02) together with (03) and, since the $\mathbf{v}$'s are linearly independent, equate the coefficient of each $\mathbf{v}_i$. This yields $b_i = -a_i/k\lambda_i$; a solution which evidently works for all i other than zero but fails in that case because λ_0 is zero. This failure is not a surprise – $\mathbf{v}_0$ as well as $\mathbf{v}_0 t$ are homogeneous solutions so we must add on the bump-up factor of t^2 to find the particular solution. Our revised guess for $\mathbf{v}_0$ is thus $b_0 t^2 \mathbf{v}_0$; plugging into (02) and using that $\mathbf{A}\mathbf{v}_0 = 0$ immediately shows $b_0 = a_0/2$. Thus the particular solution is

$$\mathbf{x}_p = b_0 t^2 \mathbf{v}_0 + \sum_n b_n \mathbf{v}_n$$

, where

$$b_0 = \frac{a_0}{2}, \quad b_n = -\frac{a_n}{k\lambda_n} \quad . \tag{04}$$

The t^2 solution attaching to $\mathbf{v}_0$ of course represents the acceleration which the train as a whole embarks on as a result of the loco traction. For this problem, we let the unlabelled summation over n run from 1 to 20 only, since that rather than infinity is the system size in this case. Adding this to (03) yields the general solution

$$\mathbf{x}_G = \mathbf{v}_0\left(b_0 t^2 + c_0 + d_0 t\right) + \sum_n \mathbf{v}_n\left(b_n + c_n \cos\omega_n t + d_n \sin\omega_n t\right) \quad . \tag{05}$$

Before entering into more details, let's fit the general solution (05) to arbitrary initial conditions $\mathbf{x} = \mathbf{x}_T$ and $\dot{\mathbf{x}} = \mathbf{u}_T$ at arbitrary time T. For this, we rewrite (05) as

$$\mathbf{x}_G = \mathbf{v}_0\left(b_0 (t-T)^2 + c_0 + d_0 (t-T)\right) + \sum_n \mathbf{v}_n\left(b_n + c_n \cos\omega_n (t-T) + d_n \sin\omega_n (t-T)\right) \quad . \tag{06}$$

Here the c_n's etc stand for different quantities from (05) but so long as they are arbitrary constants I won't change the variable names. For the boundary match, we'll also need the derivative of $\mathbf{x}_G$, so I'll display it explicitly :

$$\dot{\mathbf{x}}_G = \mathbf{v}_0\left(2b_0 (t-T) + d_0\right) + \sum_n \mathbf{v}_n\left(-\omega_n c_n \sin\omega_n (t-T) + \omega_n d_n \cos\omega_n (t-T)\right) \quad . \tag{07}$$

We expand the ICs $\mathbf{x}_T$ and $\mathbf{u}_T$ as

$$\mathbf{x}_T = \tilde{x}_0 \mathbf{v}_0 + \sum_n \tilde{x}_n \mathbf{v}_n \qquad \mathbf{u}_T = \tilde{u}_0 \mathbf{v}_0 + \sum_n \tilde{u}_n \mathbf{v}_n \quad .$$

Then, we ignore $\mathbf{v}_0$ (justification next paragraph) and fit the expanded ICs to (06) for every n; this yields

$$c_n = \tilde{x}_n - b_n \quad , \tag{08a}$$

$$d_n = \frac{\tilde{u}_n}{\omega_n} \quad . \tag{08b}$$

The mechanical part of the calculation is now complete.

