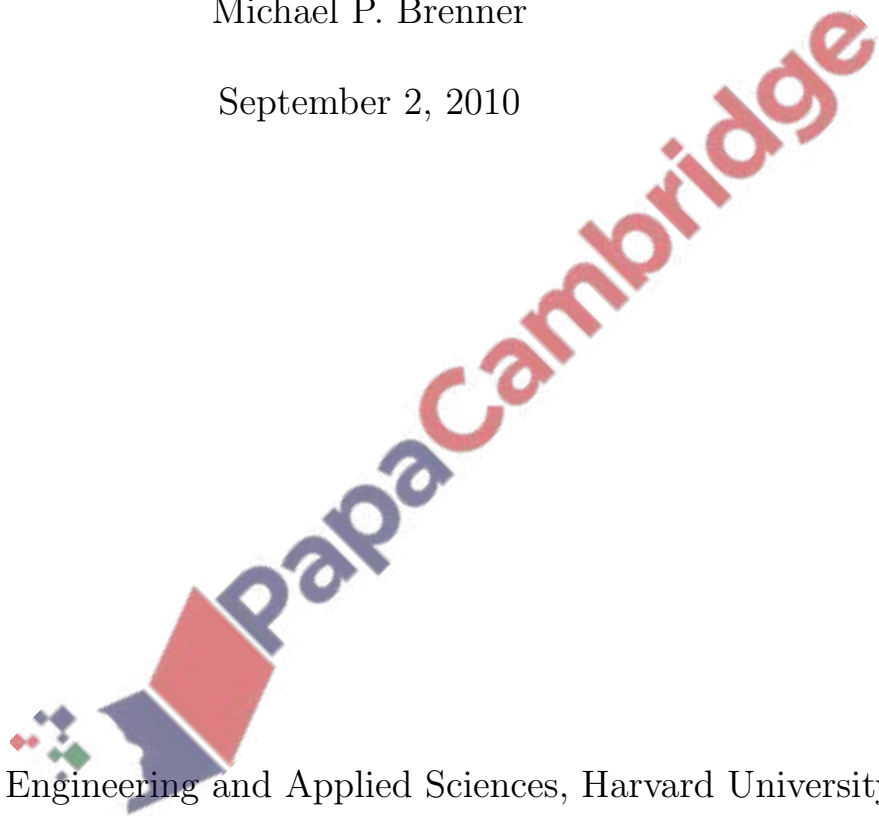


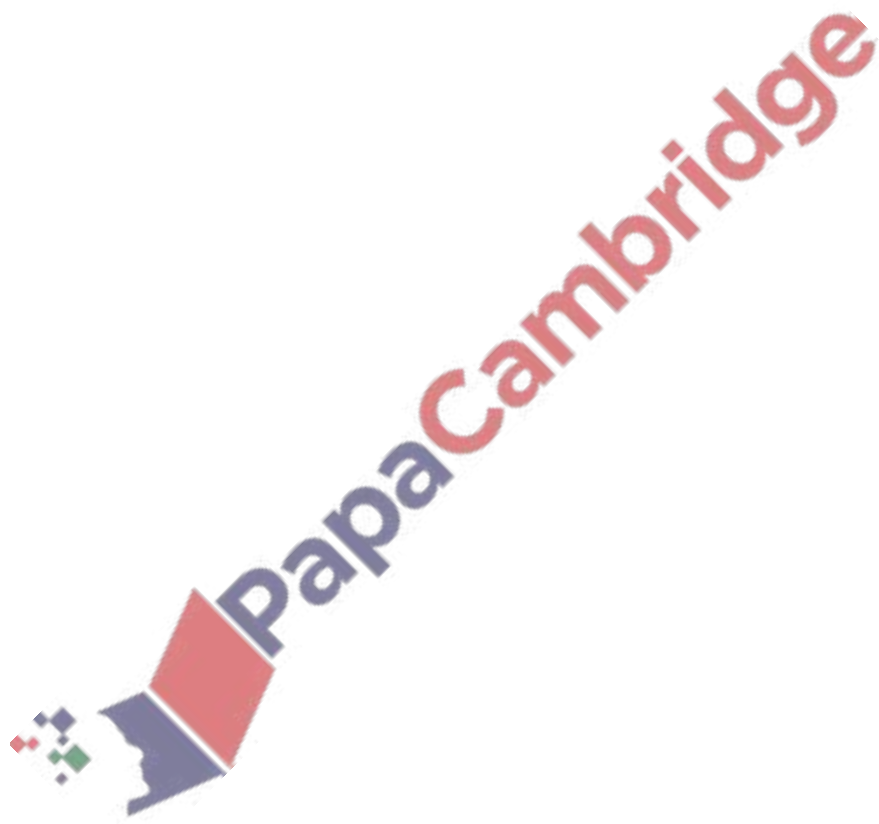
Physical Mathematics

Michael P. Brenner

September 2, 2010



School of Engineering and Applied Sciences, Harvard University



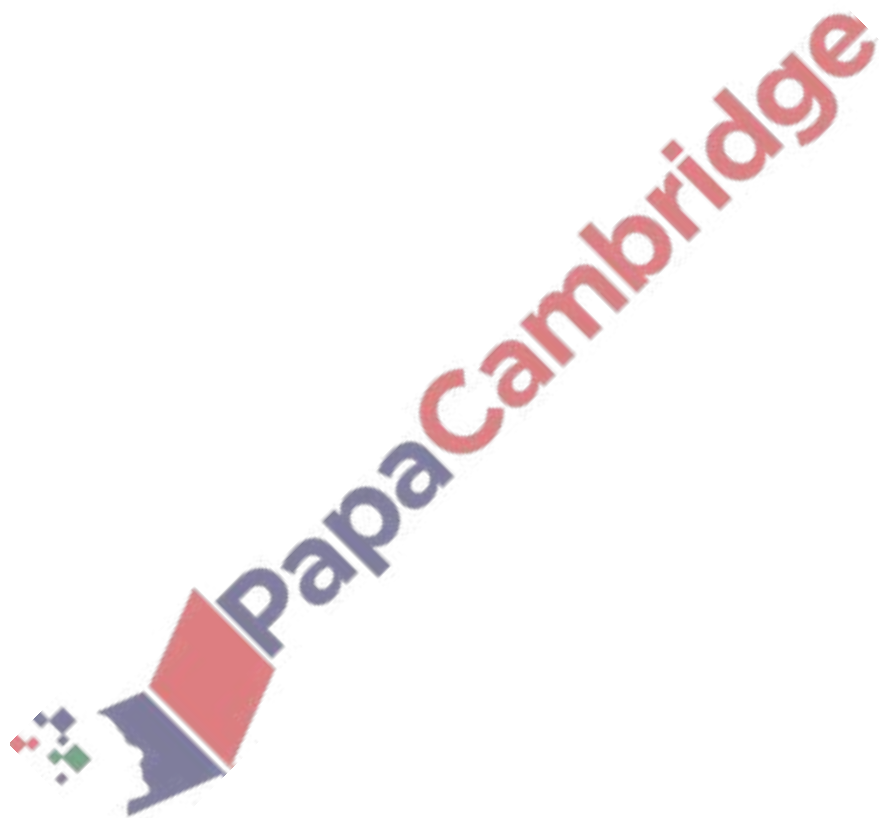
Contents

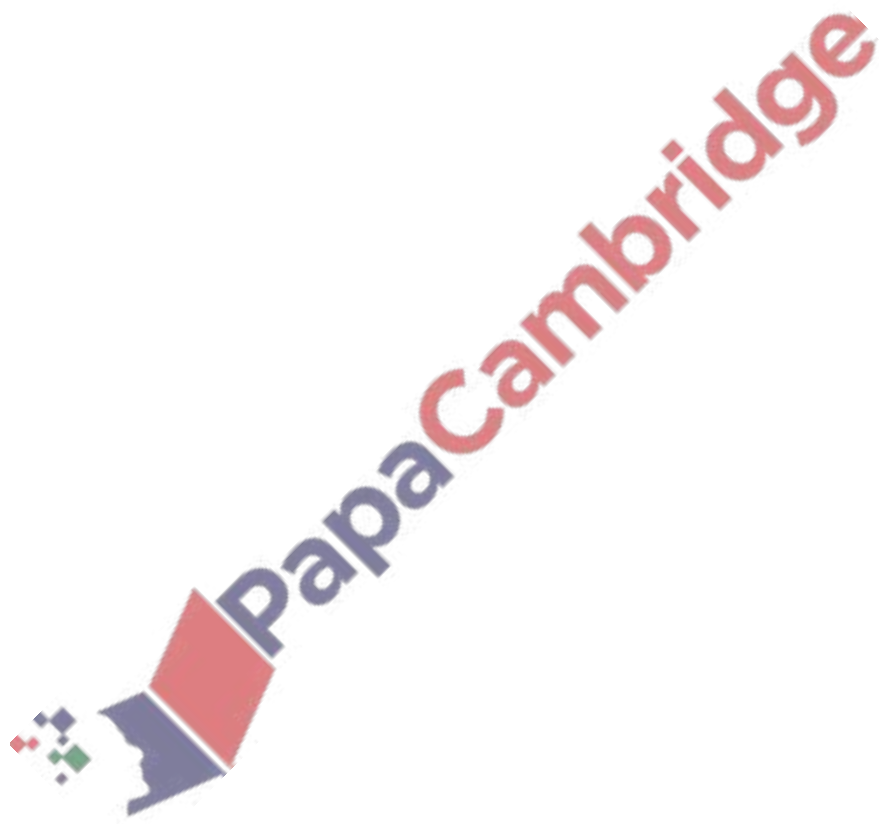
| | | |
|----------|---|-----------|
| 1 | Introduction | 9 |
| 1.1 | Computer Graphics and Mathematical Models | 9 |
| 1.2 | Calculating while computing | 10 |
| 1.3 | What is an analytical solution? | 11 |
| 2 | Solving Unsolvable Problems: An Introductory Example | 15 |
| 2.1 | Pig farming and Polynomials | 15 |
| 2.2 | Analytical Solutions to Polynomial Equations | 15 |
| 2.3 | Dominant Balance and Approximate Solutions | 17 |
| 2.4 | Nondimensionalization | 18 |
| 2.4.1 | A Dominant Balance: $\epsilon \rightarrow 0$ | 19 |
| 2.4.2 | The Second Dominant Balance : $\epsilon \rightarrow \infty$ | 21 |
| 2.5 | Testing the theory with numerical simulations | 21 |
| 2.6 | A bifurcation (phase transition) | 23 |
| 2.7 | Lessons Learned | 24 |
| 3 | Dimensions and Dimensional Analysis | 25 |
| 3.1 | Introduction | 25 |
| 3.2 | Buckingham's π Theorem | 26 |
| 3.3 | Examples | 27 |
| 3.3.1 | The Pendulum | 27 |
| 3.3.2 | Planetary orbits | 28 |
| 3.3.3 | The size of atoms | 29 |
| 3.3.4 | Fluid Viscosity | 29 |
| 3.3.5 | Atomic Energy Scale | 30 |
| 3.3.6 | Man's Size | 31 |
| 3.3.7 | The Radius of the Earth | 31 |
| 3.3.8 | Another Calculation of the Earth's Radius | 32 |
| 3.3.9 | Taylor's Blast | 32 |
| 3.3.10 | Pythagorean Theorem: | 33 |
| 3.3.11 | McMahon's Rowers | 34 |
| 3.3.12 | Lessons learned | 34 |
| 4 | Polynomial Equations: Random and Otherwise | 37 |
| 4.1 | Solving Polynomial Equations | 37 |
| 4.2 | Having courage: finding roots by iteration | 38 |

| | | |
|----------|--|-----------|
| 4.3 | Random Quartic Polynomials | 41 |
| 4.3.1 | Our first random polynomial | 41 |
| 4.3.2 | Our second random polynomial equation | 42 |
| 4.3.3 | n^{th} Order Random Polynomials | 46 |
| 4.4 | Having Courage | 46 |
| 4.4.1 | A Problem of Hinch | 46 |
| 4.4.2 | The Prime Number Theorem | 48 |
| 4.5 | Numerical Approaches | 50 |
| 4.5.1 | Newton's Method | 50 |
| 4.5.2 | MATLAB Implementation | 51 |
| 4.5.3 | Multidimensional Newton's Method | 53 |
| 5 | Ordinary Differential Equations | 55 |
| 5.1 | A Very Simple Example | 55 |
| 5.1.1 | A Simpler Way to See the Same Thing | 56 |
| 5.2 | A Harder Problem | 56 |
| 5.3 | Example 2 | 58 |
| 5.3.1 | Doing a Better Job at Large x | 61 |
| 5.3.2 | Something Disturbing | 65 |
| 5.4 | A Nonlinear Example | 67 |
| 5.4.1 | Negative Initial Conditions | 68 |
| 5.5 | Higher Order Nonlinear Ordinary Differential Equations | 78 |
| 5.5.1 | The First Balance | 79 |
| 5.5.2 | The Second Balance | 79 |
| 5.5.3 | The Third Balance | 80 |
| 5.5.4 | Testing the Theory | 80 |
| 5.6 | Numerical Solution of Ordinary Differential Equations | 81 |
| 5.6.1 | Remarks | 85 |
| 6 | "Simple" Integrals | 87 |
| 6.1 | What is a simple integral | 87 |
| 6.1.1 | A more sensible definition | 88 |
| 6.2 | A very easy simple integral | 89 |
| 6.2.1 | What if you are on a desert island without an arctan table | 91 |
| 6.2.2 | Back to the Integral | 92 |
| 6.3 | A harder integral | 93 |
| 6.3.1 | More accurate answers | 94 |
| 6.3.2 | Practical Implementation in MATLAB | 97 |
| 6.4 | Stirling's Formula | 98 |
| 6.4.1 | An Estimate | 98 |
| 6.4.2 | The Derivation | 99 |
| 6.5 | Laplace's Method | 99 |
| 6.5.1 | Proceeding more carefully | 101 |
| 6.5.2 | Moving on to $N!$ | 102 |

| | | |
|----------|---|------------|
| 6.6 | The Error Function | 103 |
| 6.6.1 | Having Courage, once again | 106 |
| 6.6.2 | Optimal Truncations | 107 |
| 6.7 | Another example of an asymptotic series | 110 |
| 6.7.1 | Large x behavior | 114 |
| 6.7.2 | Asymptotic Series | 115 |
| 6.8 | Final Remarks: Looking ahead | 115 |
| 6.8.1 | Complex valued functions and essential singularities | 115 |
| 7 | Convergence and Its Uses | 117 |
| 7.1 | What is Convergence? And is It Useful? | 117 |
| 7.2 | Singularities in the Complex Plane | 119 |
| 7.2.1 | Some Important Facts about Functions in the Complex Plane | 120 |
| 7.2.2 | Multi-valuedness | 120 |
| 7.2.3 | Differentiation and Integration | 121 |
| 7.2.4 | Types of Singularities in the Complex Plane | 123 |
| 7.2.5 | Summary | 124 |
| 7.3 | Analytic Continuation | 125 |
| 7.3.1 | Approximate Analytic Continuation | 126 |
| 7.4 | Pade Approximants | 132 |
| 7.5 | Appendix: The MATLAB Code Generating Complex Function on the Whole Complex Plane | 143 |
| 8 | The Connection Problem | 147 |
| 8.1 | Matching different solutions to each other | 147 |
| 8.2 | Connection problems in Classical Linear equations | 147 |
| 8.2.1 | Bessel Functions | 148 |
| 8.2.2 | A Less famous linear second order ordinary differential equation | 154 |
| 8.3 | A nonlinear boundary value problem | 161 |
| 8.4 | Spatially dependent dominant balances | 167 |
| 8.4.1 | An Example of Carrier | 167 |
| 8.5 | Matched Asymptotic Expansions | 169 |
| 8.5.1 | When it works | 170 |
| 8.5.2 | An Example from Bender and Orszag | 171 |
| 8.5.3 | Some remarks on numerical Methods | 177 |
| 9 | Introduction to Linear PDE's | 181 |
| 9.1 | Random walkers | 181 |
| 9.1.1 | Random walk on a one-dimensional lattice | 181 |
| 9.1.2 | Derivation #1 | 182 |
| 9.1.3 | A final remark | 183 |
| 9.1.4 | Derivation #2 | 183 |
| 9.1.5 | Random walks in three dimensions | 184 |
| 9.1.6 | Remark about Boundary Conditions | 185 |

| | | |
|-----------|--|------------|
| 9.1.7 | Simulating Random Walkers | 185 |
| 9.2 | Solving the Diffusion Equation | 186 |
| 9.2.1 | Long-time Limit of the Diffusion Equation | 191 |
| 9.3 | Disciplined Walkers | 192 |
| 9.3.1 | Disciplined Walkers Moving in More complicated ways | 194 |
| 9.4 | Biased Random Walkers | 195 |
| 9.5 | Biased Not Boring Random Walkers | 196 |
| 9.6 | Combining different classes of effects | 196 |
| 9.6.1 | Combining Diffusion and Advection | 196 |
| 9.6.2 | Fokker planck equations | 197 |
| 9.6.3 | Combining Diffusion and Growth | 197 |
| 10 | Integrals derived from solutions of Linear PDE's | 201 |
| 10.1 | Integral Transforms | 201 |
| 10.2 | Contour Integration | 202 |
| 10.3 | Asymptotics of Fourier-type integrals | 202 |
| 10.4 | Rainbows and Oscillatory Integrals | 204 |
| 10.4.1 | Colors | 207 |
| 10.4.2 | Deflection angle of the rainbow and the Method of Stationary Phase | 207 |
| 10.4.3 | Stationary Phase | 209 |
| 10.4.4 | Back to the Rainbow | 210 |
| 10.5 | Saddle Points | 211 |
| 10.5.1 | Elementary Examples | 213 |
| 10.5.2 | Example 1: Finite Fourier Transform | 213 |
| 10.5.3 | Example 2: An example of Carrier | 213 |
| 10.6 | A terrible Integral | 216 |
| 10.7 | Some Applications of Stationary Phase Ideas | 219 |
| 10.7.1 | The Front of a Wavetrain | 220 |
| 10.7.2 | Free particle Schrodinger Equation | 221 |
| 10.7.3 | Waves behind a barge | 221 |
| 10.7.4 | Waves behind a boat | 222 |
| 10.7.5 | Dispersive Electromagnetic Waves | 223 |
| 10.7.6 | Absolute and Convective Instability | 224 |
| 11 | Nonlinear Partial Differential Equations | 227 |
| 11.0.7 | Solving Nonlinear Pde's using <i>Matlab</i> | 227 |
| 11.1 | The diffusion equation, and nonlinear diffusion | 229 |
| 11.1.1 | A nonlinear diffusion equation | 230 |
| 11.1.2 | Radial Nonlinear Diffusion Equation | 234 |
| 11.2 | A reaction diffusion equation | 234 |
| 11.3 | An advection diffusion equation | 243 |
| 11.4 | Burger's Equation | 250 |
| 11.5 | Pattern Formation | 250 |





1 Introduction

The goal of this course is to give a modern introduction to mathematical methods for solving hard mathematics problems that arise in the sciences — physical, biological and social. The toolbox of applied mathematics has changed dramatically over the past fifteen years.

There are two major factors that have contributed to this change. First, the dramatic increases in inexpensive computational speed have made large scale computation much more prevalent. Computers are now sufficiently fast that algorithms with minimal sophistication can perform once unthinkable large computations on a laptop PC. The consequence of this is a dramatic increase in numerical computations in the scientific literature; it is an understatement to say that most theoretical papers in the engineering sciences contain numerical computations. Even in the biological sciences it is becoming more and more fashionable to supplement traditional arguments with simulations of one kind or another.

The second major change in the toolbox of applied mathematics is the advent of fast, reliable and easy to use packages for routine numerical and symbolic computations (Matlab, Mathematica and Maple). These packages have cut the time for writing small scale computer codes dramatically, and likewise have dramatically increased the size and accuracy of analytic computations that can be carried out.

Additionally, they have, as it were, lowered the bar of required knowledge for carrying out numerical calculations. Armed with knowledge of how to run a computer package, it is possible to carry out numerical calculations solving for example a set of coupled highly nonlinear partial differential equations. Although the computer will readily spit out answers, the question then is what do these answers mean?

1.1 Computer Graphics and Mathematical Models

The most dramatic version of this question is to ask what is the difference between the numerical solution to a mathematical model, and a computer graphics animation of the same phenomenon. A computer graphics animation of fire aims to reproduce the salient features of the combustion process and visualize it so that it looks as realistic as possible. But a combustion scientist simulating this same fire does not care if the solution visually looks like fire: she is interested instead in whether the chemical and transport mechanisms are reliably represented, in order to refine understanding of why and how burning occurs. Such a scientist will be more interested in understanding for



Figure 1.1. A real picture of fire (left, iStockphoto.com) compared with an animation of fire (right, quaife.us).

example the nature of the flame front — what is burning; which chemicals are leading to the color variation; what sets the characteristic scale of the flame, and of the small features in the flame — than in making a simulation that visibly looks like fire.

Somewhat remarkably, a computer graphicist trying to simulate fire might use exactly the same mathematical structure as the scientist, despite their completely different ends. Modern algorithms for computer graphics often solve nonlinear equations motivated by physics.

Another famous example comes from the wonderful movie *Finding Nemo*. The motions of the fishes in this movie greatly resemble those of real fishes. To achieve this the animators no doubt needed to learn and study fish physiology. However, the actual mechanisms of fish locomotion are quite different than those underlying the animation. Scientists who try to study the mechanisms for fish locomotion also carry out numerical simulations of the motion; instead of focusing on their visual appeal they instead try to create as faithful a representation of the motion as possible, in order to discover how it works.

Whereas a mathematical model aims to understand something animation aims to emulate it. The difference is easy to ascertain when going to the movies. But suppose in the course of your research into some phenomenon you write a computer program to simulate it. You worked hard on your simulation and are proud of it. Is your simulation computer graphics, or does it actually teach you something about the phenomenon in question. And how do you know?

1.2 Calculating while computing

The answer to this question depends both on the model that you have formulated, and the way that you have analyzed it. Creating good models is in itself a fascinating subject, but this is not the topic of the present course. The topic of this course is how to *analyze* the output of a computer simulation to *understand why the output is what it is*. If you solve a horribly complicated mathematics problem, whether a nonlinear partial differential equations, a set of coupled differential equations, the eigenvalues of a large

matrix, etc. it is the contention of this course that you should nonetheless be able to understand in explicit terms why the solution is the way that it is. It may be difficult, it may require some approximation, but our contention is both that it is possible in general to do this, and moreover that without doing it you do not really understand what you have done.

Indeed, in recent years, an emerging trend is that while there is more and more interest in inventing algorithms for doing fast computation, or doing them on the computer architectures that have emerged, etc, there is correspondingly less interest in both learning and teaching analytical methods. After all, why should one learn how to carry out a difficult and possibly tedious approximate analytic calculation when there exist all purpose computer programs for solving all problems?

Our answer to this question is that you cannot understand the output of a computer simulation, however sophisticated it may seem, without some analysis to back up the calculation. You must convince yourself that the calculation is correct, and moreover you must understand its essential feature. It is not enough to say "Look, my computer simulation looks like the ocean." You need to explain why it looks like the ocean, which numbers you put into the computer mapped into which numbers characteristic of the ocean, etc. Without being able to do this you simply aren't doing good science and additionally you have little basis to explain why you believe the answer that the computer simulation has given.

Our aim therefore is to teach, within a broad suite of examples, how computer simulations and analytical calculations can be effectively combined. In this course, we will begin with problems that are simple-polynomial equations and first order differential equations – and slowly march our way towards the study nonlinear partial differential equations. We will show that a set of simple ideas provides a framework for developing an understanding all of these problems.

1.3 What is an analytical solution?

By and large, the analytic computations that we will emphasize in these notes are quite different than those that are usually taught in mathematical methods text books. Our focus will be on introducing methods which are *structurally stable*, in the sense that they work equally well when the mathematics problem is changed. This is opposite to the type of understanding that is usually taught. For example, in calculus, students are taught how to carry out a series of integrals which can be carried out exactly. For example, every calculus student learns that

$$\int \frac{1}{1+x^2} dx = \tan^{-1}(x), \quad (1.1)$$

so that

$$I_1 = \int_0^{\infty} \frac{1}{1+x^2} dx = \frac{\pi}{2}. \quad (1.2)$$

Calculus students pride themselves on learning these formulas; but in truth we must admit that they only have content if you happen to know how to compute $\tan^{-1}(x)$, which is after all defined by the integral!

To drive this point home, here is another example of an integral that has long been taught to (advanced) calculus students, the so-called *elliptic integral*.

$$E(x; k) = \int_0^x \frac{\sqrt{1 - k^2 t^2}}{\sqrt{1 - t^2}} dt. \quad (1.3)$$

Probably many of you have never heard of the elliptic integral. But in many circles people still say that a problem has an analytical solution if it can be reduced to elliptic integrals!

Historically, this view of what it meant to derive an analytical solution to a problem was quite a reasonable one. Since computers for carrying out direct solutions to mathematics problems did not exist, the only way to solve a problem was to reduce it to a problem whose solution had already been tabulated numerically. Tables of functions (like the arctangent, the elliptic integral, the logarithm, and whatnot) were collected together—and indeed copies of these tables were included in the back of mathematics textbooks. For example, the classical two-volume series *Methods of Theoretical Physics*, Morse and Feshbach included tables of the following functions:

1. Trigonometric and Hyperbolic Functions
2. Hyperbolic tangent of complex numbers
3. Logarithms and Inverse Hyperbolic functions
4. Spherical Harmonic Functions
5. Legendre functions for large arguments
6. Legendre functions for imaginary arguments
7. Legendre functions of half integral degree
8. Bessel functions for cylindrical coordinates
9. hyperbolic bessel functions
10. spherical bessel functions
11. Legendre functions in spherical coordinates
12. Cylindrical bessel functions
13. Mathieu functions

In 1953 when this book was published, learning the properties of these functions and how to reduce an arbitrary problem you are confronted with to a form that one of these functions can be used was the critical skill required to do calculations. In contrast, most of you have never heard of these functions and will have no need for them at any point in your life. Today, computer calculations have completely replaced the use of tables of special functions and hence the analytical manipulations that accompany their use.

So what does it mean to develop an analytical understanding to a mathematics problem? Leaving aside its outdated pragmatic nature (and the fact that no human being can compute elliptic integrals in their head), the main problem with the historical point of view is that it obscures *why* a value of an integral has the value that it does. Remembering the name of a function does not give any intuition about its properties — why it behaves one way and not another.

For example, what if we perturb our nice arctangent integral to be:

$$I_2(a) = \int_0^{100} \frac{1}{1 + x^2 + ax^7} dx \quad (1.4)$$

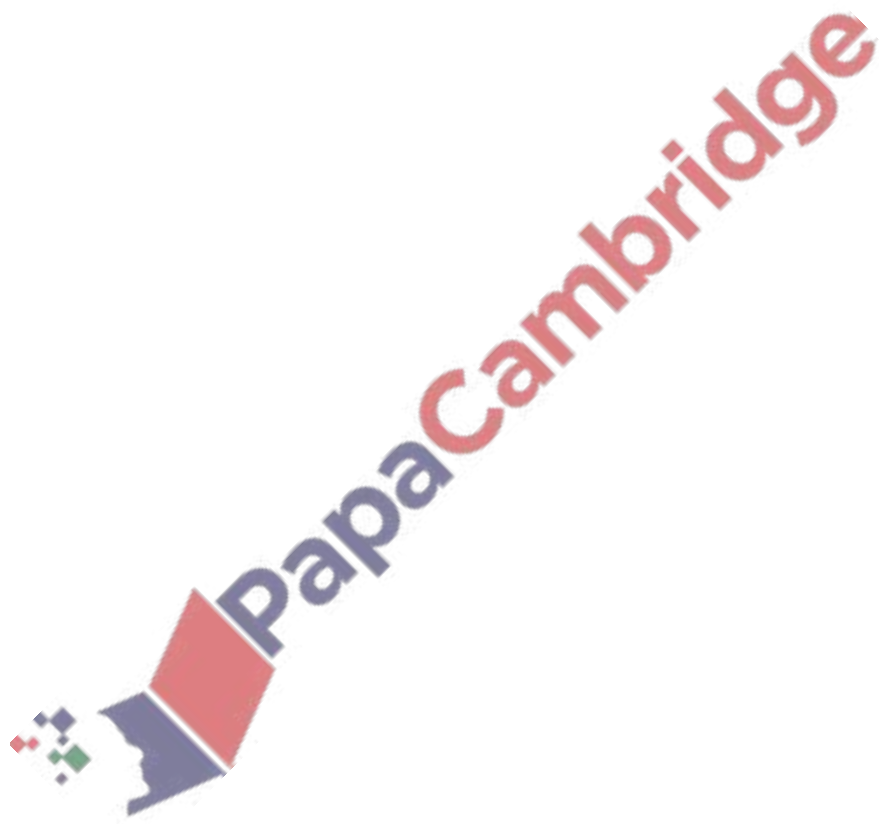
There is no special function with this name. Thus within the current system of mathematics education there is only one way to solve it — with a computer. A simple computation with Matlab yields that $I_2(a = 0.1) \approx 1.03$. Although this is a very efficient way to solve the problem it does not lead to any real understanding of why the integral has the number that it does. But on the other hand, we note that $I_2(a = 0.1)$ is close to $\pi/2$ (≈ 1.57). Is it an accident or is it a coincidence? Under what circumstances can the integral have a much different value? For example, examining the difference between I_2 and I_1 , it seems evident that $I_2(a = 0)$ should be very close to I_1 . Indeed, a numerical computation shows that $I_2(0) - \pi/2 \approx -0.01$. What determines the magnitude of this difference? What determines the rate at which $I_2(a) \rightarrow I_1$ as $a \rightarrow 0$?

Maybe you would like to name it after yourself!

Moreover how do you know that the answer is correct?

The goal of this course is to develop methods and ideas for answering this type of question in the context of the variety of different problems that arise in applications. The ideas we will discuss are weighted roughly equally between learning the mathematics and learning how to use a computer to expose and discover the mathematics. Whereas in 1953 students asking the question "how does $I_2(a)$ depend on a ?" were forced to rely solely on their wits, today computers can be used to help prod one's wits into understanding a problem.

It is our belief that developing skill in thinking about mathematics this way is crucial for educating modern students in applying mathematical and numerical methods to the sciences. Despite the relative ease of producing plausible answers to hard problems, learning numerical computation by itself is not enough. First, without having any understanding of *why* a problem has the answer that it does, one does not understand how the answer will change when the problem changes. Simply producing a graph with numbers to be compared to a phenomenon does not lead to any understanding of why the phenomenon behaves as it does. Second, without having an understanding of the answer, it is extremely difficult to determine when the numerical results are erroneous. A numerical method can be erroneous for two different reasons: either the numerical method can not solve the intended equation accurately enough, *or* the equation itself could be inaccurate, due to unjustified approximation. We will illustrate herein that both of these situations can be quite subtle; only with careful understanding can it be debugged and understood. Finally, an understanding of why the answer has the value that it does allows one to design numerical algorithms for much more difficult problems than would be possible if such an understanding did not exist.



2 Solving Unsolvable Problems: An Introductory Example

2.1 Pig farming and Polynomials

We begin this course with an example. This example is in a sense a simple one, though the procedures we use to analyze it are not. We will use these same procedures throughout the course, starting with polynomial equations and ending with nonlinear partial differential equations. The mathematics will be more complicated but the general logic remains the same.

Suppose that in the course of your research on pig farming you learn that the number of pigs that will be on a farm in a given year are related to the solutions of the following polynomial equation

$$a_0x^5 - a_1x + a_2 = 0. \quad (2.1)$$

This equation arises as the balance between the cost of pigs, which is $a_0x^5 + a_2$, and the profit that is made from the pigs a_1x . The cost contains two terms the first a_0x^5 represents the fact that pigs eat more food when there are many pigs around since they tend to chase each other around a lot; whereas a_2 is the fixed cost of maintaining the pig farm.

Since this is a quintic equation, there are five solutions; this is because of a theorem from Gauss in 1799, that showed that an n^{th} order polynomial equation has n complex roots. We need to determine how all of these solutions depend on the parameters a_0, a_1, a_2 .

How do we proceed?

Making funny comments is good.

2.2 Analytical Solutions to Polynomial Equations

At this point you might remember that at some point in your life you learned how to solve polynomial equations analytically. For example, the quadratic equation $ax^2 + bx + c = 0$ has the solution

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \quad (2.2)$$

Evidently, the definition of an 'analytic' solution used here is that the solution can be reduced to a single formula requiring only the evaluation of square roots. Indeed, fifty years ago the predominant computational device was the slide rule, which did allow the user to compute square roots. This formula was useful, because it was a mechanism to use the slide rule to solve an arbitrary quadratic equation.

It is therefore natural to wonder whether such formulas exist for arbitrary polynomial equations.

Indeed, you will be heartened to know that the cubic equation

$$y^3 + py^2 + qy + r = 0 \quad (2.3)$$

also has such an analytic solution. Substituting $y = x - p/3$, an arbitrary cubic can be reduced to

$$x^3 + ax + b = 0 \quad (2.4)$$

with $a = (3q - p^2)/3$ and $b = (2p^3 - 9pq + 27r)/27$. Defining

$$A = \left(-\frac{b}{2} + \sqrt{\frac{b^2}{4} + \frac{a^3}{27}} \right)^{1/3} \quad (2.5)$$

$$B = -\left(\frac{b}{2} + \sqrt{\frac{b^2}{4} + \frac{a^3}{27}} \right)^{1/3} \quad (2.6)$$

then the solutions are $x = A + B$ and $x = -(A + B)/2 \pm (A - B)/2\sqrt{-3}$. The price to pay for solving the cubic is that first, more arithmetic operations are required; and second, it is necessary to be able to efficiently calculate cube roots in addition to square roots. Happily, slide rules can do that too.

Even less illuminating is the exact solution of the quartic: An arbitrary quartic equation

$$x^4 + ax^3 + bx^2 + cx + d = 0 \quad (2.7)$$

can be solved exactly through the following procedure. First solve the cubic equation

$$y^3 - by^2 + (ac - 4d)y - a^2d + b^4d - c^2 = 0. \quad (2.8)$$

If y is any root of this equation and $R = \sqrt{a^2/4 - b + y}$ is nonzero then if we let

$$D = \sqrt{3a^2/4 - R^2 - 2b + (4ab - 8c - a^3)/4R} \quad (2.9)$$

and

$$E = \sqrt{3a^2/4 - R^2 - 2b - (4ab - 8c - a^3)/4R} \quad (2.10)$$

then the four roots of equation (2.7) are $x = -a/4 + R/2 \pm D/2$ and $x = -a/4 - R/2 \pm E/2$. The formula is not pretty; as a computational device for a person who can compute roots (with a slide rule!) it is quite useful though.

The complexity of these solutions is sufficiently daunting that you will probably be left with mixed emotions on learning the theorem due to Abel and Galois that there is no such formula which allows solving quintics or any higher order polynomial equation exactly with a finite number of arithmetic operations. Any disappointment at the lack of such a solution is mitigated by the realization that even the formulas for the cubic and the quartic are of marginal use for computing or understanding the solutions. The fact that the method of solution breaks down for quintics and higher order equations reflects that the method of solution is not robust.

In ancient times, the lack of solutions to such equations was problematic because it literally meant that there was not a simple method for finding the roots to higher order polynomial equations.

2.3 Dominant Balance and Approximate Solutions

So what are we to do? The history of applied mathematics has given two different approaches for finding solutions that *are* structurally stable, in that they allow finding solutions to arbitrary equations, without regard to the equation's order:

1. Numerical solutions. These generate numbers to any desired degree of accuracy. Modern computers allow solving for finding all the roots of arbitrary polynomial equations quite easily.
2. Alternatively, one can relax the requirement that the analytic solutions are exact (with a finite number of arithmetic operations) and instead seek analytic formulas which approximate the solution to a (perhaps arbitrarily) high degree of accuracy.

We will see in what follows that these two approaches complement each other well: numerical solutions robustly produce arbitrary answers with arbitrary accuracy, whereas approximate solutions produce simple formulae which capture the solution (with some error) over a wide parameter space. In some cases we will see that approximate solutions give crucial input to generate improved numerical calculations, whereas in other cases the numerical calculations will give us confidence that the approximations we have made are good ones.

In what follows, we will solve our quintic using a very simple but exceedingly powerful idea — that we will return to again and again during our course — the so-called **Method of Dominant Balance**. The idea is as follows:

The Method of Dominant Balance

Suppose you are given an equation in the form

$$A + B + C + D = 0,$$

where A, B, C, D are different terms in the equation. These terms could represent elements of a polynomial equation (e.g. x^5) or could represent terms in an ordinary differential equation or for that matter terms in a nonlinear partial differential equation. It is invariably the case (unless you are extraordinarily unlucky) that **two of the terms are larger than all of the others**. For example, it could be that A, D are larger in size than B, C . If this were the case we can then "approximate" the original equation by the neglecting B, C entirely and instead solving

$$A + D = 0.$$

We can then check that this reduced equation is "consistent", by taking the solution to the reduced equation, plugging it back into the original equation, and verifying that the neglected terms (B, C) are indeed smaller than the two terms we have kept. If they are, our hypothesized solution is "consistent"; if not the solution is "inconsistent" and we must find another dominant balance.

We will employ the method of dominant balance continuously in the rest of this course.

2.4 Nondimensionalization

Let us now return to our quintic. The first step in making progress is perhaps the easiest: we are interested in how the solution depends on the parameters a_0, a_1, a_2 . Let us write $x = a_2/a_1 y$, and substitute this into (3.2). This substitution yields

$$a_0 \left(\frac{a_2}{a_1} \right)^5 y^5 - a_2(y - 1) = 0 \quad (2.11)$$

Now dividing through by a_2 yields

$$\frac{a_0}{a_2} \left(\frac{a_2}{a_1} \right)^5 y^5 - y + 1 = 0. \quad (2.12)$$

If we now define

$$\epsilon = \frac{a_0 a_2^4}{a_1^5} \quad (2.13)$$

the equation is

$$\epsilon x^5 - x + 1 = 0. \quad (2.14)$$

(where we have taken the liberty to rename y again by x).

At first sight these might seem like mere algebraic manipulation. But we have learned something already that is very important about the solutions to this equation: namely they depend only on a single combination of the parameters a_0, a_1, a_2 . Hence if we determine how the solutions depend on the single parameter ϵ we will have learned their dependence on all of the parameters of the problem.

We will now proceed to discuss how to find out the dependence of the solutions on ϵ . Our goal here is twofold: first, we want to answer a qualitative question about the roots to the equation (2.14): How many of them are real? We will see that a rational guess for this can be developed without having to compute the values of the roots explicitly. Then we will become more adventurous, and develop accurate analytic expressions that approximate each of the roots to this equation. We will see that with enough work (without the help of a computer), one can develop formulae that are accurate to any desired accuracy.

We begin by seeking expressions for the roots in the two limits $\epsilon \rightarrow 0$ and $\epsilon \rightarrow \infty$, and then try to match the expressions in the middle to find the answer to our question.

2.4.1 A Dominant Balance: $\epsilon \rightarrow 0$

First we consider the limit where $\epsilon \rightarrow 0$. There are three terms in equation (2.14), ϵx^5 , $-x$ and 1. For the equation to be satisfied, these terms must sum to zero; typically in a distinguished limit (like that of $\epsilon \rightarrow 0$) two of the terms will be much larger than the other. To discover formulae for the roots we need to examine each of the possible balances and decide which ones are consistent.

Balancing $-x \approx 1$

The first natural possibility is the balance between $-x$ and 1. This balance implies that to a first approximation, $x \approx 1$. For this balance to be self consistent we need $\epsilon x^5 \ll 1$, which is satisfied as long as ϵ is small. Hence, this balance is *consistent*.

We can develop an improved formula for this root by noting that the error we are making in satisfying the equation when we set $x = 1$ is $O(\epsilon)$. This motivates the expansion

$$x = 1 + \sum_{n=1}^{\infty} a_n \epsilon^n. \quad (2.15)$$

To find the values of the a_n 's, we plug equation (2.15) into equation (2.14), and choose the a_n 's to satisfy the equation order by order in ϵ . Plugging in, we obtain

$$\epsilon \left(1 + \sum_{n=1}^{\infty} a_n \epsilon^n \right)^5 - 1 - \sum_{n=1}^{\infty} a_n \epsilon^n = 0. \quad (2.16)$$

At leading order $O(\epsilon)$ this gives $\epsilon - a_1\epsilon = 0$, implying $a_1 = 1$. At $O(\epsilon^2)$ we have $5a_1\epsilon^2 - a_2\epsilon^2 = 0$, implying $a_2 = 5a_1 = 5$. At $O(\epsilon^3)$ we have that $\epsilon^3 5a_2 + \epsilon^3 10a_1^2 - \epsilon^3 a_3 = 0$, or $a_3 = 35$. Hence we have shown that

$$x = 1 + \epsilon + 5\epsilon^2 + 35\epsilon^3 + \dots \quad (2.17)$$

With enough energy one could of course compute this to any order. A bit later we will discuss how such calculations can be done with symbolic manipulation packages (such as *Matlab* or *Maple*).

Before proceeding to find the other roots, note one important feature of the expansion in equation (2.17): the magnitude of the coefficients a_n are increasing with n . The consequence of this is that we anticipate that the radius of convergence is going to be less than 1, implying that this formula is not going to work all the way to $\epsilon = 1$ (our goal)! Roughly speaking, the radius of convergence of the series in equation (2.17) can be estimated as $\left| \frac{a_{n+1}\epsilon^{n+1}}{a_n\epsilon^n} \right| < 1$, therefore $\epsilon < \frac{a_2}{a_3} = 1/7$.

Balancing $x^5 \sim -1$

We have so far only found one solution to our equation (2.14), but we know there must be five! We must therefore turn to other possible leading order balances. Consider the balance $\epsilon x^5 \sim -1$. This implies that $x \sim \frac{(-1)^{1/5}}{\epsilon^{1/5}}$. For this balance to be consistent the neglected term $x \sim -\epsilon^{-1/5}$ must be smaller than the terms we have kept. But it is not! In the limit $\epsilon \rightarrow 0$, the size of the neglected term diverges, whereas the two terms we have kept are order unity. Therefore this is not a consistent balance.

Of course, if this *had* turned to be a consistent balance we would have been in real trouble. There are five roots to $1^{1/5}$ which would have led to a grand total of 6 solutions to our quintic, which we know is impossible!

Balancing $x^5 \sim x$

The only possible other balance is $\epsilon x^5 \sim x$. This balance leads to the four roots $x \sim \frac{1^{1/4}}{\epsilon^{1/4}}$. The size of the neglected term 1 is much smaller than the size of the terms that we have kept $O(\epsilon^{-1/4})$, so this balance is self consistent. These four solutions together with the solution we derived above make up the four roots of our quintic!

To develop better approximations to these roots, it is convenient to write $x = \frac{y}{\epsilon^{1/4}}$. Plugging this into our quintic equation (2.14) leads to the equation for y

$$y^5 - y + \epsilon^{1/4} = 0, \quad (2.18)$$

so we anticipate that there will be an expansion in $\epsilon^{1/4}$. We write

$$y = 1^{1/4} + \sum_{n=1}^{\infty} b_n (\epsilon^{1/4})^n, \quad (2.19)$$

and plug into equation (2.18). To leading order in $\epsilon^{1/4}$ we find that $b_1 = \frac{-1}{4}$, and so on.

Let us now reflect on what we have found. In the limit $\epsilon \rightarrow 0$ we have found five roots, 3 of which are real, and 2 of which are purely imaginary. Based on the coefficients in the perturbation series for the first real root we expect something to happen to this root for $\epsilon \sim 1/7$. Thus we are not able yet to make a statement about the number of real roots at $\epsilon = 1$.

2.4.2 The Second Dominant Balance : $\epsilon \rightarrow \infty$

The other natural limit to consider is $\epsilon \rightarrow \infty$. Considering the balance arguments we made above, the only consistent possibility in this limit is that $\epsilon x^5 \sim -1$, implying that $x \sim \frac{(-1)^{1/5}}{\epsilon^{1/5}}$. Letting $x = y/\epsilon^{1/5}$ we have

$$\epsilon^{1/5}(y^5 + 1) - y = 0. \quad (2.20)$$

Letting $y = (-1)^{1/5} + \sum_{n=1}^{\infty} c_n(\epsilon^{1/5})^n$ gives a set of equations (order by order in $\epsilon^{1/5}$) for the c_n 's.

The most important qualitative lesson from this limit is that since the roots at $\epsilon \rightarrow \infty$ are fifth roots of unity, there is only one purely real root, and four complex roots. This should be contrasted with the situation for $\epsilon \ll 1$ where there are three real roots and two complex roots.

2.5 Testing the theory with numerical simulations

We have thus developed an analytical hypothesis for the nature of the roots in both the limit of small ϵ and large ϵ . We now turn to test the theory, to see how well it works. Figure 2.1 plots the numerically calculated positive real roots of equation (2.14) (solid lines), as well as the analytic expressions derived above (dashed lines). Note that the figure is a log-log plot, expressing the logarithm of the root as a function of the logarithm of ϵ .

The great advantage of double log plots (which we will use frequently in these notes) is that if $x = \epsilon^p$ for some power p , then $\log(x) = p \log(\epsilon)$. Hence the power law is a straight line on a log-log plot and the power can be read off as the slope.

You can see that the two positive real roots merge at $\epsilon \approx 0.08$ after which they no longer exist. The dashed lines follow the roots very accurately up to this merger point. The dotted line shows the $\epsilon^{-1/4}$ law, that is valid to leading order at small ϵ . Figure 2.2 shows the negative real root. This real root exists for all ϵ . The plot shows that the root diverges as $\epsilon^{-1/4}$ at small ϵ and decreases as $\epsilon^{-1/5}$ at large ϵ , just as predicted.

All in all, the theory works extremely well.

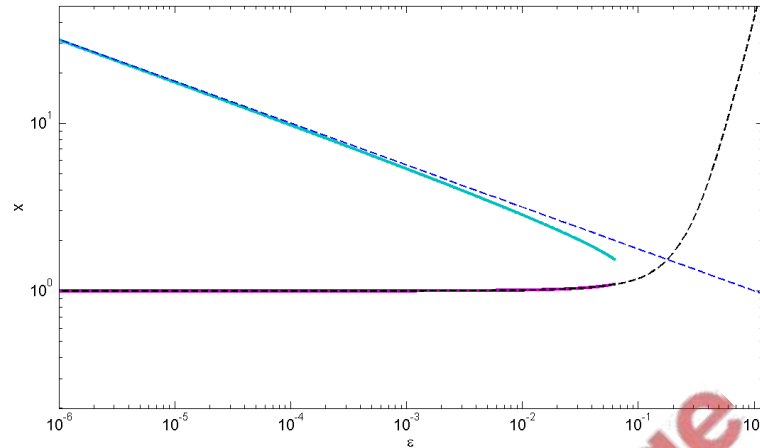


Figure 2.1. Plot of the numerically calculated negative real roots of equation (2.14) (blue line), as well as the analytic expressions derived above (black lines).

Program 1 MATLAB code used to create figure 2.1

```

1  %Calculate the numerical roots
2
3  nn=-6:0.001:-1.2;
4  eps=10.^nn;
5  for i=1:length(eps)
6      c=[eps(i) 0 0 0 -1 1];
7      dum(:,i)=roots(c);
8  end
9
10 loglog(eps,dum,'linewidth',3)
11 xlabel('\epsilon','fontsize',16);
12 ylabel('x','fontsize',16);
13 hold on;
14
15 %Plot the analytic expressions
16
17 eps=10^-6:0.001:1.2;
18 y1=1+eps+5*eps.^2+35*eps.^3;
19 y2=(1./eps).^(1/4);
20 plot(eps,y1,'-k',eps,y2,'--','linewidth',2);
21
22 set(gca,'fontsize',16)
23 axis([10^(-6),1.2,10^(-0.7),10^(1.7)]);

```

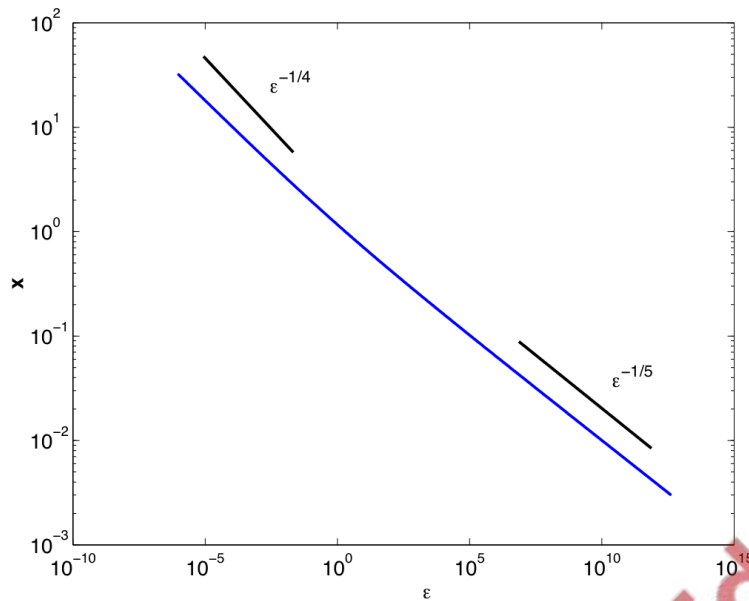


Figure 2.2. Plot of the numerically calculated negative real roots of equation (2.14) (blue line), as well as the analytic expressions derived above (black lines). At small ϵ the root diverges as $\epsilon^{-1/4}$, whereas at large ϵ it decreases as $\epsilon^{-1/5}$.

2.6 A bifurcation (phase transition)

One thing that is evident from both the analytical and theoretical treatment is that at small ϵ there are two complex roots, whereas at large ϵ there are four complex roots. There is a value of ϵ where the transition between these two regimes occurs. Where is it? What is the nature of the transition?

This transition in the behavior of the solution is an example of a mathematical bifurcation. An important question in understanding the qualitative structure of solutions to equations is to understand **how** and why bifurcations occur. Here we can understand what is going on by borrowing **some** intuition from quadratic equations: from equation (2.2) for the roots of a quadratic equation, we see that the solutions can transition from two real roots to two complex roots when the discriminant $b^2 - 4ac$ switches sign. This would be a simple mechanism for two of the real roots to annihilate each other and form two complex roots at an intermediate value of ϵ , and would occur if the two roots had the behavior $x \sim \pm\sqrt{\epsilon_* - \epsilon}$ for $\epsilon \approx \epsilon_*$. When $\epsilon < \epsilon_*$ there are two real roots, while when $\epsilon > \epsilon_*$ there are two complex roots.

There is a simple analytical way of testing this theoretical idea. At the value of ϵ when the number of complex roots changes from 2 to 4, there has to be a transition value of ϵ where the two roots that change are both real, and where they are identical. This is what happens in the mechanism from quadratic equations outlined above.

This can be expressed in mathematical terms as follows: At the critical ϵ , $F(y) = \epsilon y^5 - y + 1$ can be factored in the form $F(y) = (y_* - y)^2 G(y)$. The consequence of this is that at the critical ϵ , $dF/dy = 0$ is also satisfied! We can use this to our advantage. Since $dF/dy = 5\epsilon y^4 - 1$, we have $y = 1/(5\epsilon)^{1/4}$. Using this in the equation $F = 0$ gives a formula for ϵ , namely

$$\epsilon_* = R = \left(\left(5^{1/4} \right)^{-1} - \left(5^{5/4} \right)^{-1} \right)^4 \quad (2.21)$$

or $\epsilon = 0.0819\dots$

Note that this transition occurs precisely at this value. The singularity that the radius of convergence is pointing to is precisely the one that causes the number of real roots to switch. Thus from examining the behaviors of the roots near both $\epsilon \sim 0$ and $\epsilon \sim \infty$ we have a rather complete qualitative picture of the solutions!

2.7 Lessons Learned

Let us now summarize the lessons from this example.

1. The first step in our analysis was nondimensionalization. By this, we mean the rescaling that occurred when we transitioned the equation $a_0 x^5 - a_1 x + a_2 = 0$ to the equation for x depending only on ϵ . This step demonstrated that the solution set should only depend on a single parameter combination, and set us up for the idea that the solutions might have two regimes, one at small ϵ and another at large ϵ .
2. The second step was to consider the various possible limits when our parameter ϵ became large or small.

In these limits we could use the idea of *dominant balance* — namely of the three terms in the equation, only two dominate. Assuming different dominant balances we were able to understand quantitatively why the solutions were what they are. We developed a theory for how many real and complex roots there are and how it depends on ϵ .

3. We then tested our theory with numerical simulations and demonstrated it to be quantitatively accurate. Indeed any lack of confidence you might have felt from the analysis should have been completely assuaged with the numerical simulations

In the rest of this course, we will apply this procedure to examples of increasing mathematical complexity. First we will continue with more on polynomial equations to give you more experience and flavor for these arguments; then we will move on to ordinary differential equations, integrals and eventually to nonlinear partial differential equations. Although solving the individual mathematical problems that arise will become harder, the logical framework will remain the same.

3 Dimensions and Dimensional Analysis

3.1 Introduction

Before proceeding with more mathematical examples, we want to take a moment and think about the "nondimensionalization" step in the introductory example. Recall that we made the substitution $x = (a_2/a_1)y$, and substituted this into (3.2). The substitution ultimately led to our favorite quintic, namely $\epsilon x^5 - x + 1 = 0$.

Why did we call this dimensionalization? Every mathematical equation, including this one, arises in applications as relationships between different quantities of interest. For example, one might wish to formulate an equation for the price of a stock as a function of time; or the wind velocity as a function of position in the middle of Boston Harbor; or the energy levels of a molecule.

A simple statement can be made about the structure of the equation without any knowledge of the underlying system: Namely, the equation relates quantities to each other that have different *dimensions*. For this reason, the relationship must contain certain constants which allow the equation to convert between the different dimensions. This observation is so simple that one might expect it to have very little content: however, the opposite is the case! One can learn a lot about a problem by just thinking through the dimensions.

There are two consequences of the idea: first, one can often guess the form of an answer just by thinking about what the dimensions of the answer should be, and then expressing the answer in terms of quantities that are known to have those dimensions. Second, the scientifically interesting results are *always* expressible in terms of quantities that are *dimensionless*, not depending on the system of units that you are using.

In our polynomial problem, the different terms in the quintic a_0x^5 , $-a_1x$ and a_2 all have the same dimensions. Suppose x is the number of pigs, and the equation represents different dollar amounts associated with owning the pigs. For example:

- a_0x^5 could represent the amount of money spent on food required to house x pigs
- a_2 denotes the cost of owning a barn and a_1x denotes the money made on selling x pigs.

Hence in terms of dimensions,

- a_0 has the units of *dollars/(number of pigs)⁵*.
- a_1 has the units of *dollars/(number of pigs)*.

Pigs!! Can't Harvard professors be more creative than this?

- a_2 has the units of *dollars*.

The only dimensionless number is the combination

$$\frac{a_0 a_2^4}{a_1^5}, \tag{3.1}$$

which indeed our expression for ϵ .

In the present polynomial problem, there is only a single dimensionless parameter. If we had considered the more general quintic $a_0 x^5 + a_1 x^4 + a_2 x^3 + a_3 x^2 + a_4 x + a_5 = 0$ the number of dimensionless parameters will be larger .