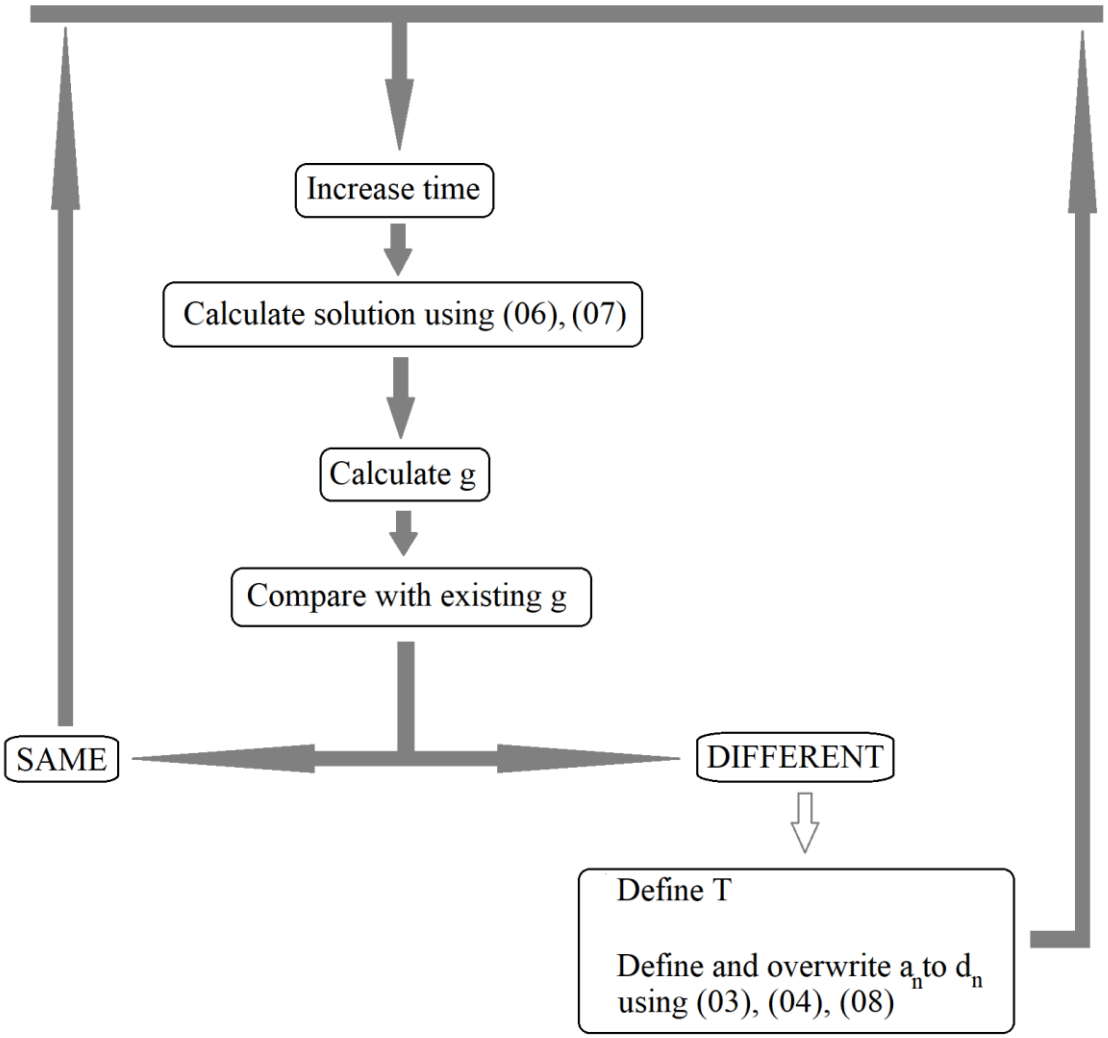
Prior to starting the boundary matching, we argue that the friction cannot be affecting the overall position or velocity of the train. Since it acts in a pairwise manner between coaches, it cannot be exerting a net force on the train as a whole. More mathematically, the dot product of the vector $\mathbf{g}_2$ with the constant vector $\mathbf{v}_0$ is zero [for a concrete example, go back to the toy train equation in Part (a)], so $\mathbf{g}_2$ lies in the orthogonal complement of $\mathbf{v}_0$. Hence the coefficient $\mathbf{a}_0$ in (03) must be coming from $\mathbf{g}_1$ (loco traction) only. Then, the corresponding

component of the particular solution shall remain invariant through all the changes of the frictional force and we can find it outright for all time by fitting (05) to zero ICs at $t = 0$. This readily yields $c_0 = d_0 = 0$.

For the remaining components of the solution, we must of course use the boundary matching technique. We can see that the friction force as given in the problem is undefined if two coaches have exactly the same velocity. We remedy this, and also make the definition more machine-friendly, by specifying the force to be zero if the relative velocity between the two coaches is less than some small number ε in magnitude. Otherwise, the force is plus and minus C accordingly as the relative velocity is less or greater than minus and plus ε. At $t = 0$, all the coaches are stationary so $\mathbf{g}_2$ is identically zero. Setting $T = 0$, we fit (06) to trivial ICs and start crawling forward in time. Here I have chosen ε as well as the time increment step to have the numerical value 10^{-5}. After every time step, we calculate the relative velocity between the coaches as per (07), and basis that, calculate the friction vector $\mathbf{g}_2$. As soon as $\mathbf{g}_2$ becomes different from its starting value (zero), we set T as the time when the transition occurred, $\mathbf{x}_T$ and $\mathbf{u}_T$ as the position and velocity at this time, and calculate a_n, b_n, c_n and d_n using (03), (04) and (08). Then, we again use (06) to crawl forward, continuously checking for a change in the friction vector. Again upon hitting the change we recalculate the coefficients, and thus keep advancing slowly but inexorably.

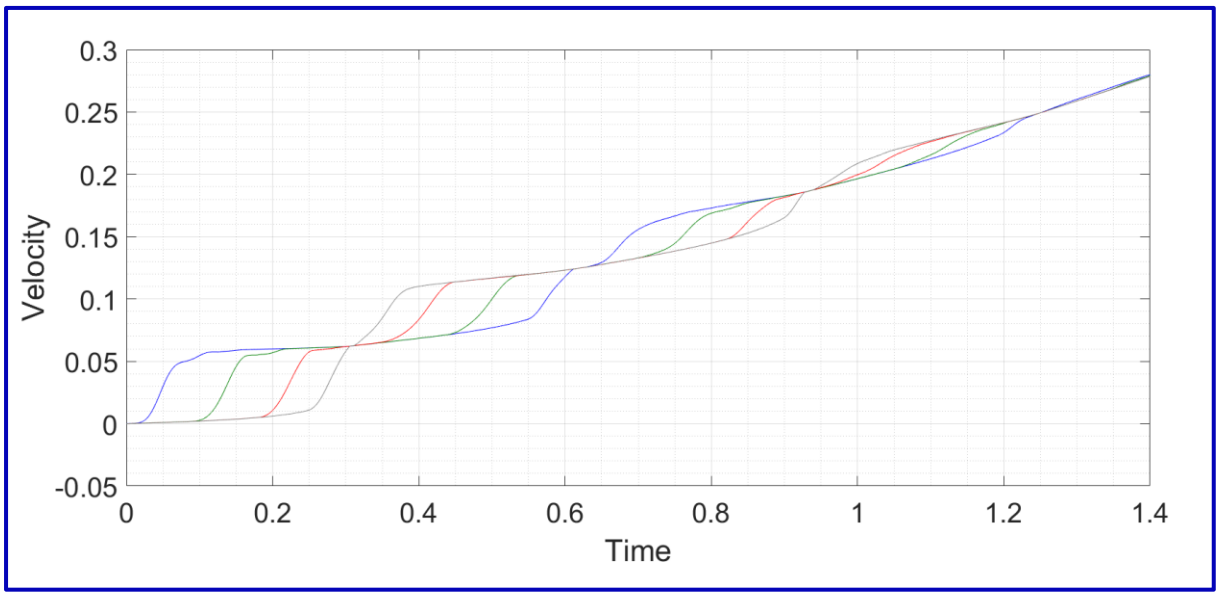
Here is the flowchart which implements the algorithm.

(A lovely place indeed for a largish Figure)



If you've already read 7.4.4, you'll immediately recognize its heavily watered-down variant. If you tried that problem and found it too difficult (or just too intimidating), it might be a good idea to go back there after you're through here.