There is a famous theorem that determines how many dimensionless parameters are needed to specify a given problem. The theorem summarizing this is called Buckingham's theorem, the general principle of dimensional analysis. We will present this theorem, and then present some simple examples of dimensional analysis to illustrate the power of the approach. And finally we will demonstrate in a few explicit examples how to put an equation into dimensionless form, thus illustrating how " ϵ " arises in some concrete cases.

3.2 Buckingham's π Theorem

This theorem was discovered/formalized by E. Buckingham *Phys. Rev.* , 4, 345-376, 1914. It is amusing to note that Buckingham worked at the U.S. Bureau of Standards!, building upon an earlier idea that (he credits to) Fourier that all of the terms of a meaningful equation must have the same dimensions.

The idea is the following: suppose you are given a problem which is characterized by some number of parameters $\{Q_i\}, i = 1 \dots N$. These parameters in general have dimensions, and for this reason they are not all independent. Buckingham's theorem states that any meaningful statement about the system

$$\Phi[\{Q_i\}] = 0 \tag{3.2}$$

is equivalent to another statement

$$\Psi[\{\Pi_n\}] = 0 \quad n = 1 \dots N - r, \tag{3.3}$$

where the variables Π_n are dimensionless. The main point is that the second relation contains r fewer variables than the first relation.

The basic reason behind Buckingham's theorem is that any problem has some number of "fundamental units" that must be specified for the problem to make sense. For example, in a problem involving Newton's laws of motion, we must specify the units of mass, length and time. Equation (3.2) contains quantities specifying this list of units, whereas equation (3.3) does not, since the variables Π_n are dimensionless. The number r is therefore just the number of fundamental units that need to be specified.

Actually, any function of this dimensionless number is also dimensionless.

How many are there?

3.3 Examples

The main point of Buckingham's theorem is that the best way to write a mathematical relation between variables is in dimensionless form. Sometimes, by writing the problem in dimensionless form one can learn everything one wants to know about it. This lucky situation sometimes happens, and sometimes doesn't; when it doesn't, calculation is needed, hence requiring the methods of this course!

In what follows we will go through several examples of dimensional analysis, exposing both when it works and when it doesn't work.

3.3.1 The Pendulum

This is a problem that you all know quite well. Consider a pendulum with length L and mass m , hanging in a gravitational field of strength acceleration of gravity g . What is the period of the pendulum? We need to construct a quantity with units of time involving these numbers. Mathematically we are seeking for the form $T = f(m, L, g)$, to balance the dimensions in both sides, the mass must not appear in the function. Therefore the only possible way of doing this is the combination $\sqrt{L/g}$. Therefore, we know immediately that

$$T_{\text{pendulum}} = c\sqrt{\frac{L}{g}}. \quad (3.4)$$

Here c is a dimensionless constant. Thus, we arrive at the result that doubling the length of the string increases the period by $\sqrt{2}$.

Before celebrating too much about this result, let's think about the assumptions and limitations. In writing equation (3.4) we have assumed that various parameters were not important, including (a) damping; and (b) the amplitude of the oscillation. The damping parameter ν has the dimensions of time^{-1} , whereas the amplitude A has the dimensions of length. Again we are looking for the function $T = f(m, L, g, \nu, A)$. There are 3 fundamental units: time, mass and length. According to Buckingham's theorem, it has 3 independent dimensionless parameters in this problem. Therefore if we account for these effects, the best formula we can write for the period is

$$T_{\text{pendulum}} = \sqrt{\frac{L}{g}}\Phi\left(\sqrt{\frac{\nu^2 L}{g}}, \frac{A}{L}\right). \quad (3.5)$$

Here, $\Phi(\alpha, \beta)$ is function of the two dimensionless variables, and $\Phi(0, 0) = c$. Thus we have learned that as long as $\Phi(\alpha, \beta)$ is not singular as $\alpha, \beta \rightarrow 0$ our original result (3.4) holds; however in general we expect the relationship between period and these parameters to be more complicated.

The only way to uncover this relationship is to solve the equation of motion. For a pendulum displaced an angle θ from the vertical, according to force balance, the equation of motion is given by

$$m \frac{d^2(L\theta)}{dt^2} + m\nu \frac{d(L\theta)}{dt} + mg \sin(\theta) = 0. \quad (3.6)$$

It is simplified as

$$L \frac{d^2\theta}{dt^2} + \nu L \frac{d\theta}{dt} + g \sin(\theta) = 0. \quad (3.7)$$

To proceed further and compute Φ we would like to rewrite this equation so that only the dimensionless variables appear: to do this, we need to rescale time in terms of $\sqrt{L/g}$. Therefore we introduce the new time variable $t = \sqrt{L/g}T$. If we now rewrite equation (3.7) in terms of T using the chain rule

$$\frac{d}{dt} = \frac{dT}{dt} \cdot \frac{d}{dT} \quad (3.8)$$

we obtain

$$\frac{d^2\theta}{dT^2} + \alpha \frac{d\theta}{dT} + \sin(\theta) = 0. \quad (3.9)$$

The equation therefore now apparently involves a single parameter, $\alpha = \sqrt{\frac{\nu^2 L}{g}}$.

In order to fully specify a solution to equation (3.9) we must also give initial conditions. If these conditions are given as $\theta(T=0) = \theta_0$ and $\dot{\theta}(T=0) = 0$, we then have specified the second dimensionless parameter $\beta = \theta_0$. The solution to equation (3.9) encodes the dependence of the period on these parameters. Clearly we see the dimensionless form $T = \sqrt{\frac{L}{g}}\phi(\alpha, \theta_0)$ is recovered

Now, to find out more about $\Phi(\alpha, \beta)$ we need to actually solve the equations—this is where 'physics' ends and mathematics begins! If this were a course in pure mathematics we would try to rigorously demonstrate the properties of the function Φ . The approach of this course will instead be to take the approach of examining the *limiting cases* $\alpha, \beta \rightarrow 0, \infty$, and then try to fit the limiting cases together.

We will not carry out a study of the limiting cases here, other than to note certain general features. The first natural limits to assume are $\beta, \alpha \rightarrow 0$. Here it is natural to start by assuming that the limit is a regular one, so that nothing singular happens to the function Φ . In this case $\Phi(\alpha, \beta \rightarrow 0) \rightarrow \Phi(0, 0)$. We expect the corrections to this limit to be $O(\alpha, \beta)$ (i.e. linearly proportional to α and β .) It is straightforward to verify that in the present problem the limit is indeed regular, though we will see in this course that this is not generally the case.

3.3.2 Planetary orbits

Present a derivation of Kepler's laws, starting from Newtonian mechanics and the r^{-2} law.

3.3.3 The size of atoms

The size of an atom is given by the solution to the Schrodinger equation, namely

$$i\hbar \frac{\partial}{\partial t} \Psi(\mathbf{r}, t) = -\frac{\hbar^2}{2m} \nabla^2 \Psi(\mathbf{r}, t) + V(\mathbf{r}) \Psi(\mathbf{r}, t), \quad (3.10)$$

where $V(r)$ is the interaction potential between the electrons and the nucleus. For the hydrogen atom this is particularly simple and the solution to this equation gives the probability distribution that the electron will be a distance r away from the nucleus. But we can solve for the characteristic scale even without solving the equation. The parameters in the equation are \hbar , Planck's constant, the mass of the electron m_e and the electric charge e . Planck's constant has the dimension of Energy-Time, whereas we must recall that e^2/Length is also an energy scale (recall Coulomb's law). The only way to make a length scale out of these numbers is through

$$a_0 = \frac{\hbar^2}{m_e e^2} = 0.53 \times 10^{-8} \text{cm.} = 0.053 \text{ nm} \quad (3.11)$$

This is the famous Bohr radius.

We can also extend this argument to estimate the density of matter. Roughly we say there is one proton for every bohr radius: remember that the number of protons and electrons increase together to maintain electrical neutrality. Hence the density is given by

$$\rho_0 = \frac{m_p}{a_0^3} = 1.4 \text{g/cm}^3 \quad (3.12)$$

Try to nondimensionalize the Schrodinger equation for the Hydrogen atom using this length scale!

3.3.4 Fluid Viscosity

Let us compute the viscosity of a fluid by dimensional analysis. The viscosity is a number ν with units of $\text{Length}^2/\text{Time}$ that characterizes the momentum transport in a fluid. Momentum transport is roughly characterized by the molecules in a fluid bumping into each other. What numbers characterize these interactions? For a fluid at temperature T , we have the sound velocity (itself given by dimensional analysis as $c \propto \sqrt{k_B T/m}$, with m the mass of a particle, and the intermolecular scale a . Now by dimensional analysis, the viscosity must have the functional form

$$\nu \sim c \times a. \quad (3.13)$$

Now, the speed of sound in water (and indeed in most fluids) is about 1000m/sec ; the characteristic distance between water molecules is about $a = 10^{-9} \text{m}$. Hence the viscosity of water should be about

$$\nu \sim 10^3 \text{m/sec} \times 10^{-9} \text{m} = 10^{-6} \text{m}^2/\text{sec} = 10^{-2} \text{cm}^2/\text{sec}. \quad (3.14)$$

Indeed, this is essentially exactly the viscosity at room temperature! Note that our little theory also gives the temperature and molecular weight dependence of the fluid viscosity.

More serious theories of viscosity (based e.g. on the so-called Chapman Enskog expansion of the Boltzmann equation) solve an integral equation for ν : since the parameters in the equation are necessarily those that we have used here c and a , the theory computes our answer up to a prefactor, typically of order unity.

A thought question If the argument we have given here is correct then really all simple fluids should have the same viscosity: after all, the molecular size is essentially constant, and the sound velocity is also rather constant. However if you look in the CRC handbook, you will note that there are some fluids e.g. glycerol that have much higher viscosity than that predicted by our simple argument. Why?

The size of animals

A famous essay in biology was written by the evolutionary biologist J.S. Haldane, *On Being the Right Size*. In it he wrote

The most obvious differences between different animals are differences of size, but for some reason the zoologists have paid singularly little attention to them. In a large textbook of zoology before me I find no indication that the eagle is larger than the sparrow, or the hippopotamus bigger than the hare, though some grudging admissions are made in the case of the mouse and the whale. But yet it is easy to show that a hare could not be as large as a hippopotamus, or a whale as small as a herring. For every type of animal there is a most convenient size, and a large change in size inevitably carries with it a change of form.

He then goes on to ask what are the constraints that lead to things to be the size they are. He points out several basic facts: Let us call the size of an animal by R .

1. The weight of the animal increases like R^3 , assuming that the density is constant. Indeed you can assume the density is roughly that estimated above in terms of the fundamental constants of nature!
2. The larger that something is, the easier it is to break when it falls over.

This leads to an interesting question, that I'm hoping one of you will fill in the answer to: What is the critical size of an animal so that it breaks when it falls over?

3.3.5 Atomic Energy Scale

What is the characteristic energy of electrical binding in atoms? We know that the interactions are electrostatic, and we have already calculated the size of an atom a_0 . Therefore the electrostatic energy is

$$\frac{e^2}{a_0} = \frac{e^4 m_3}{\hbar^2}. \quad (3.15)$$

Evaluating this number, it is about 27eV.

3.3.6 Man's Size

Bill Press, when teaching a physics class at Harvard, took this argument a step further. He asked whether he could express the size of a human in terms of fundamental physical constants, and argued that the size of man is given by

$$\left(\frac{\hbar^2}{m_e e^2}\right) \left(\frac{e^2}{G m_p^2}\right)^{1/4}, \quad (3.16)$$

where e is the electron charge, m_e, m_p is the mass of electrons and protons, G is Newton's gravitational constant.

3.3.7 The Radius of the Earth

Press also gives a very interesting argument for the radius of the earth. He argues that the earth's atmosphere does not contain hydrogen, and therefore the thermal velocity of hydrogen in the atmosphere must be above the escape velocity.

For a body to escape the earth's gravitational field, it needs a velocity

$$v^2 = \frac{GM_{earth}}{R_{earth}}. \quad (3.17)$$

On the other hand, the thermal velocity of hydrogen is given by $v^2 = \frac{k_B T}{m_p}$, where k_B is Boltzmann's constant, and T is the temperature of our atmosphere (300K). Thus we need

$$\frac{k_B T}{m_p} = \frac{GM_{earth}}{R_{earth}}. \quad (3.18)$$

This gives one relationship between the mass of the earth M_{earth} and its radius. Another relationship comes from our claim that the density of matter is given by the expression above. Namely

$$\frac{M_{earth}}{R_{earth}^3} = \frac{m_p}{a_0^3}. \quad (3.19)$$

If we combine these two equations, we arrive at the following expression for the radius of the earth

$$R_{earth} = a_0 \left(\sqrt{\frac{k_B T}{G m_p^2 / a_0}} \right) \quad (3.20)$$

Note the number in brackets is dimensionless: this is the (square root of the) ratio of a thermal energy to the gravitational energy between two protons separated by the Bohr

radius! Assuming $T = 300K$, the thermal energy $k_B T = 4.1 \times 10^{-21}$ J, whereas the gravitational attraction between two protons separated by the Bohr radius is $Gm_p^2/a_0 = 3.5 \times 10^{-54}$ J. Hence we have that

$$R_{earth} = 3.5 \times 10^{16} a_0, \quad (3.21)$$

which is 1.8×10^8 cm.

This is to be compared with the actual answer of 6.4×10^8 !

3.3.8 Another Calculation of the Earth's Radius

I was very enthusiastic about the elegance of this formula, and once enthused about it to Prof. P. Chaikin, from NYU. Chaikin pointed out to me that the problem with this argument is that it assumes that one has made a measurement of the atmosphere of the earth, to check that there is no hydrogen. He argued that if we were going to make such a measurement we might as well just measure g , the gravitational acceleration of a body, and use that to make our calculation. So let's try this: We now combine

$$\frac{GM_{earth}}{R_{earth}^2} = g, \quad (3.22)$$

with

$$\frac{M_{earth}}{R_{earth}^3} = \frac{m_p}{a_0^3}. \quad (3.23)$$

These give *another* beautiful formula, namely

$$R_{earth} = a_0 \left(\frac{g}{Gm_p/a_0^2} \right), \quad (3.24)$$

so the earth's radius is the Bohr radius multiplied by the ratio of the measured gravitational acceleration, and the acceleration that a mass feels one Bohr radius away from the center of a proton!

Evaluating numbers we find that the ratio of these accelerations is 2.28×10^{17} , so that

$$R_{earth} \sim 1.17 \times 10^9 \text{ cm}. \quad (3.25)$$

This overestimates the radius by a factor of two!

3.3.9 Taylor's Blast

This is a famous example, of some historical and fluid mechanical importance. The myth goes like this: In the early 1940's, there appeared a picture of the atomic blast on the cover of Life magazine. GI Taylor, an applied mathematician at Cambridge, wondered what the energy of the blast was. When he called his friends at Los Alamos and asked, they

informed him that it was classified information. So he resorted to dimensional analysis. Let's imagine that the energy of the blast is E_0 . The blast starts from a spatial point. The dynamics basically just pushes the air out of the way. The speeds of this process are enormous so viscosity isn't important. The only important material parameter is the density of air ρ_0 . So, let's ask: what is the radius $R(t)$ of the blast as a function of time t from the detonation time? We need to create an object with units of length out of E_0 , ρ_0 and t .

Now E_0/ρ_0 has the dimensions of L^5/T^2 . Thus, $(E_0/\rho_0 t^2)^{1/5}$ has the units of length. Therefore

$$R(t) = c(E_0/\rho_0 t^2)^{1/5}. \quad (3.26)$$

Now suppose you want to estimate the energy of the blast. c is a constant we don't know, but it is probably ≈ 1 . We can measure the radius of the explosion from the cover of Life magazine. We know how long it has been since the blast, since the limits of strobe photography are around 1μ sec. We know the density of air. Thus, one can solve for E_0 .

The story is that Taylor called up his friends at Los Alamos and told them the correct number. They were deeply worried that security had been breached.

The mathematical basis for Taylor's calculation requires finding a particular solution to the equations of gas dynamics, assuming that the gas is initially at very high density.

3.3.10 Pythagorean Theorem:

Now we try to prove the Pythagorean theorem by dimensional analysis. Suppose you are given a right triangle, with hypotenuse length L and the smallest acute angle ϕ . The area A of the triangle is clearly $A = A(L, \phi)$. Since ϕ is dimensionless, it must be that

$$A = L^2 f(\phi), \quad (3.27)$$

, where f is some function that we don't know.

Now, the right triangle can be divided into two little right triangles by dropping a line from the other acute angle which is perpendicular to the hypotenuse. These two right triangles have hypotenuses which happen to be the two other sides of our original right triangle, let's call them a and b . So we know that the area of the little right triangles are $a^2 f(\phi)$ and $b^2 f(\phi)$ (where, elementary geometry shows that the acute angle

ϕ is the same for the two little triangles as the big triangle.) Moreover, since these are all right triangles, the function f is the same for each! Therefore, since the area of the big triangle is just the sum of the areas of the little ones, we have

$$L^2 f = a^2 f + b^2 f \quad (3.28)$$

or,

$$L^2 = a^2 + b^2. \quad (3.29)$$

3.3.11 McMahon's Rowers

Professor T. McMahon from Harvard asked the following simple question in 1970. Consider the crew boats which row on the Charles. How does the speed of the boats, scale with the number of rowers in a boat? I.e. If a crew boat existed with 128 oarsmen, how much faster would it be than a crew boat with only 2 oarsmen?

The way that the boat works is that the power being imparted to the water from the rowers is balanced by the drag on the boat. If there are N oarsmen, the power that is provided is proportional to the number of oarsmen: $P = kN$. What is the drag? The drag force depends on both the velocity of the boat V , and also the area of the surface of the boat. If the length of the boat is approximately L , then dimensional analysis says that the drag force is

$$\rho V^2 L^2, \quad (3.30)$$

where ρ is the density of water.

The power required from friction is therefore

$$\rho V^3 L^2. \quad (3.31)$$

Equating these two gives

$$\rho V^3 L^2 = kN. \quad (3.32)$$

To finish this off, we need a relation between L and N . Roughly speaking, the volume of the boat is proportional to the number of oarsmen. Therefore we expect

$$L^3 \propto N. \quad (3.33)$$

Combining this in the above formula gives

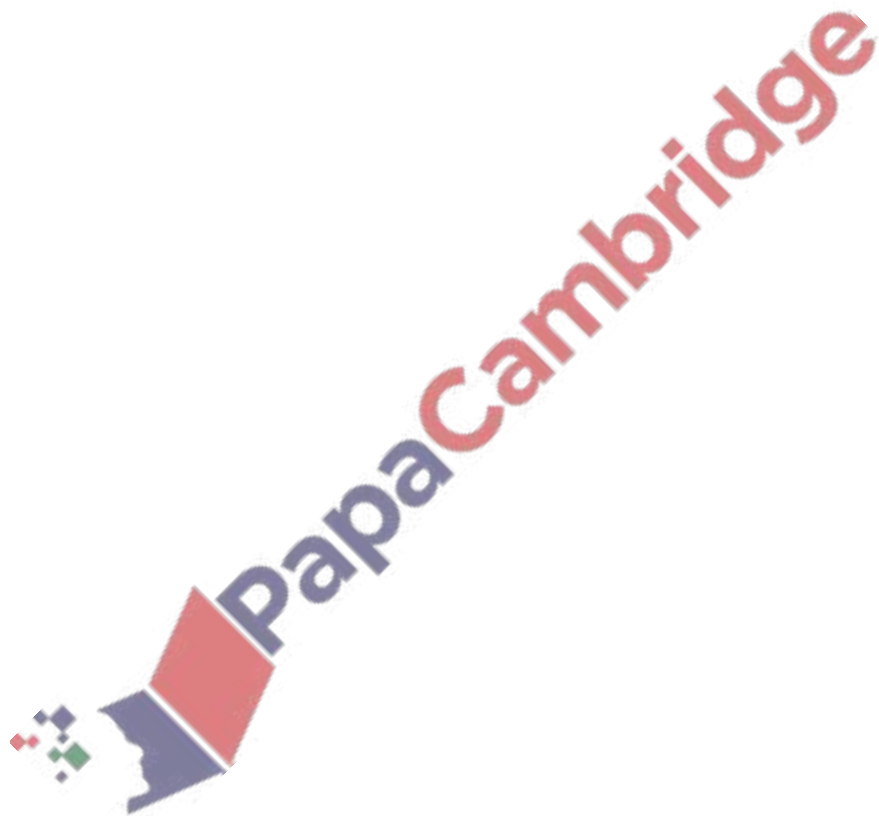
$$V \propto N^{1/9}. \quad (3.34)$$

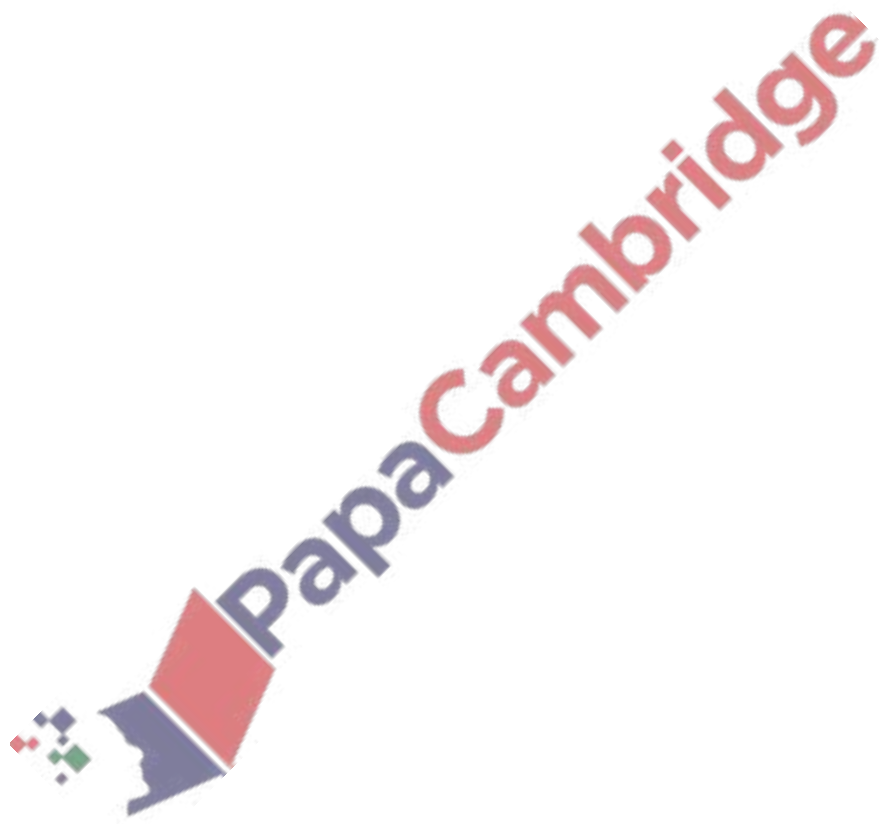
Thus, this would say that a 128 person boat is exactly $64^{1/9} = 1.6$ times faster than a 2 person boat!

3.3.12 Lessons learned

Dimensional analysis is a powerful method to understand the relation between physical quantities in a given problem. The basic idea is, all scientifically interesting results are expressed in terms of dimensionless quantities, independent on the system units you are using. There are two consequences of this idea: first, one can often guess a reasonable form of an answer just by thinking of the dimensions in complex physical situations, and test the answer by experiments or more developed theories. Second, dimensional analysis is routinely used to check the plausibility of derived equations or computations.

A general principle of dimensional analysis is formalized by Buckingham, known as Buckingham's π theorem. It states that every physically meaningful equation involving n variables can be equivalently rewritten as an equation of $n - m$ dimensionless parameters, where m is the number of fundamental dimensions used. Furthermore, it provides a method for computing these dimensionless parameters from the given variables.





4 Polynomial Equations: Random and Otherwise

4.1 Solving Polynomial Equations

The introductory example described how to solve a quintic polynomial equation. We introduced the method of dominant balance, and combined it with computer simulations to develop analytical approximations for all of the roots of the polynomial. Although we implied that our methods were very robust, it must be admitted that the analysis was carried out on polynomials with a quite well defined structure. A perfectly reasonable objection to our claim of generality is that we have provided no evidence for whether the mode of analysis we employed works on arbitrary problems. Perhaps the choice of a quintic was contrived.

To address this head on, here we consider what happens when we analyze truly random problems: namely, we will consider (in this case) polynomials, in which we choose the coefficients of the polynomials to be anything at all—e.g. we will choose the coefficients from a probability distribution.

Consider the roots of a general quartic polynomial

$$p(x) = a_0x^4 + a_1x^3 + a_2x^2 + a_4x + a_5. \quad (4.1)$$

Let us first review the method of dominant balance applied to such polynomials. The method consists of two steps—first the assertion that in general two of the terms in this polynomial will be larger than the others and will represent the *dominant balance*. Second we need to test whether the solution to the simplified equation for the dominant balance indeed implies that the neglected terms in the equation are indeed smaller than those that we have kept.

Unless you are unlucky, but by definition this is unlikely!

For example suppose that the two terms are the first two, so that the dominant balance is $a_0x^4 + a_1x^3 \approx 0$. The roots of this equation are either $x = 0$ or $x = -a_1/a_0$. We can find out if either of these possibilities is *self consistent* by plugging the solutions into the original equation and requiring that the terms are small. Now the $x=0$ root has no chance unless all of the other coefficients are zero. As for $x = -a_1/a_0$, for this to be self consistent we need the size of the terms we have kept in the dominant balance $a_0 \left(\frac{a_1}{a_0}\right)^4$ to be much larger than the other terms that we have neglected.

$$\text{For example, } a_0 \left(\frac{a_1}{a_0}\right)^4 \gg a_2x^3 = a_2 \left(\frac{a_1}{a_0}\right)^3 \text{ and } a_0 \left(\frac{a_1}{a_0}\right)^4 \gg a_3x^2 = a_3 \left(\frac{a_1}{a_0}\right)^2$$

Indeed each of the possible dominant balances correspond to a series of inequalities of this type. For each fourth order polynomial each of the roots presumably obeys some dominant balance or other. The question is how to discover what the dominant balance is for a given root. When discovered, the dominant balance will give understanding into the value of a root, and how the value will shift when the coefficients of the polynomial are changed.

4.2 Having courage: finding roots by iteration

Before moving on to more general polynomials, we first revisit our now famous quintic and discuss another method for finding roots. This method is intuitively quite clear, but has a slightly different spirit.

Recall our quintic:

$$\epsilon x^5 - x + 1 = 0. \quad (4.2)$$

We argued before that as $\epsilon \rightarrow 0$ there was a root near $x = 1$. We then were interested in trying to find the corrections to this root and to this end posited the perturbation series

$$x = 1 + \sum_{n=1}^{\infty} a_n \epsilon^n. \quad (4.3)$$

By plugging this back into the original equation and equating powers of ϵ we were able to find that $x = 1 + \epsilon + 5\epsilon^2 + 35\epsilon^3 + \dots$; we then argued that the radius of convergence of this series is of order 0.08. Hence we concluded that approximating a root as close to $x = 1$ isn't a very good idea when $\epsilon > 0.08$ or so.

This argument is of course a very good one and has a mathematical foundation that you have experienced before. However here we outline a different approach that shows that with a little courage one can make more progress.

We need to find the correction to the root at $x = 1$. Instead of writing a series lets just write $x = 1 + \delta$, where δ is the as yet unknown error, and then plug this ansatz directly into the quintic. We then obtain

$$\epsilon(1 + \delta)^5 - (1 + \delta) + 1 = 0, \quad (4.4)$$

or

$$\epsilon(1 + 5\delta + 10\delta^2 + 10\delta^3 + 5\delta^4 + \delta^5) - \delta = 0. \quad (4.5)$$

Rewriting we have

$$1 + \left(5 - \frac{1}{\epsilon}\right)\delta + 10\delta^2 + 10\delta^3 + 5\delta^4 + \delta^5 = 0. \quad (4.6)$$

So far our treatment is exact. Now, instead of constructing a perturbation series in ϵ , let us reapply dominant balance on this equation. We might think that if ϵ is small the balance will be

$$1 + \left(5 - \frac{1}{\epsilon}\right)\delta \approx 0, \quad (4.7)$$

or

$$\delta = \frac{1}{\frac{1}{\epsilon} - 5}. \quad (4.8)$$

If we take $\epsilon = 1/2$, for example, our formula gives $\delta = -1/3$. Was it consistent to neglect the other terms in the equation? If $\delta = -1/3$, then $10\delta^2 \sim 10/9$ and $10\delta^3 \sim 10/27$. Thus in this case it is OK to neglect the δ^3 and higher terms but not the δ^2 term. If we take $\epsilon = 1/8$, then $\delta = 1/3$ and we have the same conclusion.

Thus, it seems to make sense to consider δ to be the solution to the quadratic equation

$$1 + \left(5 - \frac{1}{\epsilon}\right)\delta + 10\delta^2 = 0. \quad (4.9)$$

The answer is (The second root is neglected because we need when $\epsilon \rightarrow 0$, $\delta \rightarrow 0$)

$$\delta = \frac{-(5 - \frac{1}{\epsilon}) - \sqrt{(5 - \frac{1}{\epsilon})^2 - 40}}{20}. \quad (4.10)$$

Figure 4.1 compares the approximate solution to the balance above with the numerical solution for the real root. Note that this agrees extremely well with the numerics, much better than the more tedious perturbation series. Indeed, also note that the quadratic equation also predicts that at a critical ϵ there will be imaginary roots. This occurs when the following condition is met:

$$\left(5 - \frac{1}{\epsilon}\right)^2 = 40. \quad (4.11)$$

Solving this for ϵ gives that

$$\epsilon = \frac{1}{5 + \sqrt{40}} = 0.0883, \quad (4.12)$$

amazingly close to the actual root!

For fun, let's see how accurate the root is predicted at $\epsilon = 1/2$. Our quadratic then predicts the roots $x = 1 + \frac{-3 \pm \sqrt{31}i}{20}$. Numerically evaluating this gives $x = 0.85 \pm 0.2784i$

On the other hand, the exact roots are $0.8752 \pm 0.3519i$. At $\epsilon = 1/5$ our formula predicts that $x = 1 \pm \sqrt{40}/20i = 1 \pm 0.3162i$, compared with the exact roots $1.04 \pm 0.305i$

Note that the entire approach is quite different from our expansions from before: Indeed the formula for δ is equivalent to an infinite series in ϵ ; but the infinite series works well beyond the radius of convergence of the original Taylor series! One reason for this improved performance is that the current method easily allows transitions from real to complex roots, which the straightforward series expansion does not allow. The added flexibility in the answer surely improves the convergence properties of the solutions.

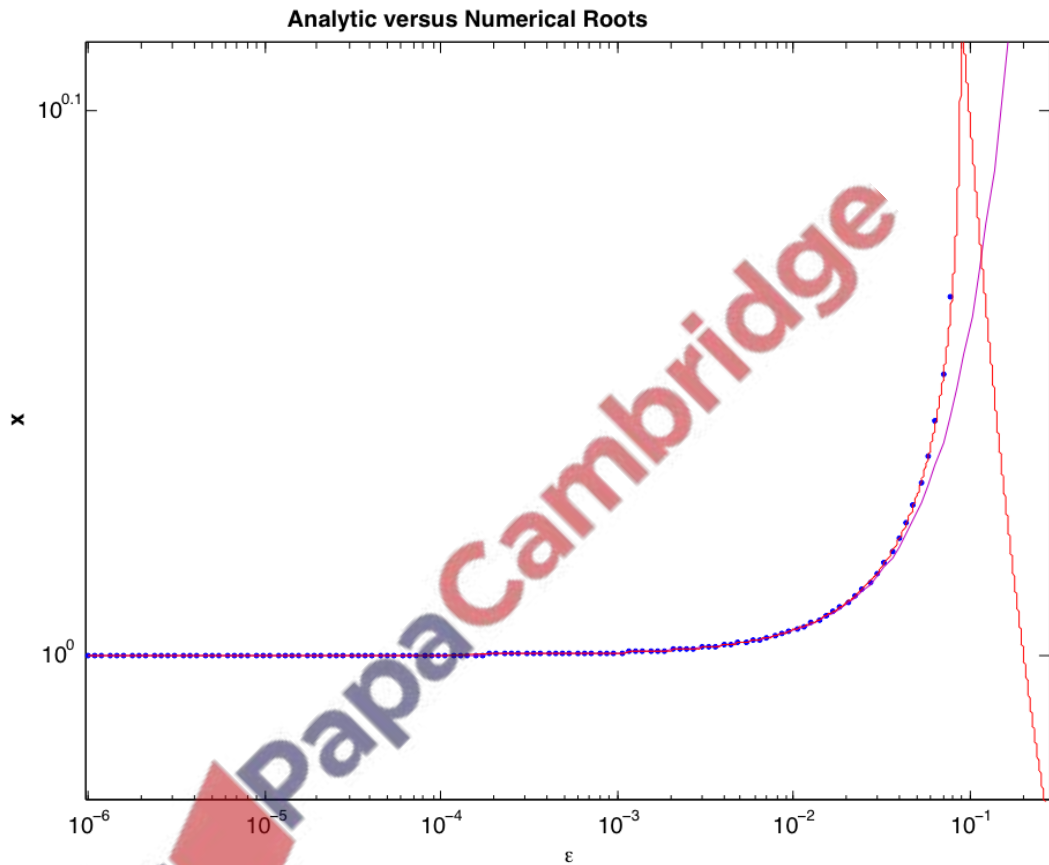


Figure 4.1. A comparison between the numerical roots to the polynomial equation (dots), the analytical approximation we invented in the last class (pink line) and the solution to the quadratic equation above (red line).

4.3 Random Quartic Polynomials

We now apply our methodologies to quartic polynomials; to demonstrate the power of the approach we will draw the coefficients randomly from probability distributions, and ask to what extent can the solutions to the resulting equation be understood in a simple way. We will use the computer to discover dominant balances, and simplified equations, for each of the different roots.

4.3.1 Our first random polynomial

We choose the coefficients for our polynomial from a normal distribution with unit variance, giving

$$p(x) = 0.1387x^4 - 0.8595x^3 - 0.7523x^2 + 1.2296x + 1.1508 \quad (4.13)$$

Using matlab's 'roots' command, we find the roots of this polynomial to be

$$x_1 = 6.7793 \quad (4.14)$$

$$x_2 = 1.2963 \quad (4.15)$$

$$x_{3,4} = -0.9386 \pm 0.2523i \quad (4.16)$$

We now want to *understand* why these are the roots, and even discover simple formulas for each of them. The easiest way to proceed is to take each roots, one by one, and evaluate the sizes of the various terms in the polynomial, to see if they are all equally important.

Let's start with the root $x = 6.7793$ the sizes of $a_1x^4, a_2x^3, a_3x^2, a_4x, a_5$ can be evaluated to be

$$p(6.7793) = 292.8880 - 267.8022 - 34.5724 + 8.3359 + 1.1508. \quad (4.17)$$

We therefore see that the first two terms in the equation are most relevant; the dominant balance is therefore

$$a_1x^4 + a_2x^3 = 0, \quad (4.18)$$

or

$$x = -\frac{a_2}{a_1} = 6.19. \quad (4.19)$$

This is fairly close to the root we started with! To obtain a better calculation for the root, we could include the first three terms in our dominant balance, namely

$$a_1x^4 + a_2x^3 + a_3x^2 = 0. \quad (4.20)$$

It is readily shown that this quadratic equation has the two roots 6.97 and -0.77 . The positive root is quite close to our root.

What about the negative root? If we evaluate the sizes of the terms in the equation for $x = -0.77$ we find that $p(-0.77) = 0.0487 + 0.3924 - 0.4460 - 0.9468 + 1.1508$. Thus, for $x = -0.77$, the largest terms in $p(x)$ are the last two terms—but in deriving this as a solution we have assumed that the first three terms are largest. This is a contradiction and shows that this solution is not correct.

One can continue this line of argumentation to uncover the reasons for the other three roots of the equation: For the root at $x = 1.296$, the sizes of the various terms are given by

$$p(1.296) = 0.3917 - 1.8722 - 1.2642 + 1.5939 + 1.1508. \quad (4.21)$$

The last four terms are the same order of magnitude; nonetheless we continue by taking the largest terms and looking at the resulting equation. The largest two terms are

$$a_2x^3 + a_4x = 0. \quad (4.22)$$

This balance gives that $x = 1.43$. An even better estimate is obtained using the three terms $a_2x^3 + a_3x^2 + a_4x = 0$, which gives $x = 1.31$.

The two complex roots are even more interesting: One can see that the real part of the complex roots follows directly from the last two terms: i.e. $a_4x + a_5 = 0$, which implies that $x = -a_5/a_4 = -0.936$. Direct computation shows that with this balance, the other terms have magnitudes $[0.1 \ 0.69 \ -0.65]$, respectively, so that this is a consistent balance.

What about the complex part of the root? Let's write $x = \alpha + \delta$, with $\alpha = -0.936$, and solve for δ . Plugging into the equation we obtain that

$$a_1\delta^4 + (4a_1\alpha + a_2)\delta^3 + (6\alpha^2a_1 + 3\alpha a_2 + a_3)\delta^2 + (4a_1\alpha^3 + 3\alpha^2a_2 + 2\alpha a_3 + a_4)\delta + (a_1\alpha^4 + a_2\alpha^3 + a_3\alpha^2 + a_4\alpha + a_5) = 0. \quad (4.23)$$

The numerical values of the coefficients of this equation are $[0.13 \ -1.3 \ 2.39 \ -0.07 \ 0.1522]$.

We now need to do a dominant balance argument on this equation. A little tinkering shows that the right balance is between the third and fifth terms: this gives $\delta = \sqrt{-0.1522/2.39} = \pm 0.2505i$.

4.3.2 Our second random polynomial equation

To demonstrate that the previous example was no fluke, we consider another random polynomial, this time with coefficients obtained from a uniform distribution between in $[-5 \ 5]$. The polynomial is

$$p(x) = 0.1551x^4 + 2.8333x^3 + 3.7437x^2 - 1.7996x + 3.3924 \quad (4.24)$$

Using Matlab's 'roots' command, we find the roots of this polynomial to be

$$x_1 = -16.784 \quad (4.25)$$

$$x_2 = -2.1315 \quad (4.26)$$

$$x_{3,4} = 0.32379 \pm 0.71172i \quad (4.27)$$

As before, we now want to discover why each root has the value that it does, and develop simple formulas for each of the roots. We begin with x_1 ; evaluating each term of this polynomial, by substituting this value of the root back into the original polynomial equation gives

$$p(x) = 12307 - 13395 + 1054.6 + 30.204 + 3.3924 \quad (4.28)$$

From this, we can see that the dominant balance is between the first 2 terms, implying the dominant balance

$$0 = 0.1551x^4 + 2.8333x^3 \quad (4.29)$$

or

$$x_1 = -\frac{2.8333}{0.1551} = -18.2676 \quad (4.30)$$

To improve upon this approximation, we include the 3rd term, which gives

$$0 = 0.1551x^4 + 2.8333x^3 + 3.7437x^2 \quad (4.31)$$

The solution to the quadratic equation gives two roots, $x = -16.8337, -1.4339$. The first root is quite close to the numerical value for x_1 .

Does the second root correspond to x_2 ? We can either investigate this by asking whether the neglected terms for a root at $x \approx -1.4339$ are larger than the terms we kept, or we can proceed as in our first example and evaluate the sizes of the various terms of the equation for x_2 . Both arguments give the same answer, that the neglect of terms is not allowed: For the second argument, we see that $p(x = -2.1315) = 3.2 - 27.4 + 18 + 3.8 + 3.4$. Thus the second and third terms are the largest. The first term which we have kept is smaller than the fourth and fifth terms that are deleted—and hence this argument is formally not admissible. Instead we get an approximation for x_2 by equating the second and third terms, namely $a_2x^3 + a_3x^2 = 0$, or $x = -a_3/a_2 = -1.3211$. We can improve this solution by letting

$$x = -1.32 + \delta \quad (4.32)$$

Substituting this ansatz into our original polynomial equation, we have

$$0 = 0.1551(-1.3211 + \delta)^4 + 2.8333(-1.3211 + \delta)^3 + 3.7437(-1.3211 + \delta)^2 - 1.7996(-1.3211 + \delta) + 3.3924 \quad (4.33)$$

Expanding in δ gives the equation $0.1551\delta^4 + 2.0137\delta^3 - 5.8613\delta^2 + 1.7126\delta + 6.24$.

Neglecting the terms $0.1551\delta^4, 2.0137\delta^3$, the resulting quadratic equation has two roots $\delta = -0.8961, 1.18$. The first root implies $x = -1.32 + \delta = -2.2161$, rather close to the exact root.

Finally we turn to the approximations for the two imaginary roots. Substituting x_3 and x_4 into the original polynomial equation we find that the terms are

$$p(x_3) = -0.0079193 - 0.057432i - 1.2979 - 0.38722i - 1.5039 + 1.7255i - 0.58269 - 1.2808i + 3.3924 \quad (4.34)$$

$$p(x_4) = -0.0079193 + 0.057432i - 1.2979 + 0.38722i - 1.5039 - 1.7255i - 0.58269 + 1.2808i + 3.3924 \quad (4.35)$$

To proceed we need to consider the dominant balance of both the real and imaginary parts. We note that the real terms are larger than their imaginary counterparts, and that for the real part the 3rd and 5th terms are the largest. This dominant balance is thus given by

$$0 = 3.7437x^2 + 3.3924 \quad (4.36)$$

This implies that

$$x_{3,4} = 0 \pm 0.9519i \quad (4.37)$$

This solution thus obtained is therefore purely imaginary; we can improve upon it by also including the 3rd largest term. This gives us

$$0 = 3.7437x^2 - 1.7996x + 3.3924 \quad (4.38)$$

or

$$x_{3,4} = 0.2404 \pm 0.9211i \quad (4.39)$$

This approximation captures reasonable approximations for both the real and complex part. To improve upon these approximations further, we write

$$x = Re(x) + \delta_r + Im(x) + \delta_i \quad (4.40)$$

where δ_r (δ_i) stands for the real (imaginary) part of the correction term. Substituting this expression back into the original polynomial, we have

$$\begin{aligned}
 0 = & -8.41267 - 2.1735i + \delta_i(-7.09092 + 10.2752i) + \delta_r(-7.09092 + 10.2752i) \\
 & + \delta_i\delta_r(10.1026 + 16.4828i) + \delta_i^2(5.05131 + 8.24139i) + \delta_r^2(5.05131 + 8.24139i) \\
 & + \delta_i^2\delta_r(8.94733 + 1.71435i) + \delta_i\delta_r^2(8.94733 + 1.71435i) + \delta_i^3(2.98244 + 0.57145i) \\
 & + \delta_r^3(2.98244 + 0.57145i) + \delta_i^3\delta_r(0.6204) + \delta_i^2\delta_r^2(0.9306) \\
 & + \delta_i\delta_r^3(0.6204) + \delta_i^4(0.1551) + \delta_r^4(0.1551) \quad (4.41)
 \end{aligned}$$

Assuming that the correction terms are indeed small, we keep terms only up to $O(\delta)$. We require that both the imaginary the real parts vanish, implying

$$O(\delta) : \quad (4.42)$$

$$0 = -1.62787 + 10.2752i\delta_i - 7.09092\delta_r \quad (4.43)$$

$$0 = -1.86998i - 7.09092\delta_i + 10.2752i\delta_r \quad (4.44)$$

so that

$$\delta_r = 0.049219 \quad (4.45)$$

$$\delta_i = -0.192393i \quad (4.46)$$

Thus, the approximation for x_3 with the correction term is

$$x_3 = \tilde{x}_3 + \delta \quad (4.47)$$

is

$$x_3 = .2404 + .049219 + .9211i - .192393i \quad (4.48)$$

$$= 0.28259 + 0.728707i \quad (4.49)$$

The approximations are thus very close to the actual root!

4.3.3 n^{th} Order Random Polynomials

Hopefully these examples has provided fodder to convince you that the ideas we just explored also work quite well (surprisingly well!) on random examples. We have shown that with **two term** perturbation expansions, we have obtained expressions for the roots that are within two percent of the actual answer, for every random root!

I have taught random polynomials in this fashion during this course for several years, and ever year assigned random polynomials as homework problems. Each year, many people insist initially that their random polynomials are not as simple as these examples, that somehow they were “unlucky” in the assignment of their random numbers. Each year, I insist that being unlucky is, by definition, quite improbable and that therefore they should go back to work. For the fourth order examples we assign for homework, there has yet to be a truly unlucky person—ie one for whom these methods do not lead to accurate approximations to the polynomial roots.

This brings up a very interesting research question that I would like to know the answer to: namely what is the probability of getting this level of accuracy with (say) two term expansions for an n^{th} order polynomial equation with random coefficients. A former student in this class (Bryan Chen) studied this question and made some progress, showing that for polynomials up to 15^{th} order, convergent expansions for the roots can be obtained about $\sim 90\%$ of the time.

4.4 Having Courage

One of the main lessons of this course will be to try to convince you to *have courage* and not be afraid to try things that might seem at first sight to be quite crazy. Here we will consider several examples in which there is really no rigorous procedure for developing approximate solutions, but by just trying things and seeing how it works is very effective.

4.4.1 A Problem of Hinch

What about finding solutions to non-polynomial equations? We consider solutions to

$$xe^{-x} = \epsilon, \tag{4.50}$$

a problem due to Hinch. For this problem, the large ϵ limit is easy: when ϵ is above $1/e$, there are no solutions to this equation. When ϵ is below $1/e$ there are two solutions. Let's characterize these solutions in the limit $\epsilon \rightarrow 0$, and see how well it works. Figure (4.2) plots xe^{-x} , demonstrating that there is a maximum occurring near $x = 1$.

As $\epsilon \rightarrow 0$, one solution occurs at small x , and the other at large x . The small x solution can be characterized by Taylor-expanding the function. We have

$$xe^{-x} = x(1 - x + x^2/2 + \dots) = \epsilon. \tag{4.51}$$

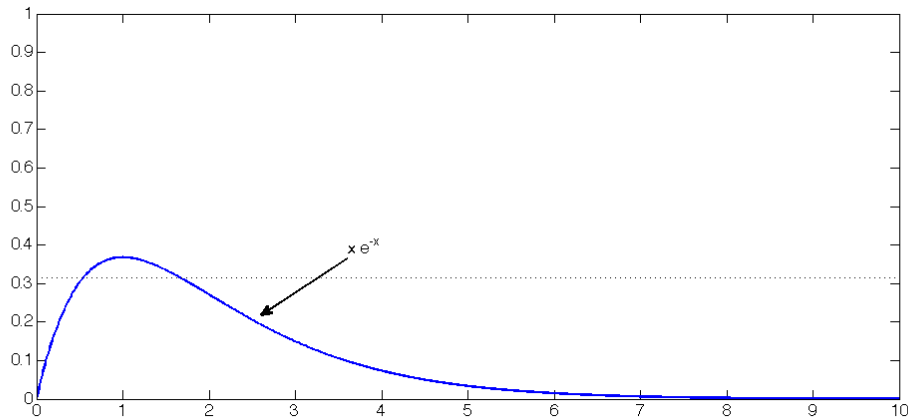


Figure 4.2. Plot of $x e^{-x}$.

The dominant balance in this equation is clearly

$$x = \epsilon \tag{4.52}$$

. Forming the series

$$x = \sum_{n=1}^{\infty} a_n \epsilon^n, \tag{4.53}$$

,the first few terms of the solution are

$$x = \epsilon + \epsilon^2 + 3\epsilon^3/2 + \dots \tag{4.54}$$

We expect this solution should have a radius of convergence close to one, reflecting the disappearance of this root at $x = 1$; verification of this is left for homework.

The harder part is finding a good approximation to the larger root. Whereas the small x root involved the balance between $x \sim \epsilon$, the large x root involves the balance between e^{-x} and ϵ . To leading order we therefore guess that

$$x \approx \log\left(\frac{1}{\epsilon}\right). \tag{4.55}$$

What is the error in this result? If we write $x = \log(1/\epsilon) + y$, and plug this into the equation (4.50), we obtain $(\log(\frac{1}{\epsilon}) + y) e^{-y} = 1$. Since by ansatz, $y \ll \log(1/\epsilon)$, the leading order balance of this equation is

$$y = \log(\log(1/\epsilon)). \tag{4.56}$$

We now have that

$$x = \log(1/\epsilon) + \log(\log(1/\epsilon)). \tag{4.57}$$

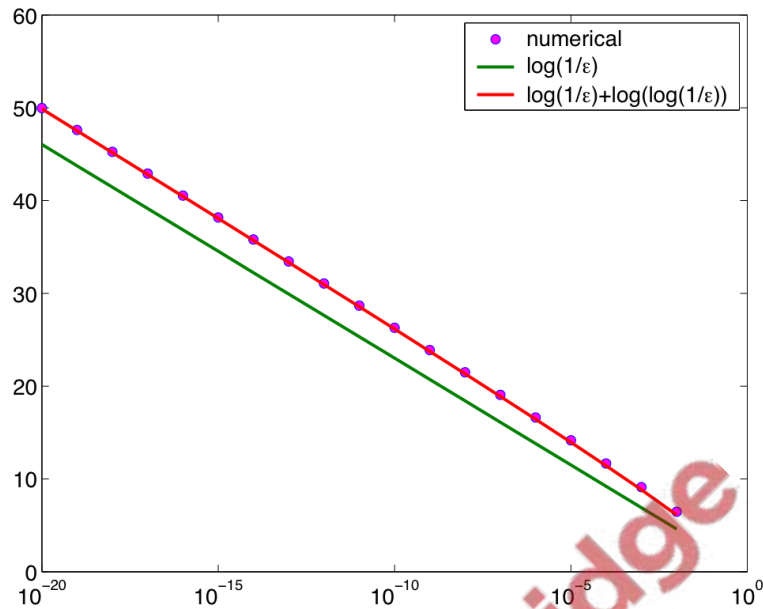


Figure 4.3. Comparison of analytical and numerical solutions.

What about finding a better approximation? We can repeat the above procedure, writing

$$x = \log(1/\epsilon) + \log(\log(1/\epsilon)) + z. \quad (4.58)$$

Inserting this into equation (4.50) and finding the same type of leading order balance, we obtain

$$z = \log\left(\frac{\log(1/\epsilon) + \log(\log(1/\epsilon))}{\log(1/\epsilon)}\right). \quad (4.59)$$

Figure 4.3 shows a comparison between our approximation and answers generated numerically.

4.4.2 The Prime Number Theorem

The Prime Number Theorem states that the number of prime numbers $\pi(x)$ less than x is approximately

$$\pi(x) \sim \frac{x}{\log(x)} + \frac{x}{(\log(x))^2} + \frac{2!x}{(\log(x))^3} + \frac{3!x}{(\log(x))^4} + \dots \quad (4.60)$$

If we invert this formula, letting $\pi = N$, and computing $x = x(N)$ we will have a formula for the N^{th} prime number!

Let us here take the first steps towards carrying out this inversion, and we will then compare our formula with actual experimental data for the N^{th} prime number.

$$\pi(x) = N = \frac{x}{\log(x)}. \quad (4.61)$$

To solve for $x(N)$, take the logarithm of both sides. We then have

$$\log(N) = \log(x) - \log(\log(x)). \quad (4.62)$$

Now when x is large $\log x \gg \log(\log x)$, so that we can heuristically delete the second term on the right hand side, obtaining

$$\log x = \log N, \quad (4.63)$$

or $x = N$.

Now, to get a better approximation let us write $x = N + \delta$, where $\delta \ll N$ is the deviation of the N^{th} prime number from N . Inserting this into our equation we have

$$N = \frac{N + \delta}{\log(N + \delta)} \approx \frac{N + \delta}{\log N + \delta/N - \delta^2/2N^2}. \quad (4.64)$$

Writing this out gives

$$N \log N - \delta^2/2N = N, \quad (4.65)$$

or $\delta^2 = 2N^2(\log N - 1)$. Unfortunately in starting this calculation we assumed $\delta \ll N$ which is violated by our answer! Thus we know that our answer is not of the form that we assumed.

What is going on? We need to try another guess for the approximation. Let us try $x = N\delta$, where we anticipate that δ varies slowly with N . Inserting this into our equation gives

$$N = \frac{N\delta}{\log(N\delta)} \approx \frac{N}{\log N} \delta, \quad (4.66)$$

or $\delta = \log N$. This is a perfectly consistent result, and we have therefore shown that

$$x \approx N \log(N). \quad (4.67)$$

Let us now test this formula on data for prime numbers. Figure 4.4 compares our formula with the first 9592 prime numbers. The agreement is not bad, but there is a quantitative deviation even at the largest values of N .

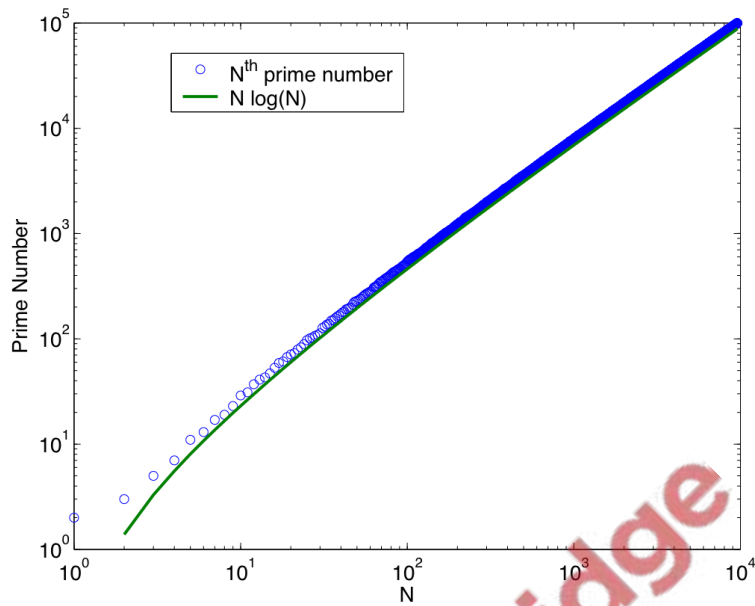


Figure 4.4. Comparison of the n^{th} prime number with our formula.

4.5 Numerical Approaches

Finally, it is worth discussing briefly the methodologies that are usually used to find the zeros of a set of (perhaps coupled) polynomial equations numerically. The workhorse of all methods is Newton's method. This is not really the best way to compute the zeros of polynomials¹ but it is both readily generalizable for computing zeros of arbitrary sets of functions, and also provides a glimpse into how approximate calculations can inspire numerical algorithms. The major limitation of this method is that it requires a good guess to guarantee convergence.

4.5.1 Newton's Method

Suppose we must solve $f(x) = 0$, and that we have a good guess for the root x_0 . The idea of Newton's method is to approximate the function $f(x)$ by a linear function near x_0 , and then find the zero of this linear function. Although this zero will not exactly solve $f(x) = 0$ (unless the function f is itself linear), the hope is that it will generate a better guess for the root than the initial guess.

¹The best method (used by MATLAB) uses the fact that finding the roots of polynomials is equivalent to finding the eigenvalues of a matrix. Therefore, sophisticated techniques for computing matrix eigenvalues can be used to compute the zeros of polynomials. The magic of these techniques is that first, the method is "global", requiring no guesses to guarantee convergence; and second, the method simultaneously computes *all* of the roots. The disadvantage of this method is however that it only works for finding the roots of polynomials.

To implement this method, we approximate $f(x)$ by $f(x) \approx f(x_0) + (x - x_0)f'(x_0)$ near x_0 . The zero of this linear function occurs at

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}. \quad (4.68)$$

Iterating this method implies the general formula

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad (4.69)$$

at the n^{th} iteration.

How quickly does this method converge to the correct root x^* ? Let us denote $e_n = x_n - x^*$ as the error in the n^{th} iteration. Substituting $x_n = x^* + e_n$ into equation (4.69), we have that

$$\begin{aligned} e_{n+1} &= e_n - \frac{f(x^* + e_n)}{f'(x^* + e_n)} \\ &= e_n - \frac{f(x^*) + f'(x^*)e_n + f''(x^*)e_n^2/2 + \dots}{f'(x^*) + f''(x^*)e_n + \dots} \\ &= e_n - e_n \frac{f'(x^*) + f''(x^*)/2e_n}{f'(x^*) + f''(x^*)e_n} \\ &= \frac{f''(x^*)}{2f'(x^*)} e_n^2 \end{aligned} \quad (4.70)$$

Hence, the error e_n converges *quadratically*.

As an illustration of this method, figure 4.5 plots the error e_{n+1} as a function of e_n , for finding a root of the quintic in the previous section. We have taken an initial guess $x_1 = 10^3\epsilon^{-1/4}$ for $\epsilon = 10^{-5}$. Eventually, the solution converges to the correct value $\approx \epsilon^{-1/4}$, with $e_n \rightarrow 0$ quadratically. However, the initial guess is sufficiently far off the correct answer that the initial convergence rate is *linear* (with $e_{n+1} \sim e_n$), instead of quadratic. This is because initially the guess is so far from the correct answer that the approximations implicit in (4.70) do not apply.

4.5.2 MATLAB Implementation

Here we provide a MATLAB program that implements Newton's method: The main program (used to generate figure 4.3) is

```

1 global eps
2
3 eps=0.01;
4
5 for i=1:20
6 x=log(1/eps); % initial guess for x—for the Hinch example

```

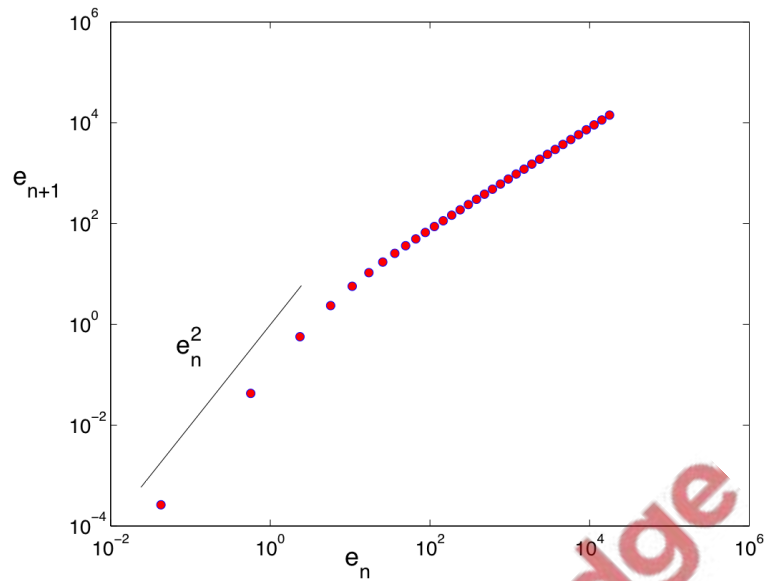


Figure 4.5. Plot of the error in the step of Newton's method, against the error in the step. The specific example chosen is the quintic of the previous section, with the initial guess for. The convergence rate is initially linear, and then eventually quadratic.

```

7           % this is a good initial guess
8   for j=1:10
9       dmyfunc=(myfunc(x+1e-6)-myfunc(x-1e-6))/2e-6;
10      % approximation for the derivative of myfunc
11      x=x-myfunc(x)/dmyfunc;
12      % essential algorithm for Newton's method
13   end
14   sepx(i)=eps; % save values of eps
15   xx(i)=x; % save values of x
16   eps=eps/10; % decrease eps for the next iteration
17   end

```

The program calls the function "myfunc", which should be saved in the folder "myfunc.m", in the same directory that MATLAB is launched from. This function is given as follows:

```

1   function y=myfunc(x)
2   global eps
3   y=x*exp(-x)-eps;

```

This program loops through different values of ϵ , to generate numerically the solutions to $xe^{-x} = \epsilon$. The inner loop is the Newton's step: note that instead of a numerical derivative, we have used a simple finite difference $f'(x) \approx (f(x + \delta) - f(x - \delta))/(2\delta)$. The parameter δ should be taken to be a small number, but not too small that roundoff

is a problem². The Newton iteration starts from the guess $x = \log(1/\epsilon)$, which we chose in order to ensure that the iteration converges on the large root. Every time ϵ is updated, the guess is reinitialized. Finally, we are doing 10 Newton iterations per ϵ . Since Newton's method converges quadratically, this should be more than sufficient. However, it is safest to check that the *residual* error is indeed small after 10 iterations to make sure there are no spurious numerical artifacts.

One way to do this is to rewrite the inner Newton's method loop as

```

1  while (myfunct (x) > 1e-10)
2    x=x-myfunct (x) / (myfunct (x+1e-6)-myfunct (x-1e-6)) *2e-6;
3  end

```

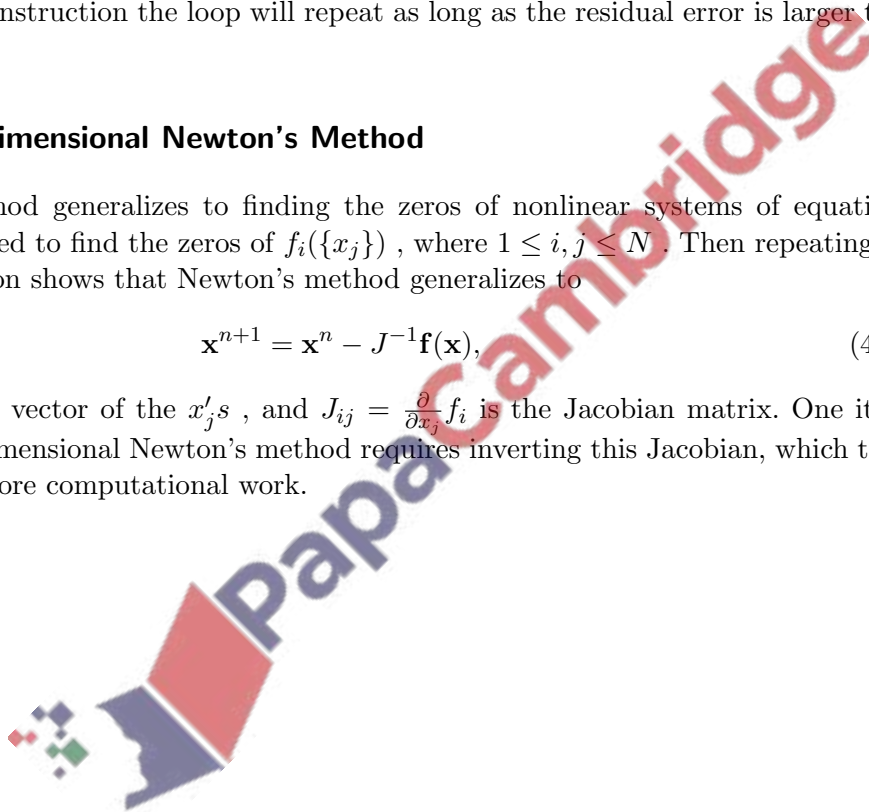
With this construction the loop will repeat as long as the residual error is larger than 10^{-10} .

4.5.3 Multidimensional Newton's Method

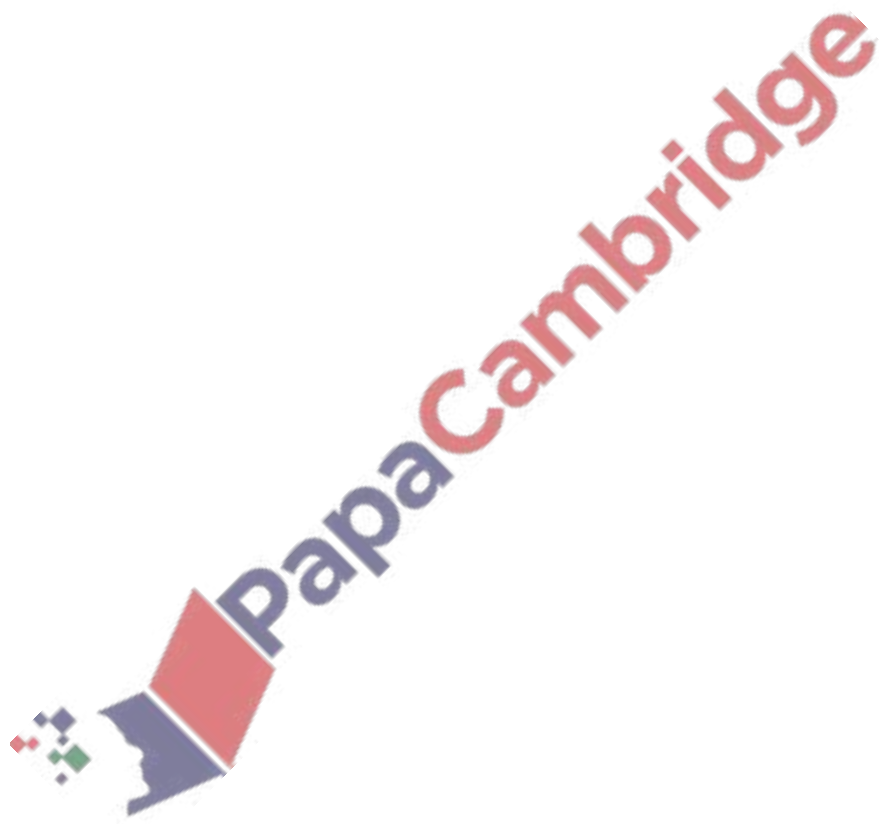
Newton's method generalizes to finding the zeros of nonlinear systems of equations. Suppose we need to find the zeros of $f_i(\{x_j\})$, where $1 \leq i, j \leq N$. Then repeating the above derivation shows that Newton's method generalizes to

$$\mathbf{x}^{n+1} = \mathbf{x}^n - J^{-1}\mathbf{f}(\mathbf{x}), \quad (4.71)$$

where \mathbf{x} is the vector of the x_j 's, and $J_{ij} = \frac{\partial}{\partial x_j} f_i$ is the Jacobian matrix. One iteration of multidimensional Newton's method requires inverting this Jacobian, which takes significantly more computational work.



²Using a numerical derivative was only a matter of convenience—obviously things would work just as well if we defined another function with the analytical formula for the derivative.



5 Ordinary Differential Equations

Approximate analysis is also an effective way to solve differential equations, even those that are ordinarily claimed to *have no solution*. To give you a flavor of how this works we will begin by focusing on a few examples of first order ordinary differential equations, which expose the essential ideas.

5.1 A Very Simple Example

Consider the first order differential equation

$$\frac{dy}{dx} + 2xy = x, \quad (5.1)$$

with the initial condition $y(x = 0) = 1$. This equation is often given as an example problem in advanced calculus courses. There, people are taught that the equation can be solved with the integrating factor e^{x^2} . Multiplying the equation by e^{x^2} gives

$$\frac{d}{dx} \left(e^{x^2} y \right) = e^{x^2} x. \quad (5.2)$$

Integrating both sides, and applying the initial condition $y(x = 0) = 1$ gives the exact solution

$$y(x) = \frac{1}{2} \left(1 + e^{-x^2} \right). \quad (5.3)$$

Now, note from the solution that as $x \rightarrow \infty$ the function $y \rightarrow \frac{1}{2}$. Does every solution of this differential equation (for every initial condition) approach $1/2$ as $x \rightarrow \infty$? Or are there other behaviors?

One can easily investigate this question by simply writing down the solution for a general initial condition: if we impose $y(x_0) = y_0$, then repeating the steps above gives the general solution

$$y(x) = \frac{1}{2} + \left(y(x_0) - \frac{1}{2} \right) e^{x_0^2 - x^2}. \quad (5.4)$$

Indeed, every solution has $y \rightarrow \frac{1}{2}$ as $x \rightarrow \infty$!

5.1.1 A Simpler Way to See the Same Thing

Using the methodology we have been discussing in this class, there is a simpler way to see that all solutions must asymptote to $\frac{1}{2}$, instead of writing down the exact solution. Namely let's look at the original equation: the equation has three terms in it: $\frac{dy}{dx}$, $2xy$ and x . Now, as x becomes very large one might guess that the biggest terms in the equation will be $2xy$ and x , since both terms apparently grow linearly with x .

If this were the case, and these two terms were (say) orders of magnitude larger than the other terms in the equation then the equation would effectively become

$$2xy = x. \quad (5.5)$$

This equation implies that $y = \frac{1}{2}$!

In addition we can check that our assumption about $\frac{dy}{dx}$ being small is consistent: If $y = \frac{1}{2}$, then $\frac{dy}{dx} = 0$, so it was perfectly legitimate to neglect $\frac{dy}{dx}$ with respect to the other two terms in our equation. Thus, we have seen that the solution $y = \frac{1}{2}$ is a *consistent* solution to the equation as $x \rightarrow \infty$.

This is of course the same idea that we spent the last two lectures studying in the context of polynomial equations!

5.2 A Harder Problem

The beauty of this way of thinking about problems in terms of dominant balances is that it allows one to make deductions about harder problems where exact solutions are not possible. We saw this before in the case of polynomials but now let's consider it for ordinary differential equations.

For example, consider the equation

$$\frac{dy}{dx} + x \tan(y) = x. \quad (5.6)$$

From the standpoint of a typical discussion of first order differential equations this nonlinear equation cannot be solved in closed form. It is therefore considered that no progress on it can be made, short of numerical simulation.

On the other hand, making the same type of argument as above, we would guess that as $x \rightarrow \infty$, we would have $\tan(y) = 1$, or $y = \pi/4$. To test this, Figure 5.1 shows a numerical solution of this equation¹ with the initial condition $y(0) = 15$.

¹The numerical solution was generated with the `ode23s` command in MATLAB. MATLAB offers a suite of ODE solvers, including `ode45`, `ode23`, `ode113`, `ode15s`, `ode23s`, `ode23t` and `ode23tb`. Sometimes one of these solvers is better for the problem at hand than others. However, the fact that there are so many solvers should be an indication that you should never trust any individual solver completely!

Program 2 MATLAB code used to create figure 5.1

```
1 function figure51
2 % type "figure51" in the command window to generate the figure
3
4 % define the starting point and the end point of the integration
5 xinit = 0;
6 xfinal = 100;
7
8 % define the initial condition
9 yinit = 15;
10
11 % integrate the equation
12 [x,y] = ode23s(@fun56, [xinit, xfinal], yinit);
13
14 % plot the result
15 plot(x, tan(y), 'linewidth', 2)
16 set(gca, 'fontsize', 16, 'fontname', 'Helvetica', 'fontweight', 'b')
17 xlabel('x')
18 ylabel('tan(y)')
19
20
21 function dy = fun56(x,y)
22 % define the ODE
23
24 dy(1) = -x*tan(y(1))+x;
```

5.3 Example 2

The next example is

$$\frac{dy}{dx} + y = \frac{1}{1+x^2}, \quad (5.7)$$

and we will impose the initial conditions $y(2) = y_0$. This example is often taught in a beginning chapter of introductory differential equations classes.

The usual advice that we give to students is to use again an integrating factor. The integrating factor for this equation is e^x ; hence multiplying the equation by e^x gives

$$\frac{d}{dx} \left(ye^x \right) = \frac{e^x}{1+x^2}. \quad (5.8)$$

Integrating both sides, imposing the initial conditions gives the solution

$$y(x) = y_0 e^2 e^{-x} + e^{-x} \int_2^x ds \frac{e^s}{1+s^2}. \quad (5.9)$$

The integral cannot be done in closed form. Often in advanced calculus classes, students are taught how to carry out such manipulations. In our view, the preceding manipulation explains nothing about the nature of solutions to the equation: it simply transforms one unsolved problem (the original ODE) to another unsolved problem (the integral).

On the other hand, using the ideas we described above, we can make a simple argument for how the solution looks as $x \rightarrow \infty$. Again, looking at the equation there are three terms: $\frac{dy}{dx}$, y and $1/(1+x^2)$. As $x \rightarrow \infty$, $1/(1+x^2) \approx 1/x^2$. Thus, by analogy with what we argued before, we can guess that as $x \rightarrow \infty$, the largest terms in the equation will be y and x^{-2} , giving that

$$y = \frac{1}{x^2}, \quad (5.10)$$

as $x \rightarrow \infty$. For this to be a consistent argument, we need

$$\left| \frac{dy}{dx} \right| \ll |y|, \quad (5.11)$$

when x is large. But if $y = 1/x^2$, then $\frac{dy}{dx} = -2/x^3$. When x is large, we then have $\frac{dy}{dx} \ll y$, so our solution is consistent.

Figure 5.2 shows a log-log plot of the solution to the differential equation with $y_0 = 3$. Taking the logarithm of $y = x^{-2}$ we see that it gives

$$\log y = -2 \log x, \quad (5.12)$$

so that when one plots $\log y$ versus $\log x$ the result is a straight line with slope -2 . The plot clearly shows that the solution asymptotes to this behavior when x is large.

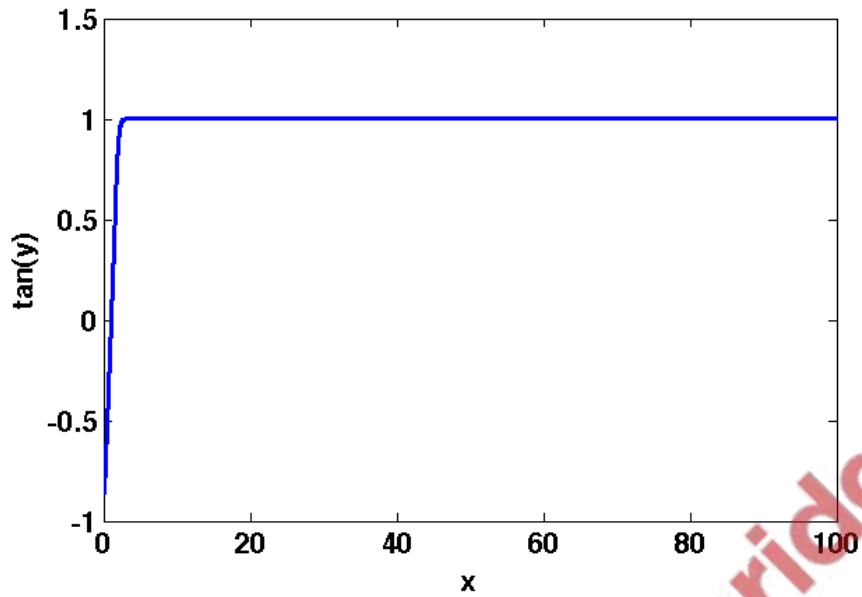


Figure 5.1. Plot of solution to Eqn. (5.6), with $y(0) = 15$. As promised, when $x \rightarrow \infty$ the solution approaches $\pi/4$.

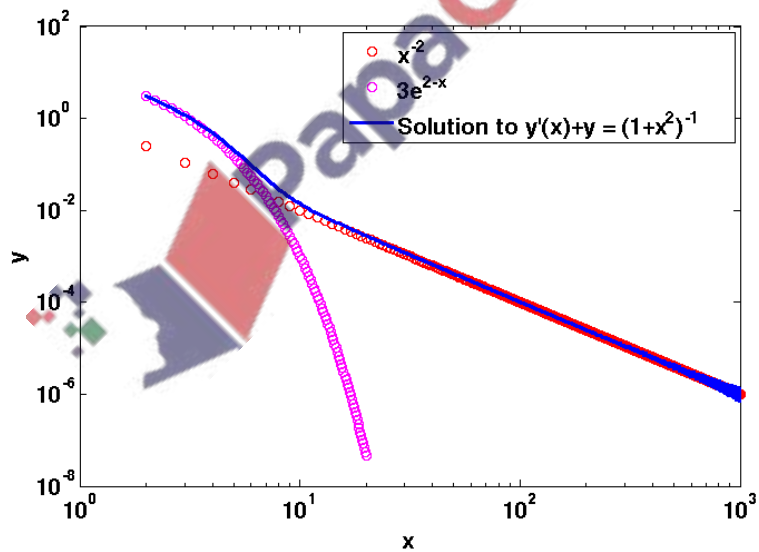


Figure 5.2. Double log plot of the solution to $y' + y = (1 + x^2)^{-1}$ demonstrating that the solution asymptotes to $y = x^{-2}$ as $x \rightarrow \infty$.

Program 3 MATLAB code used to create figure 5.2

```
1 function figure52
2 % type "figure52" in the command window to generate the figure
3
4 % define the starting point and the end point of the integration
5 xinit = 2;
6 xfinal = 1000;
7
8 % define the initial condition
9 yinit = 3;
10
11 % integrate the equation
12 [x,y] = ode45(@fun57, [xinit, xfinal], yinit);
13
14 % plot the result of the integration along with the asymptotic solutions
15 t = 2:1000;
16 loglog(t, t.^(-2), 'ro', 'markersize', 7)
17 hold on
18 t = 2:0.2:20;
19 loglog(t, 3*exp(2-t), 'mo', 'markersize', 7)
20 loglog(x, y, 'b-', 'linewidth', 2)
21
22 set(gca, 'fontsize', 16, 'fontname', 'Helvetica', 'fontweight', 'b')
23 xlabel('x')
24 ylabel('y')
25 legend('x^{-2}', ...
26        '3e^{2-x}', ...
27        'Solution to y''(x)+y = (1+x^2)^{-1}''')
28
29
30 function dy = fun57(x,y)
31 % define the ODE
32
33 dy(1) = -y(1)+1/(1+x^2);
```

On this plot we have also included the solution to $\frac{dy}{dx} + y = 0$ which satisfies the initial condition $y(2) = 3 : y = 3e^{-x}$. You should notice that for small x this solution agrees quite well with the exact solution. In fact, taken together, our approximate solutions almost describe the behavior over the entire range!

Note an important feature of our solution. For *any* initial condition y_0 the behavior of the solution is identical, approaching the asymptote $y = 1/x^2$. Figure 5.3 documents this explicitly. We simulate the equations for various initial conditions with $y_0 = 10, 8, 6, 4, 2, 0, -2, -4, -6, -8, -10$. Figure 5.3(A) shows the behavior on a linear scale: the solutions all approach $y = 1/x^2$ as they are predicted. Figure 5.3(B) shows this (plotting $|y|$ versus x) on a logarithmic scale. Note the solutions with negative initial conditions necessarily change sign in order to eventually follow $y = x^{-2}$. In the logarithmic plot (Figure 5.3(B)) this shows up as a pronounced dip in the solution, reflecting the behavior of $\log(|y|)$ as $|y|$ passes through zero.

Our analysis has clearly captured the essence of this problem.

5.3.1 Doing a Better Job at Large x

Can we improve on our proposed solution $y = x^{-2}$? Our argument for this behavior was derived assuming x is sufficiently large. Therefore, it makes sense to try to improve this formula, and see if we can make it work over a wider range. To do this, we make the variable change $x = \frac{1}{t}$. This variable change maps the large x regime to the small t regime. By the chain rule

$$\frac{dy}{dx} = \frac{dt}{dx} \frac{dy}{dt} = -t^2 \frac{dy}{dt}, \tag{5.13}$$

so in the new variables, our equation becomes

$$-t^2 \frac{dy}{dt} + y = \frac{t^2}{1+t^2}. \tag{5.14}$$

We now seek a series expansion of this equation, in powers of t . Rewriting our equation gives

$$\frac{dy}{dt} - \frac{y}{t^2} = -\frac{1}{1+t^2} = -\sum_{n=0}^{\infty} (-1)^n t^{2n}. \tag{5.15}$$

The last expansion of $(1+t^2)^{-1}$ is valid for $|t| < 1$, and thus we expect our solution to work in this range. If we now use the power series ansatz

$$y = \sum_{n=0}^{\infty} a_n t^n, \tag{5.16}$$

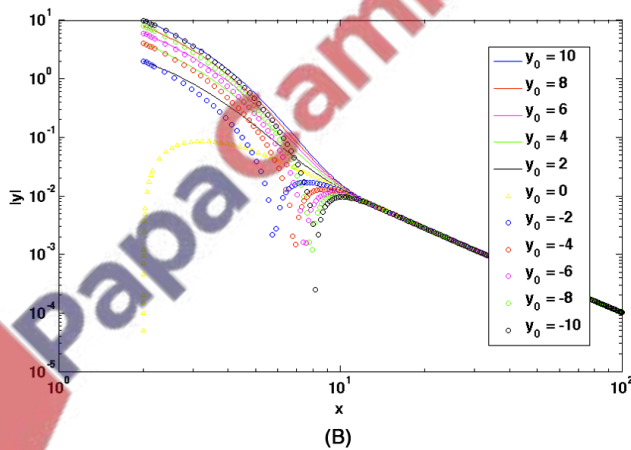
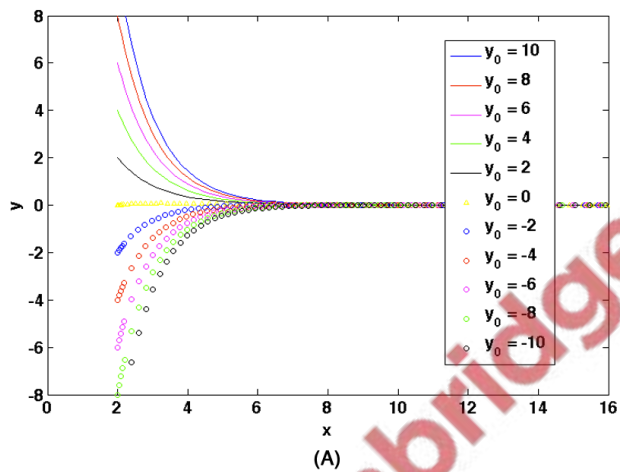


Figure 5.3. Solutions for various initial conditions in (A) linear scale and (B) double log plot. Note that different initial conditions all converge on the same final behavior, even those who start out negative end up approaching $y = 1/x^2$.

Program 4 MATLAB code used to create figure 5.3

```
1 function figure53
2 % type "figure53" in the command window to generate the figure
3
4 % define the starting point and the end point of the integration
5 xinit = 2;
6 xfinal = 100;
7
8 % define a vector containing various initial conditions
9 yinit = 10:-2:-10;
10
11 % define a cell to store solutions
12 sol = cell(length(yinit),2);
13
14 for i = 1:length(yinit)
15     % integrate the equation
16     [x,y] = ode45(@fun57, [xinit, xfinal], yinit(i));
17     sol{i,1} = x; sol{i,2} = y;
18 end
19
20 % plot the result in linear scale; changing "plot" to "loglog" generates
21 % the double log plot in (B)
22 plot(sol{1,1}, sol{1,2}, 'b-')
23 hold on
24 plot(sol{2,1}, sol{2,2}, 'r-')
25 plot(sol{3,1}, sol{3,2}, 'm-')
26 plot(sol{4,1}, sol{4,2}, 'g-')
27 plot(sol{5,1}, sol{5,2}, 'k-')
28 plot(sol{6,1}, sol{6,2}, 'y^')
29 plot(sol{7,1}, sol{7,2}, 'bo')
30 plot(sol{8,1}, sol{8,2}, 'ro')
31 plot(sol{9,1}, sol{9,2}, 'mo')
32 plot(sol{10,1}, sol{10,2}, 'go')
33 plot(sol{11,1}, sol{11,2}, 'ko')
34
35 axis([0,16,-8,8])
36 set(gca,'fontsize',16,'fontname','Helvetica','fontweight','b')
37 xlabel('x')
38 ylabel('y')
39 legend('y_0 = 10','y_0 = 8','y_0 = 6','y_0 = 4','y_0 = 2','y_0 = 0',...
40        'y_0 = -2','y_0 = -4','y_0 = -6','y_0 = -8','y_0 = -10')
41
42
43 function dy = fun57(x,y)
44 % define the ODE
45
46 dy(1) = -y(1)+1/(1+x^2);
```

and insert this into the equation we derive a series of equations for the coefficients a_n :

$$\sum_{n=1}^{\infty} na_n t^{n-1} - \sum_{n=0}^{\infty} a_n t^{n-2} = - \sum_{n=0}^{\infty} (-1)^n t^{2n}. \quad (5.17)$$

We now need to equate coefficients of the various powers of t . The lowest power is t^{-2} (from the second term on the left hand side). Setting the coefficient of this power to zero implies

$$a_0 = 0. \quad (5.18)$$

Setting the coefficient of the t^{-1} term to zero implies

$$a_1 = 0. \quad (5.19)$$

Similarly, the coefficient of t^0 term yields

$$-a_2 = -1, \quad (5.20)$$

while the coefficient of the t term gives

$$2a_2 - a_3 = 0. \quad (5.21)$$

The coefficient of t^2 term yields

$$3a_3 - a_4 = 1, \quad (5.22)$$

while the coefficient of t^3 term gives

$$4a_4 - a_5 = 0. \quad (5.23)$$

Putting these all together we arrive at

$$a_2 = 1, a_3 = 2, a_4 = 5, a_5 = 20, \dots \quad (5.24)$$

Thus we have the formula

$$y = t^2 + 2t^3 + 5t^4 + 20t^5. \quad (5.25)$$

Note: As we mentioned in Chapter 2, we can also use Maple to calculate the coefficients of this series. Input the following code line by line into Maple (The first line of the code has been split into three lines for readability reason):

```

1 poly:=- (sum((-1)^n*t^(2*n), n=0..10))
2         -(sum(n*a[n]*t^(n-1), n=1..10))
3         +sum(a[n]*t^(n-2), n=0..10)
4 coll:=series(poly, t=0, 9)
5 eqs:=seq(coeff(coll, t, n), n=-2..8)
6 aeqs:=seq(a[n], n=0..10)
7 sol:=solve({eqs}, {aeqs})

```


The final output is:

$\{a_0 = 0, a_1 = 0, a_2 = 1, a_3 = 2, a_4 = 5, a_5 = 20, a_6 = 101, a_7 = 606, a_8 = 4241, a_9 = 33928, a_{10} = 305353\}$

In terms of the original variables this is

$$y = \frac{1}{x^2} + \frac{2}{x^3} + \frac{5}{x^4} + \frac{20}{x^5}. \quad (5.26)$$

Figure 5.4 shows a numerical solution with the initial conditions $y(0) = 1$ comparing to the solution we just derived. The dashed line shows the x^{-2} solution we derived before. Note the agreement is excellent, over a much wider range than before!

5.3.2 Something Disturbing

Lest you leave this discussion feeling that all is now well in hand, let me point out something that is very disturbing about the apparently successful procedure we have just outlined. Namely, examining the series solution above, we see that the coefficients of the terms in the series grow. This occurs because the equation for n^{th} term in the series is

$$(n-1)a_{n-1} - a_n = 0, \quad (5.27)$$

if n is odd and

$$(n-1)a_{n-1} - a_n = \pm 1, \quad (5.28)$$

if n is even.

In any case the ratio

$$\frac{a_n}{a_{n+1}} \approx \frac{1}{n} \quad (5.29)$$

when n is large. Therefore, the series has a radius of convergence equal to zero! Thus, it should not work. But we have seen that it does.

This is very disturbing. I will let you be disturbed by it for a while, and we will come back and talk about this issue a little later. For now, let this be the first example you have seen which dispels the myth that convergence is a necessary property for a series to have for it to be useful.

Program 5 MATLAB code used to create figure 5.4

```
1 function figure54
2 % type "figure54" in the command window to generate the figure
3
4 % define the starting point and the end point of the integration
5 xinit = 0;
6 xfinal = 100;
7
8 % define the initial conditions
9 yinit = 1;
10
11 % integrate the equation
12 [x,y] = ode45(@fun57, [xinit, xfinal], yinit);
13
14 % plot the result of the integration along with the asymptotic solutions
15 loglog(x, y, 'bo', 'markersize', 7)
16 hold on
17 t = 0:100;
18 loglog(t, t.^(-2)+2*t.^(-3)+5*t.^(-4)+20*t.^(-5), 'r-', 'linewidth', 2)
19 loglog(t, t.^(-2), 'm-', 'linewidth', 2)
20
21 set(gca, 'fontsize', 16, 'fontname', 'Helvetica', 'fontweight', 'b')
22 xlabel('x')
23 ylabel('y')
24 legend('Numerical Solution', 'Series Solution', 'x^{-2}')
25
26
27 function dy = fun57(x,y)
28 % define the ODE
29
30 dy(1) = -y(1)+1/(1+x^2);
```

5.4 A Nonlinear Example

Before leaving this example, let us consider one more related problem, namely

$$\frac{dy}{dx} + y^5 = \frac{1}{1+x^2}. \quad (5.30)$$

One is immediately tempted to try the same dominant balance argument that worked before: Let's try to balance the y^5 term with $(1+x^2)^{-1}$. Doing this, we obtain at large x the law

$$y = \frac{1}{x^{2/5}}. \quad (5.31)$$

Is this consistent? For this solution we have that $\frac{dy}{dx} \sim x^{-7/5}$. When x is large this is actually bigger than the two terms that we have kept, so the balance is inconsistent!

So what is going on? We have two other possible balances: either

$$\frac{dy}{dx} = (1+x^2)^{-1} \quad (5.32)$$

or

$$\frac{dy}{dx} + y^5 = 0. \quad (5.33)$$

The first balance gives

$$y = A - x^{-1} \quad (5.34)$$

for large x and some constant A that depends on the initial conditions. For nonzero A this is not a consistent balance because y^5 asymptotes to a constant as $x \rightarrow \infty$.

The second balance leads to

$$y = \pm(4(x+c))^{-1/4} \quad (5.35)$$

for a constant c . For this balance both of the terms we are keeping are in fact larger than the terms we are deleting, so this is also consistent. Note that this dominant balance can be either positive or negative. We would expect that which of these solutions is selected depends on the initial condition. More on this below.

Let us test this theory for a positive initial condition. Figure 5.5 shows a numerical solution to the equation when $y(1) = 17$. Indeed, as expected at large x the solution agrees with our expected answer. We also see that there is a regime at intermediate x where the solution $y \sim x^{-2/5}$ works quite well.

Why is that? Let us recall that the $x^{-2/5}$ solution broke down because $\frac{dy}{dx} = -2/5x^{-7/5}$, which decays more slowly than the kept terms x^{-2} at large x . At which value of x does our deleted term start to lag behind the kept terms? We can find this by setting

$x^{-2} = 2/5x^{-7/5}$. This then gives $x^{3/5} = 5/2$, or $x = 4.6$. We therefore expect this solution to remain somewhat faithful up this point, and it does.

What about at small x ? We have chosen $y(1) = 17$ and so initially $y^4 \gg (1 + x^2)^{-1}$. Therefore we expect the small x dominant balance to be $\frac{dy}{dx} + y^5 = 0$. The solution to this is that $y = (4x - 4 + (17)^{-4})^{-1/4}$. Indeed, this agrees quite well with the simulation.

5.4.1 Negative Initial Conditions

Now let us consider initial conditions where $y < 0$. We consider $y(1) = y_0$, and integrate the equation outwards from $x = 1 \rightarrow \infty$. Here in principle we have two possible dominant balances: we have

$$y = -(4(x + c))^{-1/4}, \quad (5.36)$$

and also

$$y = A - 1/x. \quad (5.37)$$

We stated above that if $A = 0$ then this dominant balance is consistent. Note that setting an integration constant to zero is equivalent to setting an initial condition on the solution. Thus we expect that there is precisely one initial condition that leads to the solution $y = -x^{-1}$, whereas most negative initial conditions will lead the solutions to end up on $y = -(4x)^{-1/4}$.

We take note that when we assume a scaling of $y = -\frac{1}{x}$ for large x , we're assuming the integration constant, A , is zero for all x . In order to estimate the value of y_0 such that A equals zero, we should choose an x approximately equal to the x where the scaling of y changes from the scaling developed for small x to the scaling developed for large x . From our previous analysis, we know that the difference between small x and large x is defined as whether we neglect the "1" in the denominator of $\frac{1}{1+x^2}$, so for a first order guess, we assume the transition will occur near $x = 1$, and we solve for y given our scaling of $y = A - \frac{1}{x}$, which yields $y_0 = -1$. Recall from class that $y_0 \approx -0.97625$,

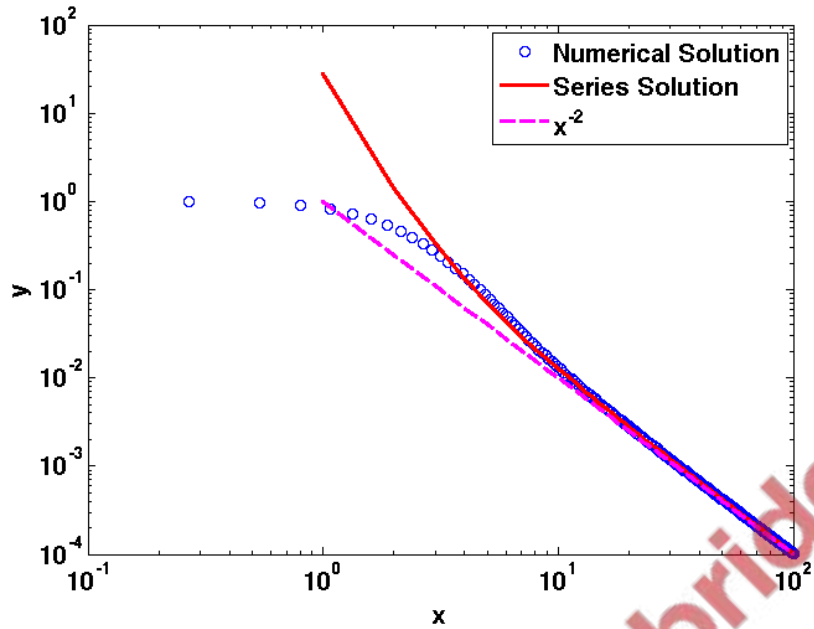


Figure 5.4. Comparison between the numerical solution to the ordinary differential equation and the formula we have derived.

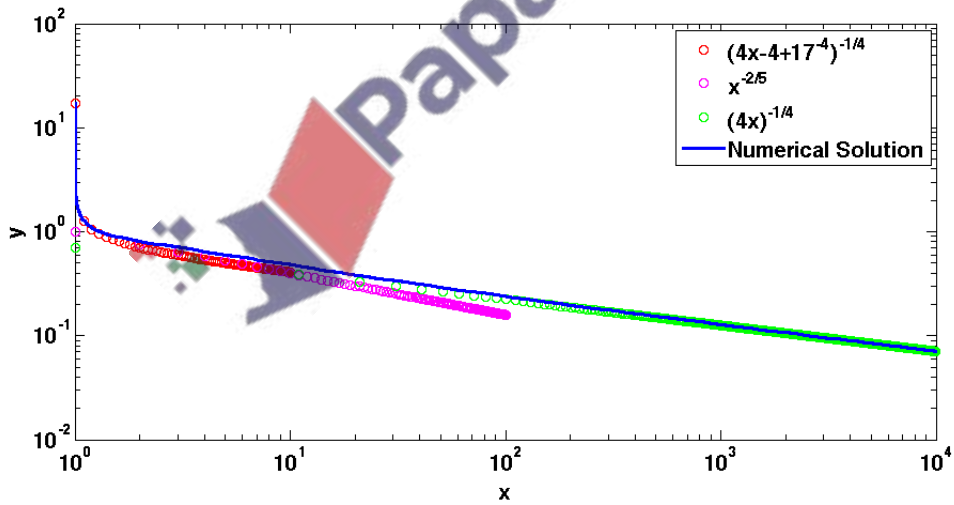


Figure 5.5. Double log plot of the solution to (5.30) comparing with various regimes.

Program 6 MATLAB code used to create figure 5.5

```
1 function figure55
2 % type "figure55" in the command window to generate the figure
3
4 % define the starting point and the end point of the integration
5 xinit = 1;
6 xfinal = 10^4;
7
8 % define the initial condition
9 yinit = 17;
10
11 % integrate the equation
12 [x,y] = ode45(@fun530, [xinit, xfinal], yinit);
13
14 % plot the result of the integration along with the asymptotic solutions
15 t = 1:0.1:10;
16 loglog(t, (4*t-4+17^(-4)).^(-1/4), 'ro', 'markersize', 7)
17 hold on
18 t = 1:100;
19 loglog(t, t.^(-2/5), 'mo', 'markersize', 7)
20 t = 1:10:10^4;
21 loglog(t, (4*t).^(-1/4), 'go', 'markersize', 7)
22 loglog(x, y, 'b-', 'linewidth', 2)
23
24 set(gca, 'fontsize', 16, 'fontname', 'Helvetica', 'fontweight', 'b')
25 xlabel('x')
26 ylabel('y')
27 legend('(4x-4+17^{-4})^{-1/4}', ...
28        'x^{-2/5}', ...
29        '(4x)^{-1/4}', ...
30        'Numerical Solution')
31
32
33 function dy = fun530(x,y)
34 % define the ODE
35
36 dy(1) = -y^5+1/(1+x^2);
```

so if we had chosen $x = 1.0243277848911652$ to evaluate our constant, we would have calculated the correct value of y_0 .

To explore this, we plot in Figure 5.6(A) the absolute value of y as a function of x for different initial conditions in a double log plot. The blue green line in the figure is the $-1/x$ solution whereas the red line represents both the $\pm(4x)^{-1/4}$ solutions—note that both solutions are represented by this line because we are plotting $|y|$. We see that for initial conditions $y_0 > -0.97625$ the solution ultimately ends upon the $(4x)^{-1/4}$ asymptote, though these solutions have a kink in them—it occurs because in this regime the solution is actually changing sign, as shown in the semilog plot in Figure 5.6(B); when $y_0 < -0.97625$, the solution converges on the $(4x)^{-1/4}$ asymptote in smooth fashion—these solutions are *not* changing sign.

What is really interesting about this figure is that it demonstrates that at the borderline between these two behaviors, namely when $y_0 = -0.97625$, the solution agrees with our $-1/x$ solution! This is the value of the initial condition where the integration constant $A = 0$. Interestingly the initial condition that corresponds to $A = 0$ separates two qualitatively different classes of initial conditions: those that have a solution ending up on $(4x)^{-1/4}$ from those who end up on $-(4x)^{-1/4}$, as you can see in the magnified window of Figure 5.6(B).

Information Lost?

The analysis described above found a relationship between the initial conditions and the final behavior, as $x \rightarrow \infty$. Whereas the final solution was one of a discrete number of possibilities, the initial condition was given by a continuous number. We have thus *lost* one parameter worth of degree of freedom in the final solution. This is interesting and peculiar: *where did the extra parameter go?*

The answer to this is kept in the full solution that is valid at large x ,

$$y = \pm(4(x + c))^{-1/4}. \quad (5.38)$$

The constant c must include memory of the initial condition. Now if the $(1 + x^2)^{-1}$ term were not present, the solution $y = \pm(4(x + c))^{-1/4}$ would be the exact solution. However with this extra term it is no longer exact.

Nonetheless, the dominant balances as $x \rightarrow \pm\infty$ still hold. With some courage (or some luck), we could try to completely ignore the term $(1 + x^2)^{-1}$, then we could fix c using the initial condition—namely, substitute (x_0, y_0) into the solution, we have that

$$y_0 = \frac{1}{(4(c + x_0))^{1/4}}, \quad (5.39)$$

or

$$c = -x_0 + \frac{1}{4y_0^4}. \quad (5.40)$$

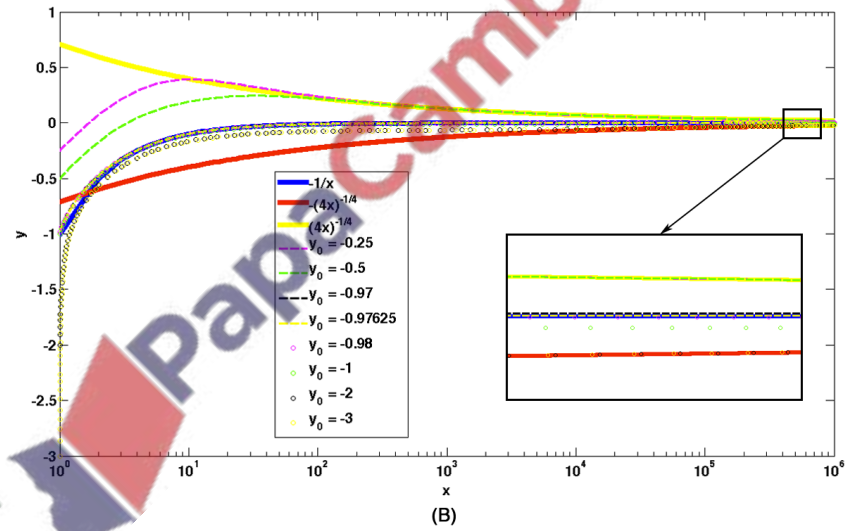
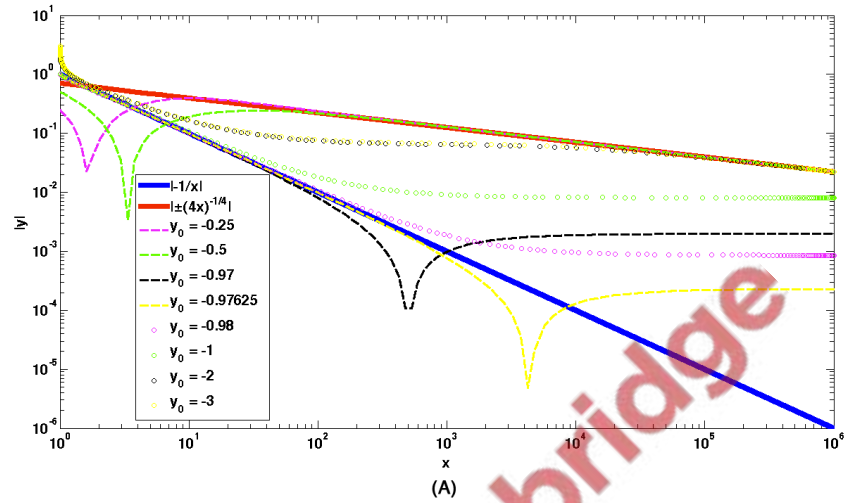


Figure 5.6. (A) Double log plot of the absolute value of the numerical and asymptotic solutions to eqn. (5.30) for a variety of initial conditions; (B) Semilog plot of the numerical and asymptotic solutions to eqn. (5.30) for a variety of initial conditions.

Program 7 MATLAB code used to create figure 5.6

```
1 function figure56
2 % type "figure56" in the command window to generate the figure
3
4 % define the starting point and the end point of the integration
5 xinit = 1;
6 xfinal = 10^6;
7
8 % define a vector containing different initial conditions
9 yinit = [-0.25, -0.5, -0.97, -0.97625, -0.98, -1, -2, -3];
10
11 % define a cell to store the solutions
12 sol = cell(length(yinit),2);
13
14 % integrate the equation and store solutions
15 for i = 1:length(yinit)
16     [x,y] = ode45(@fun530, [xinit, xfinal], yinit(i));
17     sol{i, 1} = x; sol{i, 2} = abs(y);
18 end
19
20 % plot the result of the integration along with the asymptotic solutions;
21 % change "loglog" to "semilogx" can get (B)
22 t = xinit:xfinal;
23 loglog(t, t.^(-1), 'b-', 'linewidth', 6)
24 hold on
25 loglog(t, (4*t).^(-1/4), 'y-', 'linewidth', 6)
26 loglog(sol{1, 1}, sol{1, 2}, 'm-', 'linewidth', 2)
27 loglog(sol{2, 1}, sol{2, 2}, 'g-', 'linewidth', 2)
28 loglog(sol{3, 1}, sol{3, 2}, 'k-', 'linewidth', 2)
29 loglog(sol{4, 1}, sol{4, 2}, 'y-', 'linewidth', 2)
30 loglog(sol{5, 1}, sol{5, 2}, 'mo', 'markersize', 5)
31 loglog(sol{6, 1}, sol{6, 2}, 'go', 'markersize', 5)
32 loglog(sol{7, 1}, sol{7, 2}, 'ko', 'markersize', 5)
33 loglog(sol{8, 1}, sol{8, 2}, 'yo', 'markersize', 5)
34
35 set(gca, 'fontsize', 16, 'fontname', 'Helvetica', 'fontweight', 'b')
36 xlabel('x')
37 ylabel('|y|')
38 legend('|-1/x|', '|pm(4x)^{-1/4}|', ...
39         'y_0 = -0.25', 'y_0 = -0.5', 'y_0 = -0.97', 'y_0 = -0.97625', ...
40         'y_0 = -0.98', 'y_0 = -1', 'y_0 = -2', 'y_0 = -3')
41
42
43 function dy = fun530(x,y)
44 % define the ODE
45
46 dy(1) = -y^5+1/(1+x^2);
```

Note that this predicts that the solution will diverge when

$$x = x^* = x_0 - \frac{1}{4y_0^4}. \quad (5.41)$$

To test this we consider the complete solution for $y(1) = 2$. In addition to integrating in the positive direction towards $x \rightarrow \infty$ we have also integrated backwards towards $x \rightarrow -\infty$. Our theory states we expect the divergence to occur a distance $(4y_0^4)^{-1}$ from the starting location.

Figure 5.7 is the numerical solution to the equation integrated in two directions. Note that the solution diverges at $x = 0.9826$, and our theory predicts $63/64 = 0.9844$. The divergence is well captured by our dominant balance!

Let's now find out if our theory for the location of the blowup point (x^*) works for general initial conditions. We fix $x_0 = 1$ and vary y_0 . Figure 5.8 shows the blowup location ($x_0 - x^* = 1 - x^*$) as a function of y_0 ; the blue dots denote the numerical simulations and the green line shows our theory. We see that our theory matches the simulation very well *in the limit when y_0 is large*.

Below $y_0 \approx 1$, the theory does not work as well. We can rationalize this by dominant balance: in assuming that the solution with the integration constant c is given by our theory we have completely neglected $(1 + x^2)^{-1}$. For the initial conditions this requires that

$$y_0^5 \gg \frac{1}{1 + x_0^2}. \quad (5.42)$$

For our initial condition, this requires $y_0 \gg 1/2^{1/5} = 0.8706$. Indeed, this is just where our theory starts to fail!

What happens when the theory fails? In this regime the y^5 term is evidently negligible, so the differential equation is then effectively

$$\frac{dy}{dx} = \frac{1}{1 + x^2}, \quad (5.43)$$

and the solution to this is just

$$y = \tan^{-1}(x) + y_0 - \tan^{-1}(x_0). \quad (5.44)$$

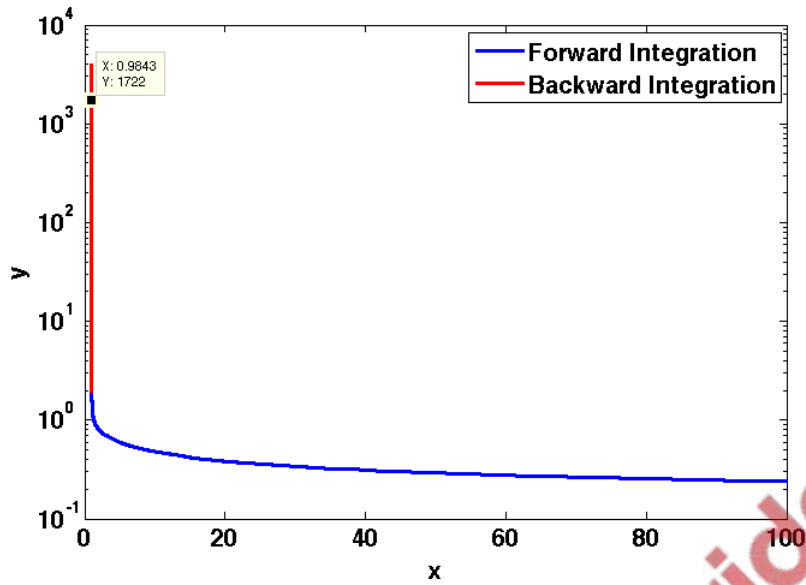


Figure 5.7. For the initial condition $y(1) = 2$, this shows the solution $y(x)$ integrating both in the forward direction and backward direction. In the backward direction (the red line) the solution blows up as expected.

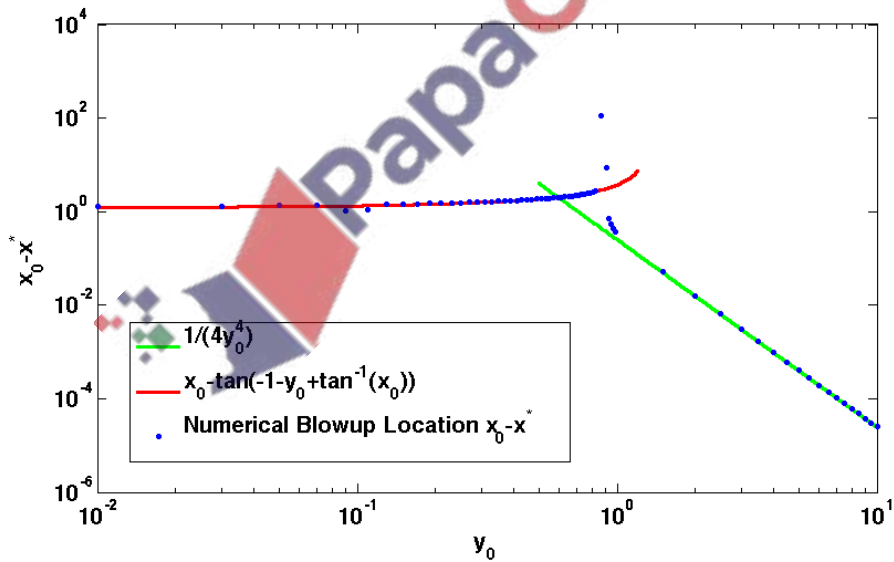


Figure 5.8. Blowup location $x_0 - x^*$ as a function of y_0 ; the blue dots denote the numerical simulations and the green and red curves show our theory.