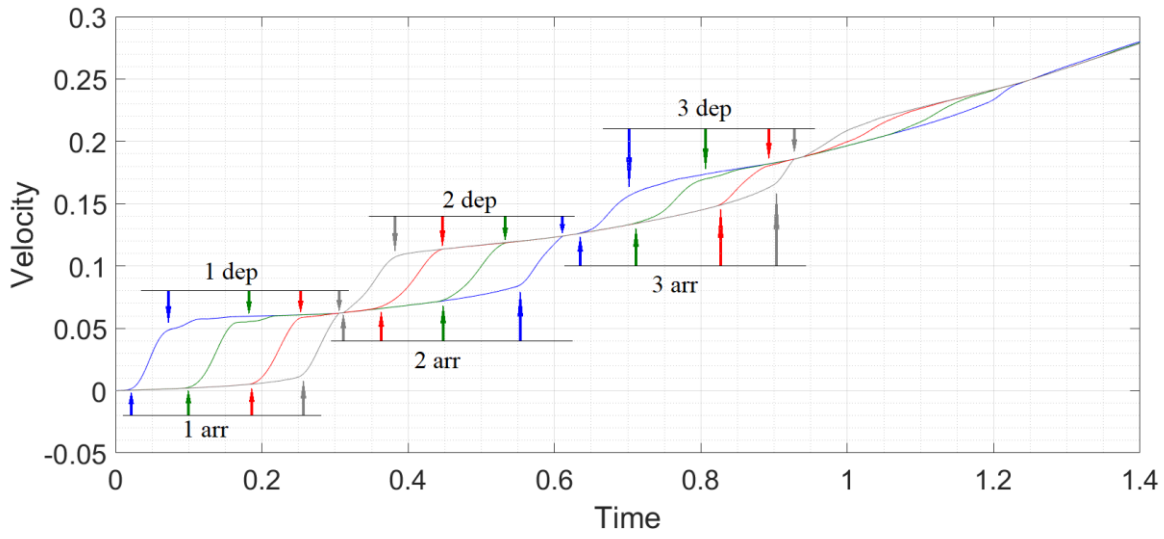
(c) The required plot is below, in standard livery with the time traces for the 2nd, 8th, 14th and 18th coaches being in blue, green, red and grey respectively. Note that these correspond to $n = 3,9,15,19$ since $n = 1$ is the big bad Wappie !



The first thing to do is check it for plausibility – after all, the algorithm which went into it is non-trivial and the results might well be garbage. (You know as well as I do that in the public version of the book, I WON'T be showing results which I know to be garbage – those all have gone into the bin together with the countless draft versions. But still, we'll do the plausibility analysis so that you are more convinced of the result's validity.)

The asymptotics clearly match up – after about 1.3 seconds, the oscillations are dying out completely and all velocities increasing together in a straight line (I've shown this line for a small region but it continues until at least $t = 4$, beyond which I didn't extend the run). The slope of this line is about 0.2 m/s^2 which is consistent with a traction of 220 kN being applied to a 1100 ton train. Then, in the initial part of the curves we can clearly see the disturbance propagating from the loco to the coaches behind it. Coach 2 jerks up almost immediately after $t = 0$, while coaches 8, 14 and 18 show the first major pulse at roughly equal intervals of time behind each other. This is consistent with the linear part of the governing equation being a discrete wave equation. As we saw in 7.2.2, any discrete version of wave equation admits some acausal responses, while 7.1.4 showed that adding damping also generates these responses. These explain the small increase in velocity before the major pulses arrive. Thus, we have run some plausibility tests at both small and large times, and can conclude that the plot is correct.

Now, from this plot we have to qualitatively estimate the jerk. This is the second derivative of velocity i.e. the degree of kinkiness in the curve. Here's the velocity plot again, with some markups. These are the instants of arrival and departure of the first, second and third pulses for each of the coaches. By arrival of pulse I mean the instant when the velocity curve transitions from a shallow slope to a steep slope; by departure I mean the reverse transition. A zoomed-up version of this plot helps so I'm putting it up on the book's website : www.shayak.in/Shayakpapers/DELTA/T130_03.png.



We can see that 1 arr has about the same degree of kinkiness for all four coaches. For 1 dep however, the kinkiness clearly increases in order blue, green, red, grey i.e. from front to back. For 2 arr, grey is the most kinky while the other three are smoother – possibly, green and red more so than blue. Moreover, the time interval between 1 dep and 2 arr decreases from front to back. The smaller this interval, the more unpleasant the ride since two back-to-back pulses shall appear as a nasty double jolt. Hence, the first and second pulses clearly cause deteriorating ride quality as one moves backwards down the train.