Program 8 MATLAB code used to create figure 5.7

```
1 function figure57
2 % type "figure57" in the command window to generate the figure
3
4 % define the starting point and the end point for two directions
5 xinit = 1;
6 xfinal_1 = 100;
7 xfinal_2 = -100;
8
9 % define the initial condition
10 yinit = 2;
11
12 % integrate the equation in two directions
13
14 [x_1, y_1] = ode45(@fun530, [xinit, xfinal_1], yinit);
15 [x_2, y_2] = ode45(@fun530, [xinit, xfinal_2], yinit);
16
17 % plot the result of the integration in two directions
18 semilogy(x_1, y_1, 'b-', 'linewidth', 2)
19 hold on
20 semilogy(x_2, y_2, 'r-', 'linewidth', 2)
21
22 set(gca, 'fontsize', 16, 'fontname', 'Helvetica', 'fontweight', 'b')
23 xlabel('x')
24 ylabel('y')
25 legend('Forward Integration', 'Backward Integration')
26
27
28 function dy = fun530(x,y)
29 % define the ODE
30
31 dy(1) = -y^5+1/(1+x^2);
```

Program 9 MATLAB code used to create figure 5.8

```
1 function figure58
2 % type "figure58" in the command window to generate the figure
3
4 % define the starting point and the end point for two directions
5 xinit= 1;
6 xfinal_1 = 10^6; xfinal_2 = -10^6;
7
8 % define the initial conditions in a vector
9 yinit = [0.01:0.02:1, 1.5:0.5:10];
10
11 % define a vector to store the numerical blowup locations
12 blocation = zeros(1, length(yinit));
13
14 % integrate the equation in two directions
15 for i = 1:length(yinit)
16     [x_1, y_1] = ode45(@fun530, [xinit, xfinal_1], yinit(i));
17     [x_2, y_2] = ode45(@fun530, [xinit, xfinal_2], yinit(i));
18
19     % get the blowup location
20     x = [flipud(x_2); x_1]; y = [flipud(y_2); y_1];
21     bpoint = x(abs(y)==max(abs(y)));
22     bpoint = bpoint(1);
23     blocation(i) = xinit-bpoint;
24 end
25
26 % plot the numerical simulations along with analytic predictions
27 t = 0.5:0.01:10;
28 loglog(t, 1./(4*t.^4), 'g-', 'linewidth', 2)
29 hold on
30 t = 0.01:0.01:1.2;
31 loglog(t, xinit-tan(-1-t+atan(1)), 'r-', 'linewidth', 2)
32 loglog(yinit, blocation, 'b.', 'linewidth', 2)
33
34 set(gca, 'fontsize', 16, 'fontname', 'Helvetica', 'fontweight', 'b')
35 xlabel('y_0')
36 ylabel('x_0-x^4')
37 legend('1/(4y_0^4)', ...
38        'x_0-tan(-1-y_0+tan^{-1}(x_0))', ...
39        'Numerical Blowup Location x_0-x^*')
40
41
42 function dy = fun530(x,y)
43 % define the ODE
44
45 dy(1) = -y^5+1/(1+x^2);
```

We expect the y^5 term to become important when $y \approx -1$, so we expect the blowup point to be at about

$$-1 = \tan^{-1}(x) + y_0 - \tan^{-1}(x_0), \quad (5.45)$$

or

$$x^* \approx \tan(-1 - y_0 + \tan^{-1}(x_0)). \quad (5.46)$$

The prediction of this theory is given by the red curve in Figure 5.8. Indeed, dominant balance has served us very well!

5.5 Higher Order Nonlinear Ordinary Differential Equations

A logical structure for how to think about differential equations is now starting to emerge. Equations consist of different terms. Depending on both the initial conditions and where you are in the solution, these terms can be important in some regions and not important in others. One can understand quantitatively the structure of the solutions to equation by figuring out which terms are important at where and piecing the different solutions together.

We now continue this discussion with an example of higher order differential equation.

Consider the following nonlinear ordinary differential equation:

$$\frac{d^2y}{dx^2} + \frac{dy}{dx} - y^4 = \frac{1}{x^2}. \quad (5.47)$$

Our goal is to find the behavior of the solution as $x \rightarrow \infty$, with the initial conditions $y(1) = 1$, $y'(1) = 0$.

Now we have *four* terms in the equation, and we must find the *consistent* dominant balance. To proceed, we have two choices:

1. Experiment with dominant balances and try to guess the largest term. Verify the conjecture with a computer simulation.
2. Simulate the equation to find out what the behavior is; then try to backout from the equation that the observed equation indeed is a valid balance, and *posteriorly* rationalize why it is the correct one.

From our point of view, either of these methodologies is perfectly legitimate. What is important is to, in the end, derive a final answer which contains both analytic and numerical components, and shows that they agree. Although each argument by itself is not completely compelling (the computer calculation might be wrong, as might the approximation you decide to make), together the two approaches *are* compelling, as long as they agree with each other.

Here we will follow the first approach, and experiment with the possible balances. Note that the term $1/x^2$ converges to 0 quadratically when $x \rightarrow \infty$, so our intuition is that we should not keep this term in our balances. Therefore we will try the three balances developed from the rest three terms.

5.5.1 The First Balance

The first balance we try is:

$$\frac{d^2y}{dx^2} + \frac{dy}{dx} = 0. \quad (5.48)$$

This dominant balance implies that

$$y = A + Be^{-x}. \quad (5.49)$$

With simple calculation, it is easy to figure out that, when $x \rightarrow \infty$, both the kept terms converge exponentially to 0, whereas one of the neglected term, $1/x^2$, converges to 0 quadratically and the other one, y^4 , converges to constant A^4 . So the kept terms are smaller than neglected terms at large x , and this balance is inconsistent.

If $A = 0$, y^4 is much smaller than the kept terms

5.5.2 The Second Balance

The second balance we try is:

$$\frac{dy}{dx} - y^4 = 0. \quad (5.50)$$

The solution to this balance is

$$y = (-3(x+c))^{-1/3}, \quad (5.51)$$

where c is a constant. Note that for this balance, x cannot go all the way to infinity, because the solution blows up at $x = -c$. Thus, we should use $x \rightarrow -c$ to check the consistency of this balance instead of $x \rightarrow \infty$.

The kept terms are

$$\frac{dy}{dx} = y^4 = (-3(x+c))^{-4/3} \sim (x+c)^{-4/3}. \quad (5.52)$$

They diverge as $x \rightarrow -c$. One of the neglected terms, $1/x^2$, converge to the constant $1/c^2$ as $x \rightarrow -c$, so it is negligible. However, for the other neglected term $\frac{d^2y}{dx^2}$, we have

$$\frac{d^2y}{dx^2} \sim (x+c)^{-7/3} \gg \frac{dy}{dx} \quad (5.53)$$

as $x \rightarrow -c$. Thus, this balance is inconsistent.

5.5.3 The Third Balance

The third balance we try is:

$$\frac{d^2y}{dx^2} - y^4 = 0. \quad (5.54)$$

This brings up the balance where $\frac{d^2y}{dx^2}$ plays a role. Here, having been schooled in school of mathematical tricks, one is tempted to solve this dominant balance exactly by multiplying the equation through by $\frac{dy}{dx}$, and then integrating:

$$\frac{d^2y}{dx^2} \frac{dy}{dx} - y^4 \frac{dy}{dx} = \frac{d}{dx} \left(\frac{1}{2} \frac{d(y^2)}{dx} - \frac{y^5}{5} \right) = 0. \quad (5.55)$$

However, this then reduces to a quadrature that cannot be solved.

We are better off by just using our insight from the previous balance to realize that y is going to diverge at finite x , as long as $y > 0$. Since our initial condition assumes that $y > 0$, this is a reasonable assumption to make for all x . So what is the functional form near the divergence? Here we make a very general claim:

Divergences nearly always have a power law form

$$y(x) = A(x^* - x)^p.$$

Plugging this into our assumed dominant balance implies that

$$\frac{d^2y}{dx^2} = p(p-1)A(x^* - x)^{p-2} = A^4(x^* - x)^{4p}. \quad (5.56)$$

This only works if $p - 2 = 4p$, or $p = -2/3$, which implies that $A^3 = 10/9$, or $A = (10/9)^{1/3} \approx 1.03574$. Hence we have concluded that

$$y(x) = A(x^* - x)^{-2/3}. \quad (5.57)$$

This balance is consistent—both our neglecting of the $\frac{dy}{dx}$ and the $1/x^2$ terms are clearly justified.

5.5.4 Testing the Theory

Now, let us see if our theorizing actually works. Figure 5.9 plots a numerical solution for $y(x)$ given the initial conditions $y(1) = 1$ and $y'(1) = 0$.

As we expected, the solution diverges at finite x , roughly at $x \approx 2.4007857$. Now let us determine if the functional form we deduced is correct. In Figure 5.10 we plot $\log(y)$ as a function of $\log(x^* - x)$, where $x^* = 2.40078568328535$. The solid dots show the functional form that we anticipated, $(x^* - x)^{-2/3}$.

Rather remarkably, the numerical solution agrees almost exactly with the analytic formula over almost the entire range of behavior. Of course, in making this comparison we have input an important piece of information, namely x^* , the value of x where $y(x)$ diverges. Later on in this course, we will discuss ways of connecting x^* to the initial conditions.

5.6 Numerical Solution of Ordinary Differential Equations

Here we briefly remark on the numerical solutions of ordinary differential equations. For these introductory remarks we will focus on the simple equation

$$\dot{y} = ay. \tag{5.58}$$

Our remarks apply more generally, however.

To solve the equation numerically, the simplest idea is to discretize time, writing $t_n = n\Delta t$ and $y(n\Delta t) = y_n$. Then we can approximate

$$\dot{y} \approx (y_{n+1} - y_n)/\Delta t. \tag{5.59}$$

The error in this approximation follows directly from a Taylor series expansion:

$$y((n+1)\Delta t) = y(n\Delta t + \Delta t) = y(n\Delta t) + \dot{y}(t_n)\Delta t + O(\Delta t)^2. \tag{5.60}$$

Using this approximation we immediately arrive at the numerical method called *Euler's method*:

$$y_{n+1} = y_n + \Delta t f(y_n) + O((\Delta t)^2). \tag{5.61}$$

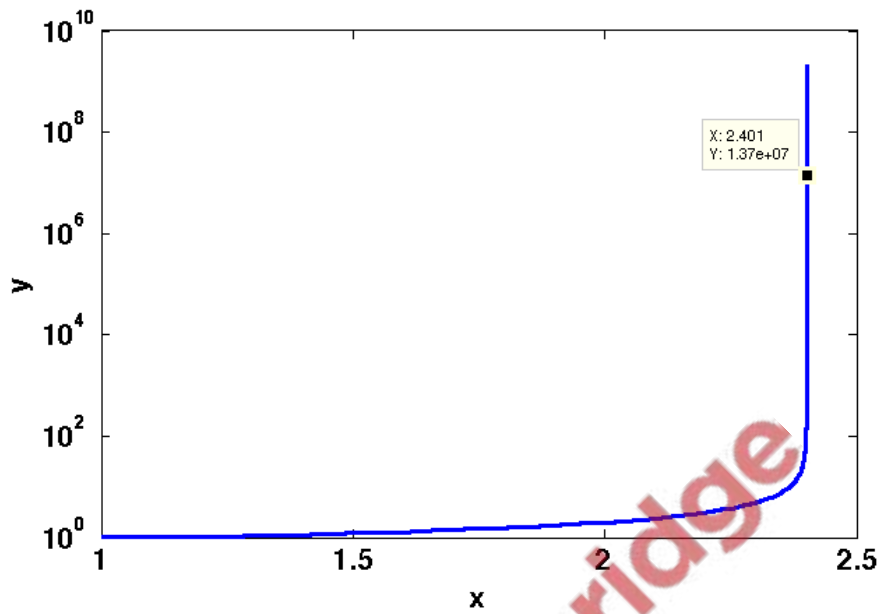


Figure 5.9. Solution of equation with $y(1) = 1, y'(1) = 0$. The solution diverges before $x \approx 2.5$.

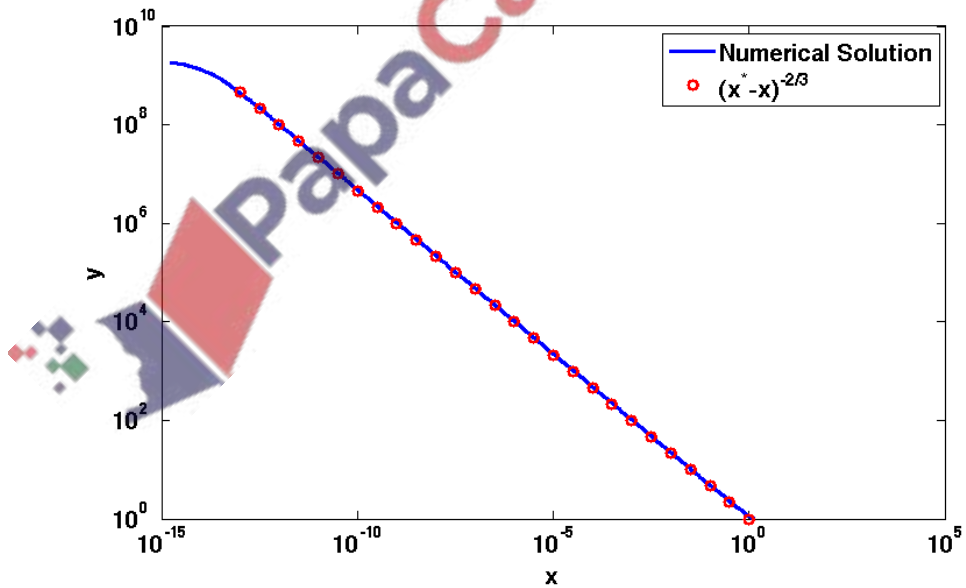


Figure 5.10. Solution of equation (5.55) with $y(1) = 1, y'(1) = 0$, plotted on a double logarithmic plot. The solid dots are the solution $y(x) = A(x^* - x)^{-2/3}$

Program 10 MATLAB code used to create figure 5.9

```
1 function figure59
2 % type "figure59" in the command window to generate the figure
3
4 % define the starting point and the end point of the integration
5 xinit = 1;
6 xfinal = 100;
7
8 % define the initial conditions
9 yinit = 1; % y(1)
10 ypinit = 0; % y'(1)
11
12 % integrate the equation
13 [x, y] = ode45(@fun548, [xinit xfinal], [yinit ypinit]);
14
15 % plot the result of the integration
16 semilogy(x, y(:,1), 'b-', 'linewidth', 2)
17
18 set(gca, 'fontsize', 16, 'fontname', 'Helvetica', 'fontweight', 'b')
19 xlabel('x')
20 ylabel('y')
21
22
23 function dy = fun548(x, y)
24 % define the ODE
25
26 dy = zeros(2,1);
27
28 dy(1) = y(2);
29 dy(2) = y(1)^4+1/x^2-y(2);
```

Program 11 MATLAB code used to create figure 5.10

```
1 function figure510
2 % type "figure510" in the command window to generate the figure
3
4 % define the starting point and the end point of the integration
5 xinit = 1;
6 xfinal = 100;
7
8 % define the initial conditions
9 yinit = 1; % y(1)
10 ypinit = 0; % y'(1)
11
12 % integrate the equation
13 [x, y] = ode45(@fun548, [xinit xfinal], [yinit ypinit]);
14
15 % get the blowup point
16 xstar = x(y(:,1)==max(y(:,1)));
17 xstar = xstar(1);
18
19 % plot the result of the integration along with the analytical prediction
20 loglog(xstar-x, y(:,1), 'b-', 'linewidth', 2)
21 hold on
22 t = 10.^(-13:0.5:0);
23 loglog(t, t.^(-2/3), 'ro', 'markersize', 7, 'linewidth', 1.1)
24
25 set(gca, 'fontsize', 16, 'fontname', 'Helvetica', 'fontweight', 'b')
26 xlabel('x')
27 ylabel('y')
28 legend('Numerical Solution', '(x*-x)^{-2/3}')
29
30
31 function dy = fun548(x,y)
32 % define the ODE
33
34 dy = zeros(2,1);
35
36 dy(1) = y(2);
37 dy(2) = y(1)^4+1/x^2-y(2);
```

To use this equation we simply specify an initial condition $y(0) = y_0 = A$, and then apply this algorithm to generate a sequence y_{n+1} . Our hope, of course, is that the sequence y_n has something to do with the solution to the differential equation $y' = f(y)$.

We remark that the error expressed in this formula is a *local error*, in that it is the error made at each time step of the method. Over many iterations, this error accumulates. Assuming that we want to integrate the ODE up to $t = 1$, we need to take $1/\Delta t$ time steps of Euler's method. This implies that the *global error* is of order $(\Delta t)^{-1}(\Delta t)^2 = \Delta t$. Hence, on cutting the timestep by a factor of two we expect a factor of two improvement in the error.

Let's try this out on our simplest example—equation (5.58). Euler's method is just $y_{n+1} = y_n + a\Delta t y_n$. The solution is

$$y_n = y_0(1 + a\Delta t)^n. \quad (5.62)$$

Writing $t = n\Delta t$ implies $y_n = y_0(1 + at/n)^n$. This converges to y_0e^{at} as $n \rightarrow \infty$. At any finite n there is an error as given above.

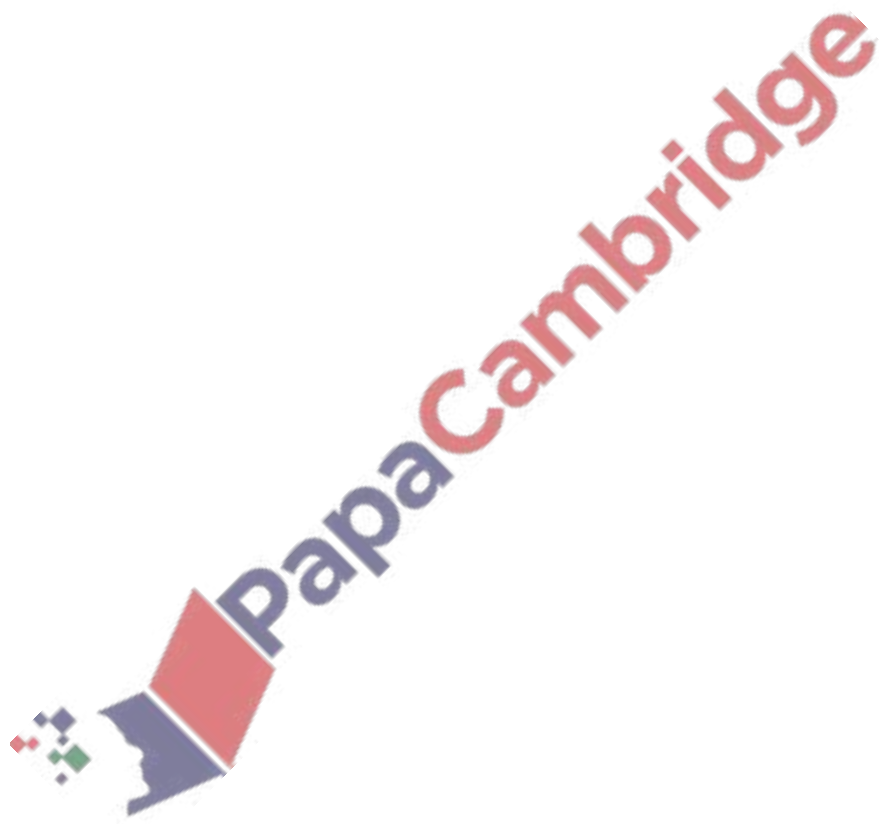
5.6.1 Remarks

- Another source of error in using a computer is round off error. The size of this depends on the type of arithmetic you are doing (i.e. double precision has accuracy 10^{-15}). This means that every time step you are effectively adding an error r to the equation. For $N = (\Delta t)^{-1}$ time steps, this is an error of $(\Delta t)^{-1}r$. Hence the total error is

$$c\Delta t + \frac{r}{\Delta t}. \quad (5.63)$$

This function has a minimum when $\Delta t \sim \sqrt{r}$. This is the smallest time step that is reasonable with double precision arithmetic.

- The central goal is to get the answer as accurate as possible while doing as little work as possible. What this means depends on the situation. Often, a more accurate scheme than Euler is *needed*. How can one construct a more accurate method? Simple, use a better *approximation* to the derivative! As an example, we present the *trapezoidal rule*. Let us take $y_{n+1} = y_n + \Delta t(f(y_n) + f(y_{n+1}))/2$. It is easy to show that the local truncation error for this method is of order $(\Delta t)^3$. So, in a sense, this is a better method. One downside of this method is that it is *implicit*, that is, if f is a nonlinear function, then to use this method it is necessary to solve a nonlinear equation at each time step.
- There are a hierarchy of more sophisticated methods that are both more accurate and have other nice properties. The detailed study of these methods is beyond the framework of this class. On the course web site, we have included the technical article describing the various choices of ODE solvers available in MATLAB, all of which are more sophisticated than our simple constructions above. I would encourage you to look through this material to get a better idea of what is used in practice.



6 “Simple” Integrals

We now move on to use dominant balance ideas to explicitly evaluate integrals. As always our goal is to learn how to find formulae that represent the value of arbitrary integral, over some range of parameter space; these formulae will be tested with numerical simulations.

6.1 What is a simple integral

Before beginning it is worth reflecting on the question—what is a simple integral? When are integrals complicated? Typical mathematics problems define simple integrals as those which can be evaluated with analytical formulae, whereas complicated integrals are those that cannot be evaluated in closed form.

For example, here is an integral to try your skills on. Show that:

$$\int \frac{x^{1/4}}{1 + \sqrt{x}} = 4x^{3/4} - 4x^{1/4} + 4 \tan^{-1}(x^{1/4}). \quad (6.1)$$

Now this is not a simple integral in that it is easy to evaluate, but it is simple in that the solution can be expressed in closed form in terms of well known functions. On the other hand it must be admitted that the function $\tan^{-1}(x)$ is only simple if you happen to have a way to evaluate it. To some extent, the same thing is true of $x^{3/4}$, etc.

Sometimes integrals are defined to be simple because the integral is given a name. An example of this is the elliptic integral, given by

$$E(x; k) = \int_0^x \frac{\sqrt{1 - k^2 t^2}}{\sqrt{1 - t^2}} dt. \quad (6.2)$$

where the elliptic function $E(x; k)$ is a so-called special function. Another famous integral is the so-called Sine integral, given by

$$\text{Si}(x) = \int_0^x \frac{\sin t}{t} dt \quad (6.3)$$

which is a special case of the exponential integral

$$E_1(z) = \int_1^\infty \frac{\exp(-zt)}{t} dt, \quad (\text{Re}(z) \geq 0) \quad (6.4)$$

where z is a complex number.

More complete lists of integrals can be found at

http://en.wikipedia.org/wiki/Table_of_integrals#Table_of_Integrals

or in the classical book by Gradshteyn and Ryzhik

<http://www.mathtable.com/gr/>

Within the classical definition, any integral that cannot be evaluated in closed form is complicated. We completely disagree with these definitions. In point of fact, the phrase *closed form* refers to functions that have been named and tabulated. Before the computer revolution this was a quite reasonable point of view, because the only way that one could evaluate the numerical value of an integral was to convert it to a known tabulated form and then evaluate it using the table. But we no longer need tables, and in general, computational methods can be used equally well for any integrals.

The one caution with this is that as we will see numerical evaluation of integrals can become quite challenging; in principle the algorithms that have been designed should work on arbitrary integrals given sufficient computer power.

6.1.1 A more sensible definition

A more sensible definition of simple integral asks how easy is it to invent an analytical approximation to the integral. There is a simple criterion that can be used to evaluate this:

Integrals are easy to evaluate when the integrand is essentially of a single sign.

Integrals are hard to evaluate when the integrand oscillates rapidly.

Why? When an integral has a single sign we need to figure out the volume (area) enclosed by the integrand; when it oscillates there are cancellations between the positively signed quantities and the negatively signed ones and this makes things much trickier to evaluate rapidly.

For example, here is a truly difficult integral:

$$I(x) = \int_0^{10} \frac{e^{-x(4t^2+5t)} \sin(13x(t+3t^3))}{1+8t^3} dt. \quad (6.5)$$

The integral is highly oscillatory as the accompanying figure demonstrates for $x = 50$. To calculate this integral we need to figure out exactly how much cancellation occurs.

Our definition completely contradicts the classical one. Whereas the classical definition would define the Sine integral as easy (being defined as a special function) by our own definition the integral is hard because the integrand rapidly oscillates. In contrast we think single signed integrands are simple whereas the classical definition does not recognize this.

For fun invent a positive definite integrand that Matlab or Mathematica cannot evaluate in closed form.

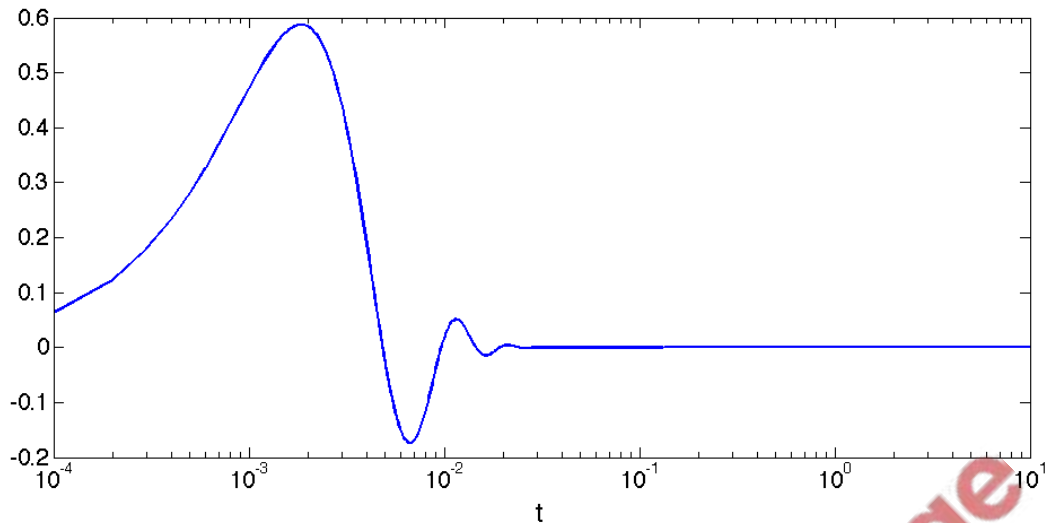


Figure 6.1. A plot of the integrand of the complicated integral.

In this class we will discuss both simple and complicated integrals. They both come up frequently in practice. Here we will discuss the case of simple integrals. Complicated integrals require further mathematical developments, in particular the ability to extend and perform integrals in the complex plane. As a precursor to how powerful such arguments can be, we claim that the complicated integral can be shown to obey the very simple law

$$I(x) \approx \frac{13}{194} \frac{1}{x}. \quad (6.6)$$

Figure ?? compares numerical computation of $I(x)$ to this exceedingly simple formula.

6.2 A very easy simple integral

To get us into the spirit, let us consider a very easy simple integral, indeed one that you have done before. Consider

$$\int_0^1 \frac{dx}{\epsilon + x^2}. \quad (6.7)$$

This integral can be done exactly but it will serve a purpose.

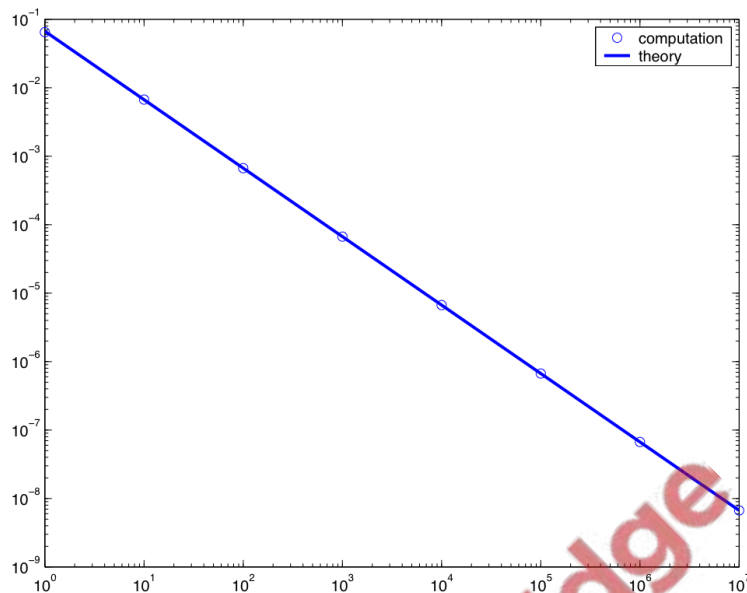


Figure 6.2. Comparison of the formula $I(x) = 15/194x^{-1}$ with the numerical simulations

Basic Principle for Evaluating Simple Integrals

If an integral does not oscillate, so that the integral is simple, we can always estimate the value of an integral using the formula

$$\text{Integral} = \text{Height} \times \text{Width}.$$

In the limit where $\epsilon \rightarrow 0$, the maximum value of the integrand occurs at $x = 0$, and is

$$\text{Height} = 1/\epsilon. \tag{6.8}$$

What about the width? A simple way to calculate the width over which an integrand oscillates is simply to ask at what distance from the location where the integrand is a maximum does it decrease to a factor of 2 of this maximum. In the present case the maximum occurs at $x = 0$, so that the width satisfies the equation

$$\frac{1}{\epsilon + \text{Width}^2} = \frac{1}{2\epsilon}. \tag{6.9}$$

Hence

$$\text{Width} = \sqrt{\epsilon}. \tag{6.10}$$

Thus, using our basic principle for evaluating simple integrals, we have that

$$\text{Height} \times \text{Width} = \sqrt{\epsilon} \frac{1}{\epsilon} \sim \frac{1}{\sqrt{\epsilon}} \quad (6.11)$$

How do we turn this into a more exact formula?

Motivated by the above discussion, let's make the change of variable $x = y\sqrt{\epsilon}$. The integral then becomes

$$\frac{1}{\sqrt{\epsilon}} \int_0^{1/\sqrt{\epsilon}} \frac{1}{1+y^2} dy. \quad (6.12)$$

Now students of calculus pride themselves with the ability to evaluate this integral, as

$$= \frac{1}{\sqrt{\epsilon}} \arctan\left(\frac{1}{\sqrt{\epsilon}}\right). \quad (6.13)$$

This formula is a very good thing unless you are unfortunate enough to find yourself needing the answer without a calculator that stores the values of $\arctan(x)$. If you remember something about the arctan function, you will also remember that in the limit $\epsilon \rightarrow 0$, there is an analytical limit, namely

$$\frac{1}{\sqrt{\epsilon}} \int_0^\infty \frac{1}{1+y^2} dy = \frac{\pi}{2\sqrt{\epsilon}}. \quad (6.14)$$

Thus we have an approximate formula for our integral!

6.2.1 What if you are on a desert island without an arctan table

Of course the above derivation requires either an arctan table or knowledge about the integral identity about $\pi/2$. Even living in Cambridge, MA, personally I never remember these things. There is a simple way however to figure out which actually well illustrates the approach we will follow henceforth.

We write

$$\int_0^\infty \frac{1}{1+x^2} dx = \int_0^1 \frac{1}{1+x^2} dx + \int_1^\infty \frac{1}{1+x^2} dx. \quad (6.15)$$

For the first integral, we can expand

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \dots \quad (6.16)$$

whereas for the second we can write

$$\frac{1}{1+x^2} = \frac{1}{x^2} \frac{1}{1+x^{-2}} = x^{-2} - x^{-4} + x^{-6} + \dots \quad (6.17)$$

Now the expansion for the first integral is convergent since $|x| < 1$, whereas the expansion for the second integral is convergent since $|x| > 1$. We thus write

$$\int_0^1 \frac{1}{1+x^2} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots, \quad (6.18)$$

whereas

$$\int_1^\infty \frac{1}{1+x^2} = \frac{1}{2} - \frac{1}{4} + \dots \quad (6.19)$$

We therefore have that the whole integral

$$\int_0^\infty \frac{1}{1+x^2} dx = \left(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots\right) + \left(\frac{1}{2} - \frac{1}{4} + \dots\right) \quad (6.20)$$

This is of course a series approximation for $\pi/2$. But even if we didn't know this we can for example keep the first several terms. Not a bad answer.

6.2.2 Back to the Integral

Lets now go back to the original integral and ask how to generate corrections to this formula? One simple way to go about this is to observe that

$$\frac{1}{\sqrt{\epsilon}} \left(\int_0^\infty \frac{1}{1+y^2} - \int_0^{1/\sqrt{\epsilon}} \frac{1}{1+y^2} \right) = \frac{1}{\sqrt{\epsilon}} \int_{1/\sqrt{\epsilon}}^\infty \frac{1}{1+y^2}. \quad (6.21)$$

This last integral only covers the range where y is very large (when $\epsilon \rightarrow 0$). Therefore we can approximate

$$\int_{1/\sqrt{\epsilon}}^\infty \frac{1}{1+y^2} \approx \int_{1/\sqrt{\epsilon}}^\infty \frac{1}{y^2} = \frac{1}{y} \Big|_{1/\sqrt{\epsilon}}^\infty = \sqrt{\epsilon}. \quad (6.22)$$

Thus we have shown that our integral is actually

$$\frac{\pi}{2\sqrt{\epsilon}} - 1 + \dots \quad (6.23)$$

All of this can of course be directly verified by solving the integral exactly (giving an arctangent, as above), and then Taylor-expanding the arc tangent. The advantage of this example is that the general methodology is much more useful than any knowledge of specific special functions, as we will see in the next example.

6.3 A harder integral

Now consider the integral

$$I(\epsilon) = \int_0^{100} \frac{1}{\epsilon + x^2 + x^5} dx, \quad (6.24)$$

as a function of ϵ . This integral is not exactly solvable in closed form; however we will see that applying the above ideas we will derive arbitrarily accurate approximations to it.

The basic principle is the same as above, we use our basic principle to estimate

$$\text{Integral} = \text{Height} \times \text{Width}.$$

Now in the present case the Height is just the maximum value of the integrand $1/\epsilon$. What is the width? We follow the same procedure as above and define the width as the distance over which the integrand decreases from its maximum value by a factor of 2. This is

$$\frac{1}{\epsilon + \text{Width}^2 + \text{Width}^5} = \frac{1}{2\epsilon}. \quad (6.25)$$

Hence the width obeys the equation

$$x^2 + x^5 = \epsilon \quad (6.26)$$

where we are now denoting the width by x . Hence to find the width we need to solve an algebraic equation—something you are now quite talented at!

Let us proceed and look for a dominant balance.

1. Let us suppose that the width is determined by the x^2 term, so that the integral falls off when $\epsilon \sim x^2$; in this case we have that

$$x \sim \sqrt{\epsilon} \quad (6.27)$$

so that $x^5 \sim \epsilon^{5/2}$. This term is negligible only when $\epsilon < 1$ (as only in this case is $x^5 \ll \epsilon$).

2. On the other hand it is also possible that the x^5 term determines the width. If we assume that $\epsilon \sim x^5$, then $x^2 \sim \epsilon^{2/5}$. This is negligible only when $\epsilon > 1$.

We have therefore found that the width of the integral is $\sqrt{\epsilon}$ when $\epsilon \ll 1$ and $\epsilon^{1/5}$, when $\epsilon \gg 1$. Hence if $\epsilon \ll 1$, then the integral is approximately

$$\epsilon^{-1} \times \sqrt{\epsilon} = \epsilon^{-1/2}. \quad (6.28)$$

On the other hand, if $\epsilon \gg 1$ the integral is approximately

$$\epsilon^{-1} \times \epsilon^{1/5} = \epsilon^{-4/5}. \quad (6.29)$$

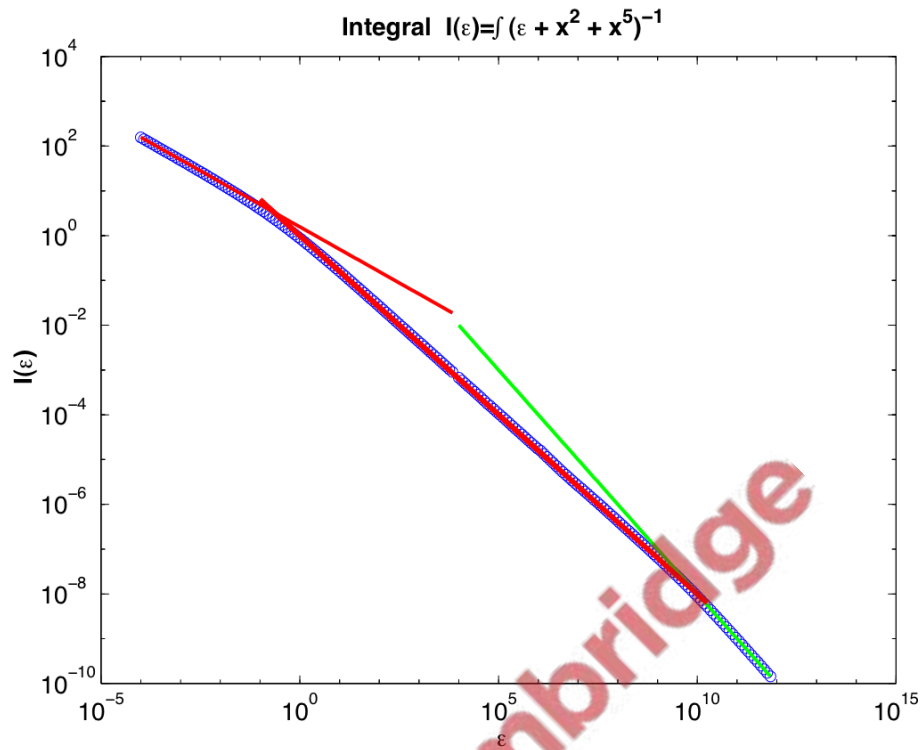


Figure 6.3. Integral as a function of ϵ . Note the existence of three scaling regimes, as argued above. The thin red line is the law $\pi/2\epsilon^{-1/2}$; the thick red line is the law $\epsilon^{-4/5}$; the green line is the law $100/\epsilon$

There is one other limit that we need to consider. Note that the integrand is over the range $0 \leq x \leq 100$. If the width is larger than 100 then the width is actually determined by the range of integration. Hence if the width

$$\epsilon^{1/5} > 100 \tag{6.30}$$

we expect that the integral will be approximately

$$\epsilon^{-1} \times L = 100/\epsilon \tag{6.31}$$

The accompanying figure shows all three of these limits:

6.3.1 More accurate answers

The scaling estimates we have developed work rather well. Lets now make these scaling estimates into more accurate predictions for the value of the integral. Here we carry this out for the case where ϵ is small. In this case, the integrand varies over a scale $\sqrt{\epsilon}$. We

are thus interested in studying the integral over this scale. To do this we introduce the scaled variable

$$y = \frac{x}{\sqrt{\epsilon}}, \tag{6.32}$$

so that our integral becomes

$$I(\epsilon) = \frac{1}{\sqrt{\epsilon}} \int_0^{100/\sqrt{\epsilon}} \frac{1}{1 + y^2 + \epsilon^{3/2}y^5} dy \tag{6.33}$$

As long as the integral

$$J = \int_0^{100/\sqrt{\epsilon}} \frac{1}{1 + y^2 + \epsilon^{3/2}y^5} dy \tag{6.34}$$

is independent of ϵ the scaling prediction that $I \sim 1/\sqrt{\epsilon}$ will be correct. We can rewrite J as

$$J = \int_0^\infty \frac{1}{1 + y^2 + \epsilon^{3/2}y^5} dy - \int_{100/\sqrt{\epsilon}}^\infty \frac{1}{1 + y^2 + \epsilon^{3/2}y^5} dy. \tag{6.35}$$

The second integral is easily estimated as $\sqrt{\epsilon}/10^8/4 + O(\epsilon)^2$. This follows because the smallest value of the denominator occurs when $y = 100/\sqrt{\epsilon}$ and there the bigger term is the $\epsilon^{3/2}y^5$ term in the denominator. Hence

$$\int_{100/\sqrt{\epsilon}}^\infty \frac{1}{1 + y^2 + \epsilon^{3/2}y^5} dy \approx \int_{100/\sqrt{\epsilon}}^\infty \frac{1}{\epsilon^{3/2}y^5} dy = \frac{1}{\epsilon^{3/2}} y^{-4} \Big|_{100/\sqrt{\epsilon}}^\infty = \frac{\sqrt{\epsilon}}{4 \cdot 10^8}. \tag{6.36}$$

Thus, we are left with the first integral: we expand its denominator as

$$\frac{1}{1 + y^2 + \epsilon^{3/2}y^5} = \frac{1}{1 + y^2} - \frac{\epsilon^{3/2}y^5}{(1 + y^2)^2} + \dots \tag{6.37}$$

This expansion of the denominator is legitimate as long as $\epsilon^{3/2}y^5/(1 + y^2) < 1$. This is true when $y < 1/\sqrt{\epsilon}$. Thus, when expanding the denominator we must write

$$J = \int_0^{1/\sqrt{\epsilon}} \frac{1}{1 + y^2} \left(1 - \frac{y^5 \epsilon^{3/2}}{1 + y^2} + \dots \right) dy + \int_{1/\sqrt{\epsilon}}^\infty \frac{1}{1 + \epsilon^{3/2}y^5 + y^2} dy = J_1 + J_2. \tag{6.38}$$

The first term of the integral J_1 can be approximated

$$\int_0^{1/\sqrt{\epsilon}} \frac{dy}{1 + y^2} = \int_0^\infty \frac{dy}{1 + y^2} - \int_{1/\sqrt{\epsilon}}^\infty \frac{dy}{1 + y^2}. \tag{6.39}$$

The first integral on the right hand side is $\pi/2$. The second integral can be evaluated by noting that since y is large

$$\int_{1/\sqrt{\epsilon}}^\infty \frac{dy}{1 + y^2} = \int_{1/\sqrt{\epsilon}}^\infty \frac{dy}{y^2} + O(y^{-4}) = \sqrt{\epsilon} + O(\epsilon^{3/2}). \tag{6.40}$$

To finish off J_1 at $O(\sqrt{\epsilon})$ we also need to consider the higher order terms in the series expansion of the denominator: Let's try to get an approximation that works to this order. Note that

$$\int_0^{1/\sqrt{\epsilon}} \frac{1}{1+y^2} \left(\frac{y^5 \epsilon^{3/2}}{1+y^2} - \left(\frac{y^5 \epsilon^{3/2}}{1+y^2} \right)^2 + \dots \right) dy \quad (6.41)$$

is dominated at large values of y , since when $y < 1$, the integral is $O(\epsilon^{3/2})$. At large y , we can replace $1+y^2 \rightarrow y^2$, and only the upper limit of integration matters. Then, the expansion becomes

$$\int^{1/\sqrt{\epsilon}} (\epsilon^{3/2} y - \epsilon^3 y^4 + \epsilon^{9/2} y^7 + \dots) = \sqrt{\epsilon} \left(\frac{1}{2} - \frac{1}{5} + \frac{1}{8} + \dots \right). \quad (6.42)$$

Thus, putting this together we have that $J_1 = \pi/2 - \sqrt{\epsilon}(1 + \bar{c}) + O(\epsilon^{3/2})$, where $\bar{c} = \sum_{n=0}^{\infty} (2+3n)^{-1} (-1)^n$.

The second integral J_2 follows from approximating

$$J_2 = \int_{1/\sqrt{\epsilon}}^{\infty} \frac{1}{1 + \epsilon^{3/2} y^5 + y^2} = \int_{1/\sqrt{\epsilon}}^{\infty} \frac{1}{\epsilon^{3/2} y^5} - \int_{1/\sqrt{\epsilon}}^{\infty} \frac{1}{\epsilon^3 y^8} + \dots \quad (6.43)$$

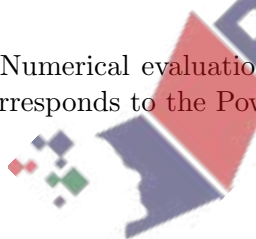
This series becomes

$$J_2 = \sqrt{\epsilon} \left(\frac{1}{4} - \frac{1}{7} + \frac{1}{10} + \dots \right) = \bar{\bar{c}} \sqrt{\epsilon} \quad (6.44)$$

Putting this all together implies that the integral (at small ϵ) is given by

$$I = \frac{\pi}{2\sqrt{\epsilon}} - 1 - \bar{c} + \bar{\bar{c}} + O(\epsilon). \quad (6.45)$$

Numerical evaluation of the sums (can be done directly, or by noting that the sum corresponds to the Power series of



$$\int_0^1 x(1+x^3)^{-1}, \quad (6.46)$$

and evaluating this integral numerically) gives $\bar{c} \approx 0.375$, $\bar{\bar{c}} = 0.163$ so that $I \sim \frac{\pi}{2\sqrt{\epsilon}} - 1.21$.

Let's compare this prediction to the numerical results: Figure 6.4 plots $I - \pi/2/\sqrt{\epsilon}$ as a function of ϵ . As predicted the error saturates to this constant at small ϵ .

Note that this expansion procedure could be applied to the other regimes of interest in the original integral (ie the $\epsilon^{-4/5}$ regime and the ϵ^{-1} regime) as well.

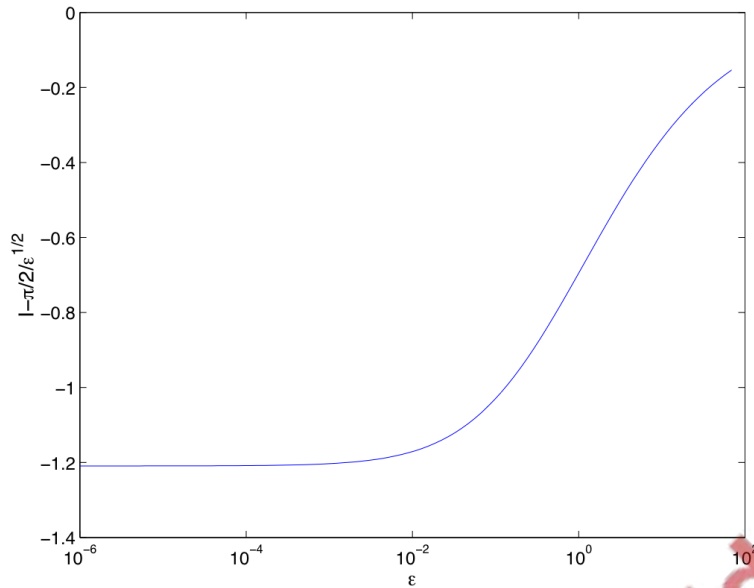


Figure 6.4. Error in the leading order term for the integral, as a function of ϵ

6.3.2 Practical Implementation in MATLAB

The MATLAB routine used to produce the plots shown in this section are given here:

We have a master driver program that we use to set the initial conditions and the range of integration [xinit,xfinal]. It is as follows (You can call it driver.m and hence run directly from the Matlab prompt.)

```

1 global eps
2 eps=1e-5;
3
4 for i=1:70
5     int(i)=quad(@myintegral,0,100);
6     epsilon(i)=eps;
7     eps=eps*2;
8 end

```

You can call this file something like driver.m, and then call it by saving it in the directory from which you are running MATLAB and then typing *driver*. We start with a value of $\epsilon = 10^{-4}$ and during each step during the for loop we carry out the integral using the matlab command *quad* and then multiply ϵ by 2. This makes the spacing of the ϵ 's linear on a logarithmic scale. We store the values of the integral in the array *int* and the values of ϵ in the array *epsilon*.

You can then make the plot by typing `loglog(epsilon,int, '.')`; For the integral we considered, the myintegral.m file is as follows

```

1 function f=myintegral(x)
2 global eps
3 f=1./(eps+x.^2+x.^5);

```

Note that we have passed the value of epsilon to the function using the *global* command. Note it was also used in the driver file. This makes the variable *eps* visible to every program that is called, whether or not it is passed directly. Not elegant, but it works.

6.4 Stirling's Formula

Our next example will be a derivation of Stirling's formula, one of the greatest formulas ever derived. This was invented by James Stirling, an English mathematician who lived from 1692-1770. He discovered that for large N , one can approximate

$$N! \approx \sqrt{2\pi N} N^N e^{-N}. \quad (6.47)$$

Here we present the derivation of this formula

6.4.1 An Estimate

Note that

$$\log(n!) = \sum_{j=1}^n \log(j). \quad (6.48)$$

Now, a simple way of estimating the size of a sum is to note that heuristically

$$\sum_j f(j) = \sum_j \Delta j f(j) \sim \int dj f(j). \quad (6.49)$$

Therefore we would anticipate that

$$\log(n!) \approx \int_{j=1}^n dj \log(j) = n \log(n) - n, \quad (6.50)$$

which implies that

$$n! \approx n^n e^{-n}. \quad (6.51)$$

This result is of course essentially correct; however, with more work one can derive a better formula.

One way of deriving a better version of this formula is to ask what is the error in approximating the sum by an integral. The essence of the approximation that we have written is to use the fact that in every subinterval

$$\int_j^{j+1} dx \log(x) = (j+1)\log(j+1) - (j+1) - j\log(j) + j, \quad (6.52)$$

An exaggeration? Tell me a better one!

and then approximate

$$\int_j^{j+1} dx \log(x) \approx \log(j). \quad (6.53)$$

Clearly although this formula is roughly correct at large j it is quantitatively in error.

To do better than this, we need to more carefully consider the error in the integration formula. Although following this road would lead us into the (important) theory of numerical integration, we will instead proceed with another approach.

6.4.2 The Derivation

The derivation of $n!$ for large n starts from an integral representation. Integration by parts demonstrates that

$$n! = \int_0^\infty t^n \exp\{-t\} dt. \quad (6.54)$$

This formula for $n!$ is generalized into the definition of the Γ function, which is defined as follows:

$$\Gamma(z + 1) = \int_0^\infty t^z \exp\{-t\} dt. \quad (6.55)$$

We would like to develop a method for evaluating the factorial function (or the Gamma function) when $n(z)$ is large.

6.5 Laplace's Method

Before solving this problem, we start with a simpler and more general question. Suppose we are given an integral of the form $F(x) = \int_a^b g(t) \exp\{-\lambda f(t)\} dt$ and are asked to evaluate the integral when λ is very large. Let us assume that the function $f(t)$ is positive and has an absolute minimum at some point t_0 in the **interior** of the interval $[a, b]$, and let's start by assuming that the minimum does not occur at the boundary of the interval. The idea of Laplace's method is that the value of the integral is determined almost entirely near this point t_0 when $\lambda \rightarrow \infty$. The reason for this is best seen in a plot of a specific example. Consider the integral

$$\int_5^{10} \frac{1 + \sin(t)}{t^5 + 7x + 10e^t} \exp\{-\lambda(t - 8)^2\} dt. \quad (6.56)$$

Figure 6.5 shows the integrand as a function of x ; it is seen that as $\lambda \rightarrow \infty$ the integral becomes more and more localized near $t = 8$.

We can therefore approximate this integral by just estimating the area near the point t_0 . Following our prescription from before, the Height = $g(t_0) \exp(-\lambda f(t_0))$ (i.e. the

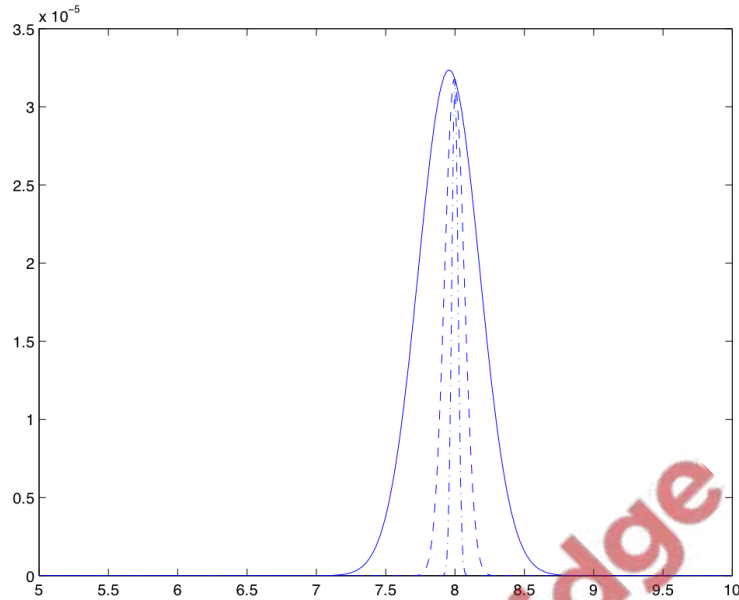


Figure 6.5. Plot of the integrand $g(x) = \frac{1+\sin(x)}{x^5+7x+10e^x} \exp -\lambda(x-8)^2$ for $\lambda = 10, 100, 1000$ (solid, dashed, dot-dashed). Note that as $\lambda \rightarrow \infty$, the integral becomes more localized near $x = 8$

integral evaluated at the maximum point t_0 . To find the width we want to determine the distance from t_0 over which the integrand decreases substantially. This decrease will be dominated by the exponential—the increase in $f(t)$ away from $t = t_0$ means $e^{-\lambda f}$ will decrease rapidly as λ increases.

To determine the width, let us find the t so that

$$e^{-\lambda f(t)} = \frac{e^{-\lambda f(t_0)}}{e}. \quad (6.57)$$

To find the width, we expand $f(t)$ in a Taylor series near $t = t_0$,

$$f(t) = f(t_0) + (t - t_0)^2 f''(t_0)/2 + \dots \quad (6.58)$$

We therefore have

$$e^{-\lambda(f(t_0)+(t-t_0)^2 f''(t_0)/2)} = \frac{e^{-\lambda f(t_0)}}{e}, \quad (6.59)$$

or

$$\text{width} = \sqrt{\frac{2}{f''(t_0)\lambda}}. \quad (6.60)$$

then the integral is approximately

$$\text{height} \times \text{width} = \left[g(t_0) \exp\{-\lambda f(t_0)\} \right] \times \left[\sqrt{\frac{2}{f''(t_0)\lambda}} \right]. \quad (6.61)$$

6.5.1 Proceeding more carefully

One can do this a bit more carefully as follows. We expand both $f(t), g(t)$ around $t = t_0$, and insert these into the integral. This gives

$$\begin{aligned} I &= \int_a^b g(t) \exp\{-\lambda f(t)\} dt \\ &= \int_a^b \left(g(t_0) + (t - t_0)g'(t_0) + (t - t_0)^2/2g''(t_0) + \dots \right) \\ &\quad \times \exp\left\{-\lambda \left(f(t_0) + f''(t_0)/2(t - t_0)^2 + f'''(t_0)(t - t_0)^3/6 + \dots \right)\right\} dt \end{aligned} \quad (6.62)$$

Now this is just

$$\begin{aligned} &\exp\{-\lambda f(t_0)\} \int_a^b \exp\{-\lambda f''(t_0)(t - t_0)^2/2\} \\ &\quad (1 - \lambda f'''(t_0)(t - t_0)^3/6 + \dots)(g(t_0) + (t - t_0)g'(t_0) + \dots) dt. \end{aligned} \quad (6.63)$$

If we change variables so that $\frac{\lambda f''(t_0)}{2}(t - t_0)^2 = u^2$ then the integral becomes

$$\exp\{-\lambda f(t_0)\} \int_{-(a-t_0)^2\lambda f''/2}^{(b-t_0)^2\lambda f''/2} \sqrt{\frac{2}{f''(t_0)\lambda}} \exp\{-u^2\} (g(t_0) + \text{corrections}) du. \quad (6.64)$$

Now if we let $\lambda \rightarrow \infty$, the limits of the integration are well approximated by $\pm\infty$. Hence the integral becomes

$$g(t_0) \exp\{-\lambda f(t_0)\} \sqrt{\frac{2}{f''(t_0)\lambda}} \int_{-\infty}^{\infty} e^{-u^2} du, \quad (6.65)$$

or

$$\sqrt{2\pi} g(t_0) \exp\{-\lambda f(t_0)\} \sqrt{\frac{2}{f''(t_0)\lambda}}, \quad (6.66)$$

in the limit $\lambda \rightarrow \infty$. There is an interesting question about higher order corrections that we will turn to later on. (It turns out that the formula above is the first term in a divergent series.)

$f(t)$ does not have a global minimum on $a \leq x \leq b$

The above analysis works as long as $f(t)$ has a global minimum on the interval $[a, b]$. Often, the minimum value of the integral occurs not in the middle of the region, but on one of the end points. In this case the integral is dominated by the behavior near the endpoints. Lets assume that the minimum occurs at $t_0 = a$. Then we can expand $f(t) = f(a) + f'(a)(t - a) + \dots$. It seems like we have the maximum value of the integral in mind (rather than the minimum as stated above), which occurs at the minimum of $f(t)$.

Since a is a minimum it is necessarily the case that $f'(a) > 0$. If we now expand the integrand near $t = a$ we obtain

$$I = \int_a^b \left(g(a) + (t - a)g'(a) + \dots \right) \times \exp\{-\lambda(f(a) + f'(a)(t - a) + \dots)\} dt, \quad (6.67)$$

or changing variables so that $\lambda f'(a)(t - a) = s$ we have that

$$I \approx g(a)e^{-\lambda f(a)} \int_0^{(b-a)f'(a)\lambda} e^{-s} \frac{ds}{f'(a)\lambda}, \quad (6.68)$$

so that

$$I \approx \frac{g(a)e^{-\lambda f(a)}}{f'(a)\lambda}. \quad (6.69)$$

We see that the major difference between the two cases (the minimum of $f(t)$ is in the interior versus the endpoints of the interval) is the decay with λ : in the former case the decays is like $\lambda^{-1/2}$ whereas in the latter it is λ^{-1} .

6.5.2 Moving on to $N!$

With these ideas in mind we can now develop an approximate formula for the integral representation of $n!$. As above we have that

$$n! = \int_0^\infty t^n \exp\{-t\} = \int_0^\infty \exp\{n \log(t) - t\}. \quad (6.70)$$

Actually it is pretty close to the form of the second case, where the minimum is on the endpoint. But there is no large parameter multiplying the decaying exponential!

This is not exactly in the form we have manipulated above. One can readily show that $f(t) = n \log(t) - t$ has its minimum when $t = n$. Hence as $n \rightarrow \infty$ the location of the minimum shifts towards ∞ . Additionally, the integral is broad, and not particularly localized.

We get around this in two steps: first, write

$$t = n\tau. \quad (6.71)$$

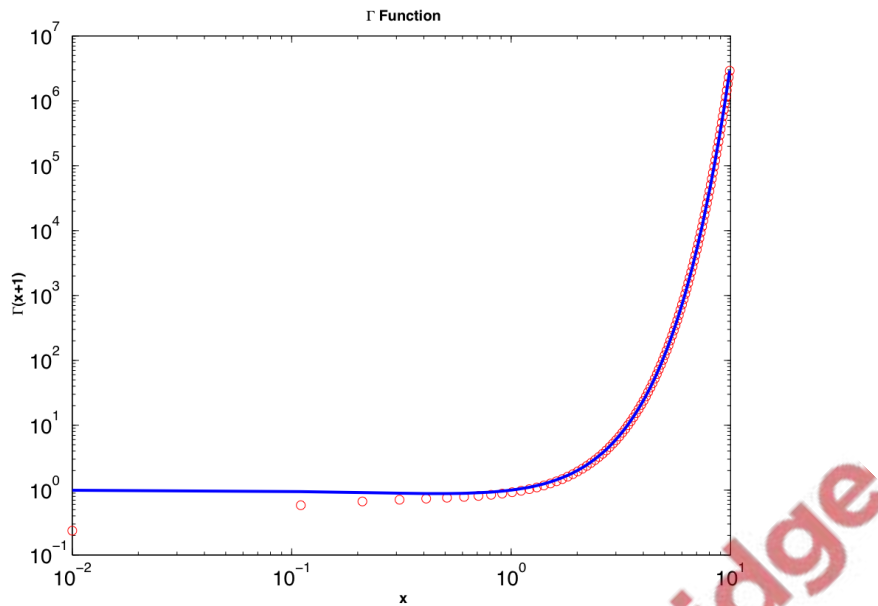


Figure 6.6. Plot comparing the Γ function to the asymptotic formula we have derived.

Then

$$f = n \log(t) - t = n \log(n) + n(\log(\tau) - \tau). \quad (6.72)$$

The function $\tau - \log(\tau)$ has a peak at $\tau = 1$, so the n dependence of the location of the maximum is now gone. We can therefore apply Laplace's method to the integrand, and therefore write the expansion

$$\tau - \log(\tau) = 1 + \frac{1}{2}(\tau - 1)^2 + \dots \quad (6.73)$$

Substituting this expansion into the integral gives

$$\begin{aligned} \int_0^\infty \exp\{n \log(n) - n(1 + (\tau - 1)^2/2 + \dots)\} &\approx n^{n+1} \exp -n \int_0^\infty \exp\{-n(\tau - 1)^2/2 \\ &= \sqrt{2\pi n} n^n \exp\{-n\} \end{aligned} \quad (6.74)$$

Figure 6.6 compares the Γ function to the asymptotic formula. As usual it works much beyond where you might expect.

6.6 The Error Function

Let us now consider one further integral. Although the spirit of our derivation will be quite similar to that which has come before, we will use this example to pay closer attention to the accuracy of our formula.

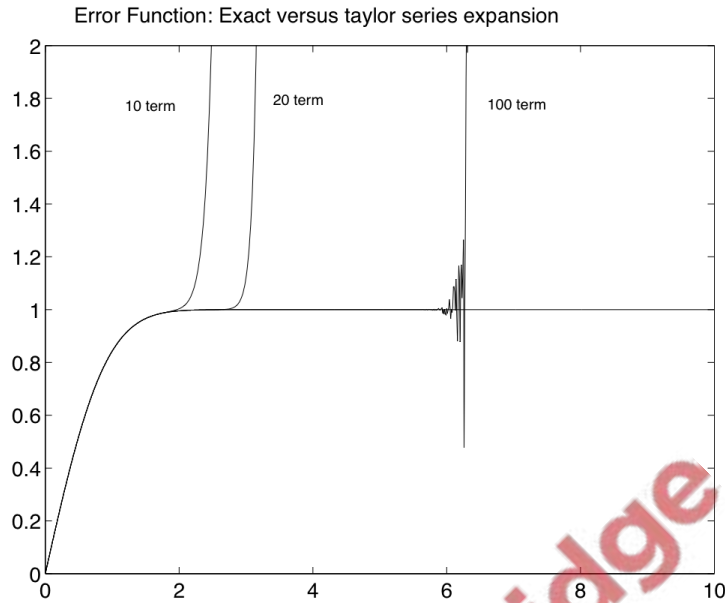


Figure 6.7. Comparison of the 10, 20 and 100 term Taylor series expansion for the error function with the exact result.

Consider the "error function"

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp(-t^2) dt. \quad (6.75)$$

We seek an estimate for this function. Now, since the error function is expressed as the integral over an exponential, which has a Taylor series about $z = 0$ with an infinite radius of convergence, there is a Taylor series expansion of the error function with an infinite radius of convergence. Namely

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{(2n+1)n!} \quad (6.76)$$

Let's try it out. Figure 6.7 compares a 10, 20 and 100 term Taylor series expansion of the error function with the exact result:

Although increasing the number of terms in the expansion does increase the range where the Taylor series gives reasonable answers, the convergence is slow—the last summation even demonstrates that it may not work at all: the round off error incurred from adding and subtracting large negative numbers eventually becomes prohibitive.

There is however another way to estimate the integral, which is deceptively subtle. We know of course that since

$$\int_0^{\infty} \exp(-t^2) dt = \frac{\sqrt{\pi}}{2}, \quad (6.77)$$

$\operatorname{erf}(\infty) = 1$

We therefore write

$$\operatorname{erf}(z) = 1 - \frac{2}{\sqrt{\pi}} \int_z^\infty \exp(-t^2) dt, \quad (6.78)$$

and try to develop a good approximation for the second integral. We can do this by integrating by parts. Note that

$$\int_z^\infty \exp(-t^2) dt = \int_z^\infty \frac{d(e^{-t^2})}{-2t} = -\frac{\exp(-t^2)}{2t} \Big|_z^\infty + \int_z^\infty \frac{\exp(-t^2)}{2t^2} dt. \quad (6.79)$$

Continuing in this way after integrating by parts several times we obtain

$$\operatorname{erf}(z) = 1 - \frac{2}{\sqrt{\pi}} \frac{\exp(-z^2)}{2z} \left(1 - \frac{1}{(2z^2)} + \frac{1 \cdot 3}{(2z^2)^2} - \frac{1 \cdot 3 \cdot 5}{(2z^2)^3} \right) + R, \quad (6.80)$$

where

$$R = 1 \cdot 3 \cdot 5 \cdot 7 \int_z^\infty \exp(-t^2) / (16t^8) dt \quad (6.81)$$

So far everything we have written is exact. What would clearly be good, is if we could delete the remainder integral R , and use the series to evaluate the function. Now as $z \rightarrow \infty$, R decreases more rapidly than the terms in the series.

In particular we know that

$$\begin{aligned} R &= 1 \cdot 3 \cdot 5 \cdot 7 \int_z^\infty dt \frac{\exp(-t^2)}{(16t^8)} = \frac{105}{32} \int_z^\infty \frac{d(e^{-t^2})}{t^9} \\ &\leq \frac{105}{32z^9} \int_z^\infty d(e^{-t^2}) = \frac{105e^{-z^2}}{32z^9}. \end{aligned} \quad (6.82)$$

Here the inequality is caused by the fact that $t^{-9} < z^{-9}$ over the integration range of the integral (where $t > z$).

The consequence of this is that the remainder term is the smallest term in the series as $z \rightarrow \infty$. Thus we expect this to be a good approximation when z gets very large. (In fact, this same argument implies that the first term itself in the series will also be a very good approximation as z gets large.)

On the other hand, let's consider the radius of convergence of the series multiplying the e^{-z^2} : Following the pattern laid out in the equation, the n^{th} term in the series is

$$\frac{1 \cdot 3 \cdot \dots \cdot (2n+1)}{(2z^2)^{n+1}}. \quad (6.83)$$

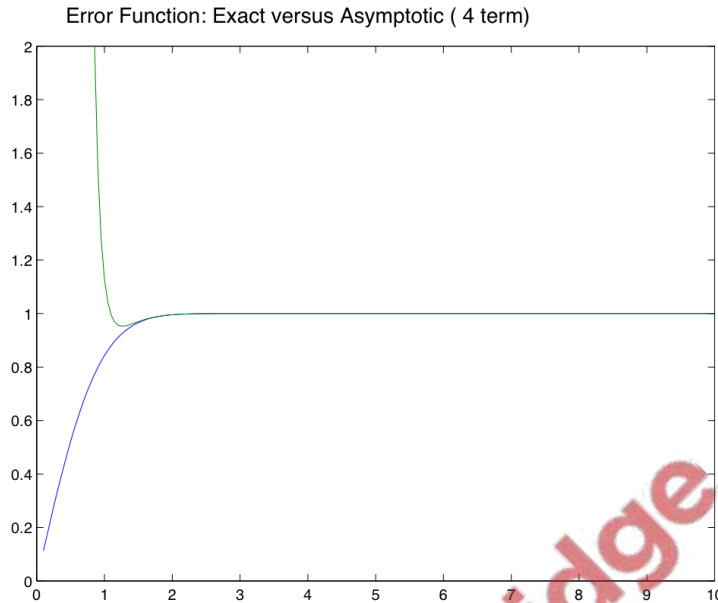


Figure 6.8. Comparison of the 4 term asymptotic result with the exact result.

Hence taking the ratio of the n^{th} and the $n + 1^{\text{st}}$ terms, we have that

$$\text{Ratio} = \frac{2z^2}{2n + 3}. \tag{6.84}$$

Therefore, if we fix z and send $n \rightarrow \infty$, the radius of convergence of the series vanishes! The series is divergent.

All of this is bad news. But let's have courage, and see if the series provides an accurate representation of the function (6.6). Figure 6.8 compares the four term expansion with the exact formula. It agrees beautifully!

Thus, we have discovered an example of a series with the following properties:

- As $n \rightarrow \infty$, the series does not converge.
- However, at fixed n with $z \rightarrow \infty$ the formula is very accurate.

This type of situation happens a lot, and we will come to understand later on that this generically happens when one is doing an expansion about an essential singularity (here we are effectively doing an expansion of e^{-z^2} about $z = \infty$.) Expansions of this type are called "asymptotic expansion. The hallmark of such an expansion is that the remainder is smaller than the last term kept.

6.6.1 Having Courage, once again

Let us reconsider this problem, this time having courage.

Consider the most interesting part of the error function, namely

$$\int_x^\infty e^{-s^2} ds. \quad (6.85)$$

We would like to estimate this integral by using our simple prescription, namely that the integral is the Height multiplied by the width.

What is the height? Clearly the maximum value of the integrand occurs at the lower endpoint of the integrand, namely

$$\text{Height} = e^{-x^2}. \quad (6.86)$$

What about the width? For this we use our simple prescription: the width is the distance from the maximum such that the integrand decreases by e^{-1} . Namely

$$e^{-(x+W)^2} = e^{-1}e^{-x^2} \quad (6.87)$$

or

$$(x + W)^2 = x^2 + 1. \quad (6.88)$$

This implies that

$$W = \sqrt{x^2 + 1} - x. \quad (6.89)$$

The figure compares the height times the width calculated in this way with the asymptotic formulae we have derived above (!)

6.6.2 Optimal Truncations

Given the fact that the terms multiplying the e^{-z^2} are themselves divergent, it is worth wondering how many of the terms it is actually worth keeping. Let's check this out by evaluating the error function at $x = 2$ and then comparing it numerically to the series expansion, truncated at various places. First the error function $erf(2) = 0.99532226501895$. Figure 6.10 shows the series evaluated as a function of truncation: you see that at a truncation of order $n = 10$ the evaluated number starts diverging dramatically from the answer.

Let's examine the result of the first 10 truncations: we find that our series gives the numbers

0.99547909695379
 0.99523690571917
 0.99538827524081
 0.99525582690938

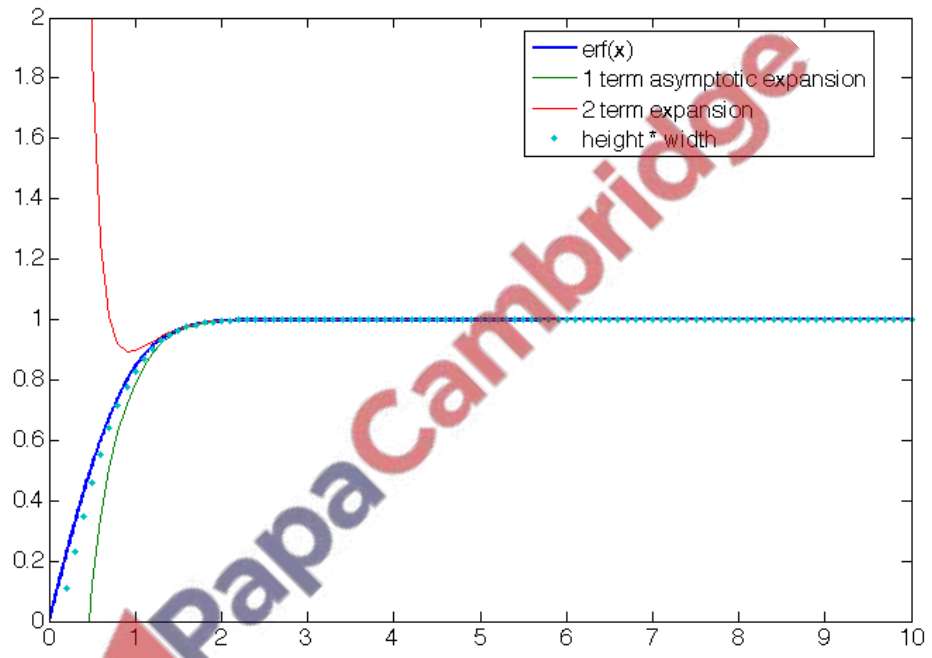


Figure 6.9. Comparison of having courage formula with asymptotic approximations.

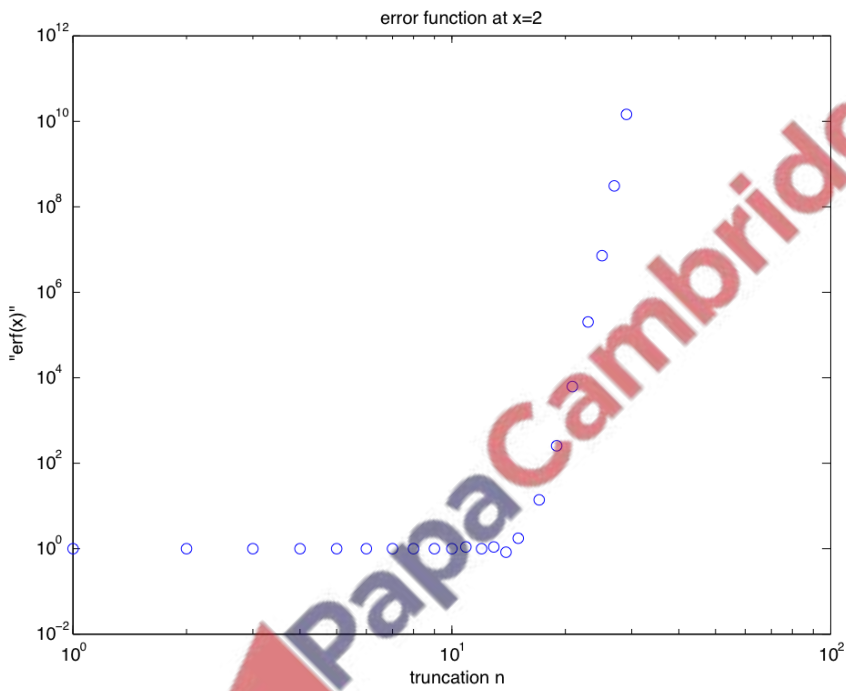


Figure 6.10. Asymptotic series for error function as a function of the truncation.

0.99540483128224
 0.99519995026956
 0.99553288191517
 0.99490863507965
 0.99623515960513
 0.99308466385711
 1.00135471519566
 0.97757831759733
 1.05187956009209
 0.80111286667229
 1.71014213031905

You can see that for the first few truncations, the answer improves a bit, and then it starts diverging. Figure 6.11 shows the dependence of the error in the formula as a function of the truncation. You can see that the best answer is at about $n = 4$

Could we have guessed that the best formula occurs at $n = 4$? If we go back to the argument concerning the radius of convergence, we see the ratio of the n^{th} term to the $n + 1^{\text{st}}$ term is

$$\text{Ratio} = \frac{2z^2}{2n+3} = \frac{8}{2n+3}. \quad (6.90)$$

At $n = 3$, this ratio becomes smaller than 1, indicating that the terms start increasing in magnitude. Hence we expect things to get worse around the 3rd term, and indeed they do. For a given z we therefore expect the optimal truncation to occur after around \sqrt{n} terms.

To illustrate this, figure 6.12 compares the asymptotic series with the exact formula for

$$(erf(x) - 1) \exp(x^2) \quad (6.91)$$

The index N labels the number of terms that were kept in the asymptotic expansion. It is seen that beyond a certain x dependent N , the series no longer converges to the exact result (given by the solid line).

Note that for $x = 2, 3, 5$ the error deviates at $n = 4, 15, 60$.

6.7 Another example of an asymptotic series

This is not the first series of this form we have encountered: you will recall the difficulty we had when we addressed the solution to the differential equation

$$\frac{dy}{dx} + y = \frac{1}{1+x^2}. \quad (6.92)$$

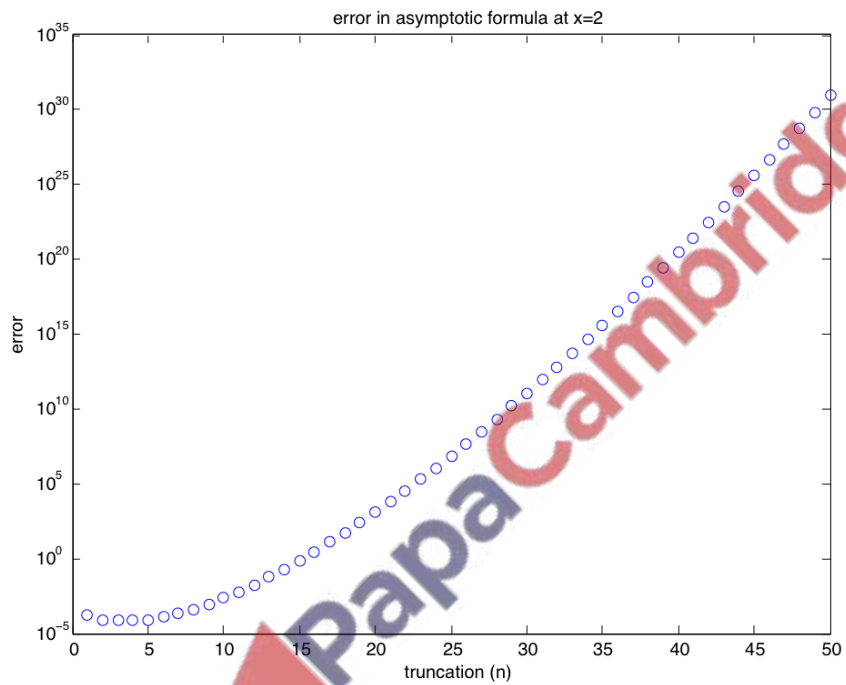


Figure 6.11. Error in the asymptotic formula as a function of truncation.

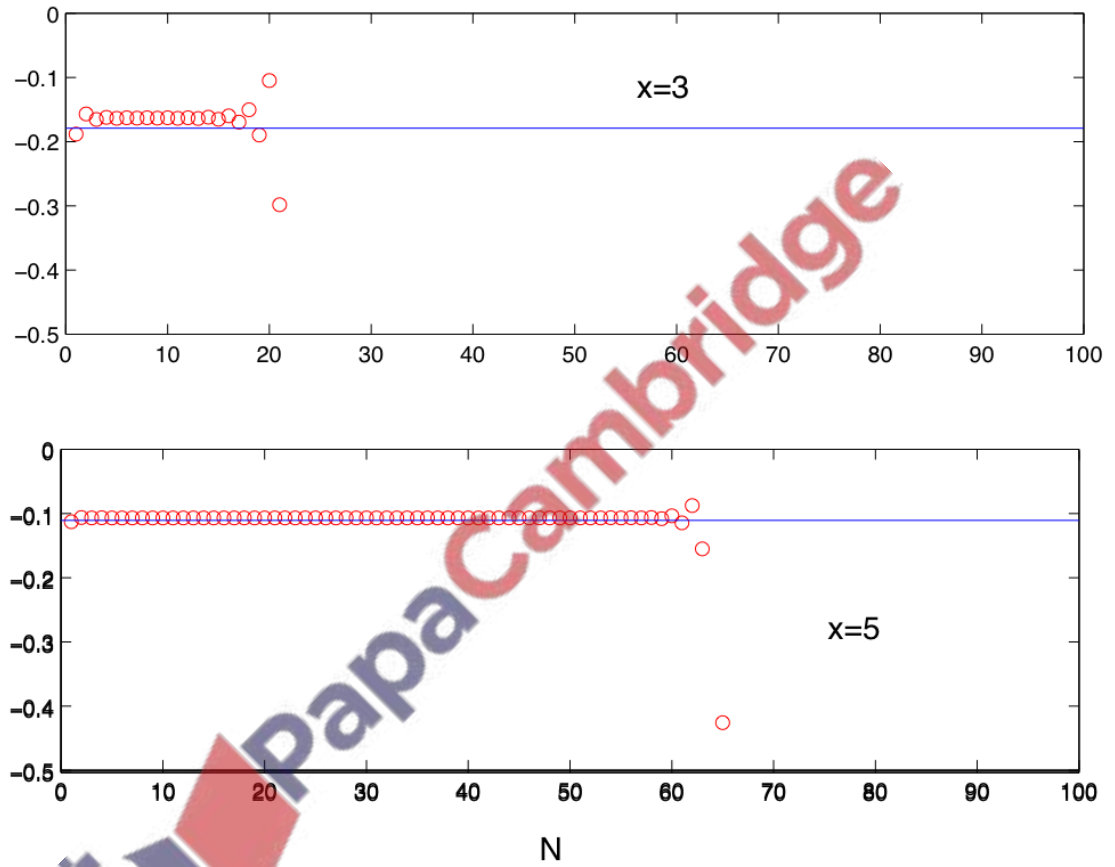


Figure 6.12. Plot of $(\operatorname{erf}(x)-1)*\exp(x)$ evaluated at $x = 3$ and $x = 5$, compared with different truncations of the asymptotic approximation to the error function. It is seen that beyond a certain x dependent on n , the series no longer converges.

We found a series solution to the equation that was valid when x is large, namely

$$y = \frac{1}{x^2} + \frac{2}{x^3} + \frac{5}{x^4} + \frac{20}{x^5}. \quad (6.93)$$

We saw by comparing this solution with numerical solutions to the differential equation that it worked very well in describing the solutions. However, we then studied the radius of convergence of the series and found that the radius of convergence vanished. Hence, by usual mathematical principle, there is no way this series should be useful.

With a little effort we can convince you that the phenomenon here is exactly the same as that we uncovered for the error function. To proceed, let's first write down the "exact" solution to our differential equation, using the fact that there is an integrating factor e^x . We therefore arrive at (as usual)

$$y(x) = Ce^{-x} + e^{-x} \int_1^x ds \frac{e^s}{1+s^2}. \quad (6.94)$$

Let us now study this when x is large. In particular we need to examine the integral $\int_1^x ds \frac{e^s}{1+s^2}$ when x is large.

The convenience of writing the formula this way is that now we can easily study the behavior of the integral at large x . To do this, we use

$$\frac{1}{1+x^2} = \frac{1}{x^2} \frac{1}{1+x^{-2}}. \quad (6.95)$$

When $x > 1$ this is just

$$\frac{1}{x^2} \frac{1}{1+x^{-2}} = \frac{1}{x^2} \left(1 - \frac{1}{x^2} + \dots \right) \quad (6.96)$$

Thus, the integral we need to evaluate can be written

$$\int_1^x ds \frac{e^s}{1+s^2} = \int_1^x \frac{e^s}{s^2} \left(1 - \frac{1}{s^2} + \dots \right) ds. \quad (6.97)$$

This is an infinite series, the largest term in the series is clearly the first term. We therefore now need to evaluate the integral

$$\int_1^x \frac{e^s}{s^2} ds \quad (6.98)$$

when x is large. It turns out that this type of integral comes up very frequently. It is a special case of the so-called "incomplete gamma function" (Note the resemblance to the Gamma function we described in class.)

6.7.1 Large x behavior

Our trouble is now to figure out how to evaluate this integral when x is large. To do this, we will integrate by parts: Note that

$$\int_1^x \frac{e^s}{s^2} ds = \int_1^x d(e^s) \frac{1}{s^2} = \int_1^x \left(d\left(\frac{e^s}{s^2}\right) + 2\frac{e^s}{s^3} ds \right) \quad (6.99)$$

$$= \frac{e^x}{x^2} - e + 2 \int_1^x \frac{e^s}{s^3} ds. \quad (6.100)$$

Similarly, integrating by parts again one can see that

$$\int_1^x \frac{e^s}{s^3} ds = \frac{e^x}{x^3} - e + 3 \int_1^x \frac{e^s}{s^4} ds. \quad (6.101)$$

Continuing this many times we find that

$$\int_1^x \frac{e^s}{s^2} ds = e^x \left(\frac{1}{x^2} + \frac{2}{x^3} + \frac{6}{x^4} + \cdots + \frac{(n-1)!}{x^n} \right) + n! \int_1^x \frac{e^s}{s^{n+1}} ds. \quad (6.102)$$

We now need to carry this procedure out for each of the terms in our expansion in equation (6.97). The second term in the expansion is

$$\int_1^x \frac{e^s}{s^4} ds, \quad (6.103)$$

Using the same type of procedure on this integral (integrating by parts) one can show

$$\int_1^x \frac{e^s}{s^4} ds = e^x \left(\frac{1}{x^4} + \frac{4}{x^5} + \frac{20}{x^6} + \cdots + \frac{(n-1)!}{6x^n} \right) + \frac{n!}{6} \int_1^x \frac{e^s}{s^{n+1}} ds. \quad (6.104)$$

If we now use combine these approximate formula in the solution to our differential equation, we find:

$$y(x) = \bar{C}e^{-x} + \left(\frac{1}{x^2} + \frac{2}{x^3} + \frac{5}{x^4} + \cdots + \frac{\bar{c}}{x^n} \right) + e^{-x} C_n \int_1^x \frac{e^s}{s^{n+1}} ds. \quad (6.105)$$

Here, C_n is a constant which is the sum of the prefactors from the expansions of all of the integrals in equation (6.97). The only important property of this integral for us is that it grows rapidly with n : One can show that

$$C_n = n!(1 + 1/6 + \dots) \quad (6.106)$$

Note that at this point, everything we are writing down is exact. Also note that we can now see what the approximation we made in writing down equation (1) corresponds to. This corresponds to neglecting both the first term in equation (6.105), and the last term. The nonconvergence of the series in equation (1) must therefore be connected somehow to this approximation.

6.7.2 Asymptotic Series

Note that equation (6.105) has the following remarkable property: Neglecting the exponentially small term

$$y(x) - \left(\frac{1}{x^2} + \frac{2}{x^3} + \frac{5}{x^4} + \cdots + \frac{\bar{c}}{x^n} \right) = e^{-x} C_n \int_1^x \frac{e^s}{s^{n+1}} ds \leq \frac{C_n}{x^{n+1}}. \quad (6.107)$$

Here n is one higher power than the last term in the series. Thus, the *error in the result is arbitrarily small if we take x large enough*. In fact, the error is smaller than the last term in the series!

However, the catch is that *if we consider an infinite number of terms in the series for fixed x* , then the series does not converge. Namely, as pointed out before, the series is nonconvergent. We can see this in the above formula because the error term is

$$C_n/x^{n+1}. \quad (6.108)$$

We have already argued that C_n grows very rapidly with n , roughly speaking $C_n \sim n! \sim n^n$ (the last equality follows from Stirling's formula). Thus, the error term is approximately

$$n^n/x^n = (n/x)^n. \quad (6.109)$$

When $n > x$ the error is not small anymore.

As we remarked above in our discussion of the error function, these are the hallmarks of a so-called "asymptotic series". Our normal definition of convergence asks whether the sum converges to a number for fixed x when the number of terms $n \rightarrow \infty$. Here, the series is sensible when the number of terms is fixed, and $x \rightarrow \infty$. The definition of an asymptotic series means that *it is useful in practice* (as we have shown above, the error is small!); but, one does not necessarily do better by including lots and lots of terms. In fact, in general, there is an optimal number of terms that one should choose in an asymptotic series to get the best answer.

6.8 Final Remarks: Looking ahead

Before leaving this topic, we close with a few remarks:

6.8.1 Complex valued functions and essential singularities

First, so far all of the asymptotic expansions we have discussed were for real valued functions. In the complex plane, things are a little tricky. Namely, let us consider what happens if we extend the error function to complex valued functions (This section will

be easier to read once we cover the essentials of functions in the complex plane, in the next week or so.) The analytic continuation of the error function is

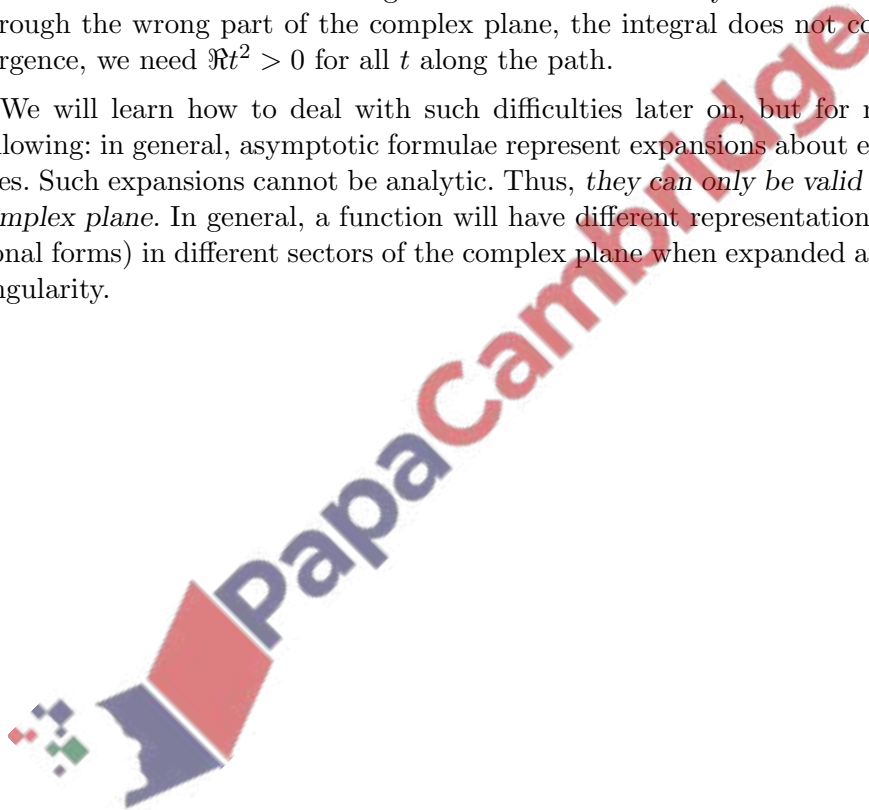
$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_C \exp\{-t^2\} dt, \quad (6.110)$$

where C is a contour in the complex plane that starts at the origin and ends at z . Owing to the analyticity of erf , the value of this integral is path independent. However, what happens if we try to apply our asymptotic trick from before? We have

$$\operatorname{erf}(z) = 1 - \frac{2}{\sqrt{\pi}} \int_{C'} \exp(-t^2) dt, \quad (6.111)$$

where C' is a contour connecting z and ∞ . The difficulty is that if this contour goes through the wrong part of the complex plane, the integral does not converge! For convergence, we need $\Re t^2 > 0$ for all t along the path.

We will learn how to deal with such difficulties later on, but for now we note the following: in general, asymptotic formulae represent expansions about essential singularities. Such expansions cannot be analytic. Thus, *they can only be valid in a sector of the complex plane*. In general, a function will have different representations (different functional forms) in different sectors of the complex plane when expanded about an essential singularity.



7 Convergence and Its Uses

7.1 What is Convergence? And is It Useful?

Thus far we have generated many different kinds of approximate formulas for solving our hard mathematics problems. We have tended to focus on only computing one or two terms in the series in the belief that we will get most important information just from this. But what if we were to study the entire series? Would this lead to a better or worse answer than we had before (especially if we include all of the terms)?

The answer to this question depends in large part on whether or not the series converges. For this reason, we now turn to a discussion of convergence—what it is, why it fails and how to think about it in general. This will lead us to spend a little time considering the properties of functions in the complex plane—this is the only context under which convergence can be understood.

Under what conditions do the series we have generated converge? What does the convergence (or divergence) of a series tell us about the properties of the underlying function it seeks to approximate?

Conservative Definitions

There are various competing definitions for determining the convergence of a series, each of which can be useful. First we review the most conservative definitions, which will be familiar to you from previous courses. The basic notion of convergence for a series $\sum_n a_n x^n$ is that the partial sums $A_N = \sum_{n=0}^N a_n x^n$ approach a limit as $N \rightarrow \infty$. Namely, given an $\epsilon > 0$ there is an M such that $|A_{n+p} - A_n| < \epsilon$ for all $p > 0$ and $n > M$. A more stringent notion of convergence is *absolute convergence*, which requires that the series converges when each term is replaced by its modulus. Absolute convergence implies convergence, as is evident from the triangle inequality $|\sum_n a_n x^n| \leq \sum_n |a_n x^n|$.

There are various tests that are typically applied to determine the convergence of a series. Clearly, the magnitude of the terms $a_n x^n$ in a series must decrease as $n \rightarrow \infty$. Demonstrating convergence or divergence typically requires comparing the series to another benchmark series whose convergence properties are known. The two most useful benchmarks are the *geometric series* $\sum_{n=0}^{\infty} r^n$, which converges to $(1 - r)^{-1}$ when $|r| < 1$, and the *harmonic series* $\sum_{n=1}^{\infty} 1/n^m$ which converges when $m > 1$. Comparing the general series to the geometric series, we find that the series converges absolutely when $|a_n x^n| \leq |r^n|$, or

$$\frac{|a_{n+1} x^{n+1}|}{|a_n x^n|} \leq |r| < 1. \quad (7.1)$$

It is a beautiful fact that this convergence criterion corresponds to points within a circle in the complex plane. Suppose that a given series converges at x_0 , then $|a_n x^n| = |a_n x_0^n| |x/x_0|^n$. Hence as long as $|x| \leq |x_0|$ the series converges. In contrast if the series diverges at x_1 then it diverges at every $|x| > |x_1|$. The consequence of this fact is that whenever something goes awry at any point in the complex plane, a series can only converge in a circular region up to this point. Conversely, when a series does not converge you can bet there is something funny going on in the complex plane! Below, we will classify the various types of singularities that can occur in the complex plane, so we will have a list of possible ways that a series can cease to converge.

Determining when a series no longer convergence often contains valuable information about the behavior of the function it seeks to approximate. For example, we will see that the perturbation series we constructed for the roots of the quintic stop converging at a singularity in the complex plane that corresponds to the change of two purely real roots into two complex roots. We will see that using such singularities intelligently will allow us to construct accurate approximations for the underlying functions.

Less Conservative Definitions

Whether or not a series converges, the coefficients contain information about the function the series seeks to approximate. Therefore it is legitimate to ask whether a divergent series contains information about the behavior of a function. One of the themes of this course will be that one can learn a great deal from divergent series. Here we give a definition that starts to hint at how this can be possible.

Our “conservative” definitions of convergence states that the partial sums $A_N(x) = \sum^N a_n x^n$ converge at a point x to a function $f(x)$ if $|A_N(x) - f(x)| \rightarrow 0$ as $N \rightarrow \infty$. In this definition the value of x is kept fixed as $N \rightarrow \infty$.

One can weaken this definition in the following way. Let’s keep N fixed, and instead send $x \rightarrow x_0$. We will say that a series is asymptotic to the function $f(x)$ at $x = x_0$ if for fixed N the partial sum $|A_N(x) - f(x)|/f(x) \rightarrow 0$ as $x \rightarrow x_0$. What this definition says is that the partial sum A_N is a good approximation to $f(x)$ (in that the numerical value of A_N is close to that of f) for x close to x_0 . Note also what this definition does not say. It does *not* say that the value of $A_N(x)$ becomes a better approximation to $f(x)$ as $N \rightarrow \infty$. Hence in principle (and as we have seen, in practice!) the agreement between the partial sums A_N and f gets worse and worse as N increases; nonetheless the A'_N s provide quite excellent approximations to the function.

At first sight this definition seems hard to believe; it implies that the notion of convergence is *not* the same as the notion that a formula provides a good approximation for the behavior of a function. At this point there is no reason you should be convinced that this definition is a valuable one; during this course we will provide numerous examples illustrating series which have precisely this property. We will also explore why and when such series work so well.

Some Examples

To whet your appetite for this subject, we will give some examples of asymptotic series.

Solution to $y' + y = (1 + x^2)^{-1}$ We saw that at large x $y(x) = 1/x^2$, but also saw that this solution was the first term in a divergent series.

Stirling's Formula We derived the first term in Stirling's formula above. One can also consider the entire series,

$$n! = \sqrt{2\pi n} \left(\frac{n}{e}\right)^n \left(1 + \frac{1}{12n} + \frac{1}{288n^2} - \frac{139}{51840n^3} + \dots\right). \quad (7.2)$$

Number of Prime Numbers A classic result of number theory is that the number of prime numbers less than n is given by

$$\pi(n) = \frac{n}{\log(n)} + \frac{n}{(\log(n))^2} + \frac{2!n}{\log(n)^3} + \frac{3!n}{(\log(n))^4} + \dots \quad (7.3)$$

This series clearly doesn't converge under a typical definition.

Harmonic Series The sum $H_M = \sum_{n=1}^M \frac{1}{n}$ is well known to diverge as $M \rightarrow \infty$. What is the approximate value at finite M ? The asymptotic answer is

$$H_M = \log(M) + \gamma + \frac{1}{2M} - \frac{1}{12M^2} + \frac{1}{120M^4} + \dots \quad (7.4)$$

Here γ is a constant of order unity, called "the Euler Mascheroni constant".

The exponential A normal power series expansion is also by default an asymptotic expansion. Thus, the exponential function has the series expansion

$$e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \dots \quad (7.5)$$

which is also an asymptotic expansion as $z \rightarrow 0$.

7.2 Singularities in the Complex Plane

The convergence of series approximations to a function $f(x)$ is entirely dictated by the singularities of the function in the complex plane. Here we review the elements of complex analysis that are necessary to understand and classify these singularities.

First we start with a simplistic example. Consider the power series

$$1 + x^2 + x^4 + x^6 + \dots \quad (7.6)$$

Application of the ratio test demonstrates that this series converges if and only if $|x| < 1$. One can understand why this divergence occurs by noting that the power series is the

Taylor expansion of $(1 - x^2)^{-1}$ around $x = 0$. Since the function diverges at $x = 1$ it is not at all alarming that the series it approximates diverges. On the other hand the function

$$\frac{1}{1 + x^2} = 1 - x^2 + x^4 - x^6 + \dots \quad (7.7)$$

also converges only when $|x| < 1$. This function does not diverge at $x = 1$, and thus this divergence does seem alarming. It was the consideration of this type of question that led to the subject of complex analysis (by Cauchy), and the notion of radius of convergence.

The consequence of this consideration is that even if you are interested in properties of real-valued functions, understanding the continuation of these functions to the complex plane is crucial—both for understanding why functions behave as they do, and in learning how to calculate them efficiently. In particular, the lack of convergence of a formula is always linked to underlying singularities of the formula in the complex plane.

7.2.1 Some Important Facts about Functions in the Complex Plane

Our goal now is to invent a complete characterization of the types of singularities in the complex plane. We would like to know: what types of singularities lead to a finite radius of convergence? What types lead to a zero radius of convergence? How can we understand the behavior of a function around a point of zero radius of convergence? In order to do this, we will need to introduce the basic elements of complex analysis. This information is described in standard text books. We will consider the following points.

- Multivaluedness of Complex Functions: Branch points and branch cuts.
- Differentiation: Cauchy-Riemann equations.
- Integration: Cauchy's Theorem, Cauchy's integral formula and principle values.
- Proof that interesting analytic functions always have singularities.
- Taylor series and Laurent series.
- Classification of singularities in the complex plane.

7.2.2 Multi-valuedness

There is another simple way in which a function in the complex plane can develop a singularity: namely, functions in the complex plane can be multivalued. In particular, if one has a function $f(z)$, it can happen that by making a loop around a particular value of $z = z_0$, the value of the function can change. Such a property is never shared by a series expansion of the form $\sum_n a_n z^n$, and for this reason, any series expansion that attempts to represent a multivalued function must diverge. This is best illustrated by an example - consider

$$f(z) = \sqrt{z}. \quad (7.8)$$

If we let $z = re^{i\theta}$, then by changing $\theta \rightarrow \theta + 2\pi$, $\sqrt{z} \rightarrow -\sqrt{z}$.

This behavior is summarized by saying that the function has two *branches*.

The typical fix for such a behavior is a “branch cut”, in which one draws a line in the complex plane through which one does not allow paths to cross. With this correction, the function is now single valued.

A particularly relevant example of a multivalued function is

$$f(z) = (1 + z)^\nu. \quad (7.9)$$

If ν is not an integer, this function has a branch point (is multivalued) at $z = -1$. To see this simply write $z = -1 + re^{i\theta}$. Sending $\theta \rightarrow \theta + 2\pi$ leads to a picking up of the phase factor $e^{i2\pi\nu}$. Now, alternatively you know that the Taylor series expansion of this function is

$$f(z) = 1 + \nu z + \frac{\nu(\nu - 1)}{2!} z^2 + \frac{\nu(\nu - 1)(\nu - 2)}{3!} z^3 + \dots \quad (7.10)$$

The radius of convergence can therefore be computed by considering as usual the ratio between the n^{th} and $(n + 1)^{\text{th}}$ terms. Since the coefficient of the n^{th} term (neglect the constant term) is

$$a_n = \frac{\nu(\nu - 1) \dots (\nu - n + 1)}{n!}, \quad (7.11)$$

we have that

$$\lim_{n \rightarrow \infty} \frac{|a_n|}{|a_{n+1}|} = \lim_{n \rightarrow \infty} \frac{n + 1}{n - \nu} = 1. \quad (7.12)$$

This also establishes the radius convergence as unity.

Now let us mention an interesting fact that will be useful to us later on: Note that

$$\frac{n + 1}{n - \nu} = \frac{1 + 1/n}{1 - \nu/n} \approx 1 + \frac{1 + \nu}{n}. \quad (7.13)$$

Hence if we look at the ratios $|a_n|/|a_{n+1}|$ the series actually contains information (namely ν) about the type of singularity that we are running into! Soon we will use this as a way to find the singularity we are running into and try to get rid of it.

7.2.3 Differentiation and Integration

Now we summarize the brief highlights of differentiation and integration in the complex plane. More details and beautiful pictures can be found in complex analysis textbooks on reserve in the library.

Complex Differentiation

First a definition: A function $f(z)$ with $z = x + iy$ is said to be *analytic* at the point z_0 if the derivative of $f(z)$ exists at z_0 . This is a nontrivial definition because since f depends on two variables (x and y) which are in principle independent there is a necessary constraint that must be satisfied for the definition to hold. Namely if we differentiate $f(z)$ in the 'x direction' and the 'y direction' we must get the same answer. Specifically

$$\frac{f(x + \Delta x + iy) - f(x + iy)}{\Delta x} = \frac{f(x + iy + i\Delta y) - f(x + iy)}{i\Delta y}. \quad (7.14)$$

If we write $f(x + iy) = u(x, y) + iv(x, y)$ this constraint is equivalent to constraints on the function u and v , known as the Cauchy Riemann equations: namely

$$\partial_x u = \partial_y v \quad (7.15)$$

and

$$\partial_x v = -\partial_y u. \quad (7.16)$$

Cauchy's Integral Formula

Let $f(z)$ be analytic in a simply connected region, and let C be any closed contour in that region. Choose any fixed point z inside C ; then we derive the fact that

$$f(z) = \frac{1}{2\pi i} \int_C \frac{f(\zeta)}{\zeta - z} d\zeta. \quad (7.17)$$

for some closed contour C in the complex plane. Furthermore we show that

$$f'(z) = \frac{1}{2\pi i} \int_C \frac{f(\zeta)}{(\zeta - z)^2} d\zeta, \quad (7.18)$$

and in general

$$f^n(z) = \frac{n!}{2\pi i} \int_C \frac{f(\zeta)}{(\zeta - z)^{n+1}} d\zeta. \quad (7.19)$$

An interesting consequence of these formulas is the following. Consider a circle of radius R and assume that f is analytic inside this circle. Then letting $C = Re^{i\theta}$ (and taking $z = 0$ for definiteness) we have that for example

$$f'(0) = \frac{1}{2\pi} \int_0^{2\pi} \frac{f(Re^{i\theta})}{R} e^{-i\theta} d\theta. \quad (7.20)$$

Now if $f(z)$ is bounded by some constant M as $z \rightarrow \infty$ then we have that

$$f'(0) \leq \frac{M}{2\pi R}. \quad (7.21)$$

Hence as $R \rightarrow \infty$, $f' \rightarrow 0$. This therefore demonstrates that any function that is *bounded at infinity and analytic is necessarily constant*. Hence any interesting behavior of a function is due to its singularities in the complex plane!

We therefore see that although (as we have previously argued) singularities are the essential cause for the nonconvergence of series expansions, it is also the case that the interesting part of a function are its singularities.

7.2.4 Types of Singularities in the Complex Plane

Thus motivated we now turn to the question of what are the types of singularities that can exist in the complex plane. We have argued above for several types:

1. Multivaluedness leads to singularities, in the form of branch points. These are points which when orbited in the complex plane lead to multivaluedness. Functions with this property do not converge when expanded around the branch point for the simple reason that the expansion is made up of a sum of powers z^n , and each of which is not multivalued. One cannot create a multivalued function by adding up single valued functions.
2. Poles. This is an actual divergence of the function of the form

$$f(z) = \frac{1}{z^n}. \quad (7.22)$$

3. Essential singularities. This is stronger than any pole and the nastiest type of singularity. A prototypical example is

$$f(z) = e^{-1/z}. \quad (7.23)$$

We saw essential singularities above when we were expanding the error function – typically essential singularities can be well approximated in sectors of the complex plane, though different formulas are required for each sector.

We now prove that these are the only types of singularities that can exist in the complex plane. To do this we pretend that our function is not multivalued, and ask what other types of singularities can occur. The proof starts with Cauchy's formula, but we ask for the behavior of a function near a point z_0 . There may be singularities or other nastiness at or near z_0 and for that reason we generalize our contour $C = C_0 + C_i$ to involve both the contour C_0 and C_i , where C_0 is the outer contour (a big counterclockwise circle around the point z_0) and C_i is the inner contour (a small clockwise circle right around z_0). We write

$$f(z) = \frac{1}{2\pi i} \int_{C_0} \frac{f(\zeta)}{\zeta - z_0 + (z_0 - z)} d\zeta + \frac{1}{2\pi i} \int_{C_i} \frac{f(\zeta)}{\zeta - z_0 + (z_0 - z)} d\zeta \quad (7.24)$$

Now consider the denominator

$$\frac{1}{\zeta - z_0 + (z_0 - z)}. \quad (7.25)$$

For the contour C_0 (the outer contour) we have that $|\zeta - z_0| > |z_0 - z|$. On the contour C_i (the inner contour) we have the opposite. This therefore motivates the expansion

$$\frac{1}{\zeta - z_0} \frac{1}{1 - (z - z_0)/(\zeta - z_0)} = \frac{1}{\zeta - z_0} \left(1 + \frac{z - z_0}{\zeta - z_0} + \left(\frac{z - z_0}{\zeta - z_0} \right)^2 + \dots \right). \quad (7.26)$$

We can use this in Cauchy's formula to show that $f(z)$ is composed of the sum of a power series around z_0 with *both positive and negative powers*, i.e.

$$f(z) = \sum_{n=-M}^{\infty} a_n (z - z_0)^n. \quad (7.27)$$

This is the *Laurent series* of $f(z)$ in the annular region between C_0 and C_i . Here M is the most negative term in the series. The negative terms come from considering the integral around the inner contour in the Cauchy's formula. If $M = \infty$ then the function has an essential singularity at z_0 . Otherwise it has a pole of order M . Because in carrying out this derivation we have only assumed that $f(z)$ is analytic except for an excised region around z_0 , we have just shown that the most general singularities of a function are as claimed.

7.2.5 Summary

In the last section, we classified the types of possible singularities that can occur in the complex plane, with a view towards understanding why power series expansions cease to converge. These are:

- Branch points. Such singularities denote points around which the function is multi-valued. Since power series are sums of integer powers z^n which are single-valued, it is impossible for a power series to accurately represent a multi-valued function. Therefore when the circle of convergence overlaps a branch point, a power series can no longer converge.
- The principal part of the Laurent series (the negative powers) contains a finite number of terms. If the last term is $a_{-N}(z - z_0)^{-N}$ then the function has a *pole of order N* .
- In the final case, the principal part contains an infinite number of terms. In this case the function has an *essential singularity*. An example of a function with an essential singularity at $z = 0$ is $e^{1/z}$; it should also be noted that e^z has an essential singularity at $z = \infty$.

Now there are two more topics of complex analysis that we will consider: first we will consider the problem of analytic continuation. This is the question of when can one improve a formula (that only works in a certain region of the complex plane) to make it work in a larger region of the complex plane. This is a very practical question – can we find anything to do with our non-converging series?

The second topic we will consider is contour integration – namely how to think about integration when it is extended to the complex plane. We have already discussed that as long as there are no singularities, the contour in the complex plane can be deformed at will; but even if there are singularities we can still calculate what the error is in Cauchy's formula due to the singularities. This is a powerful way of making complicated integrals into simple ones, as we will see.

7.3 Analytic Continuation

When a representation of a function only converges in a region of the complex plane, it is natural to ask whether there is a way of continuing the function to a larger part of the complex plane. For example, consider $\sum_n z^n$. This series only converges when $|z| < 1$; on the other hand we also know that the function $(1 - z)^{-1}$ agrees with the series within the region of convergence. Therefore, $(1 - z)^{-1}$ is an analytic continuation of the series. The formula works everywhere in the complex plane, except at the pole $z = 1$.

A common example of analytic continuation is how to take a real valued function $f(x)$ and continue it into the complex plane. For this, one just replaces $x \rightarrow z$, and then considers $f(z)$ as a complex valued function.

As a final example, consider the Riemann zeta function

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}. \quad (7.28)$$

This sum converges absolutely when $Re(s) > 1$. Can it be continued to other s ? To do this, note that

$$\begin{aligned} \sum_{n \leq x} n^{-s} &= \int_{1-}^x t^{-s} d[t] \\ &= [x]x^{-s} + s \int_1^x [t]t^{-s-1} dt \\ &= [x]x^{-s} + s \int_1^x ([t] - t)t^{-s-1} dt + \frac{s}{s-1} - \frac{sx^{1-s}}{s-1}. \end{aligned} \quad (7.29)$$

Here $[x]$ denotes the integer nearest to x . Taking $x \rightarrow \infty$ and assuming $Re(s) > 1$ we have

$$\zeta(s) = \frac{s}{1-s} + s \int_1^{\infty} ([t] - t)t^{-s-1} dt. \quad (7.30)$$

This formula shows that ζ has a simple pole at $s = 1$. Note that although we derived this formula assuming that $Re(s) > 1$, the integral actually converges when $Re(s) > 0$, since $[t] - t$ is always smaller than unity. Thus, this formula provides an analytic continuation for the $\zeta(s)$ when $Re(s) > 0$.

Of course, these examples of analytic continuation are all special. It is rare that one can write down an analytic continuation in such a straightforward fashion. The canonical method for performing analytic continuation is called the “circle-chain method”. Suppose the series $a_n z^n$ converges within the circle $|z| < R$. Somewhere on the circle $|z| = R$, there is a singularity. Consider a point close to the circle $|z| = R$. One can now carry out a new Taylor series expansion of the function about the new point. The radius of convergence of the resulting series will be the distance of the chosen point to the singularity; in general this series will extend the region of convergence beyond the original circle. By repeating this method again and again, one can “fill up” the complex plane with an analytic continuation of the function. This method is tedious, though it is guaranteed to work.

Is the resulting analytic continuation unique? Imagine we carried out the circle chain method starting from two separate points; under what conditions will we arrive at the same answer? Suppose the function $f(z)$ is analytic in the entire region comprising the two “circle-chain”s. Then the answer had better agree! Indeed it does. This is the result of the so-called *identity theorem* for analytic continuation. Suppose $f_1(z)$ and $f_2(z)$ are both analytic in a region R , and they coincide in some subregion of R . The identity theorem states that they coincide throughout R . More generally, if there are no singular points of the function in between the two paths of analytic continuation, the two continuations will agree.

The proof of this theorem is as follows. Consider $F = f_1 - f_2$. By assumption, F is analytic in the region R , and $F = 0$ in some subregion. Consider a curve connecting the subregion to R . In the subregion, all derivatives of F vanish, all the way up to the boundary of the subregion. Now consider a Taylor series expansion of the function around the intersection point. This Taylor series is clearly $= 0$ everywhere in a disk around the intersection point. Thus we have extended the region where $F = 0$. Continuing this will fill up the region R with $F = 0$.