The situation reverses if we consider 2 dep and 3 arr. For 2 dep, the kinkiness clearly decreases in order blue, green, red, grey while the separation from 3 arr increases in that order. This time, it's the front which has the worse ride. By now however, the damping has caused the overall amplitudes to reduce so the front's gain relative to the back is less than the loss incurred earlier. The next

transition 3 dep 4 arr (this last not labelled, but you can easily guess) is a scaled down version of 1 dep 2 arr, so again the back is worse off. Thereafter, the pulses are of such small size that they are impossible to identify and the ride is the same in all coaches.

Summarizing the above, the reducing interval between the two strongest pulses and the increasing kinkiness at these pulses make us conclude that the ride quality generally deteriorates from front to back.

*

COMMENTS : The LHB coaches of Rajdhani Express are well known for jerks, especially when the loco is a WAP7. The coach dependence is because the LHB coaches have a centre buffer coupler (CBC) whereas the ICF coaches have a screw type coupler with side buffers. The CBC features much stiffer springs and much lower damping than the traditional design, resulting in greater oscillation amplitude at every traction change. (In other areas, primarily safety, CBC coupler enjoys marked advantages over the screw type, which make it worthwhile despite the increased oscillations.)

The loco dependence arises from the motor control algorithm. In WAP4, the throttle is a tap changer and it takes time for the LP to work the notches and build tractive effort. But in WAP7, direct torque control ensures that the transition from zero torque to rated torque can occur in a fraction of a second. This then sends vibrations down the coaches, as we saw here. A skilful loco pilot can reduce the jerks a lot by increasing or decreasing throttle gradually. However, if the train is late-running and/or the section is tightly timed, then he often has no option but to throttle and dethrottle as quickly as possible.

It is an observed fact that the jerks at the tail of the train are significantly greater than at the head. This is well brought out by our model. In fact, it is for this that I had to go with a nonlinear model and a harder calculation than I had wanted at this point. An earlier and easier version had linear springs without friction. The ride quality analysis there was realistic but the jerks persisted ad infinitum – a big weakness. A second iteration had viscous damping proportional to the relative velocity between adjacent coaches. This was easy enough, it damped out the jerks, but it made poor predictions of the ride quality. Much of the jerks actually come from the high-frequency oscillation modes, but relative damping

killed off the higher modes with faster and faster decay constants. A third model had viscous damping proportional to the train's absolute velocity (i.e. $F_D = -C\dot{x}_i$ on the i^{th} mass). This too was easy and it preserved the high order modes but it also dragged down the speed of train as a whole – a large-scale unphysical behaviour. Finally, in this present iteration of this problem we have a model where the large-scale as well as the small-scale behaviour is captured accurately. The cost is additional calculational complexity, but I don't think that any easier model can do the job in this case. As ALBERT EINSTEIN famously said, everything should be as simple as possible but not simpler.

You might ask why we analysed the jerk by eye from the velocity plot instead of obtaining it from a calculation – surely it would have been quite easy given everything we already have. This is because actually finding the jerk would have entailed some extra numerical hassle. As a result of the Coulomb friction, in time intervals where the velocities of adjacent coaches are nearly equal, the acceleration of the coaches often has very rapid fluctuations as the force oscillates between $-C$ and $+C$ at every incremental time step. To make a meaningful estimate of jerk, we would first have to filter out these fluctuations and then take a derivative. This is an excellent problem for a numerical methods textbook, but not for here.

* * * * *

This page intentionally left blank

Differential Equations – Linear Theory and Applications (DELTA) is a text cum problem collection on ordinary and partial differential equations. For each topic, a succinct exposition of the theory is followed by a series of problems, ranging from simple to difficult, and dealing with real-world phenomena taken from classical mechanics and vibrations, elasticity theory, electromagnetism and electrical engineering, quantum mechanics, and acoustics.

B SHAYAK, a third year doctoral student of Theoretical and Applied Mechanics at the Sibley School in Cornell University at the time of the original edition, is now a fifth year doctoral student there.