On the other hand, it is equally clear that if there is a branch point type singularity in the middle of the region, analytic continuations that circle this branch point cannot be unique. For example, consider $\log(1 + z) = \sum (-1)^j z^j / j$. Circling the branch point at $z = -1$ via the circle chain method will lead to different values of the logarithm at these points.

7.3.1 Approximate Analytic Continuation

Why do we care about analytic continuation? Often one is presented with a formula representing a function which is only valid in a certain range of the argument. One would like to find a way of extending the range of validity. Analytic continuation is the technical procedure through which this is possible. We have seen from above that explicit analytic continuation is often very tedious, and does not lead to the types of compact formulas that lend physical intuition.

Of course, in the modern world, one can often perform analytic continuation by numerical computation; however, although this is a good way of producing accurate numbers, it also does not lend any intuition as to why the numbers are what they are.

In practice, what one is most often looking for is not an exact formula for the analytically continued function, but an approximate one. Here we describe a number of methods for extracting approximate analytic continuations of functions, with varying levels of sophistication.

Euler Transformation

A perhaps “hopelessly” naive way of fixing a singularity is to try to define it away. As an example of this method, let us consider the power series expansion for

$$\log(1+z) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{z^n}{n}. \quad (7.31)$$

The radius of convergence of this expansion is $z = 1$, owing to the branch point singularity at $z = -1$. Let us try to define this away as follows. Let

$$\zeta = \frac{z}{1+z}. \quad (7.32)$$

Inverting this expression gives

$$z = \frac{\zeta}{1-\zeta}. \quad (7.33)$$

We will now use this substitution to write the power series in terms of ζ instead of z . The reason this might work is that we have included a singularity in the $\zeta(z)$ transformation at exactly the point $z = -1$ where the branch point singularity occurs. Can this singularity “cancel out” the singularity in the expansion?

Making this substitution, we obtain

$$\begin{aligned} \log(1+z) &= \frac{\zeta}{1-\zeta} - \frac{1}{2} \frac{\zeta^2}{(1-\zeta)^2} + \dots \\ &= \zeta(1+\zeta+\zeta^2+\dots) - \frac{1}{2}\zeta^2(1+2\zeta+3\zeta^2+\dots) + \dots \\ &= \zeta + \zeta^2/2 + \zeta^3/3 + \dots \end{aligned} \quad (7.34)$$

The radius of convergence of this series is therefore $|\zeta| < 1$, which corresponds to $1/2 < z < \infty$!

Note 1: The coefficients of this series can be computed using Maple, as we mentioned in Chapter 2. Input the following code line by line into Maple:


```

1 zp:=s/(1-s)
2 yp:=sum((-1)^(n+1)*z^n/n, n=1 .. 10)
3 dum:=subs(z=zp,yp)
4 stuff:=series(dum,s=0,10)

```

In the code above, the first and second lines are used to define the substitution expression and the original series; the third line does the substitution; the fourth line expands the series after substitution. The final output is:

$$s + \frac{1}{2}s^2 + \frac{1}{3}s^3 + \frac{1}{4}s^4 + \frac{1}{5}s^5 + \frac{1}{6}s^6 + \frac{1}{7}s^7 + \frac{1}{8}s^8 + \frac{1}{9}s^9 + O(s^{10}).$$

Note 2: There is a nifty command in Maple for converting symbolic expressions to latex, for use in technical documents! The command `latex(stuff)` converts the symbolic expression `stuff` to latex.

The conversion of the series from a series converging in $|z| < 1$ to converging on the entire positive real axis is remarkable; but does it work?

Figure 7.1 compares $\log(1+z)$ (solid line) with the normal power series (open circles) and the Euler transformed series (squares) on positive real axis. For each of the series we have summed 299 terms. The Euler transformed series exactly captures the function, whereas the normal power series is terrible outside of its radius of convergence.

You may also wonder what the function/series look like on the whole complex plane, which is what Figure 7.2 visualizes. The hue encodes the phase angle, whereas the lightness encodes the magnitude (from dark to light the value increases from zero to infinity). As you can see, the normal series (B) only captures the original function (A) inside the radius of convergence - it diverges outside the circle, whereas the Euler transformation series (C) captures the original function in a much larger area on the complex plane. The MATLAB code generating this figure is shown in the Appendix at the end of this chapter.

Domb Sykes Plot

Another way of improving the convergence of a series is to *find* the form of the singularity which is creating the divergence, and subtracting it out from the series expansion. Without the singularity, the radius of convergence of the series will be larger, proceeding at least to the next singularity.

How does one find the singularity, given only the representation of a function in terms of a series? Recall that given a series $f(x) = \sum_n a_n x^n$, the radius of convergence is given by

$$R = \lim_{n \rightarrow \infty} \frac{a_n}{a_{n+1}}. \quad (7.35)$$

One can get more information about the singularity by studying the *rate* at which the ratio a_n/a_{n+1} asymptotes to a constant. Namely, suppose that the dominant singularity causing the radius of convergence has the form $(R+x)^\beta$; this singularity has the Taylor series expansion

$$\begin{aligned} R^\beta \left(1 + \frac{x}{R}\right)^\beta &= R^\beta \left(1 + \beta \frac{x}{R} + \frac{\beta(\beta-1)}{2} \left(\frac{x}{R}\right)^2 + \dots \right. \\ &\quad \left. + \frac{\beta(\beta-1)\dots(\beta-n+1)}{n!} \left(\frac{x}{R}\right)^n + \dots\right). \end{aligned} \quad (7.36)$$

Hence, for this series, when n is sufficiently large, the ratio

$$\frac{|a_n|}{|a_{n+1}|} = R \frac{n}{n-\beta-1} \approx R \left(1 + \frac{\beta+1}{n}\right). \quad (7.37)$$

Thus, if we were to plot a_n/a_{n+1} as a function of $1/n$, the intercept would give the radius of convergence, whereas the slope would give the type of singularity we have! This trick was popularized by Domb and Sykes, in the context of determining singularities in partition functions of statistical mechanics, from a low temperature expansion of the partition function. Such singularities physically reflect phase transitions in the underlying system.

Let us carry out an example: We consider our very first series, the perturbation series for the root near $y = 1$ as $\epsilon \rightarrow 0$ for our quintic. We have already established that this series has a finite radius of convergence, as discussed in class last week. Now let's plot a_n/a_{n+1} , and see what we can conclude from a Domb-Sykes Plot. Such a plot is shown in Figure 7.3.

Carrying out a linear regression on this data, we find that $a_n/a_{n+1} = 0.0826 + 0.1183/n$; this implies that the radius of convergence is 0.0826, whereas the index of the singularity is $0.1183/0.0826 - 1 = 0.4322$.

Program 12 MATLAB code used to create figure 7.1

```
1 % generate log(1+z)
2 z = 0:100;
3 y = log(1+z);
4
5 % generate the normal series
6 z1 = [0:0.1:1,1.005:0.005:1.025];
7 y1 = zeros(1,size(z1,2));
8 for i = 1:size(z1,2)
9     for n = 1:299
10        y1(i) = y1(i)+(-1)^(n+1)*z1(i)^n/n;
11    end
12 end
13
14 % generate the euler transf. series
15 z2 = 0:4:100;
16 y2 = zeros(1,size(z2,2));
17 s = z2./(1+z2);
18 for i = 1:size(s,2)
19     for n = 1:299
20        y2(i) = y2(i)+1/n*s(i)^n;
21    end
22 end
23
24 % plot the result
25 figure
26 plot(z,y,'r-','linewidth',2)
27 hold on
28 plot(z1,y1,'mo')
29 plot(z2,y2,'bs')
30 set(gca,'fontsize',16,'fontname','Helvetica','fontweight','b')
31 legend('log(1+z)','Normal Series','Euler Transf. Series')
32 xlabel('z')
33 ylabel('Function/Series Values')
```

Is this correct? Since the singularity leading to this radius of convergence occurs at the merger of two real roots, forming two complex roots, we expect that the singularity should be of the form $(R - \epsilon)^{1/2}$, since this is the generic way in which two real roots become two complex roots. Thus, the index of the singularity from this analysis is off of what we expect.

Moreover, there is a simple analytical way of analytically computing the radius of convergence of this series. At the value of ϵ corresponding to the radius of convergence, two distinct real roots become identical. Thus, $F(y) = \epsilon y^5 - y + 1$ can be factored in the form $F(y) = (y_* - y)^2 G(y)$. The consequence of this is that at the critical ϵ , $dF/dy = 0$ is also satisfied! We can use this to our advantage. Since $dF/dy = 5\epsilon y^4 - 1$, we have $y = 1/(5\epsilon)^{1/4}$. Using this in the equation $F = 0$ gives a formula for ϵ , namely

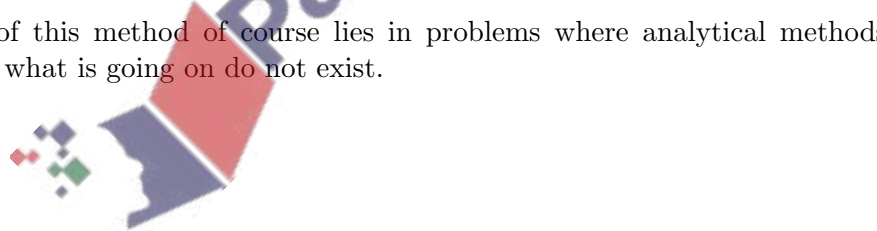
$$\epsilon_* = R = \left(\left(5^{1/4} \right)^{-1} - \left(5^{5/4} \right)^{-1} \right)^4 = 0.0819. \quad (7.38)$$

Thus we see that the Domb Sykes plot has given wrong answers for both the radius of convergence and the strength of the singularity. The reason for this is of course that the result of the Domb Sykes plot is only valid as $n \rightarrow \infty$, and we have only let $n = 16$. One can get around this by noting that at finite n ,

$$\frac{a_n}{a_{n+1}} = R \left(1 + \frac{\beta + 1}{n} + O\left(\frac{1}{n^2}\right) + \dots \right). \quad (7.39)$$

Thus, if we fit a_n/a_{n+1} to a higher order polynomial in $1/n$ we will get a more accurate answer. Doing a quadratic fit to our data gives $a_n/a_{n+1} = 0.0817 + 0.1258/n + \dots$, implying the radius of convergence is 0.0817, and $\beta = 0.1258/0.0817 - 1 = 0.53$; a cubic fit gives $a_n/a_{n+1} = 0.0819 + 0.1236/n + \dots$, implying the radius of convergence is 0.0819, and $\beta = 0.1236/0.0819 - 1 = 0.509$. Thus we are converging on the answer we know to be true.

The power of this method of course lies in problems where analytical methods of understanding what is going on do not exist.



7.4 Pade Approximants

Let us consider a function $f(z)$ whose only singularities are poles. If these poles occur at the points $\{z_i\}$, and the strength of the i^{th} pole is α_i , then the function $g(z) = f(z)\prod_i(z - z_i)^{\alpha_i}$ is analytic. At worst, $g(z)$ can have a singularity at ∞ , but (by assumption) this singularity must be a pole. Therefore $g(z)$ is a polynomial! The consequence of this is that any function whose singularities are poles can be written as the ratio of two polynomials,

$$f(z) = \frac{P(z)}{Q(z)}. \quad (7.40)$$

Of course, most functions have singularities which are not poles. But, it is natural to ask the question: how good of a job can we do of approximating the behavior of an arbitrary function if we *assume* that its only singularities are poles? This idea is called *Pade approximation*.

We will try to represent a function $f(z)$ by a ratio of two polynomials. Let

$$P_M^N(z) = \frac{\sum_{n=0}^N a_n z^n}{\sum_{n=0}^M b_n z^n}. \quad (7.41)$$

How do we find the coefficients (a_n, b_n) ? Without loss of generality we can choose $b_0 = 1$. The rest of the coefficients can be computed in many different ways, among them:

- Compute the Taylor series of $f(z)$ around $z = 0$. Demand that the first $N + M + 1$ Taylor series coefficients match the first $N + M + 1$ coefficients of the Pade approximant.
- Or, the $M + N + 1$ unknown coefficients can be constructed in any other systematic way. For example, we could match the Taylor series of $f(z)$ with that of the approximant around 2 points, or three points, or... A natural choice of two points ♦ is to match the expansions around $z = 0$ and $z = \infty$.

We now discuss the implementation of this, fitting the coefficients of the Pade approximant with the Taylor series of $f(z) = \sum c_n z^n$ expanded around $z = 0$. We have

$$\sum_{n=0}^{N+M+1} c_n z^n = \frac{\sum_{n=0}^N a_n z^n}{\sum_{n=0}^M b_n z^n}. \quad (7.42)$$

Multiplying out the denominator we have

$$\left(\sum_{n=0}^{N+M+1} c_n z^n \right) \left(\sum_{n=0}^M b_n z^n \right) = \sum_{n=0}^N a_n z^n. \quad (7.43)$$

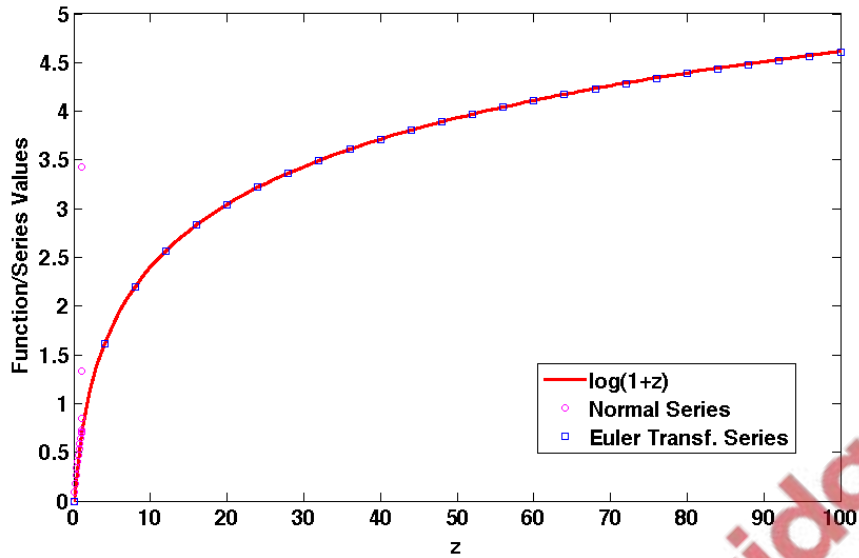


Figure 7.1. Plot comparing $\log(1+z)$ (solid line) with the normal power series (open circles) and the Euler transformed series (squares) on the positive real axis

Program 13 MATLAB code used to create figure 7.3

```

1 % generate a series of the reciprocal of n
2 n = 1:16;
3 n_reci = 1./n;
4
5 % store the ratio of a_n and a_{n+1} got from the Maple calculation in
6 % chapter 2
7 ratio = [0.2000000000, 0.1428571429, 0.1228070175, ...
8          0.1126482213, 0.1065218307, 0.1024279800, ...
9          0.9950031003e-1, 0.9730315062e-1, 0.9559364912e-1, ...
10         0.9422578901e-1, 0.9310652435e-1, 0.9217376797e-1, ...
11         0.9138450968e-1, 0.9070801459e-1, 0.9012173702e-1, ...
12         0.8960876413e-1]
13
14 % plot the Domb Sykes plot
15 plot(n_reci,ratio,'bo')
16 set(gca,'fontsize',16,'fontname','Helvetica','fontweight','b')
17 xlabel('1/n')
18 ylabel('a_n/a_{n+1}')

```

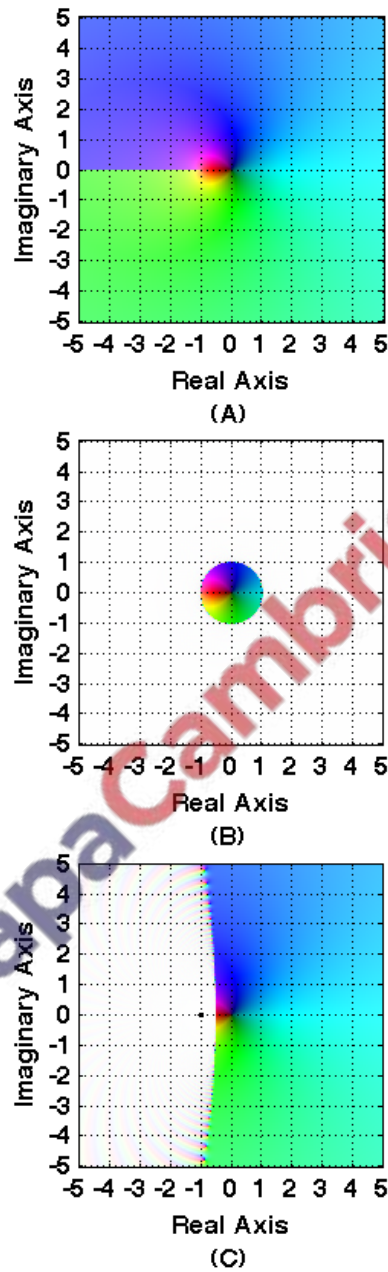


Figure 7.2. Plot comparing (A) $\log(1+z)$ with (B) the normal power series and (C) the Euler transformation series in the area $[-5, 5] \times [-5, 5]$ of the complex plane. The hue encodes the phase angle, while the lightness encodes the magnitude (value increasing from zero to infinity is encoded as color going from dark to light).

From this equation we need to extract $N + M + 1$ equations for the $N + M + 1$ unknowns. We do this in two steps: first consider that when the power z^q has $q > N$, the second term does not contribute at all. Since by definition $b_0 = 1$, we therefore have

$$c_{N+1} + \sum_{j=1}^M c_j b_{N+1-j} = 0, \quad (7.44)$$

as the coefficient of z^{N+1} . This is a matrix equation $\mathbf{Bc} = -\bar{c}$, where $B_{ij} = b_{N+i-j}$. The lower powers z^q with $0 < q < N$ satisfy the equation

$$a_n = \sum_{j=0}^N b_{n-j} c_j. \quad (7.45)$$

With these equations, we can solve for the coefficients b and c of the Pade approximant.

The *MATLAB* function below implements the algorithm we outlined above for Pade approximants.

```

1 function [A,B] = pade(y,N,M)
2 % This function calculates the coefficients of Pade approximant
3 % P = \sum_{n=0}^N a_n z^n / \sum_{n=0}^M b_n z^n
4 % Input variables:
5 % y - the symbolic expression of the original function
6 % N - the highest order of the numerator
7 % M - the highest order of the denominator
8 % Output variables:
9 % A - a row vector containing the coefficients of the numerator from low
10 %   order to high order
11 % B - a row vector containing the coefficients of the denominator from low
12 %   order to high order
13
14 % get the coefficients of the Taylor series
15 c = zeros(1,N+M+1);
16 c(1) = subs(taylor(y,1),1);
17 for i = 2:N+M+1
18     c(i) = subs(taylor(y,i)-taylor(y,i-1),1);
19 end
20
21 % calculate the coefficients of the denominator
22 for i=1:N
23     for j=1:M
24         mat(i,j)=c(N+i-j+1);
25     end
26 end
27 rhs = -c(N+2:N+M+1)';
28 bb = inv(mat)*rhs;
29 B(1) = 1;
30 B(2:M+1) = bb;
31
32 % calculate the coefficients of the numerator

```

```

33 A(1) = c(1);
34 for i = 2:N+1
35     A(i) = c(i)*B(1);
36     for j = 2:i
37         A(i) = A(i)+c(i+1-j)*B(j);
38     end
39 end
40
41 end

```

Let's try this out on a specific example. Consider

$$\sqrt{1+x} = 1 + \frac{1}{2}x - \frac{1}{8}x^2 + \dots \quad (7.46)$$

The radius of convergence of this expansion is $|x| < 1$, due to the branch point singularity at $x = -1$. How do the Pade approximants fare?

Figure 7.4 shows Pade approximation to this function on positive real axis. Even the 2nd order Pade approximant gives good agreement with the square root beyond the radius of convergence of the normal Taylor series expansion. The 4th order Pade approximant agrees with the function almost over the entire range shown.

It is interesting to examine the zeros and poles of the Pade approximant: something must go awry *somewhere* in the complex plane; after all, the initial function we were trying to approximate was multivalued! Figure 7.5 shows the zeros and poles of the Pade approximant. The zeros and poles align themselves along the branch cut along the negative real axis! The zeros and the poles alternate along this axis.

Figure 7.6 shows the error in the Pade approximant at $x = 200$ (defined as $|y_{pade} - y_{exact}|$) as a function of the order of the approximant. It is seen that the Pade approximants have exponential decaying error with increasing order N .

As we did perviously, in Figure 7.7, we also plot the original function, 8th order Taylor series and 4th order Pade approximant in the area $[-10, 10] \times [-10, 10]$ of the complex plane. As you can see, the Taylor series diverges outside the convergent circle, while the Pade Approximant captures the original function in a much larger area. MATLAB code generating this figure is similar to the one generating Figure 7.2

Before closing we give one more example. Figure 7.8 and Figure 7.9 shows Pade approximants to the exponential function e^x on positive real axis and the complex plane.

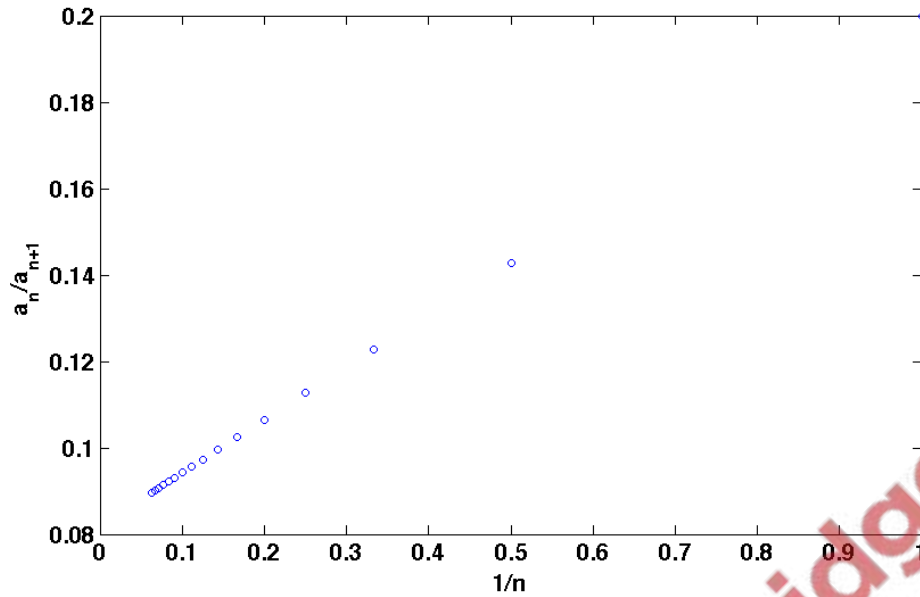


Figure 7.3. Domb Sykes Plot for the root of $\epsilon y^5 - y + 1 = 0$ near $y = 1$ for small ϵ . Note that here we only expand the perturbation series to ϵ^{17} term ($n = 16$ in the figure); as $n \rightarrow \infty$, the ratio a_n/a_{n+1} will approach 0.0819, the radius of convergence.

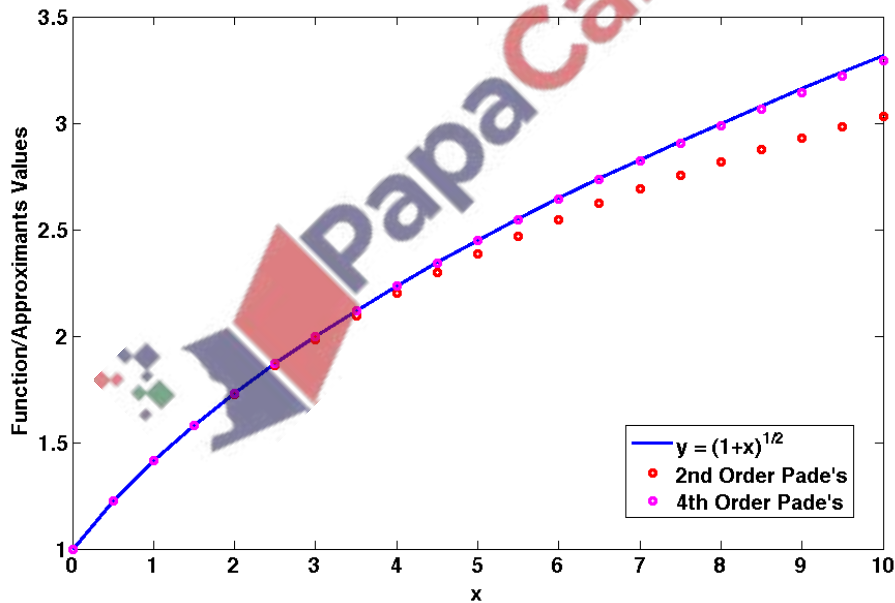


Figure 7.4. Pade approximants to $\sqrt{1+x}$ on positive real axis. The solid line gives the exact result. The red circles show a 2nd order Pade approximant (P_2^2), whereas the magenta circles show a 4th order Pade approximant (P_4^4).

Program 14 MATLAB code used to create figure 7.4

```
1 % define the original function
2 syms x
3 y = sqrt(1+x);
4
5 % get the coefficients of the Pade approximants and flip them horizontally
6 % to decreasing order
7 [A2,B2] = pade(y,2,2); [A4,B4] = pade(y,4,4);
8 A2 = fliplr(A2); B2 = fliplr(B2);
9 A4 = fliplr(A4); B4 = fliplr(B4);
10
11 % plot the original function and the Pade approximants
12 x = 0:0.5:10;
13 original = sqrt(1+x);
14 p22 = polyval(A2,x)./polyval(B2,x);
15 p44 = polyval(A4,x)./polyval(B4,x);
16 plot(x,original,'b-','linewidth',2)
17 hold on
18 plot(x,p22,'ro','linewidth',2)
19 plot(x,p44,'mo','linewidth',2)
20 set(gca,'fontsize',16,'fontname','Helvetica','fontweight','b')
21 xlabel('x')
22 ylabel('Function/Approximants Values')
23 legend('y = (1+x)^{1/2}', ...
24        '2nd Order Pade's', ...
25        '4th Order Pade's')
```

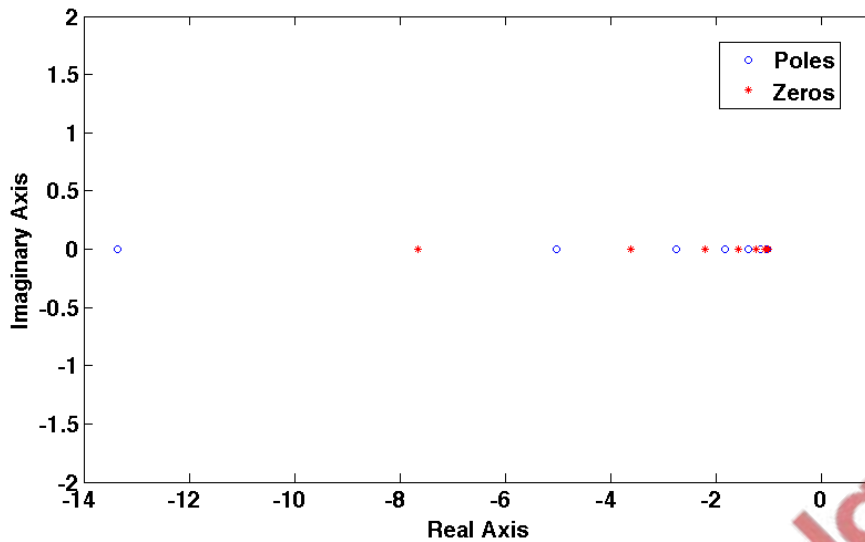


Figure 7.5. Poles and zeros of the Pade approximant - the circles show the poles, the asterisks show the zeros. Note that the zeros and poles alternate, lining up along the branch cut $(-\infty, -1]$!

Program 15 MATLAB code used to create figure 7.5

```

1 % define the original function
2 syms x
3 y = sqrt(1+x);
4
5 % get the coefficients of the Pade approximant and flip them horizontally
6 % to decreasing order
7 [A,B] = pade(y,8,8);
8 A = fliplr(A);
9 B = fliplr(B);
10
11 % solve poles and zeros
12 zero = roots(A);
13 pole = roots(B);
14
15 % plot poles and zeros on the complex plane
16 plot(real(pole),imag(pole),'bo')
17 hold on
18 plot(real(zero),imag(zero),'r*')
19 axis([-14 1 -2 2])
20 set(gca,'fontsize',16,'fontname','Helvetica','fontweight','b')
21 xlabel('Real Axis')
22 ylabel('Imaginary Axis')
23 legend('Poles','Zeros')

```

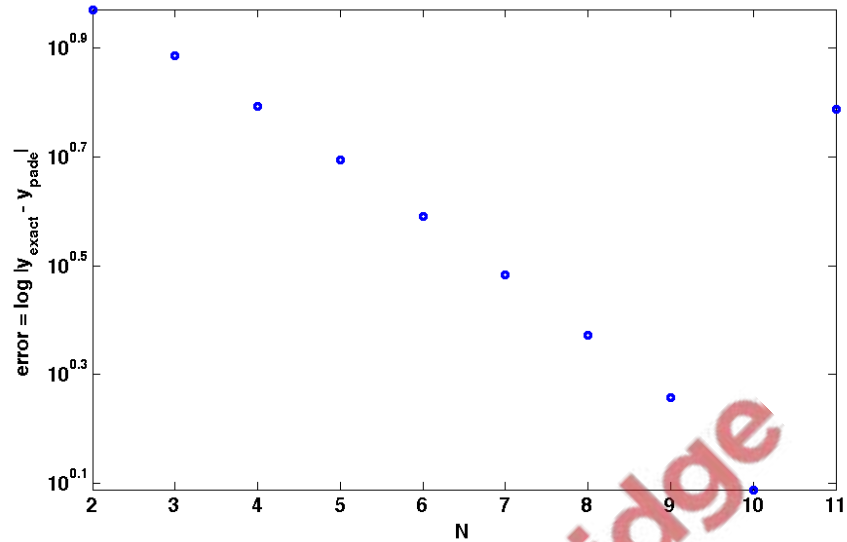


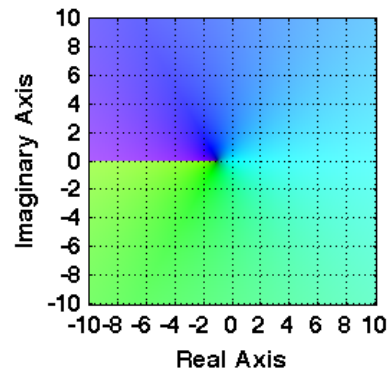
Figure 7.6. Error in the Pade approximant of $\sqrt{1+x}$ at $x = 200$ as a function of the order N .

Program 16 MATLAB code used to create figure 7.6

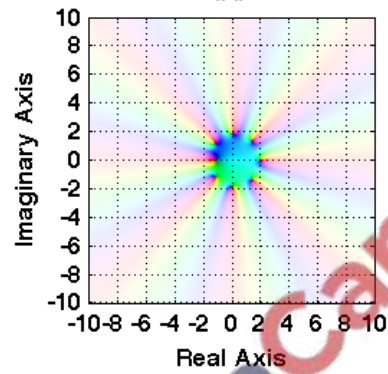
```

1 % define the original function and other variables
2 syms x
3 y = sqrt(1+x);
4 z = 200;
5 N = 2:11;
6 error = zeros(1,length(N));
7
8 % get the errors
9 for i = 1:length(N)
10     [A,B] = pade(y,N(i),N(i));
11     A = fliplr(A);
12     B = fliplr(B);
13     error(i) = abs(subs(y,z)-polyval(A,z)./polyval(B,z));
14 end
15
16 % plot the error function
17 semilogy(N,error,'o','linewidth',2)
18 set(gca,'fontsize',16,'fontname','Helvetica','fontweight','b')
19 xlabel('N')
20 ylabel('error = log |y_{exact} - y_{pade}|')

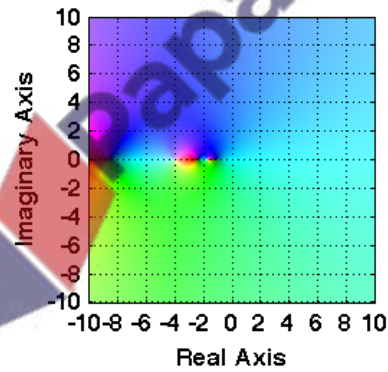
```



(A)



(B)



(C)

Figure 7.7. Plot comparing (A) $\sqrt{1+x}$ with (B) the 8th order Taylor series and (C) the 4th order Padé approximant in the area $[-10, 10] \times [-10, 10]$ of the complex plane. The hue encodes the phase angle, whereas the lightness encodes the magnitude.

Program 17 MATLAB code used to create figure 7.8

```
1 % define the original function
2 syms x
3 y = exp(x);
4
5 % get the coefficients of the Pade approximants
6 [A2,B2] = pade(y,2,2); A2 = fliplr(A2); B2 = fliplr(B2);
7 [A3,B3] = pade(y,3,3); A3 = fliplr(A3); B3 = fliplr(B3);
8 [A4,B4] = pade(y,4,4); A4 = fliplr(A4); B4 = fliplr(B4);
9 [A5,B5] = pade(y,5,5); A5 = fliplr(A5); B5 = fliplr(B5);
10 [A6,B6] = pade(y,6,6); A6 = fliplr(A6); B6 = fliplr(B6);
11 [A7,B7] = pade(y,7,7); A7 = fliplr(A7); B7 = fliplr(B7);
12 [A8,B8] = pade(y,8,8); A8 = fliplr(A8); B8 = fliplr(B8);
13
14 % plot the original function and the Pade approximants
15 x = 0:0.05:10;
16 original = exp(x);
17 plot(x,original,'b-','linewidth',5)
18 hold on
19 p = polyval(A2,x)./polyval(B2,x); plot(x,p,'k-')
20 p = polyval(A3,x)./polyval(B3,x); plot(x,p,'g-')
21 p = polyval(A4,x)./polyval(B4,x); plot(x,p,'c-')
22 p = polyval(A5,x)./polyval(B5,x); plot(x,p,'r-')
23 p = polyval(A6,x)./polyval(B6,x); plot(x,p,'m-')
24 p = polyval(A7,x)./polyval(B7,x); plot(x,p,'y-')
25 p = polyval(A8,x)./polyval(B8,x); plot(x,p,'ro','markersize',8)
26 axis([0 10 -3000 10000])
27 set(gca,'fontsize',16,'fontname','Helvetica','fontweight','b')
28 xlabel('x')
29 ylabel('Function/Approximants Values')
30 legend('y = e^x', ...
31        '2nd Order Pade''s', ...
32        '3rd Order Pade''s', ...
33        '4th Order Pade''s', ...
34        '5th Order Pade''s', ...
35        '6th Order Pade''s', ...
36        '7th Order Pade''s', ...
37        '8th Order Pade''s')
```

7.5 Appendix: The MATLAB Code Generating Complex Function on the Whole Complex Plane

Below is the MATLAB code generating Figure 7.2. Similar codes are used to generate Figure 7.7 and Figure 7.9.

```
1 % generate the complex plane [-5 5]*[-5 5] with the resolution 500*500;
2 % set it smaller to accelerate the calculation if needed
3 res = 500;
4 X = linspace(-5,5,res); X = repmat(X,[res 1]);
5 Y = linspace(5,-5,res); Y = repmat(Y',[1 res]);
6 z = X+Y*i;
7
8 % apply the function/series to the complex plane to get the function values
9 y = log(1+z);
10
11 y1 = zeros(res,res);
12 for n = 1:299
13     y1 = y1+(-1)^(n+1)*z.^n/n;
14 end
15
16 y2 = zeros(res,res);
17 s = z./(1+z);
18 for n = 1:299
19     y2 = y2+1/n*s.^n;
20 end
21
22 % map function values to the HSL color model, where the phase angle is
23 % mapped to hue (H), the magnitude is mapped to lightness (L) through a
24 % special function, and the saturation (S) is set to 100% so that the
25 % outcome looks beautiful
26 h = (angle(y)+pi)*180/pi/360;
27 h1 = (angle(y1)+pi)*180/pi/360;
28 h2 = (angle(y2)+pi)*180/pi/360;
29 l = atan(log(abs(y).^0.6))*1/pi+0.5;
30 l1 = atan(log(abs(y1).^0.6))*1/pi+0.5;
31 l2 = atan(log(abs(y2).^0.6))*1/pi+0.5;
32 s = ones(res,res);
33
34 hsl = zeros(res,res,3);
35 hsl1 = zeros(res,res,3);
36 hsl2 = zeros(res,res,3);
37 hsl(:, :, 1) = h; hsl(:, :, 2) = s; hsl(:, :, 3) = l;
38 hsl1(:, :, 1) = h1; hsl1(:, :, 2) = s; hsl1(:, :, 3) = l1;
39 hsl2(:, :, 1) = h2; hsl2(:, :, 2) = s; hsl2(:, :, 3) = l2;
40
41 % use the function hsl2rgb(), downloadable at
42 % http://www.mathworks.se/matlabcentral/fileexchange/3360-rgb-to-hsl,
43 % to convert HSL model to RGB model so that we can plot them in Matlab
44 rgb = zeros(res,res,3);
45 rgb1 = zeros(res,res,3);
```

```

46 rgb2 = zeros(res, res, 3);
47 for m = 1:res
48     for n = 1:res
49         temp = hsl2rgb([hsl(m, n, 1), hsl(m, n, 2), hsl(m, n, 3)]);
50         temp1 = hsl2rgb([hsl1(m, n, 1), hsl1(m, n, 2), hsl1(m, n, 3)]);
51         temp2 = hsl2rgb([hsl2(m, n, 1), hsl2(m, n, 2), hsl2(m, n, 3)]);
52         rgb(m, n, 1) = temp(1);
53         rgb(m, n, 2) = temp(2);
54         rgb(m, n, 3) = temp(3);
55         rgb1(m, n, 1) = temp1(1);
56         rgb1(m, n, 2) = temp1(2);
57         rgb1(m, n, 3) = temp1(3);
58         rgb2(m, n, 1) = temp2(1);
59         rgb2(m, n, 2) = temp2(2);
60         rgb2(m, n, 3) = temp2(3);
61     end
62 end
63
64 % plot three figures together
65 subplot(3, 1, 1)
66 image([-5 5], [5 -5], rgb)
67 axis image
68 set(gca, 'ydir', 'normal', 'xtick', -5:5, 'ytick', -5:5)
69 set(gca, 'fontsize', 14, 'fontname', 'Helvetica', 'fontweight', 'b')
70 title('A', 'position', [0 -8.5])
71 xlabel('Real Axis'); ylabel('Imaginary Axis')
72 grid
73
74 subplot(3, 1, 2)
75 image([-5 5], [5 -5], rgb1)
76 axis image
77 set(gca, 'ydir', 'normal', 'xtick', -5:5, 'ytick', -5:5)
78 set(gca, 'fontsize', 14, 'fontname', 'Helvetica', 'fontweight', 'b')
79 title('B', 'position', [0 -8.5])
80 xlabel('Real Axis'); ylabel('Imaginary Axis')
81 grid
82
83 subplot(3, 1, 3)
84 image([-5 5], [5 -5], rgb2)
85 axis image
86 set(gca, 'ydir', 'normal', 'xtick', -5:5, 'ytick', -5:5)
87 set(gca, 'fontsize', 14, 'fontname', 'Helvetica', 'fontweight', 'b')
88 title('C', 'position', [0 -8.5])
89 xlabel('Real Axis'); ylabel('Imaginary Axis')
90 grid

```

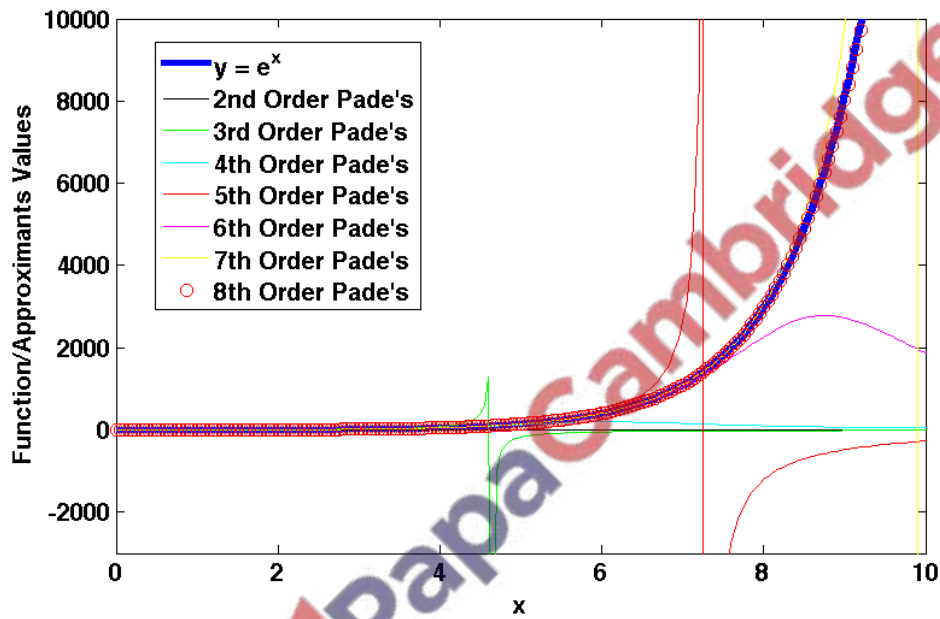



Figure 7.8. Padé approximant to e^x on positive real axis

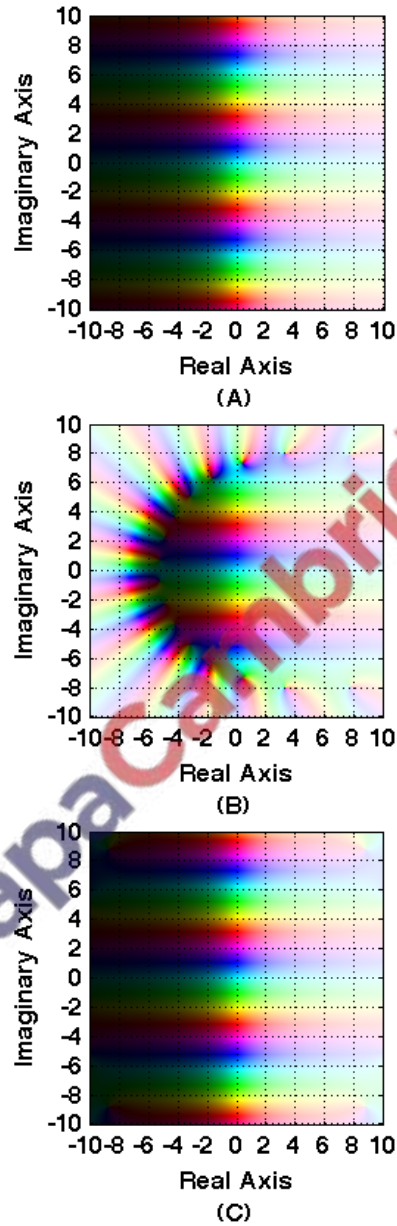


Figure 7.9. Plot comparing (A) e^x with (B) the 16th order Taylor series and (C) the 8th order Padé approximant in the area $[-10, 10] \times [-10, 10]$ of the complex plane. The hue encodes the phase angle, whereas the lightness encodes the magnitude.

8 The Connection Problem

8.1 Matching different solutions to each other

In our studies of equations heretofore, we have learned that solutions to complex equations can be understood as a patchwork: in each region there is some dominant balance that leads to a simpler equation, and the solution of the simpler equation often provides a quantitative characterization of what is going on. The complete solution to the original equation requires putting the individual solutions together.

We call the study of how to piece solutions together into a coherent whole **the connection problem**. In its simplest form, we are given two solutions each of which is valid in some region of space. Each of these solutions has one or many free parameters associated with them. The essential question is to determine the relationship between the parameters that makes it a complete solution of the equation. The essential idea behind this topic is one of the great ideas of twentieth century applied mathematics, and was invented by the great fluid mechanician Ludwig Prandtl.

In what follows we will learn to think about the connection problem. Our discussion will have three main sections.

First we will discuss connection problems that arise in the classical ordinary differential equations of mathematical physics—these are the special functions whose results used to be tabulated at the back of mathematical methods textbooks. We will see that the essential intellectual question posed by these equations (and the essential mathematical content that is described in the old textbooks) is a connection problem.

We will then discuss how to solve such connection problems numerically, first in linear examples (such as the classical special functions) and then, more importantly, in nonlinear examples.

Finally we will discuss analytical methods for solving connection problems. This topic goes under the general name of *boundary layer theory*.

8.2 Connection problems in Classical Linear equations

We demonstrate the essential features of connection problems in the context of the classical equations of mathematical physics. In traditional textbooks, discussions of these equations is surrounded by large quantities of algebraic manipulation. Our contention is that the essential idea that is exposed by these manipulations is a connection problem.

We will consider two examples: The first is the classical example of Bessel Functions: these are generalizations of sin and cos that arise in cylindrical coordinates. The second is a linear second order equation that to my knowledge has not ever been named.

8.2.1 Bessel Functions

Bessel's equation is

$$x^2 \frac{d^2 y}{dx^2} + x \frac{dy}{dx} + (x^2 - \nu^2)y = 0. \quad (8.1)$$

Here ν is a parameter characterizing the solution— ν may or may not be an integer.

To characterize the solutions, we would like to understand first the behavior of the solutions near the origin $x = 0$, and then the behavior of the solutions as $x \rightarrow \infty$. We then would like to figure out how to connect the solution at the origin to the solution at ∞ .

Solution near the origin

Near the origin, we might hope to use a Taylor series approximation; however, a straightforward exercise convinces you that this will not work. Let us use the ansatz

$$y = a_0 + a_1 x + a_2 x^2 + \dots \quad (8.2)$$

and substitute this into Bessels equation. This gives

$$x^2(2a_2 + 3a_3 x + \dots) + x(a_1 + 2a_2 x + 3a_3 x^2 + \dots) + (x^2 - \nu^2)(a_0 + a_1 x + a_2 x^2 + \dots) = 0. \quad (8.3)$$

To solve this equation we must equate the coefficient of every power of x to zero. The lowest power x^0 has the coefficient

$$-\nu^2 a_0 = 0; \quad (8.4)$$

For x we have

$$a_1 - \nu^2 a_1 = 0, \quad (8.5)$$

and for x^2 we have

$$2a_2 + 2a_2 + a_0 = 0. \quad (8.6)$$

Thus the only solution is that *all* of the coefficients are zero. A solution by Taylor series expansion does not exist. Another idea for how to solve this was invented in 1866 by Fuchs. He suggested looking for a solution of the form

$$y(x) = x^\rho F(x) \quad (8.7)$$

where ρ is a number and $F(x)$ is a function with a Taylor series expansion. To see why this works, let's guess

$$y = x^\rho(a_0 + a_1x + \dots) \tag{8.8}$$

as a solution to Bessels equation. This gives

$$a_0\rho(\rho - 1)x^\rho + \dots + a_0\rho x^\rho + \dots + (x^2 - \nu^2)a_0x^\rho = 0. \tag{8.9}$$

The lowest power of x in the equation is x^ρ —its coefficient is $a_0(\rho^2 - \nu^2)$. Hence this coefficient can vanish with $a_0 \neq 0$ as long as

$$\rho = \pm\nu. \tag{8.10}$$

We therefore have two possible solutions, namely

$$y_\pm = x^{\pm\nu}(a_0 + a_1x + a_2x^2 + \dots) \tag{8.11}$$

where a_1, a_2, \dots can be related to a_0 by looking at higher terms in the equation.

Since this is a linear equation, the general solution is given by

$$y = c_1y_+(x) + c_2y_-(x). \tag{8.12}$$

This is a second order differential equation and hence requires two independent constants to solve the initial value problem—and this is what the solution reflects.

For integer ρ one of the solutions that is usually defined is

$$y_+ = \sum_{m=0}^{\infty} \frac{(-1)^m}{m!\Gamma(m + \rho + 1)} \left(\frac{x}{2}\right)^{2m+\rho}, \tag{8.13}$$

where here $\Gamma(x)$ is the Gamma-function that we discussed in our study of Stirlings formula (ie it is the factorial function). People usually denote this function $J_\rho(x)$.

The second solution $y_-(x)$ is usually taken as

$$Y_\rho(x) = \frac{J_\rho(x) \cos(\rho\pi) - J_{-\rho}(x)}{\sin(\rho\pi)}. \tag{8.14}$$

Why this complicated linear combination? We will see this below.

Solution near infinity

Now we need to examine the solutions as $x \rightarrow \infty$. To think about what this will look like, let us rewrite Bessel's equation slightly as follows:

$$\frac{d^2y}{dx^2} + \frac{1}{x} \frac{dy}{dx} + \left(1 - \frac{\nu^2}{x^2}\right)y = 0. \quad (8.15)$$

We recognize this as the equation of a damped harmonic oscillator. The damping constant is $1/x$, and the squared frequency is $(1 - \frac{\nu^2}{x^2})$. The squared frequency approaches unity as $x \rightarrow \infty$, whereas the damping constant $1/x$ decreases. Hence we expect an oscillatory damped solution. With this in mind, we make the ansatz

$$y(x) = e^{S(x)}, \quad (8.16)$$

and insert this into Bessel's equation. Note that the ansatz is on one hand exact—we are not approximating anything, just replacing the function $y(x)$ with $S(x)$; and on the other hand this reflects our expectation for a damped oscillatory solution. Inserting this ansatz into Bessel's Equation gives

$$S'' + S'^2 + \frac{1}{x}S' + \left(1 - \frac{\nu^2}{x^2}\right) = 0. \quad (8.17)$$

This is now a nonlinear equation for $S(x)$, which at first seems worse than the initial case. But then we remember our affection for dominant balance! We thus proceed to solve this equation by dominant balance. You can experiment with the different possibilities, but I claim the dominant balance is

$$S'^2 + 1 = 0, \quad (8.18)$$

leading to

$$S' = \pm i. \quad (8.19)$$

You can see that this is self consistent because clearly $S'' = 0$, and the other neglected terms are also small.

The solution implies $S = ix + \text{constant}$, so that

$$e^S = Ae^{ix} \quad (8.20)$$

is an oscillatory solution, as we expected. But what of the damping? To derive the damping term, we must look for a correction to the dominant balance.

Let's assume

$$S' = \pm i + W(x), \quad (8.21)$$

and use this in the equation. We then derive

$$W' + 2 \pm iW + W^2 + \frac{1}{x}(\pm i + W) - \frac{\nu^2}{x^2} = 0. \quad (8.22)$$

Now since W is a correction to the full solution, we know $W \ll i$. Thus we can guess the dominant balance

$$2 \pm iW + \frac{1}{x} \pm i = 0, \quad (8.23)$$

or

$$W = -\frac{1}{2x}. \quad (8.24)$$

This balance is indeed consistent, as can be verified. Thus we have

$$S' = \pm i + \frac{1}{2x} \quad (8.25)$$

which integrates to

$$S = \pm ix + \frac{1}{2} \log(x). \quad (8.26)$$

Hence we have that

$$y = e^S = \frac{1}{\sqrt{x}} e^{\pm ix}. \quad (8.27)$$

This is the decaying exponential we were looking for! Thus these are the two independent solutions for $x \rightarrow \infty$. If you would prefer real valued solutions then just take

$$y_1 = \frac{A}{\sqrt{x}} \cos(x) \quad (8.28)$$

and for imaginary solutions,

$$y_2 = \frac{A}{\sqrt{x}} \sin(x). \quad (8.29)$$

Connections

Now we are faced with a connection problem: how do we connect an arbitrary solution at the origin with an arbitrary solution at ∞ ? We could of course proceed numerically.

Somewhat remarkably, there are explicit formulas connecting the behavior at ∞ to that at the origin for the classical special functions of applied mathematics. For the Bessel function, if we define J_ν as the solution to Bessel's equation corresponding to the series solution starting with x^ν and then define

$$Y_\nu = \frac{2}{\pi} \left(\log \frac{x}{2} + \gamma \right) J_\nu(x) - \frac{1}{\pi} \left(\frac{x}{2} \right)^{-\nu} \sum_{k=0}^{n=1} \frac{(n-k-1)!}{k!} \left(\frac{x^2}{4} \right)^k - \frac{1}{\pi} \left(\frac{x}{2} \right)^\nu \sum_k a_k \left(\frac{-x^2}{4} \right)^k \quad (8.30)$$

(where ν is assumed to be an integer and γ is the Euler-Mascheroni constant), then the asymptotic relations at large x are:

$$J_\nu \sim \left(\frac{2}{\pi x} \right)^{1/2} \cos(x - \pi\nu/2 - \pi/4), \quad (8.31)$$

and:

$$Y_\nu \sim \left(\frac{2}{\pi x} \right)^{1/2} \sin(x - \pi\nu/2 - \pi/4). \quad (8.32)$$

These are explicit formulas!!! You should now find the subject of special functions much more interesting: the question is how did people manage to find these explicit formulas? Examining the solutions from above, it is very disappointing that there is no way to see from one of the solutions how it connects to the other. This is of course also the case with the Bessel functions. But some very clever people figured out how to work this out.

Perhaps even more remarkably, Fig. 8.1 compares the asymptotic formula for the Bessel function with the actual function. Note that they agree all the way to $x \approx 1$!

Unfortunately, although the methods originally used work for all special functions (and in particular all special functions contained in the class of the hypergeometric function), they are not readily generalizable to all problems. One would really like an analytic procedure or at least an intellectual paradigm for connecting one behavior of a differential equation with another behavior that occurs in a different part of parameter space. This is what will occupy us a little later, trying to figure out whether we can do this in a robust and satisfying way.

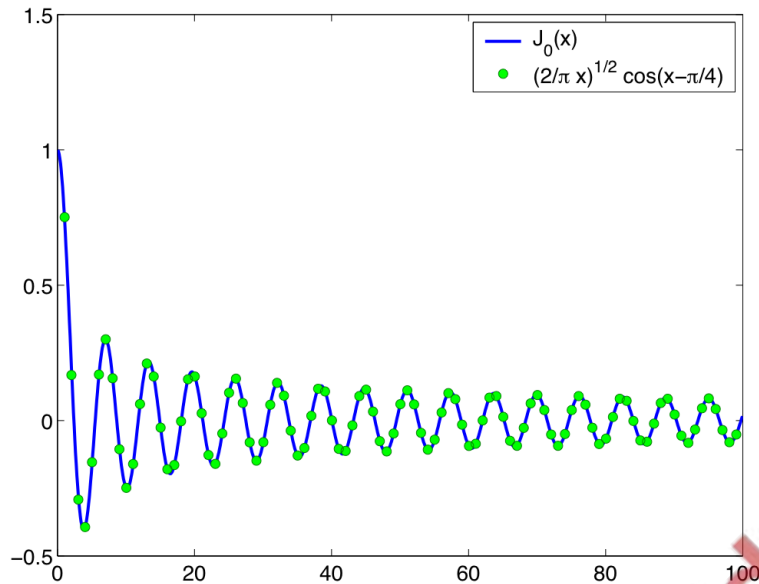


Figure 8.1. Comparison of asymptotic formula for Bessel's function with $J_0(x)$.

Numerical algorithms for Bessel Functions

Every serious mathematical package now contains a suite of special function software for computing Bessel functions. Many of the algorithms that these packages use rely heavily on the ideas we have described already for characterizing Special Functions. As an example, below we reproduce the subroutine for computing the special function $y_1(x)$ (the singular Bessel function) from *Numerical Recipes*:

```

1  FUNCTION bessy0(x)
2  REAL bessy0,x C USES bessj0
3  Returns the Bessel function Y0(x) for positive x.
4  REAL xx,z,bessj0
5
6  DATA p1,p2,p3,p4,p5/1.d0,-.1098628627d-2,.2734510407d-4,
7  * -.2073370689d-5,.2093887211d-6/, q1,q2,q3,q4,q5/-.1562499995d-1,
8  * .1430488765d-3,-.6911147651d-5,.7621095161d-6,-.934945152d-7/
9  DATA
10 r1,r2,r3,r4,r5,r6/-2957821389.d0,7062834065.d0,-512359803.6d0,
11 * 10879881.29d0,-86327.92757d0,228.4622733d0/,
12 * s1,s2,s3,s4,s5,s6/40076544269.d0,745249964.8d0,
13 * 7189466.438d0,47447.26470d0,226.1030244d0,1.d0/
14 if(x.lt.8.)then
15   y=x**2
16   bessy0=(r1+y*(r2+y*(r3+y*(r4+y*(r5+y*r6))))/(s1+y*(s2+y
17   * *(s3+y*(s4+y*(s5+y*s6))))+.636619772*bessj0(x)*log(x)
18 else
19   z=8./x y=z**2
20   xx=x-.785398164

```

```

21 bessy0=sqrt(.636619772/x)*(sin(xx)*(p1+y*(p2+y*(p3+y*(p4+y*
22 * p5))))+z*cos(xx)*(q1+y*(q2+y*(q3+y*(q4+y*q5))))
23 end if
24 return
25 END

```

Note that this routine uses both of the expansions we have already talked about, matched to an arbitrary point in between. This matching point is supposedly chosen to enforce some error tolerance criterion.

8.2.2 A Less famous linear second order ordinary differential equation

Now we consider another example: We now consider the solutions to the differential equation

$$y'' = \frac{y}{x^3}. \quad (8.33)$$

This is also a linear equation, but one without a name. You will note that there is a singularity in the solution at the origin that is even stronger than before. In fact, it turns out that the singularity here is so strong that even the trick we used above by Fuchs does not work. If we make the substitution $x = 1/t$, the equation transforms to

$$t^2 y'' + 2ty' = ty. \quad (8.34)$$

In this equation, the point $x = 0$ was mapped to $t = \infty$. You see when $t \rightarrow \infty$ that this equation has similar features to Bessel's equation, and we will indeed see below that the ansatz $y = e^S$ works well in this case.

Indeed, there is a classification scheme that is often used to understand what type of approximation is needed in a linear second order differential equation. Consider an equation of the general form

$$y'' + p(x)y' + q(x)y = 0. \quad (8.35)$$

The rule is that if $p(x)$ has a singularity that is no stronger than $1/(x - x_0)$ as $x \rightarrow x_0$ and $q(x)$ has a singularity that is no stronger than $1/(x - x_0)^2$ as $x \rightarrow x_0$, then the point at x_0 is called a *regular singular point*. In this case a technique like Fuch's expansion works well. If the singularity is stronger than this (for example q diverges like $1/(x - x_0)^3$), the point x_0 is called an *irregular singular point*, and then the ansatz $y = e^S$ is needed. Indeed our example has an irregular singular point at $x = 0$, but we see from the transformed equation that it is regular singular point at $t = 0, x = \infty$; hence we expect the Fuchs guess to work near $x = \infty$.

Try it!

The solution near the origin

To analyze the solutions near $x = 0$, we substitute $y = e^{S(x)}$ equation (8.33) thus becomes

$$S'' + S'^2 = \frac{1}{x^3}. \quad (8.36)$$

This is a nonlinear differential equation for $S(x)$, and we solve it using dominant balance. Trying the various choices demonstrates that the only consistent balance in the limit $x \rightarrow 0$ is

$$S'^2 = x^{-3}. \quad (8.37)$$

This implies

$$S = \pm 2x^{-1/2}; \quad (8.38)$$

the neglected term $S'' = O(x^{-5/2})$, which is indeed smaller as $x \rightarrow 0$ than the terms that we have kept. Thus, to leading order, the solution near $x = 0$ has the behavior

$$y \sim e^{\pm 2/\sqrt{x}}. \quad (8.39)$$

To derive corrections to this behavior, we follow the procedure from above and make the ansatz

$$S = \pm 2/\sqrt{x} + W, \quad (8.40)$$

and then substitute this into equation (8.36). We have

$$W'' + \pm \frac{3/2}{x^{5/2}} + \left(\frac{1}{x^3} + \frac{\pm 2}{x^{3/2}} W' + W'^2 \right) = \frac{1}{x^3}. \quad (8.41)$$

The dominant balance here is

$$\pm \frac{3/2}{x^{5/2}} = \frac{\pm 2}{x^{3/2}} W', \quad (8.42)$$

so that

$$W' = \frac{3}{4x}, \quad (8.43)$$

or

$$W = \frac{3}{4} \log(x). \quad (8.44)$$

This implies that we have

$$y \sim x^{3/4} e^{\pm 2/\sqrt{x}}. \quad (8.45)$$

Deriving even further corrections

At the risk of beating a dead horse, we could continue this procedure and derive further corrections. If we did, we know that the next term in the expansion for W will be smaller than $\log(x)$ as $x \rightarrow 0$, and it doesn't seem like too much to ask that this term is a constant or smaller. If so, we might as well just write

$$y = e^S W(x) \tag{8.46}$$

with

$$S = \pm \frac{2}{\sqrt{x}} + \frac{3}{4} \log(x), \tag{8.47}$$

and then derive a differential equation for W . A little algebra then shows:

$$S''W + S'^2W + 2S'W' + W'' = \frac{W}{x^3}, \tag{8.48}$$

or

$$W'' + \left(\frac{3}{2x} - \frac{2}{x^{3/2}} \right) W' - \frac{3}{16x^2} W = 0. \tag{8.49}$$

We seek a solution of this with $W = 1 + \epsilon$, where ϵ is small. This gives the differential equation for ϵ that

$$\epsilon'' - \frac{2}{x^{3/2}} \epsilon' - \frac{3}{16x^2} \epsilon = 0. \tag{8.50}$$

As usual we need a dominant balance. Some trial and error shows that the only consistent balance as $x \rightarrow 0$ (maintaining the assumption that ϵ is small), is the second and third terms. Hence

$$\epsilon' \sim -\frac{3}{32} \frac{1}{x^{1/2}}, \tag{8.51}$$

so that

$$\epsilon = -3/16\sqrt{x}. \tag{8.52}$$

In fact, it is not hard to convince yourself that the series expansion for ϵ will be a series in powers of \sqrt{x} . If we write

$$W(x) = \sum_n a_n x^{n/2}, \tag{8.53}$$

then substituting into the equation for W yields

$$\sum_n \frac{n}{2} \left(\frac{n}{2} - 1 \right) a_n x^{n/2-2} + \frac{3}{2} \sum_n a_n \frac{n}{2} x^{n/2-2} - 2 \sum_n a_n \frac{n}{2} x^{n/2-5/2} - \frac{3}{16} \sum_n a_n x^{n/2-2} = 0. \quad (8.54)$$

Equating coefficients of $x^{n/2-2}$ gives

$$\frac{n}{2} \left(\frac{n}{2} - 1 \right) a_n + \frac{3n}{4} a_n - \frac{3}{16} a_n + a_{n+1} \frac{n+1}{2} = 0. \quad (8.55)$$

From this one can easily show that $a_{n+1} \sim n/4a_n$, so the radius of convergence of this series is zero! As promised, a non-convergent series!

But, does it work? Let's set initial conditions and compare the solutions. Figure 8.2 compares a solution to

$$y'' = \frac{y}{x^3} \quad (8.56)$$

with the asymptotic formula approximating $W = 1$. We have required that $y(1) = y'(1) = 1$. It is seen that the agreement is really quite excellent, especially at small x , where the approximation was derived. It does start to deviate from the numerical solution at large x ; judging from the slope on this log-log plot, it appears as if the solution is asymptoting to a function that is linear in x .

Behavior near infinity

Let us now examine the behavior at ∞ . We have already shown that by transforming $x = 1/t$, our equation becomes

$$t^2 y'' + 2ty' = ty, \quad (8.57)$$

so there is a regular singular point at ∞ . Therefore we are motivated to use Frobenius's idea and write $y \sim t^\rho$. This gives the indicial equation $\rho(\rho - 1) + 2\rho = 0$, which has solutions $\rho = 0, -1$. Hence according to this analysis, there are two asymptotic behaviors:

$$y \sim x \quad (8.58)$$

and

$$y \sim 1. \quad (8.59)$$

Let's derive this result through another method, by directly analyzing the equation: If we write $y \sim x^\beta$, and seek a solution that works at large x , we find that

$$x^{\beta-2} \beta(\beta - 1) = x^{\beta-3}. \quad (8.60)$$

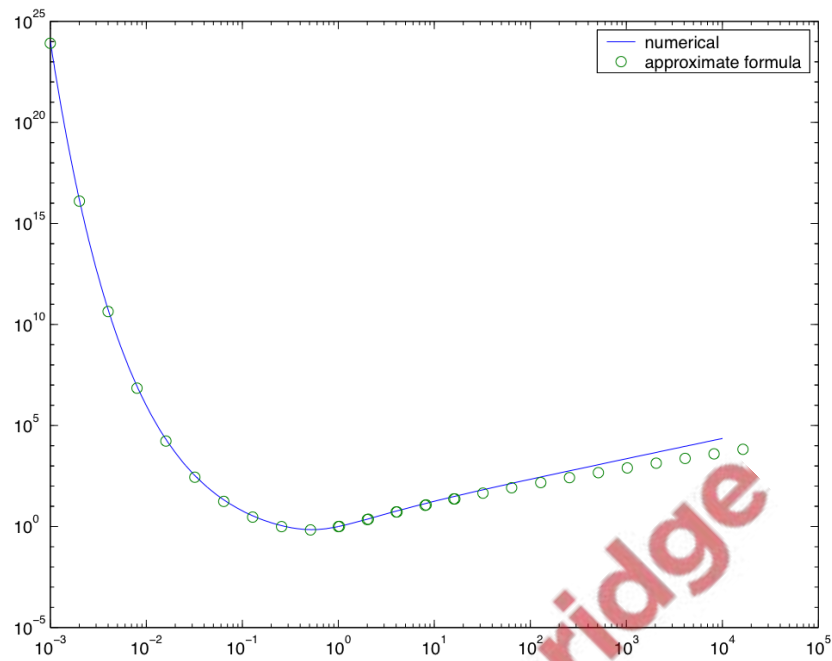


Figure 8.2. Comparison of numerical solution to $y'' = y/x^3$ with $y(1) = y'(1) = 1$ with first term in asymptotic series (solid dots).

The only solutions to this equation at leading order satisfy $\beta = 0, 1$. For these solutions to be consistent at large x it is good to check the next lower order terms: If we write

$$y = ax + z(x), \tag{8.61}$$

then $z(x)$ satisfies

$$z'' = x^{-2} \tag{8.62}$$

so that $z = \log(x)$. Since asymptotically $\log(x)$ is smaller than x , this expansion is asymptotically consistent. Similarly, for the expansion $y = 1 + z(x)$ we find that $z(x) = 1/x$. Note that the expansion for the linear function picked up a logarithm: we should have expected this. The indicial equation has two solutions which differ by an integer. This is the situation where we would expect a logarithm.

Summary: The Connection Problem

To summarize, we have now demonstrated how to compute expansions for the solution around $x = 0$ and $x = \infty$. Around $x = \infty$ the equation has a regular singular point and therefore we expect the expansion to converge for all $t = 1/x$ (there are theorems that guarantee a convergent expansion). Around $x = 0$, there is an essential singularity and

we therefore have a non-convergent expansion, which we have shown nonetheless works extremely well.

Near both $x = 0$ and $x = \infty$ we have shown that there are two solutions: as $x \rightarrow \infty$, we have that $y \sim ax + b$, whereas for $x \rightarrow 0$, we know that

$$y \sim x^{3/4}(ce^{2/\sqrt{x}} + de^{-2/\sqrt{x}}). \quad (8.63)$$

Since the equation is second order, given a, b we have a unique c, d , whereas conversely given a c, d we have a unique a, b .

This raises a fairly difficult and deep question about the structure of solutions to an equation: we call this *the connection problem*. Namely how does one connection one solution around one region to another solution around another region? We encountered this briefly before in our discussions of the solution to the differential equation

$$y' + y^5 = \frac{1}{1 + x^2}. \quad (8.64)$$

There we saw that there was the possible dominant balance between the first and third terms that was only consistent for a specific initial condition. We addressed the question of 'what this initial condition is' using the computer. This type of question turns out to be rather general, and we can see an example right here:

It is easy to see that for a *generic* initial condition at ∞ , the solution will diverge at the origin like $e^{2/\sqrt{x}}$, since this term dominates the other solution near the origin. On the other hand, for a *generic* initial condition at $x = 0$, the solution at ∞ will be dominated by $y \sim x$. These generic possibilities were indeed uncovered in the (random) example we have worked out already, Fig. 8.6.

On the other hand, it is clear that there does exist the possibility of there being a solution which asymptotes to a constant at ∞ , or a solution that vanishes at the origin. How can we find these solutions? Let us focus on finding a solution which asymptotes to a constant at ∞ . To find a solution, we first note that given any two initial conditions (be they c, d or $y(1), y'(1)$, or whatever), then clearly

$$a = a(y(1), y'(1)) \quad b = b(y(1), y'(1)). \quad (8.65)$$

To find a solution which asymptotes to a constant we must find an initial condition so that $a(y(1), y'(1)) = 0$. For any $y'(1)$, there exists a value of $y(1)$ such that this is satisfied (we hope)! The reason we have a freedom in picking $y'(1)$ is that the equation itself is linear, so that multiplying any solution by a solution gives a solution: hence we really only have one degree of freedom. Thus, the solutions that asymptote to a constant are in some real sense unique, and we must figure out how to find them

How can we find this solution? Using Newton's method! Let us pick $y(1) = 1$, and vary $y'(1) = \beta$ to find the requisite solution. We therefore choose $f(\beta) = y'(\infty)$. This

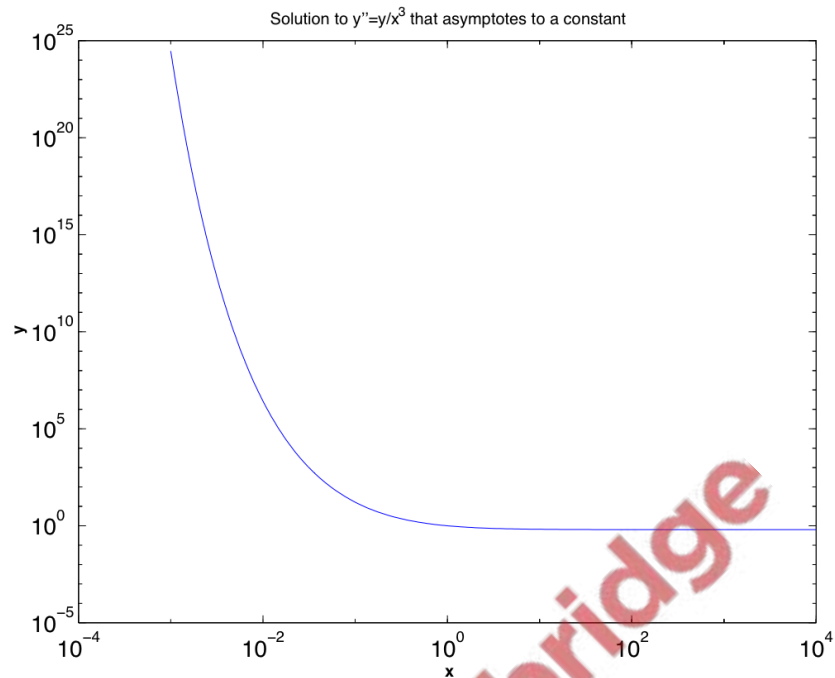


Figure 8.3. Solution of $y'' = y/x^3$ that asymptotes to a constant at ∞ .

function cannot be computed analytically (such a formula would require solving the equation exactly!) but it can be evaluated numerically. In addition, the derivative

$$f'(\beta) \approx \frac{f(\beta + \epsilon) - f(\beta)}{\epsilon} \quad (8.66)$$

can also be evaluated numerically. By taking a guess for β and then using Newton's method, we can therefore find the root.

Matlab Code for this problem

```

1
2 y0=1;
3 yp0=0;
4 eps=1e-8;
5 for i=1:10
6 [t,y]=ode45('hmkprob',[1 10000],[y0 yp0]);
7 [t,y2]=ode45('hmkprob',[1 10000],[y0 yp0+eps]);
8 nn=size(y,1);
9 n2=size(y2,1);
10 yp0=yp0-y(nn,2)/(y2(n2,2)-y(nn,2))*eps
11 end

```


8.3 A nonlinear boundary value problem

We now consider a nonlinear problem in which connecting a solution in one region to another is required.

In the linear problems described above, we were interested in solutions which obeyed particular boundary conditions at (say) $x = 0, \infty$. We constructed local solutions valid at each of these points, and then discussed how to numerically match the solutions together. In the case of linear equations, it is often possible to perform this matching explicitly by constructing an integral representation of the solution (This is the origin of the magical formulas about Bessel functions we discussed above!).

On the other hand, the great advantage of the numerical matching procedure we discussed is that it also easily generalizes to nonlinear problems. Here we discuss an example of such a problem.

Consider the equation

$$u^3 u''' + u^3 = -3/4 + 7/4u \quad (8.67)$$

satisfying the boundary conditions $u(\infty) = 1/2$ and $u(-\infty) = 1$.

Note that both $u = 1, 1/2$ are exact solutions to the equation. Thus what we are asking is whether there is a solution that approximates one exact solution at positive infinity and another exact solution at negative infinity. The solution we are constructing is not without physical significance. It arises in the study of fluid flowing down an inclined surface, as occurs when you paint a wall.

To analyze this question, we first need to ask: how many conditions do the boundary conditions correspond to? To solve this, it is necessary to *linearize* the solution around both $u = 1$ and $u = 1/2$ to find out how many degrees of freedom must be specified to get convergence to this solution. Namely, writing

$$u = 1 + \delta \quad (8.68)$$

implies that δ obeys

$$\delta''' + 3\delta = 7/4\delta, \quad (8.69)$$

so that

$$\delta''' = -5/4\delta. \quad (8.70)$$

Guessing

$$\delta \sim e^{Sx} \quad (8.71)$$

implies that

$$S^3 = -5/4. \quad (8.72)$$

There is one negative root of this equation $S = -(5/4)^{1/3}$ and two positive roots; since $u \rightarrow 1$ as $x \rightarrow -\infty$ we require the coefficient of the negative root to vanish, so this boundary condition corresponds to one condition.

Similarly, writing

$$u = 1/2 + \delta \tag{8.73}$$

implies

$$\frac{1}{8}\delta''' + \frac{3}{4}\delta = \frac{7}{4}\delta, \tag{8.74}$$

so that

$$\delta''' = 8\delta. \tag{8.75}$$

The roots of this are again of the form

$$\delta \sim e^{Sx} \tag{8.76}$$

where $S^3 = 8$. Here we want to zero the coefficient of the positive root, and there is only one positive root. Hence, this boundary condition also corresponds to one condition.

We have therefore found that the boundary conditions correspond to two conditions on the equations. Since the equation is third order there is a one parameter family of solutions left. This family is easily interpreted: since the equation is invariant under spatial translations (i.e. when $u(x)$ is a solution so is $u(x + a)$ for any a the family of solutions just corresponds to uniform translation.

There are many ways of implementing these conditions; what follows is one method. Probably the simplest thing to do is start at either large positive or negative x , implement the condition at that position, and then integrate out to the other condition. Since only one condition is then left to satisfy one should only have a one parameter family of solutions to play with. I started out at $x = -10$ with the family of initial conditions

$$u(x) = 1 + c \exp(\alpha x) \cos(\beta x), \tag{8.77}$$

where $\alpha = 0.5386$ and $\beta = 0.9329$, and then integrate out to positive infinity. There is a one parameter family of solutions of this type (for different values of c) and thus one would expect that at some special value of c one will get a solution

My first step was to integrate the equation for a bunch of different c 's, and see what came out. Figure 8.4 shows a sweep of ten solutions from $c = 0.02$ to $c = 0.12$ stepping by 0.01. It is seen there is a transition which occurs near

$c = 0.06$. The transition occurs near $u \sim 1/2$, the other asymptotic solution! Therefore I guess that the solution occurs right near there and try to bracket it more closely.

To do this, I note that at this transition, the solutions change from solutions which die immediately (at small x) to solutions which die after a long time (at large x). I therefore wrote a code which tries to bracket these two extreme cases.

In fact there are lots of possible c 's, a countable set.

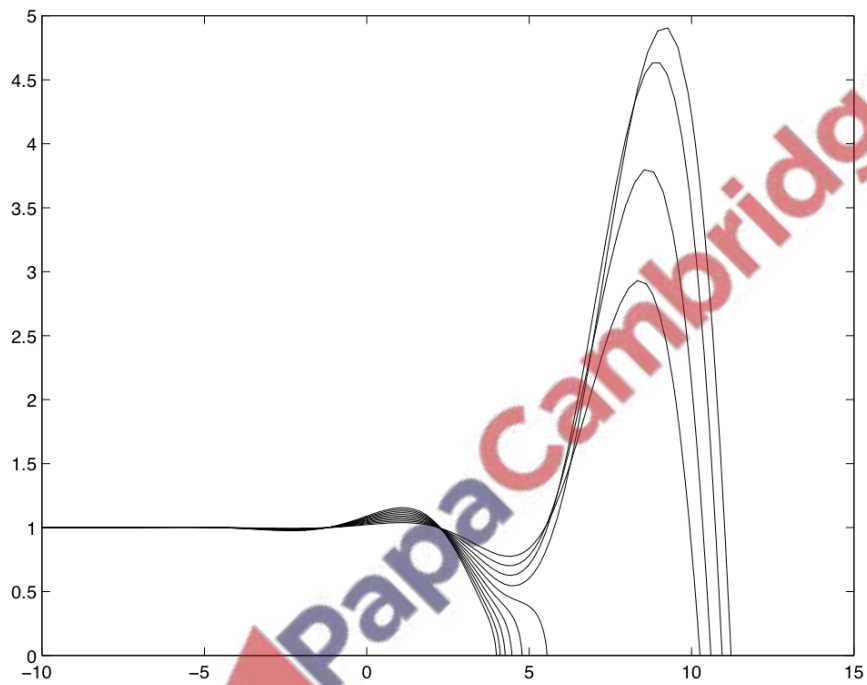


Figure 8.4. Sweeping through the different c values, as described in the text.

The bracketing works by noting that the solutions which go above 0.5 go to zero at a much larger $x(u = 0)$ than those which go below 0.5. Hence I picked one initial condition c_+ for which the solution went above 0.5 and another c_- for which the solution went below 0.5. I then considered $c = 0.5(c_- + c_+)$. If this solution had a $x(u = 0; c) > x(u = 0; c_+)$ larger than that for c_+ then we let $c = c_+$. Otherwise we let $c = c_-$. Iterating this procedure leads to convergence onto a good solution. The different iterations are shown in the figure below:

The solution extracted as the limit of the above sequence is shown in the next figure:

Finally, the matlab code used to produce these graphs is included below:

```

1 hold off;
2 fid=fopen('data.txt','w');
3 c=0.06
4 cup=0.06;
5 cdown=0.08;
6 tup=11.;
7 for i=1:100
8 c=0.5*(cdown+cup);
9 init;
10 [t,y]=ode45('hmk',[-10,40],[u up upp]);
11 plot(t,y(:,1));
12 n=size(t,1);
13 if(t(n) < tup)
14 cdown=c
15 else
16 cup=c
17 tup=t(n);
18 end
19 hold on;
20 end

```

This program calls the program `init.m`, which has the initial conditions. This program is here.

```

1 a=0.5386;
2 b=0.9329;
3 x=-10;
4 u=1+c*exp(a*x)*cos(b*x);
5 up=c*a*exp(a*x)*cos(b*x)-c*b*exp(a*x)*sin(b*x);
6 upp=c*a*a*exp(a*x)*cos(b*x)-...
7     2.0*c*b*a*exp(a*x)*sin(b*x)-c*b*b*exp(a*x)*sin(b*x);

```

The function to solve the equations is here

```

1 function dy =hmk(t,y)
2 dy=zeros(3,1);
3 dy(1)=y(2);

```

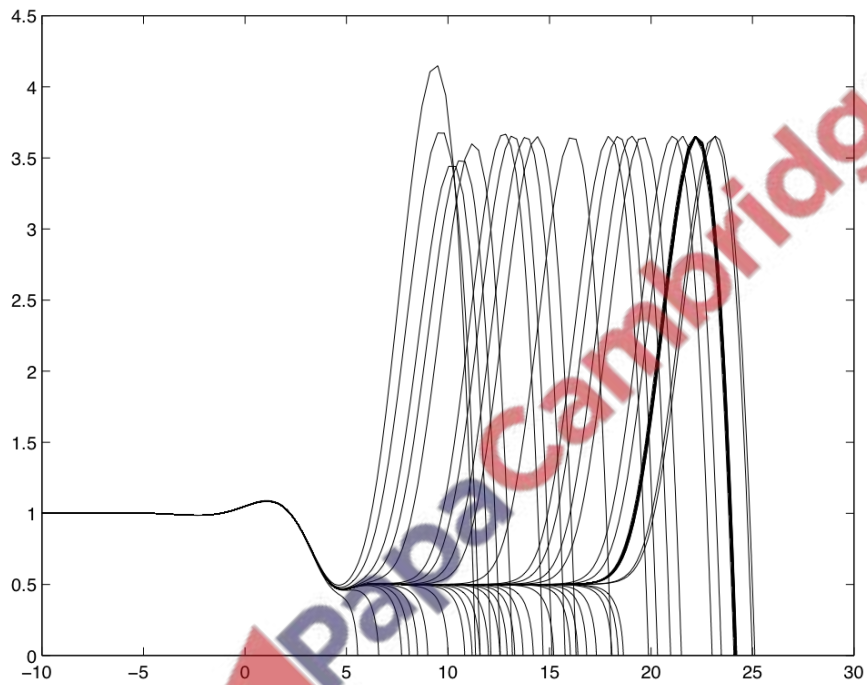


Figure 8.5. Bracketing procedure, implemented.

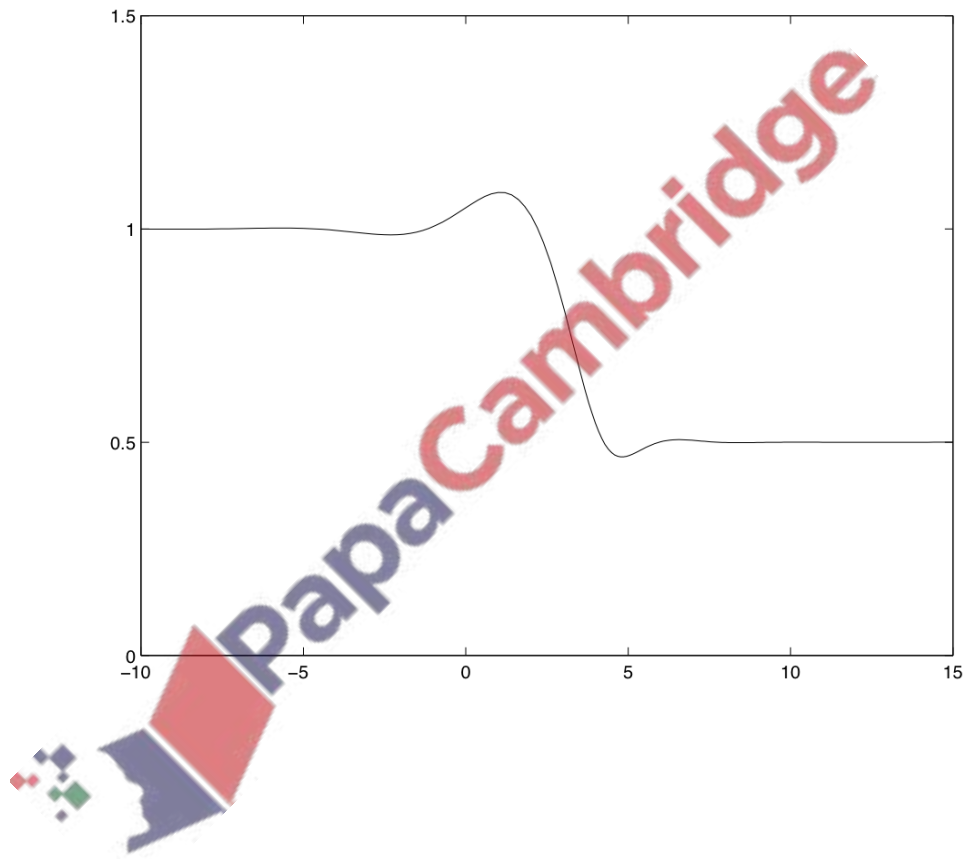


Figure 8.6. The solution!

```

4 dy(2)=y(3);
5 dy(3)=(-3/4+7/4*y(1))-y(1)^3;
6 dy(3)=dy(3)/y(1)^3;

```

8.4 Spatially dependent dominant balances

In the previous section, we discussed numerical methods for solving the connection problem. These work and are very practical but we would like analytical methods for doing this as well. It is infuriating that we cannot see from a globally convergent series expansions of special function how they in fact match on to each other. Analytical methods are often practically useful and give mechanistic understanding into how connections work.

There are however a class of problems where matching can be done explicitly and completely. Roughly speaking, one can make a matching argument for an equation that has at least three terms with (potentially) different orders of magnitude. In one region two of these term give a dominant balance, and in another region another pair of terms balance. The method of matched asymptotic expansions is a very effective method for matching these leading order solutions on to each other. We will see that doing this effectively requires introducing a small parameter, whose function is basically to cleanly separate the regions of validity of the two different leading order balances.

Historically the essential idea of the analysis we are about to outline was discovered by Prandtl in the early 1900s, in the context of understanding the lift and drag on airplane wings, and we will discuss his method in a later section.

Before outlining formal rules, it is best to start with an explicit example.

8.4.1 An Example of Carrier

Consider the equation

$$\epsilon u'' - (2 - x^2)u = -1, \quad (8.78)$$

with the boundary conditions $u(1) = u(-1) = 0$. Let us first consider the solution in the limit $\epsilon \rightarrow 0$. Naively, we expect the very small prefactor to suppress the first term in this limit. Neglecting it entirely, we arrive at an equation

$$u \equiv u_0 = \frac{1}{2 - x^2}. \quad (8.79)$$

The difficulty with this is of course that now we have that $u(\pm 1) = 1$, which contradicts our desired boundary conditions!

What does this mean? Evidently, we were not correct to ignore the $\epsilon u''$ term everywhere. This is unsurprising, since we were allowed to impose two boundary conditions upon the solution only because the equation is second order. By deleting the $\epsilon u''$ term,

we have reduced the order of the differential equation, but not the number of boundary conditions that the solution must satisfy!

To see what is going on let's first try a slightly simpler problem that has similar features:

$$\epsilon u'' - 2u = -1, \quad (8.80)$$

with the boundary conditions that $u(\pm 1) = 0$. For this problem we can make the same argument as above: in the limit where $\epsilon \rightarrow 0$, we see that roughly $u = 1/2$, but again, we have that boundary condition problem. But now we can solve the problem exactly: the general solution is

$$u = Ae^{\sqrt{2/\epsilon}x} + Be^{-\sqrt{2/\epsilon}x} + \frac{1}{2}. \quad (8.81)$$

If we impose the boundary conditions, we need $A = B$ (since the solution is symmetrical about the origin), and then we find that

$$A = B = \frac{-1}{4 \cosh(2/\sqrt{\epsilon})}. \quad (8.82)$$

If we now examine the solution, we see that it consists of three "pieces": over most of the solution region, the solution indeed satisfies $u \approx 1/2$. However, near each of the boundaries there are rapid transition layers where the solution departs from this value to satisfy the boundary conditions. The width of the transition layers is $O(\sqrt{\epsilon})$. Outside of the transition layer the correction to $u = 1/2$ are exponentially small. Inside the transition layer, the solution $u(x)$ varies by an $O(1)$ amount over an $O(\sqrt{\epsilon})$ interval, so has an $O(1/\sqrt{\epsilon})$ gradient, and an $O(1/\epsilon)$ second derivative. It follows that the $\epsilon u''$ term participates in the leading order balance within the transition layers.

Now returning to our original problem, the answer must be of the form

$$u = u_0(x) + p(X_1) + q(X_2), \quad (8.83)$$

where $u_0 = 1/(2 - x^2)$ as previously determined, and p and q are important near $x = 1$ and $x = -1$, respectively, but exponentially small away from their respective boundaries, and X_1 and X_2 are local variables tailored to the transition layer regions. We define: $X_1 = (x + 1)/\epsilon^\alpha$ and $X_2 = (1 - x)/\epsilon^\alpha$, so that $X_1 = 0$ coincides with the left boundary $x = -1$, and X_1 varies by an $O(1)$ amount, when x varies by an $O(\epsilon^\alpha)$ amount. The putative width of the boundary layer is therefore $O(\epsilon^\alpha)$, and we want to determine the exponent α . Plugging this solution form into the original equation we obtain:

$$\epsilon u_0'' - (2 - x^2)u_0 + 1 + \epsilon^{1-2\alpha}p'' - (2 - x^2)p + \epsilon^{1-2\alpha}q'' - (2 - x^2)q = 0. \quad (8.84)$$

Now, if we examine this equation near $x = -1$, we can neglect the q terms (by assumption). If we choose $\alpha = 1/2$, then to $O(\epsilon)$ we have that

$$p'' - p = 0,$$

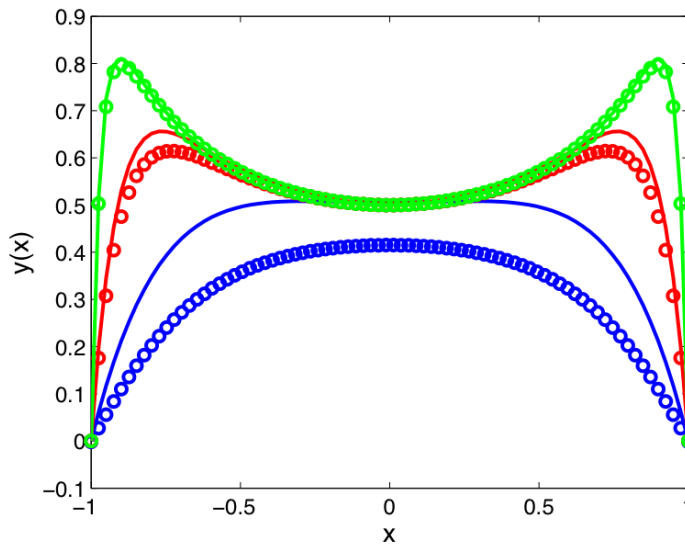


Figure 8.7. Comparison of numerical solutions with the boundary layer formulas for $\epsilon = 0.1, 0.01, 0.001$ in blue, red and green respectively. The solid curves are the numerical solutions of the differential equation, and the circles are the asymptotic solution.

so that

$$p = A_1 e^{X_1} + B_1 e^{-X_1}.$$

We have constructed the solution so that p only matters near the left boundary. Therefore, we must choose $A_1 = 0$ to eliminate the exponentially growing contribution. A similar argument near $x = 1$ yields $q = B_2 e^{-X_2}$. Thus we have

$$u = B_1 e^{-X_1} + B_2 e^{-X_2} + u_0.$$

We can now choose the constants to satisfy the boundary conditions: We have that $B_1 + u_0(-1) = 0$, and $B_2 + u_0(1) = 0$. Thus $B_1, B_2 = -1$. In Figure 8.7 we compare the asymptotic expression with the full numerical solution of the equation, showing the convergence to a form in which there is rapid variation close to two ends of the interval.

8.5 Matched Asymptotic Expansions

The example of Carrier demonstrated that with a little ingenuity one can develop an approximate solution to a hard differential equation, in the limit as a small parameter approaches zero. The highlight of the derivation was the fact that first, although the differential equation appeared complex, one could break the solution into different regions, and in each independent region the equation could be solved. Then, by matching the regions together, we arrived at a complete picture of the solution.

Although the procedure that we followed is admittedly heuristic, in the 1950s Proudman and Pearson, and a little later van Dyke, figured out how to formalize the procedure to make it work with a well defined set of rules. Here our goal is to (a) outline when and why the procedure works, and (b) set down the formal rules. We also discuss a little of the history of where this idea first arose, during the study of the drag on airplane wings. The real discoverer of this idea was Ludwig Prandtl, one of the greatest applied mathematicians of the last century.

8.5.1 When it works

To make a very rough generalization, there are two ingredients that must be satisfied for matching to work:

First, the differential equation must be composed of several independent terms, which have different orders of magnitudes in different regions of the solution. Hence, different “dominant balances” will hold in different regions of the solution. Hopefully, the dominant balances will give simple equations of the type that can be exactly solved; however even if this is not the case, the method of matched asymptotics is often useful and fruitful.

Second, a good matching argument requires that the *typical length scales* characterizing two different regions differ by orders of magnitude - that is, one of the regions is much thinner than the other. The more well-separated the length scales of the solutions are the better chance we will have of achieving a good match. Technically, the match is usually carried out in the limit where the ratio of the two length scales vanishes.

Of course, if we have a differential equation where a region of the solution completely disappears as a small parameter (the ratio of the two length scales) vanishes, we should expect this will be a rather drastic change in the nature of solutions to the equation. We should not expect that the solutions depend in a smooth way on this small parameter. Indeed, often times, the class of problems where good matching arguments can be made are called “singular perturbation” problems for that reason.

Typically these problems have a small parameter multiplying the highest derivative term in the equation, for instance consider as a prototypical problem the linear differential equation:

$$\epsilon y^{(n)} + a_1(x)y^{(n-1)} + \cdots + a_{n-1}(x)y = 0. \quad (8.85)$$

Clearly, when $\epsilon = 0$ the solutions to the equation are vastly different from the solutions for small but finite ϵ : indeed, when ϵ is finite, we need another boundary or initial condition to uniquely identify a solution. The evidence of the example studied in the previous section suggests that over most of the interval, we may delete the first term. However, the highest order derivative may be non-negligible within certain transition layer regions. If, in anticipation of there being a transition layer region at $x = x_0$ we introduce a local variable $X = (x - x_0)/\epsilon^\beta$, with ϵ^β the width of the transition layer, the

equation becomes

$$\epsilon^{1-n\beta}y^{(n)} + \epsilon^{-(n-1)\beta}a_1(x_0 + \epsilon^\beta X)y^{(n-1)} + \cdots + a_{n-1}(x_0 + \epsilon^\beta X)y = 0. \quad (8.86)$$

Clearly the two biggest terms in the equation (assuming β is positive) are the first two, and these two terms balance if $\beta = 1$. Moreover, supposing that the coefficient functions $\{a_n(x)\}$ vary on $O(1)$ length-scales, each can be approximated in the transition layer region by constants $\{a_n(x_0)\}$. Hence we have the possibility that in the transition layer region the dominant balance is

$$y^{(n)} + a_1(x_0)y^{(n-1)} = 0.$$

This equation can be solved exactly, and its solution can be matched onto the rest of the solution. Note that in order to perform this type of analysis we must correctly locate the transition layer - that is the point x_0 . This depends upon the boundary conditions, and can only in general be found, when the outer-layer solution (i.e. the solution of the equation formed by omitting the first term) is known. In the next section we will show how to locate the transition regions, and perform matching for the simplest non-trivial equation of the type (8.86).

8.5.2 An Example from Bender and Orszag

A particularly simple example of this problem is the second order equation

$$\epsilon y'' + a(x)y' + b(x)y = 0. \quad (8.87)$$

We will go through this carefully to understand how and why the matching works, and then draw general conclusions.

Let us assume to start that we want to solve the above equation in the interval $0 \leq x \leq 1$, and that neither a nor b have any singularities. We will take boundary conditions $y(0) = A$ and $y(1) = B$. In the limit $\epsilon \rightarrow 0$, the equation becomes (to leading order)

$$a(x)y' + b(x)y = 0,$$

so that

$$y = y_0(x) \equiv \exp\left(-\int^x \frac{b(x')}{a(x')} dx'\right). \quad (8.88)$$

We call this the *outer solution*, and expect it to be valid outside of any transition layers in which the solution varies over some extremely small length scale. If we assume that this is the leading order term in an expansion in powers of ϵ , then if we write $y = \sum y_n \epsilon^n$, then each y_n obeys

$$a(x)y'_n(x) + b(x)y_n(x) = -y''_{n-1}(x).$$

Note that using the leading order solution, we can satisfy only one of the two boundary conditions that must be satisfied. We will need to place a boundary layer on the opposite boundary. But which one?

Moreover, where will the transition layer occur? If we write $X = (x - x_0)/\epsilon^\beta$, then our equation becomes

$$\epsilon^{1-2\beta} y_{XX} + a(x_0 + \epsilon^\beta X) \epsilon^{-\beta} y_X + b(x_0 + \epsilon^\beta X) y = 0. \quad (8.89)$$

Choosing $\beta = 1$ to give a dominant balance between first two terms yields $y_{XX} + a(x_0) y_X = 0$, or

$$y = y_{BL} \equiv C \exp(-a(x_0)X) + D, \quad (8.90)$$

for some pair of constants C and D to be determined. As with the solution away from the transition layer, this formula is only the leading order term in an expansion in ϵ . $C = 0$ corresponds to no variation across the transition layer, and this will only arise from the matching conditions if we have mislocated the transition layer.

Where are the transition layers? First note that the transition layer solution (8.90) decays exponentially in one direction (X increasing, if $a(x_0) > 0$) but grows exponentially in the other direction. It is impossible to match to the outer solution in the direction of exponential growth, so a dominant balance of the form described is only possible if x_0 is coincident with one of the endpoints of the interval: $x_0 = 0$ or $x_0 = 1$. Because transition regions are often found at endpoints, they are often called *boundary layers*. We are now presented with four different possibilities:

- There is a boundary layer at $x = 0$, but none at $x = 1$.
- There is a boundary layer at $x = 1$, but none at $x = 0$.
- There are boundary layer at both ends.
- Boundary layers occur in the interior of the interval.

If there are boundary layers at neither end, then since we cannot in general satisfy the boundary conditions with the outer solution alone, we are going to need to place a transition region in the interior of the interval!

As already mentioned, the biggest constraint upon the location of the boundary layer is that we must *match* the boundary layer solution to the outer layer solution. Examining the boundary layer solution, we see that agreement of the two solutions is only going to be possible if $a(x_0)X > 0$ as one moves away from the boundary layer. For $x_0 = 0$, this requires that $a(0) > 0$, whereas for $x_0 = 1$ this requires $a(1) < 0$.

$a(x) > 0$ or $a(x) < 0$ over the entire of the interval.

Hence if we consider the case where $a > 0$ in the entire interval, the boundary layer can only be located at $x = 0$. We choose $y_0 = B \exp\left(\int_x^1 b(x')/a(x') dx'\right)$ in order to satisfy the boundary condition at $x = 1$. The constants C and D must be chosen so that the boundary layer solution satisfies the boundary conditions: $y_{BL}(X = 0) \equiv C + D = A$;

To complete the solution, we need to match these solutions onto each other. This means making sure that the the outer solution and transition region solution agree at

the interface between the regions in which the two dominant balances were obtained, i.e. that:

$$\lim_{X \rightarrow \infty} y_{BL}(X) = \lim_{x \rightarrow 0} y_0(x) . \quad (8.91)$$

and this should hold true if we continue the perturbation expansion to *to higher orders* in ϵ . How matching can be enforced is discussed in detail in Section ??, but the meaning of the statement (8.91) is comprehensible without going into any detail on this. Marching towards $x = x_0 = 0$ from the outer layer, we approach the outer limit of the boundary layer solution - i.e. the limiting value of the boundary layer solution, as X is made large. Note that the limit on the left hand side is $X \rightarrow \infty$ and this is not the same as $x \rightarrow \infty$ (which would take us out of the interval on which we are looking for a solution to the differential equation!): it is easier to think of the limit being realised by fixing x and letting ϵ tend to 0. In the $X \rightarrow \infty$ limit, the first term of the boundary layer solution (8.90) becomes exponentially small, leaving only the constant D . Thus, to achieve matching it suffices that:

$$e^{\int_0^1 b(x')/a(x')dx'} = D. \quad (8.92)$$

Which together with the constraint $C + D = A$ allows both C and D to be determined.

A similar argument can be applied when $a < 0$ everywhere in the domain: here the boundary layer is at $x = 1$.

$a(x)$ changes sign

The more interesting cases occur when the sign of a is not fixed in the interval. Let us suppose for the sake of simplicity that a changes sign at a single point $x = x_0$ in the interval. There are then two cases to consider:

- $a(x) > 0$ when $x < x_0$, i.e. $a'(x_0) < 0$
- $a(x) < 0$ when $x < x_0$, i.e. $a'(x_0) > 0$.

The solution can only have a interior transition layer at $x = x_0$ (since transition layers at any other interior points give solutions which grow unphysically in one of the X -directions). Additionally, in the former case we have the possibility of there being boundary layers at *both* $x = 0$ and $x = 1$, whereas in the latter, we cannot have a boundary layer at either point!

To determine which of the potential sites $x = 0, x_0, 1$ might be transition regions for the solution, we try to construct transition layer solutions in a near neighbourhood of each of these points, and then see if they can be connected by outer layer solutions. Let's start by analyzing the possible boundary layer solutions near x_0 . Here, let us approximate $a(x) \approx \alpha(x - x_0) + O((x - x_0)^2)$, introducing a constant $\alpha \equiv a'(x_0)$. We thus have

$$\epsilon y'' + \alpha(x - x_0)y' + b(x)y = 0.$$

As before we write $X = (x - x_0)/\epsilon^\beta$, and thus obtain

$$\epsilon^{1-2\beta}y_{XX} + \alpha X y_X + b(x_0 + \epsilon^\beta X)y = 0 .$$

The only nontrivial balance takes $\alpha = 1/2$, so that we have (to leading order)

$$y_{XX} + \alpha X y_X + b(x_0)y = 0 , \tag{8.93}$$

in the boundary layer. We need to determine the conditions under which this solution can be matched to the rest of the solution: clearly, this requires that $y_{BL}(X)$ is reasonably behaved as $X \rightarrow \infty$.

Now, recall our previous work on linear second order differential equations. $X = \infty$ is a irregular singular point of the equation (8.93), and therefore we expect the transition layer solution to have essential singularity there. To obtain the asymptotic behaviour of the solution as $X \rightarrow \pm\infty$ we make the substitution $y = e^S$, and rewrite the equation in terms of the new function $S(X)$, obtaining:

$$S'' + S'^2 + \alpha X S' + b(x_0) = 0 . \tag{8.94}$$

In the limit of large X , we find two consistent dominant balances for this differential equation, namely: $S' \sim -\alpha X$, and $S' = -b(x_0)/\alpha X$. Hence as $X \rightarrow \infty$, we have possible asymptotic behaviours:

$$y \sim A e^{-\alpha X^2/2} \quad \text{or} \quad y \sim B |X|^{-b(x_0)/\alpha} ,$$

as X marches away from zero in either of the directions $\pm\infty$. If $\alpha > 0$ then we expect the solution to have the second type of asymptotic behaviour, since this represents the least severe decay, and the inner solution has two free constants, the value of B at $\pm\infty$ which can be determined by matching to the outer solutions for $x < x_0$ and $x > x_0$. On the other hand, if $\alpha < 0$, then we expect to see the first type of asymptotic behaviour (which now gives super-exponential growth!). To eliminate this we must require that $A = 0$ on both sides $\pm\infty$, which imposes two boundary conditions upon the inner-layer solution, and leaves us no freedom for matching! It follows that the solution must vanish identically within the transition layer.

We summarise the cases as follows: when $\alpha < 0$ there can be boundary layers at both boundaries and the solution must vanish identically at $x = x_0$. In contrast, when $\alpha > 0$, the outer solution should apply near each boundary and there should be a single transition layer in the middle of the interval.

We demonstrate the marked difference between the two cases $\alpha \gtrless 0$ by solving numerically the differential equation:

$$\epsilon y'' \pm \left(x - \frac{1}{2}\right) y' + (1 + x^2)y = 0 . \tag{8.95}$$

Later we show how the Matlab routine `bvp4c` may be used to generate these graphs.

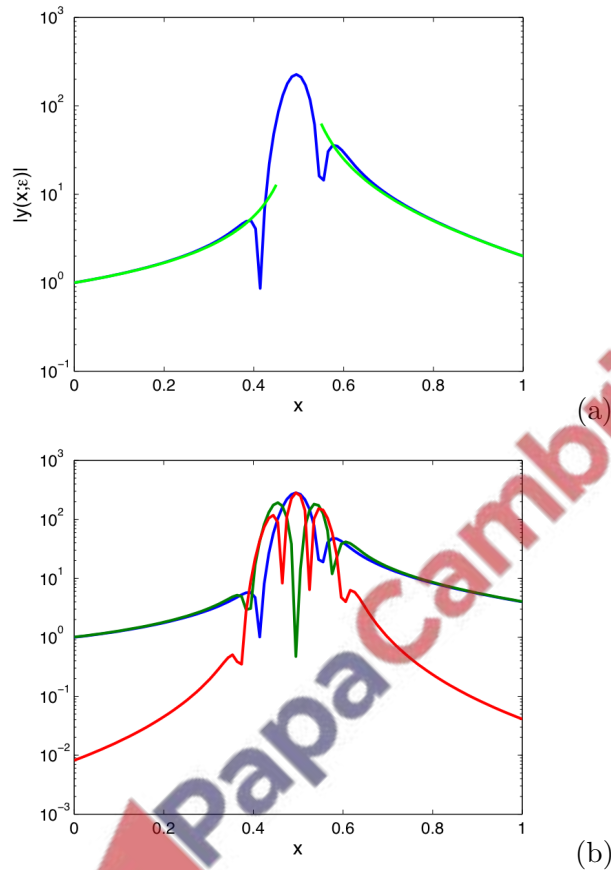


Figure 8.8. (a) Solution of $\epsilon y'' + (x - 1/2)y' + (1 + x^2)y = 0$, subject to boundary conditions $y(0) = 1$ and $y(1) = 2$, for $\epsilon = 10^{-3}$. The blue curve is the numerical solution and the green curves are outer layer solutions (8.88). (b) Plot of the individual terms of the differential equation: $|\epsilon y''|$ (red curve), $|(1 + x^2)y|$ (blue curve), $|(x - 1/2)y'|$ (green curve), showing that dominant balance is between last pair of terms in outer region, while all terms participate in the inner region.

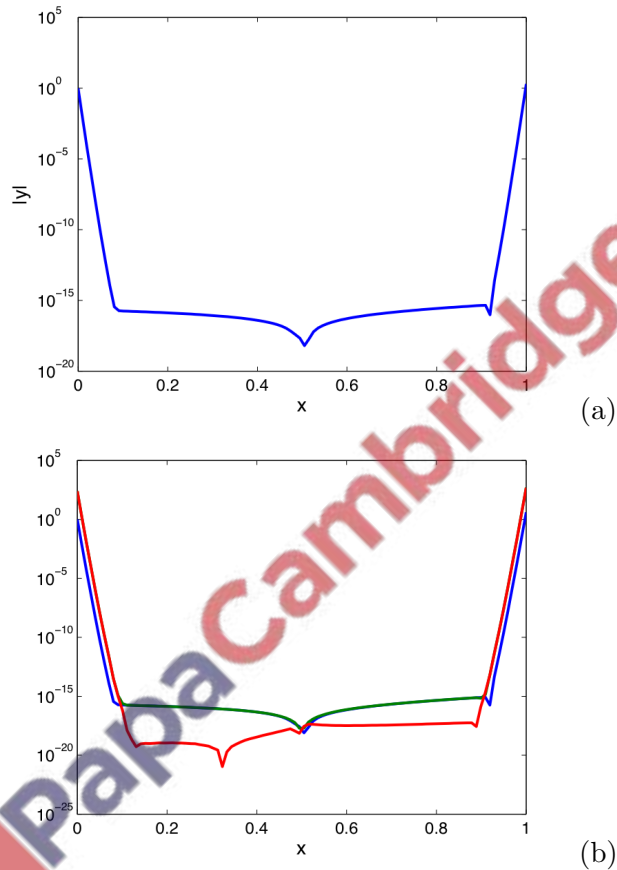


Figure 8.9. (a) Numerical solution of $\epsilon y'' - (x - 1/2)y' + (1 + x^2)y = 0$, subject to boundary conditions $y(0) = 1$ and $y(1) = 2$, for $\epsilon = 10^{-3}$. (b) Plot of the individual terms of the differential equation: $|\epsilon y''|$ (red curve), $|(1 + x^2)y|$ (blue curve), $|(x - 1/2)y'|$ (green curve), showing existence of boundary layers at $x = 0, 1$ in which first and second terms balance, outer solution for x not close to any of $0, 1/2$ or 1 , in last pair of terms balances, and exact cancellation at $x = 1/2$, in which all terms feature.

8.5.3 Some remarks on numerical Methods

We now consider some numerical methods for solving boundary value problems.

Finite Elements

The finite element method is an extremely robust way of solving nonlinear PDE's—here we will illustrate the method on a specific example. Consider the equation

: $y'' = -a(x)y' - b(x)y$. We will represent $y(x)$ with an expansion of the form $y(x) = \sum_i a_i \psi_i(x)$. Here $\psi_i(x)$ are local *expansions* of the solution, which are nonzero only in a small region of a given point, and zero outside of this region. There are many different ways of doing this—the *lowest order*/simplest method is to construct a mesh $x_i = i\Delta x$, and then to construct the basis functions so that $\psi_i(x_j) = \delta_{ij}$. With this choice, then the $a_i = y_i = y(x_i)$ is the value of $y(x)$ at the appropriate mesh point.

This *expansion* of $y(x)$ is equivalent to expanding $y(x)$ as a piecewise linear function; one could imagine higher order representations involving basis functions which included e.g. both the value of $y(x)$ and the value of $y'(x)$ at each mesh point—such a basis would be piecewise quadratic. Or, one could keep even higher order terms.

In any case—we are now left with the question as to how to find the coefficients y_i of our expansion so that we solve the differential equation. We proceed as follows. Define $\phi_j(x)$ to be *test functions*. If we multiply our differential equation by a test function and then integrate over the domain we obtain

$$\int \phi_j(x) \left(y'' + a(x)y' + b(x)y \right) = 0, \quad (8.96)$$

or $\sum_i y_i \int \phi_j(x) \left(\psi_i'' + a(x)\psi_i' + b(x)\psi_i \right) = 0$.

Typically, one chooses the test functions to be equivalent to the trial functions, i.e. $\phi_j = \psi_j$; since the ψ 's are piecewise linear functions, we then have the minor difficulty that it appears that the ψ_i'' have singularities in them. This is gotten around by integrating by parts, and rewriting our equation as : $\sum_i y_i \int \left(-\phi_j(x)'\psi_i' + a(x)\phi_j\psi_i' + b(x)\phi_j\psi_i \right) = 0$.

To the equation, we must supplement the boundary/initial conditions which specify e.g. that $y(a) = y_0$ and $y(b) = y_1$. Together, we have now reduced the solution of the differential equation to a linear algebra problem, namely an equation of the form

$\mathbf{M}\mathbf{y} = \mathbf{b}$, where \mathbf{b} is a nonzero vector resulting from the boundary conditions, and the matrix \mathbf{M} comes from the differential equation.

As an example, a common choice of finite elements are the *Tent functions*, where $T(\eta) = 1 - |\eta|$, when $-1 \leq x \leq 1$, and $T(\eta) = 0$ otherwise. If we take $\eta_i = (x - x_i)/\Delta x$, then $\psi_i(x) = T(\eta_i)$. With this choice, we can evaluate the integral $N_{ij} = \int \psi_i'(x)\phi_j'(x)dx$. The integral is only nonzero when the *tents* overlap; this only occurs when $j = i, i \pm 1$. A straightforward exercise shows that when $j = i$, $N_{ii} = 2/\Delta x$, and $N_{i,i\pm 1} = -1/\Delta x$.

Finite Differences

Another method for solving boundary value problems numerically that is as follows: suppose one is given a set of equations:

$$y'' = f(y, y'), \quad (8.97)$$

with the boundary conditions $y(0) = A$ and $y(1) = B$. Here we are taking the equation to be second order, though the method is more general. The first step is to discretize equation (8.97), so that $y(x) = y(n\Delta x) \equiv y_n$, hence

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{\Delta x^2} = f((y_{n+1} - y_{n-1})/(2\Delta x), y_n). \quad (8.98)$$

This equation is applied everywhere in the domain except on the boundaries. If $\Delta x = 1/N$, then we take $y_0 = A$ and $y_N = B$, and use equation (8.98) at all other points. In total this is N equations with N unknowns, and thus the system can be solved. For general $f(y, y')$ the equations are nonlinear and must be solved using a Nonlinear Newton's method.

The equation

$$y'' = \frac{y_{n+1} - 2y_n + y_{n-1}}{\Delta x^2} \quad (8.99)$$

of course needs to be derived: note that this same discretization arises from the finite element method described above. The finite difference derivation is different, and starts by asking for a *polynomial approximation* to the function $y(x)$, that interpolates its values on nearby mesh points. Namely, we write that $y(x) = \sum_i y_i \ell_i(x)$, where $\ell_i(x)$ is a polynomial that satisfies $\ell_i(x_j) = \delta_{ij}$. These polynomials are called *Lagrange polynomials*, and are given by: $\ell_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}$. Now, given an interpolant, one can evaluate the second derivative just by taking the second derivative of the interpolant. Namely: $y'' = \sum_i y_i \ell_i''(x_i)$. If we use a quadratic interpolant, the formula then directly gives our second derivative formula.

Note that although there is a similarity here to finite elements, there are also important differences—both in the procedure to follow to develop higher order methods, and in the form of the other terms in the equations.

A MATLAB implementation

Below are subroutines of a matlab program that implements this algorithm on the equation: $y'' + a(x)y' + b(x)y = 0$.

The main program is:

```
1 global a global b global eps a=1; b=2;
2 x=linspace(0,1,1000);
```

```

3 y=x*0; for i=1:10
4     res=residual(x,y);
5     jac=jacobian(x,y);
6
7     y=y-(inv(jac)*res')';
8
9 end

```

In this program, a, b are global variables that are passed to the subroutines containing the differential equations. The subroutines *residual* and *jacobian* compute the residual and jacobian matrix of the equations we are trying to solve. (Recall Newton's method works by finding the zeros of $R = y'' - f(y, y')$; the function

R is often called the residual.) The last equation in the loop is a step of Newton's method.

The subroutines computing the residual and jacobian are below.

```

1
2 function res=residual(x,y)
3
4 global a
5 global b
6 n=length(y);
7 dx=x(2)-x(1);
8 res(1)=y(1)-a;
9
10 for i=2:n-1
11     res(i)=eps*(y(i+1)-2*y(i)+y(i-1))/dx^2 + ...
12     coeffa(x(i))*(y(i+1)-y(i-1))/2/dx + coeffb(x(i))*y(i);
13 end
14
15 res(n)=y(n)-b;

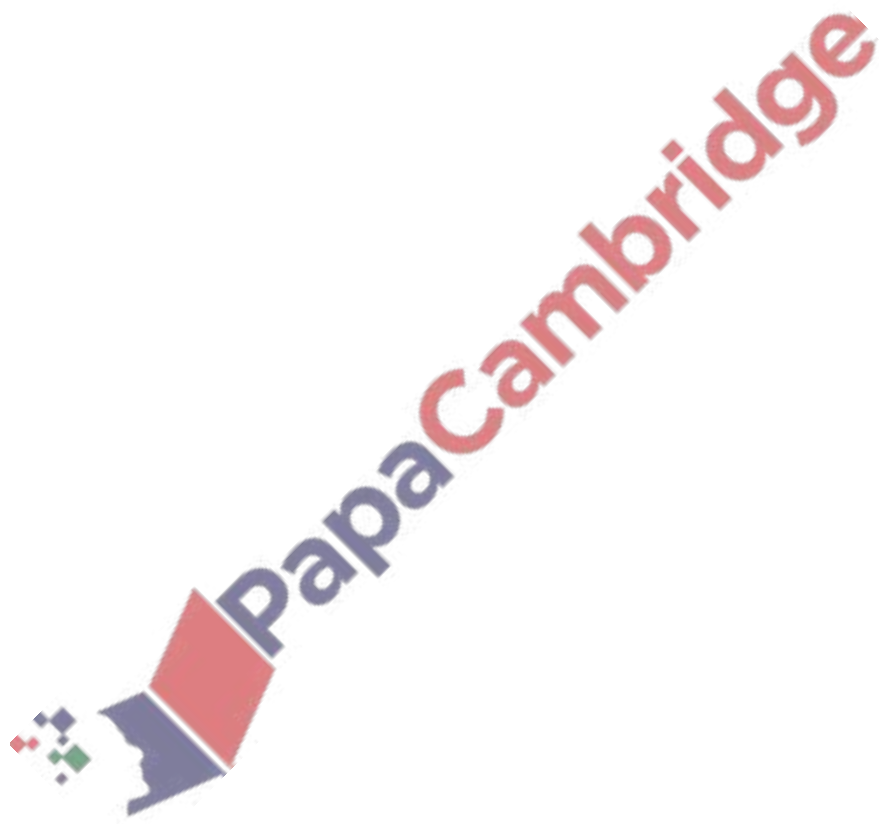
```

```

1
2 function m=jacobian(x,y)
3 n=length(y);
4 dx=x(2)-x(1);
5 m=zeros(n,n);
6 m(1,1)=1;
7 for i=2:n-1
8
9     m(i,i+1)=1/dx^2+coeffa(x(i))/dx/2;
10    m(i,i)=-2/dx^2+coeffb(x(i));
11    m(i,i-1)=1/dx^2-coeffa(x(i))/dx/2;
12
13 end
14 m(n,n)=1;

```

The functions *coeffa* and *coeffb* are user defined functions that these programs called.



9 Introduction to Linear PDE's

These notes will serve as an introduction to partial differential equations. There are three *prototypical types* of linear partial differential equations. These types are

- **Parabolic** Here the prototype is the diffusion equation

$$\partial_t \phi = D \nabla^2 \phi. \quad (9.1)$$

- **Hyperbolic** The prototype here is the wave equation,

$$\partial_{tt} \phi = c^2 \nabla^2 \phi. \quad (9.2)$$

- **Elliptic** This prototype here is Laplace's equation,

$$\nabla^2 \phi = 0. \quad (9.3)$$

This type of equation tends to represent equilibrium situations. Indeed if you delete the time derivative from the diffusion equation and the wave equation you get Laplace's equation!

These three classes of equations represent dramatically different physical situations. One of the most important goals of our studies will be for you to be able to recognize which of these possibilities represents the most natural model for a given situation.

9.1 Random walkers

In trying to understand the solutions to partial differential equations, it is particularly useful to first have a simple model to think about for where such equations come from. A particularly powerful approach is to consider the equations as arising from walkers. We start our discussion with random walkers, which are the prototypical model for diffusion.

9.1.1 Random walk on a one-dimensional lattice

Consider a *walker* who lives on a one dimensional lattice, with positions $x = ia$, where $i = \dots -3, -2, -1, 0, 1, 2, 3, \dots$ and a is the lattice spacing. We assume that in a time interval τ the particles can move either right or left; in order to decide which way to move, the walker flips a coin. If the coin comes up heads the walker moves to the left whereas if it is tails he moves to the right. We want to derive an equation for the evolution of the walker. We will present two derivations, the first based on a single walker and the second based on a cloud of walkers.

9.1.2 Derivation #1

Let $p_i^n = p(ia, n\tau)$ be the probability that the walker is at position i at time n . Then the rule for the motion of our walker translates into

$$p_i^{n+1} = \frac{1}{2} \left(p_{i-1}^n + p_{i+1}^n \right). \quad (9.4)$$

Let us transform this rule into an equation for the evolution of the probability density of the walker. To do this, we will try to find an equation for $p(x, t)$ where we will think of $x = ia$ as the position of the walker, and $t = n\tau$ the time. The fundamental approximation we will make is that of *scale separation*; namely we assume that $x \gg a$ and $t \gg \tau$. Our equation for the probability distribution will work on scales much larger than the size of a single step.

Written in our new notation, the rule for the motion of the walker translates into

$$p(x, t + \tau) = \frac{1}{2} \left(p(x - a, t) + p(x + a, t) \right). \quad (9.5)$$

To proceed, we now use a Taylor series expansion: We know that

$$p(x, t + \tau) = p(x, t) + \partial_t p(x, t)\tau + \dots \quad (9.6)$$

Similarly

$$p(x \pm a, t) = p(x, t) \pm a\partial_x p(x, t) + a^2/2\partial_{xx}p + \dots \quad (9.7)$$

Plugging these expansions into the equation for p we arrive at

$$\partial_t p = \frac{a^2}{2\tau} \partial_{xx} p + \dots \quad (9.8)$$

It is easy to verify (by continuing the Taylor series expansion of p) that the next term in the Taylor expansion is of order $a^4/\tau \partial_{xxxx} p$. If we assume that these corrections are small, we see that the probability distribution obeys the equation

$$\partial_t p = \frac{a^2}{2\tau} \partial_{xx} p. \quad (9.9)$$

This is the *Diffusion equation*, with a diffusion constant

$$D = \frac{a^2}{2\tau}. \quad (9.10)$$

In class, we emphasized the beauty of this formula for the diffusion constant. Every physical (or nonphysical) process that involves diffusion has an underlying dynamics resembling random walkers, and thus the above formula can be used for computing/estimating diffusion constants.

9.1.3 A final remark

Before moving on to the second derivation, it is worth remarking on the condition for which the error in our derivation above is small. The largest neglected term $a^4/\tau\partial_{xxxx}p$ must be smaller than the terms we have kept: namely,

$$\frac{a^4}{\tau}\partial_{xxxx}p \ll \frac{a^2}{\tau}\partial_{xx}p. \quad (9.11)$$

If we assume that $p(x, t)$

varies on the length scale L , this condition boils down to roughly

$$\frac{a^4}{\tau} \frac{p}{L^4} \ll \frac{a^2}{\tau} \frac{p}{L^2}, \quad (9.12)$$

or $a^2 \ll L^2$. Thus as long as the scale of that $p(x)$ varies on is much larger than the lattice spacing, our diffusive theory will work fine. Even if the initial condition does not obey this condition (e.g. all of the walkers are bunched at the origin) then since diffusion produces a wider and wider distribution with time, eventually the condition will be obeyed!

9.1.4 Derivation #2

The second derivation instead imagines a cloud of random walkers. Let $n(x, t)$ denote the number of particles at x at time t . The number of particles in the region from $x \rightarrow x + \Delta x$ is $N = \int_x^{x+\Delta x} n(x, t) dx$.

The rate of change of this number is given by the flux J moving out of each boundary, i.e.

$$\frac{dN}{dt} = \int_x^{x+\Delta x} \partial_t n(x, t) dx = -J(x + \Delta x) + J(x). \quad (9.13)$$

The signs here have been chosen so that the number of particles will in the region will increase when particles are moving in from the right (i.e. $J(x + \Delta x)$ is negative) or the left ($J(x)$ is positive). We can rewrite the right hand side of this equation as

$$-J(x + \Delta x) + J(x) = - \int_x^{x+\Delta x} \partial_x J. \quad (9.14)$$

Hence we have that

$$\int_x^{x+\Delta x} (\partial_t n + \partial_x J) = 0. \quad (9.15)$$

Hence since this is true for every region of space one can consider we must have

$$\partial_t n + \partial_x J = 0. \quad (9.16)$$

Now, what is the flux? Often, the fundamental law for diffusion is called *Fick's Law*, which states that the flux is given by

$$J = -D\partial_x n. \quad (9.17)$$

With this law, we have that

$$\partial_t n = D\partial_{xx} n. \quad (9.18)$$

A diffusion equation!

You should note that there is a real sense that our first derivation is better than our second. The first derivation led us to a formula for a diffusion equation in which

1. It was apparent that the equation itself is an approximation to what is really going on.
2. We were given an explicit formula for the diffusion constant.
3. No phenomenological assumptions (Fick's law) were made. Indeed, our derivation can be viewed as a derivation of Fick's law.

9.1.5 Random walks in three dimensions

All of this can be extended to three dimensions. In three dimensions, we need to erect a three dimensional lattice. Imagine the lattice is a 3d grid with lattice spacing a . Now every point has 6 nearest neighbors, instead of just two. If we think of $p = p(x, y, z, t)$, then the dynamical rule for a random walker is

$$p(x, y, z, t + \tau) = \frac{1}{6} \left(p(x - a, y, z, t) + p(x + a, y, z, t) + p(x, y - a, z, t) + p(x, y + a, z, t) + p(x, y, z - a, t) + p(x, y, z + a, t) \right). \quad (9.19)$$

I will leave it as an exercise for you to show that the first derivation leads to the diffusion equation

$$\partial_t p = \frac{a^2}{6\tau} \nabla^2 p, \quad (9.20)$$

where as usual

$$\nabla^2 = \partial_x^2 + \partial_y^2 + \partial_z^2. \quad (9.21)$$

9.1.6 Remark about Boundary Conditions

In order to solve the diffusion equation (or the random walker problem) we need to specify two things: the initial distribution of walkers (an initial condition), and the boundary conditions: Namely, is there any condition on the walkers at the boundary of the domain? As we formulated the random walker problem above there was no such condition: the walkers will keep walking forever. However, we could impose conditions. For example, we could impose that the walkers are confined between $0 \leq x \leq L$. This would require making the probability that the walker moves outside the domain vanish, or making the flux of walkers at each of these walls vanish.

In the computer, these conditions are simple to implement. For example if we wanted the flux of walkers at $x = L$ to vanish, we would simply modify the rule

$$p_i^{n+1} = \frac{1}{2}(p_{i-1}^n + p_{i+1}^n), \quad (9.22)$$

by using the fact that at the endpoint p_M , we must have $p_M - p_{M+1} = 0$. (i.e. the flux vanishes). Hence the rule at the end point will be

$$p_M^{n+1} = \frac{1}{2}(p_{M-1}^n + p_M^n). \quad (9.23)$$

If we wanted the probability of walkers at $x = L$ to vanish, we would simply write

$$p_M^{n+1} = 0 \quad (9.24)$$

for each time n . This would then imply that at the 'second to last' grid point, the equation will be

$$p_{M-1}^{n+1} = \frac{1}{2}(p_{M-2}^n). \quad (9.25)$$

In a subsequent homework you will use simulations of clouds of random walkers on bounded domains to simulate the diffusion equation.

9.1.7 Simulating Random Walkers

We have thus shown that random walkers obey the diffusion equation at long times. It is quite illuminating to see this explicitly and determine whether our theory (claiming that random walkers obey the diffusion equation) is indeed correct. Here we describe how to simulate clouds of random walkers using Matlab, and then we compare our theory for the distribution of walkers with the simulations.

The trickiest part of the program is the plotting: I use the command *hist* to create a histogram of the particle distribution, and then plot the histogram.

```

1 % initial parameters: tmax, total number of steps;
2 %nparticles=number of random walkers;
3 % r0=size of initial region
4 tmax=200; nparticles=100; r0=4;
5 % initial distribution: rand chooses numbers randomly between 0 and 1
6 x=rand(nparticles,1)*r0; y=rand(nparticles,1)*0;
7 xpoints=-20:0.5:20; % places where histogram centers the bins
8 nn=hist(x,xpoints); axis([-20 20 0 20]); plot(xpoints,nn);
9 for i=1:tmax
10 for j=1:nparticles
11 if rand(1,1)>0.5
12 x(j)=x(j)+1;
13 else
14 x(j)=x(j)-1;
15 end
16 end
17 axis manual
18 nn=hist(x,xpoints);
19 plot(xpoints,nn);
20 axis([-20 20 0 20]);
21 M(i) = getframe;
22 pause(1/10);
23 width(i)=std(x); % variance of the distribution
24 meanx(i)=mean(x); % mean of the distribution
25 end
26 movie(M); % plays the frames again as a movie!

```

I kind of don't like the way the plot command makes the figure. Alternatively, the command bar(xpoints,nn) can be used instead.

The figure shows a simulation of a cloud of random walkers; 100 walkers were started at the origin and the figure shows the distribution a time 5, 20 and 100 time units later. The distribution spreads

The second figure shows the width of the distribution as a function of time; the dots are the simulations and the solid green line is the law $width = \sqrt{t}$. The width spreads as expected!

9.2 Solving the Diffusion Equation

Although the previous discussion included a complete solution for the wave equation, we still have not solved the diffusion equation. Here we discuss two different ways of solving the diffusion equation,

$$\partial_t n = D \partial_x^2 n, \quad (9.26)$$

subject to the initial condition that $n(x, t = 0) = n_0(x)$. In order to connect to the simulations of random walkers shown above, we will employ the boundary conditions that ρ vanishes at $\pm\infty$. (This is appropriate for our random walker simulations because we just let the walkers spread out indefinitely.) Later on we will relax this and discuss how to solve the diffusion equation on a finite domain. However, although the details

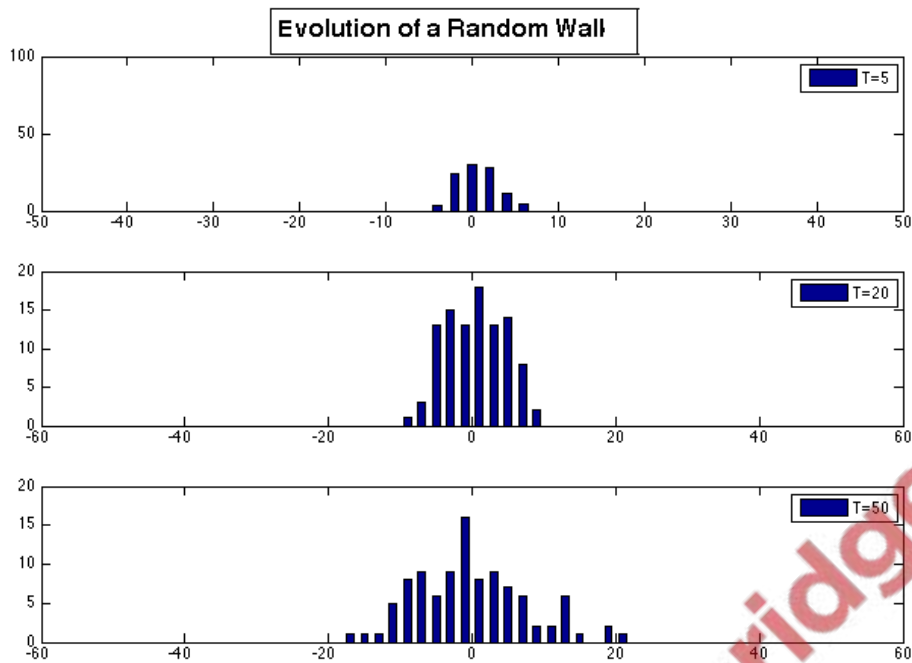


Figure 9.1. Simulation of random walkers, starting from the origin, at three subsequent times..

are different, *morally*, the basic ideas behind methods we will describe are the same for all boundary conditions.

There are two basic methods for solving this problem, each of which relies on a different method for representing the solution. In each case, the central idea is that since the equation is *linear*, it is possible to "break down" any initial state into a linear combination of simpler problems. By solving the "simpler problems" *explicitly*, it is then possible to reconstruct the general solution. The two methods are:

The Fourier Method

This method relies on the Fourier Transform representation for the density field $n(x, t)$. Namely we can write

$$n(x, t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dk e^{ikx} \hat{n}(k, t).$$

Here $\hat{n}_0(k)$ is the *fourier coefficient* of $n_0(x)$. The strength of this method is that it is simple to solve the diffusion equation for a *single plane wave*.

Let's consider the solution $n(x, t) = a(t)e^{ikx}$. Plugging this into the diffusion equation gives $\dot{a} = -k^2 a$, which has the solution $a = a_0 e^{-k^2 t}$, if $a(t = 0) = a_0$. By using the Fourier representation of the solution, and applying this trick to each wave in the fourier

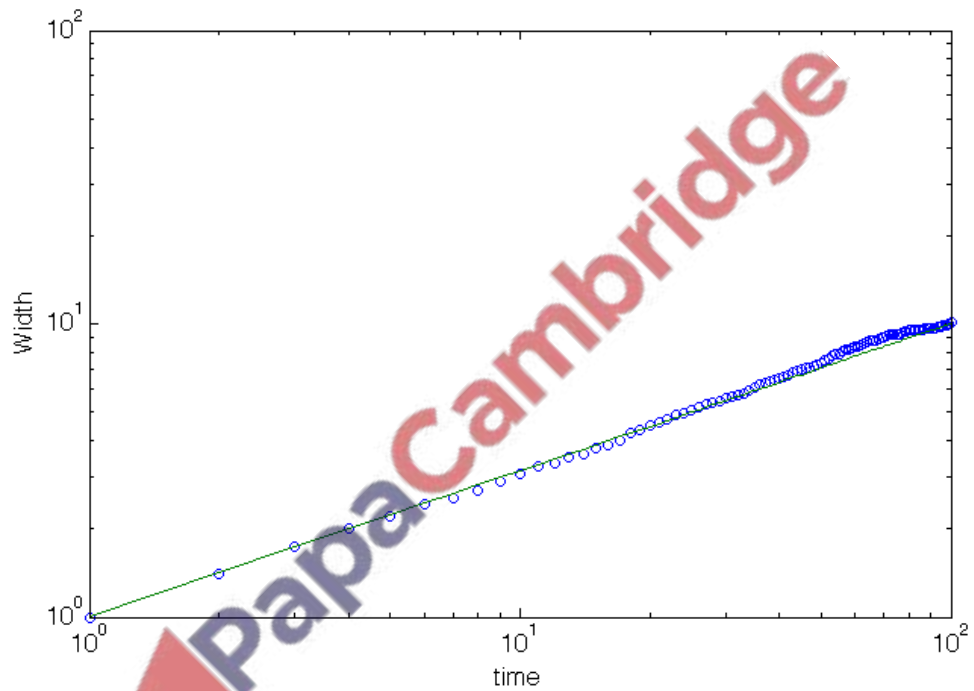


Figure 9.2. Width of distribution as a function of time. The dots are from the computer simulations and the solid line is the solution to the diffusion equation..

description, we see that the most general solution is

$$n(x, t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dk e^{ikx - k^2 t} \hat{n}_0(k),$$

where $\hat{n}_0(k)$ are the fourier coefficients of n_0 .

This formula (although correct) is not particularly illuminating in the present form, as it involves an integral that needs to be evaluated. It turns out that this integral *can* be evaluated—but we will follow a different route here.

Green's Function Method

This method relies on a simple idea for representing the solution. We will express the solution as a basis of states which are localized in *position*. This is done by using the so called Dirac Delta Function, denoted $\delta(x - x_0)$. You should think of this as a large spike of unit mass which is centered exactly at the position x_0 . The definition of δ is that given any function $f(x)$,

$$\int dx' f(x') \delta(x' - x) = f(x).$$

Now we can use δ to represent n_0 as

$$n_0(x) = \int_{-\infty}^{\infty} dx' \delta(x - x') n_0(x').$$

This formula decomposes n_0 into a continuous series of “spikes”. The idea of this method is that to understand how each “spike” individually evolves, and then to superimpose the evolution of each of the spikes to find the final density distribution. We define the Green's function $G(x - x', t)$ so that $G(x - x', t = 0) = \delta(x - x')$, and

$$n(x, t) = \int_{-\infty}^{\infty} dx' G(x - x', t) n_0(x').$$

Plugging this representation into the diffusion equation, we see that $G(x - x', t)$ *obeys the diffusion equation!*

Thus, we have reduced this problem to the mathematics problem of solving the diffusion equation for the localized initial condition $\delta(x - x')$.

There are many ways of solving this problem. The one that is in most textbooks is to actually use the Fourier decomposition of $\delta(x - x')$ to solve the equation in fourier space, and then (as above) transform back to real space. This is an advisable procedure because the Fourier Transform of δ is very very simple.

We will advocate another procedure. This procedure is more elegant, and moreover uses an idea that is both general and important. The idea is to use dimensional analysis to determine the solution. The equation contains a single parameter, D . We know from

above that $D = a^2/2\tau$; this quantity has the dimensions of a length squared divided by a time. We can also see this directly from the diffusion equation; in order for $\partial_t n$ to have the same dimensions as $D\partial_{xx}n$, D is required to have the dimensions of length squared over time.

Suppose we were to ask: How far has the pulse spread after a time t ? The answer to this question is a number with the units of length. What can this length be? If we call the width ℓ , then we know that

$$\ell = f(D, t, \text{parameters}). \quad (9.27)$$

Here D is the diffusion constant, t is time, and parameters represents any other parameters that might be represented in the initial condition. But our initial condition is a delta function and has zero width—thus there really aren't any parameters in the initial condition. The only dimensionally consistent formula we can write is

$$\ell = \sqrt{Dt}. \quad (9.28)$$

Hence the width must grow like the square root of time! We can represent this width in our green's function in the following way: The fact that the width of the solution is $\ell(t)$ means that we can write

$$G(x - x', t) = A(t)F\left(\frac{x - x'}{\sqrt{Dt}}\right). \quad (9.29)$$

Note that the scale over which F varies is exactly $\ell(t)$. We have multiplied by an arbitrary time dependent constant $A(t)$ to keep full generality.

How to proceed further: There is one other thing that we know about the solution: we know that the total number of random walkers (or the total amount of the substance that is diffusing) is constant in time. You can see this from the original derivation of the random walkers, or you can see this directly by showing that

$$\frac{d}{dt} \int n(x', t) dx' = 0.$$

But if the pulse spreads in width this means it must decrease in height in order to keep the total integral $\int n$ constant! To see this explicitly, note that

$$\int_{-\infty}^{\infty} n = \int_{-\infty}^{\infty} A(t)F(x/\sqrt{Dt})dx = A(t)\sqrt{Dt} \int_{-\infty}^{\infty} dyF(y)$$

must be constant in time. Here we have changed variables from x to $y = x/\sqrt{t}$. We therefore see that $A(t) = 1/\sqrt{Dt}$. We thus have shown that

$$G(x - x', t) = \frac{1}{\sqrt{Dt}}F\left(\frac{x - x'}{\sqrt{Dt}}\right). \quad (9.30)$$

How do we determine F ? Let's just plug in $G(x, t) = 1/\sqrt{Dt}F((x - x')/\sqrt{Dt})$ into the diffusion equation. This gives the following ordinary differential equation for $F(y)$ (with $y = (x - x')/\sqrt{Dt}$).

$$\frac{1}{\sqrt{Dt}^{3/2}}\left(-\frac{1}{2}F - \frac{1}{2}yF'\right) = \frac{1}{(Dt)^{3/2}}DF''.$$

Cancelling out the time factors, this equation is

$$-\left(\frac{1}{2}F + \frac{1}{2}yF'\right) = F''. \tag{9.31}$$

This equation can be integrated once to give and integrating this equation once gives

$$F' = -\frac{1}{2}Fy.$$

This equation can be immediately integrated to give $F(y) = F_0e^{-y^2/4}$, or

$$G(x - x', t) = \frac{F_0}{\sqrt{t}} \exp -\frac{(x - x')^2}{4Dt}, \tag{9.32}$$

where the constant $F_0 = N/\sqrt{4\pi}$ is determined by requiring that $\int G = N$, where N is the number of random walkers.

9.2.1 Long-time Limit of the Diffusion Equation

The idea underlying Laplace's method is generally useful. Given an initial condition $n(x, t = 0) = n_0(x)$ to the diffusion equation, we want to understand the behavior of the solution at long times. Let us suppose that the initial condition is localized in space so that $n_0(x) > 0$ for $-L \leq x \leq L$. The solution at long times can be written as

$$n(x, t) = \int_{-\infty}^{\infty} dx' G(x - x', t)n_0(x'). \tag{9.33}$$

Now, we can use the essential idea implicit in Laplace's method to evaluate the integral at long times. At long enough times so that $\sqrt{4Dt} \gg L$, the Green's function does not vary very much over the scale of the initial condition. Therefore we can write

$$n(x, t) = \int_{-\infty}^{\infty} dx' G(x - x', t)n_0(x') \approx G(x, t) \int_{-L}^L n_0(x') dx' = \frac{M}{\sqrt{4Dt}} \exp -\frac{x^2}{4Dt}, \tag{9.34}$$

where $M = \int n_0(x)$. Hence we have demonstrated that the solution to the diffusion equation in the long time limit approaches a gaussian.

9.3 Disciplined Walkers

Let us now consider the completely opposite case. The case of *disciplined walkers*. A disciplined walker always knows the direction in which it wants to go. For example, a disciplined walker might always move to the right. Such a walker obeys the equation

$$p_i^{n+1} = p_{i-1}^n, \quad (9.35)$$

or

$$p(x, t + \tau) = p(x - a, t). \quad (9.36)$$

Another type of disciplined walker might move only to the left. Such a walker obeys

$$p(x, t + \tau) = p(x + a, t). \quad (9.37)$$

If we Taylor series

$$p(x, t + \tau) = p(x, t) + \tau \partial_t p(x, t) + \tau^2/2 \partial_{tt} p + \dots \quad (9.38)$$

and

$$p(x \pm a, t) = p(x, t) \pm a \partial_x p + \partial_{xx} p a^2/2 + \dots, \quad (9.39)$$

we obtain for the left moving walkers

$$\partial_t L = \frac{a}{\tau} \partial_x L, \quad (9.40)$$

and for the right moving walkers

$$\partial_t R = -\frac{a}{\tau} \partial_x R. \quad (9.41)$$

Here we have written $p(x, t) = R(x, t)$ for the right moving walkers and $p(x, t) = L(x, t)$ for the left moving walkers, to keep our notation straight. Each of these equations (for R and for L) are called *advection* equations. They have the property that they simply translate the initial distribution of particles to the left or the right, respectively. Because the 'physics' is so simple, both of these problems can be solved directly: The solution for

$$R(x, t) = F(x - a/\tau t) \quad (9.42)$$

, where $R(x, 0) = F(x)$ is the initial distribution. We can verify that this solves the equation, since

$$\partial_t R = -\frac{a}{\tau} F', \quad (9.43)$$

and

$$\partial_x R = F'. \quad (9.44)$$

For the L equation, one can similarly demonstrate that the exact solution is just

$$L(x, t) = G(x + a/\tau t), \quad (9.45)$$

where $L(x, 0) = G(x)$ is the initial distribution.

There is one more aspect of this problem that deserves mention. What would happen if you had a collection of disciplined walkers, but you didn't know which ones moved to the left and which one move to the right? What would you do? In this case, you would not be able to measure R and L separately. Instead, you would only be able to measure the total number of walkers $n(x, t) = R + L$. Can we write an equation that describes the evolution of the total density? If we compute

$$\partial_t n = \partial_t(R + L) = -\frac{a}{\tau} \partial_x(R - L). \quad (9.46)$$

On the other hand, if we introduce a new variable $\sigma = R - L$, we then have

$$\partial_t \sigma = \partial_t(R - L) = -\frac{a}{\tau} \partial_x(R + L). \quad (9.47)$$

Combining these two equations we arrive at:

$$\partial_{tt} n = \frac{a^2}{\tau^2} \partial_{xx} n. \quad (9.48)$$

Hence, the number density of particles obeys a partial differential equation that is *second order* in time. This equation is called a wave equation.

Note that the wave equation requires two initial conditions, in contrast to the diffusion equation. Here we must specify the total number of walkers moving to the left and to the right; or alternatively, we must specify $n(x, t = 0)$ and $\partial_t n(x, t = 0)$ to get a unique solution. We can write the general solution for the wave equation using our results for $R(x, t)$ and $L(x, t)$: namely

$$n(x, t) = R + L = F(x - a/\tau t) + G(x + a/\tau t). \quad (9.49)$$

Given $n(x, 0)$ and $\partial_t n(x, 0)$ we can compute F and G since $n(x, 0) = F + G$ and $\partial_t n(x, 0) = a/\tau(G' - F')$. The second equation can be integrated to yield

$$G - F = \frac{\tau}{a} \int dx \partial_t n(x, 0). \quad (9.50)$$

This can then be combined with the first relation $F + G = n(x, 0)$ to solve for F and G .

We will return to other methods for solving the wave equation later on, but the ideas we have just outlined are in a real sense the most physical and direct.

Information Propagation

The defining feature of the disciplined walkers is that information is never destroyed as they are moving—in each timestep the right moving walkers move to the right and the left moving walkers move to the left—their numbers and their patterns are perfectly preserved. Information is propagated along the lines $x \pm \frac{a}{\tau}t = \text{constant}$. People often call these lines *characteristics*.

9.3.1 Disciplined Walkers Moving in More complicated ways

The above discussion assumed that the disciplined walkers moved in a very boring way: each time step everyone went the same distance. Boring indeed.

But, walkers can move by much more general rules than this. As an example one could imagine a walker that had constant acceleration; or a walker whose trajectory obeyed some set of differential equations, etc. One could write down partial differential equations for the density of walkers in all of these situations.

A short remark: You should realize the power of formulating this in terms of walkers. Many many applications can be formulated in this way and one can apply the ideas we are inventing here to these applications. As a simple example, consider cars moving on a road. They are disciplined. But they certainly do not move at constant velocity! You know from your experience driving that relatively benign changes in the car motions can lead to large consequences—traffic jams and whatnot. We will see eventually how this can come about. Another example involves the case of molecular motors walking along microtubules.

Let us suppose that our walkers all obey the following equation of motion:

$$\frac{dx}{dt} = V(x), \quad (9.51)$$

so that the walkers walk with a velocity that depends on their position. If we repeat our derivation of disciplined walkers, this law could be implemented by letting the step size $a = a(x)$, or by letting $\tau = \tau(x)$. In either case, we arrive at the exact same equation as before, namely

$$\partial_t p = -V(x)\partial_x p. \quad (9.52)$$

How to solve this equation? We expect information to still be propagated by our walkers; this time the information will just move on trajectories that are more complicated than those above. Let us use the ansatz: $p(x, t) = p(x(t), t)$. Namely, we imagine that the solution will evolve on a trajectory given by $x(t)$. If we then compute

$$\frac{dp}{dt} = \partial_t p + \frac{dx}{dt}\partial_x p, \quad (9.53)$$

we see that our equation corresponds to

$$\frac{dp}{dt} = 0, \tag{9.54}$$

as long as the association

$$\frac{dx}{dt} = V(x) \tag{9.55}$$

is made—namely, the walkers must walk along the trajectories of the velocity field!

A special and important case of the problem we are discussing occurs when

$$V(x) = \frac{-1}{\zeta} \frac{dU}{dx} \tag{9.56}$$

where here $U(x)$ is the potential, and ζ is the mobility of the walker. We will see this example subsequently when we start combining effects.

9.4 Biased Random Walkers

Intermediate between the case of *disciplined walkers* and *random walkers* is the case of biased random walkers. Let us imagine that there is a probability α of moving to the left, and a probability $1 - \alpha$ of moving to the right.

Then if we proceed as before, and let $p_i^n = p(ia, n\tau)$ be the probability that the walker is at position i at time n . Then the rule for the motion of our walker translates into

$$p_i^{n+1} = \alpha p_{i-1}^n + (1 - \alpha) p_{i+1}^n. \tag{9.57}$$

Now this translates into the equation

$$p(x, t + \tau) = \alpha p(x - a, t) + (1 - \alpha) p(x + a, t). \tag{9.58}$$

Carrying out a Taylor series expansion as before, we find the equation

$$\tau \partial_t p = (1 - 2\alpha) a \partial_x p + \frac{a^2}{2} \partial_{xx} p, \tag{9.59}$$

or

$$\partial_t p = -U \partial_x p + D \partial_{xx} p, \tag{9.60}$$

where $D = a^2/(2\tau)$ and $U = a/\tau(2\alpha - 1)$. This equation combines both advection and diffusion!

9.5 Biased Not Boring Random Walkers

The random walkers thus described are boring, in that at every step they move to the left or the right but they do the same way. The natural generalization of this is to allow the probability of moving to the right to depend on where you are—namely, that $\alpha = \alpha_i$.

Then, the equation for the probability of walkers becomes

$$p_i^{n+1} = \alpha_{i-1} p_{i-1}^n + (1 - \alpha_{i+1}) p_{i+1}^n. \quad (9.61)$$

We now carry out our Taylor series expansion, as before. To do this we must expand both $p_{i\pm 1}^n$ as well as $\alpha_{i\pm 1}$: We therefore have that

$$\tau p = \partial_x \left((1 - 2\alpha) a p \right) + \frac{a^2}{2} \partial_{xx} p. \quad (9.62)$$

9.6 Combining different classes of effects

All equations do not fall neatly into one of these classes—in fact most are combinations of these, containing some terms of one type and others of another type. We now discuss what happens when the consequences of different classes of effects are combined. First we will consider combining linear equations, and then we will begin to ask how to make them nonlinear.

9.6.1 Combining Diffusion and Advection

Consider

$$\partial_t u + V \partial_x u = D \partial_{xx} u. \quad (9.63)$$

If we write $u(x, t) = w(x - Vt, t)$, then it is straightforward to show that

$$\partial_t w = D \partial_{xx} w. \quad (9.64)$$

Therefore combining diffusion with advection leads to a diffusion equation, it is just that we must move to a frame that is co-moving with the advected velocity! The same conclusion holds even if the advection velocity $V = V(x)$. Then we must consider $u(x, t) = w(x - x_0(t), t)$. If we plug this ansatz into equation (9.63) we obtain

$$\partial_t w - \dot{x}_0(t) \partial_x w + V(x) \partial_x w = D \partial_{xx} w. \quad (9.65)$$

Hence if we choose $\dot{x}_0 = V(x_0)$ we also reduce this to a simple diffusion equation.

9.6.2 Fokker planck equations

to be included

9.6.3 Combining Diffusion and Growth

The above effect was not particularly surprising. Combining terms can however lead to quite surprising effects: The combination of diffusion and exponential growth.

The combination of these two factors is widespread in nature. The first analysis of the phenomenon that we will describe below that I am aware of was by the great population geneticist R.A. Fisher. He was interested in modelling the spread of a new phenotype into a population. The question was: how fast does it spread?

The basic mathematical argument

Our goal is to combine two different effects: exponential growth, as exemplified by

$$\frac{du}{dt} = \alpha u, \quad (9.66)$$

and diffusion, as exemplified by

$$\frac{du}{dt} = D\partial_{xx}u. \quad (9.67)$$

Hence we will begin by studying the simple system

$$\frac{du}{dt} = \alpha u + D\partial_{xx}u. \quad (9.68)$$

We will imagine that the initial condition $u(x, 0)$ is localized on the x axis, and that the boundary conditions are that u (which might represent the number of people with a particular phenotype) decays at $\pm\infty$.

We can solve this equation by writing

$$u(x, t) = e^{\alpha t} F(x, t).$$

Hence

$$\frac{du}{dt} = \alpha u + \partial_t F e^{\alpha t},$$

and plugging into our equation we obtain

$$\partial_t F = D\partial_{xx}F. \quad (9.69)$$

Hence F obeys a diffusion equation!

Now, let's take our initial condition to be very localized so that we can approximate it by a delta function. Then we know from our previous work that the Green's function for the diffusion equation is:

$$F = \frac{1}{\sqrt{4\pi Dt}} e^{-x^2/4Dt}$$

so that

$$u(x, t) = \frac{1}{\sqrt{4\pi Dt}} \exp \left(\alpha t - \frac{x^2}{4Dt} \right). \quad (9.70)$$

Now we can rewrite the argument of the exponential as

$$\alpha t - \frac{x^2}{4Dt} = \frac{4D\alpha t^2 - x^2}{4Dt} = \frac{(ct - x)(ct + x)}{4Dt},$$

where here $c = \sqrt{4D\alpha}$.

Now, let's examine this formula: When $x \gg ct$ or $x \ll -ct$ the argument is positive and hence the solution decays exponentially. The *front* which separates the region where the solution is growing and that where the solution is shrinking obeys

$$x_{front} = \pm ct = \sqrt{4D\alpha} t.$$

This is a remarkable result: it demonstrates that the front moves at constant velocity! Combining diffusion and growth leads to constant velocity "waves". The consequence of this result is profound, on a number of levels: first of all it leads to a simple prediction for how the front velocity depends on the measured parameters. Secondly, it demonstrates that there is another mechanism for things to move at constant velocity other than our 'disciplined walker' model from early on: if the walkers move randomly and grow the front also moves at constant velocity. This particular result is especially important within biological contexts—besides large scale phenomena like the spread of the phenotype in the population, there are a myriad of examples at the scales of cells. One prominent example is the amoebae *Dictyostelium*, which sends out spiral waves which help different cells communicate with each other.



Saturation

The one odd feature of the solution we constructed in the last section is that u increases exponentially. In realistic examples the population size can only reach some maximum size. For example the organisms might run out of food, or something else might come in to inhibit their growth.

At the level of ODE's, the simplest way to account for the saturation of the population size is to modify our growth equation to

$$\frac{du}{dt} = \alpha u \left(1 - \frac{u}{N} \right). \quad (9.71)$$

In this equation the fixed point at $u = 0$ is unstable, whereas the one at $u = N$ is stable. Hence the population will start out small near $u = 0$ and then saturate at $u = N$. People call N the carrying capacity of the system.

This then motivates the change in the model for the spread of the population:

$$\partial_t u = D \partial_{xx} u + \alpha u \left(1 - \frac{u}{N}\right). \quad (9.72)$$

Now this equation is nonlinear and hence we cannot use our simple trick from above to solve it. Intuitively however we should note that when $u \ll N$, the equation is essentially the one we analyzed above and we expect there to be travelling waves.

To bear this out, let us look for solutions of this equation which have the form of a travelling wave: Namely we will write

$$u(x, t) = F(x - ct).$$

This is in many ways a courageous move—we are trying to get a wave out of diffusion and growth—but given the discussion above it is the natural thing to do. We then can compute

$$\partial_t u = -cF',$$

and

$$\partial_{xx} u = F''$$

so that we derive the following ordinary differential equation for F :

$$-cF' = DF'' + \alpha F \left(1 - \frac{F}{N}\right). \quad (9.73)$$

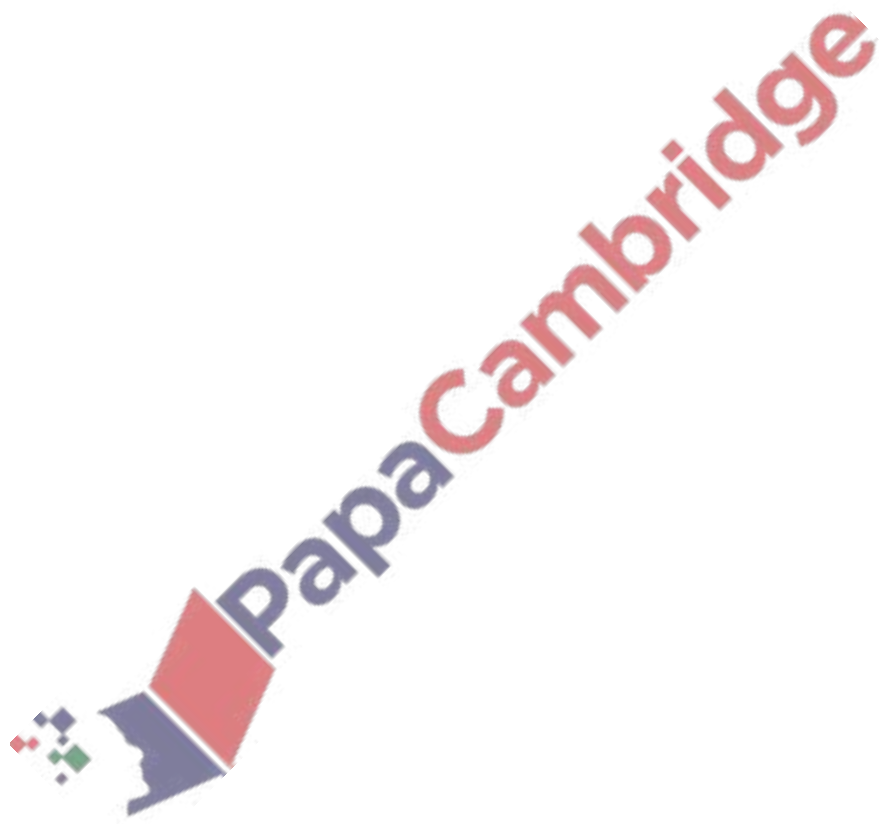
The derivatives here are with respect to the argument of F , which we will call $s = x - ct$. To make things a little simpler, let's change notation slightly: Let's write $F = Nf$. Hence $0 \leq f \leq 1$. Using this in the equation we have

$$-cf' = Df'' + \alpha f(1 - f). \quad (9.74)$$

Now we are interested in solutions of this ODE in which the population f changes from $0 \rightarrow 1$. We could seek solutions in which as $x \rightarrow -\infty$ $f \rightarrow 1$ and as $x \rightarrow \infty$, $f \rightarrow 0$; or alternatively we could seek solutions in which as $x \rightarrow -\infty$ $f \rightarrow 0$ and as $x \rightarrow \infty$, $f \rightarrow 1$.

The question then is to figure out whether such solutions exist: Now we should say from our experience in the previous section that in that case we found that the exponentially growing region was connected to the zero density region and that there is a solution that expands in both directions. In particular there is a front that moves both to the right and to the left. Hence we would rather expect that if we connect $u = 1(x = -\infty)$ to $u = 0(x = \infty)$ this will work with $c > 0$ and $u = 0(x = -\infty)$ to $u = 1(x = \infty)$ this will work with $c < 0$.

Solving this connection problem can be done using the methodology that we discussed previously in this course.



10 Integrals derived from solutions of Linear PDE's

10.1 Integral Transforms

We will now continue in the spirit that we have begun, and spend a few weeks attacking the general subject of integral transforms, in the context of various types of real examples. Integral transforms arise most often when solutions to linear partial differential equations are expressed in terms of fourier series. As a simple example, consider the one dimensional wave equation

$$\partial_{tt}u = \partial_x \left(c^2 \partial_x u \right), \tag{10.1}$$

where c is the propagation velocity, If we write

$u(x, t) = a_k(t)e^{ikx}$, then the wave equation implies that

$$\ddot{a}_k = -c^2 k^2 a_k, \tag{10.2}$$

so that $a_k = a_k^0 e^{\pm ikct}$.

Hence the most general solution to the wave equation is of the form

$$u(x, t) = \int_{-\infty}^{\infty} dk a_k^0 e^{\pm ikct + ikx}. \tag{10.3}$$

From this it is easy to see that the general solution to the wave equation is of the form $u(x, t) = G(x - ct) + F(x + ct)$.

For the wave equation, the frequency of oscillation of a wave of wavenumber k is $\omega_k = kc$. There are general classes of equation for which this frequency-wavenumber relation (dispersion relation) can be more complicated: e.g. $\omega = \omega(k)$. An example of such an equation is Schrodinger's equation where

$i\partial_t\psi = \nabla^2\psi + V(x)\psi$. Here if the potential $V = 0$, we have that $\omega = k^2$. Another example is the bending beam equation

$\partial_{tt}u = -\partial_{xxxx}u$. Here you should show by repeating the above derivation that

$\omega = k^2$. When the dispersion relation is nontrivial, the equation is said to be dispersive.

Solutions to the wave equation become more complex in multidimensions, and/or when the material is not homogeneous, so that the propagation velocity has spatial dependence. Calculations of this sort underly the theory of optics, in which one wants to know how the scattering of light interacts with refracting media (e.g. lenses) where the index of refraction (and hence the propagation velocity) changes in a localized region of space. Questions like: what is the resolution of a lens, what is the optimal size of a lens, what is the nature of aberration, require careful analysis of this type of integral.

More complicated integral transforms arise in the solution of other types of partial differential equations. The wave equation is special in that the oscillation frequency ω of the component with wavenumber k scales linearly with k , so that the medium is *nondispersive*. In general system, the oscillation frequency $\omega = \omega(k)$, and this leads to complications that we will need to evaluate.

10.2 Contour Integration

Suppose we have a function $f(z)$ in the complex plane and consider

$\int_c f(\zeta) d\zeta$ where c is a closed contour. What is the value of this integral? If the function is analytic inside c then of course the integral vanishes (Cauchy's theorem). But what if it is nonanalytic? Let us study this case by considering f to be single valued and consider integrating over the Laurent series.

Show that the only singularities that contribute to the integral are poles, and hence derive the residue theorem.

Do some examples:

$$\int_0^\infty \frac{dx}{1+x^2}$$

$$\int_0^\infty \frac{dx}{1+x^3}$$

$$\int_0^\infty \frac{dx}{z^2+3z+2}$$

10.3 Asymptotics of Fourier-type integrals

Here we will discuss the asymptotics of the Fourier transform. Given a function $f(x)$, the fourier transform is defined (up to prefactors) as $\hat{f}(k) = \int_a^b f(x)e^{ikx}$. The Fourier transform is defined with

$[a, b] = [-\infty, \infty]$, or one can consider the asymptotics of this function in general, as $k \rightarrow \infty$. To understand what happens at large k , we will integrate by parts:

$$\int_a^b f(x)e^{ikx} dx = \int_a^b f(x) \frac{d(e^{ikx})}{ik} = e^{ikx} f(x) \Big|_a^b - \frac{1}{ik} \int_a^b f'(x)e^{ikx} dx.$$

Thus, as long as either (a) $f(x)$ is periodic on $[a, b]$, or (b)

$f(x)$ vanishes at the end points, this shows that $\hat{f} \sim 1/k \int e^{ikx} f'(x) dx$. Repeating the integration by parts N

times shows that, as long as $f(x)$ is n times differentiable, and as long as the end point contributions vanish, we have that

$\hat{f}(k) \sim \frac{1}{(ik)^n} \int_a^b f^{(n)}(x) e^{ikx} dx$. Now if the $(n+1)^{st}$ derivative is discontinuous, then integrating by parts one more time gives a

$$\hat{f}(k) \sim \frac{1}{(ik)^{n+1}} \int a^b f^{(n+1)}(x) e^{ikx} \sim \frac{1}{(ik)^{n+1}}. \text{ Thus, a function with } n$$

continuous derivatives has $\hat{f} \sim k^{-n-1}$.

A corollary of this statement is that if the function is infinitely differentiable, then the Fourier transform decays faster than any power: namely it decays *exponentially* with

k . This is why representing smooth functions by a Fourier transform is so powerful. On truncating a Fourier expansion at some wavenumber k^* , the error incurred in $\hat{f} \sim O(e^{-k^*})$; this means that upon reconstructing the function

$$f(x) \sim \int \hat{f}(k) e^{-ikx} dk, \text{ the error is}$$

$O(e^{-k^*})$, uniform in x ! Contrast this with the error that is made upon truncating an ordinary power series: truncating at the n^{th} term leads to an error going like x^n , depending on

x , and decaying only like a power law.

All of this analysis brings up two questions: First, what exactly *is* the decay of a Fourier transform of a given function? We have learned that it decays exponentially, but what sets the strength of the exponential? Secondly, how do the singularities of

$f(x)$, extended to the complex plane affect the convergence rate?

To answer this, let us assume that $f(x)$ can be analytically continued to $f(z)$ everywhere in the complex plane. We will consider $[a, b] = [-\infty, \infty]$. The function $f(x)$ is infinitely differentiable when restricted to the real axis, so that we know that $\hat{f}(k)$ decays exponentially. What happens in the complex plane? We convert our transform to a contour integral using standard methods: If $x > 0$, we close the contour with a semi-circle in the upper complex plane, while if $x < 0$ we close on the lower. If there are no branch points, we can use Jordan's Lemma to guarantee that the contour can be closed. In this case the value of the transform is just the sum of residues at all of the poles. If the poles are located at $\{z_i\}$ with strength α_i , then we have that

$\int_{-\infty}^{\infty} f(x) e^{ikx} dx = 2\pi i \sum_{n=1}^N \alpha_n e^{ikz_n}$. The size of the transform is therefore set by the distance of the *closest pole* to the real axis: The pole with the smallest $\Im z_n$ dominates the sum.

What if there are branch points? We have previously discussed that with branch points, we are not allowed to close the contour without encircling the branch cuts. This is the now infamous "keyhole" contour that we described in class last week. The keyhole contour comes in from $i\infty$ along the branch cut, encircles the branch point, and goes right out again to rejoin the Jordan's lemma circular contour. If we assume that

$f(z) = (z - z_j)^\nu \sum_{p=0}^{\infty} a_p (z - z_j)^p$ near the branch point (at z_j), then if we take $z = z_j + i\tau$, the path of integration is from $\tau = \infty \rightarrow 0$ and then

$\tau = 0 \rightarrow \infty$, with a phase shift of 2π in between. Namely,

$$\int_{\infty}^0 (i\tau)^\nu e^{-k\tau} i d\tau + \int_0^{\infty} (i\tau e^{i2\pi})^\nu e^{-k\tau} i d\tau.$$

This is just

$$e^{i\pi/2(\nu+1)} (e^{i2\pi\nu} - 1) i \int_0^{\infty} \tau^\nu e^{-k\tau} = \frac{2\sin(\pi\nu)}{k^{\nu+1}} e^{ikz_j - i\pi/2\nu} a_0 \Gamma(\nu + 1) + \dots$$

Thus, we have seen that the transform again decays exponentially, but this time the exponential is dominated by the location of the branch point!

Therefore, in general, we have seen that the rate of decay of the Fourier transform gives information about the location of the nearest singularity! Were we so inclined, we could invent a “Domb Sykes like method for the asymptotics of Fourier series. Note that the information about the nature of the singularity causing the nearest singularity to the origin is also encoded into the Fourier series.

10.4 Rainbows and Oscillatory Integrals

We now move on to study a class of oscillatory integrals that are more complicated than those described above. Despite their seeming obscurity, such integrals occur all of the time in science, either explicitly or implicitly. To demonstrate this and motivate our general methods we will start by describing the problem of the rainbow. This elegant, classical problem, contains within it examples of all of the integrals we would like to learn to evaluate!

The attached image of a rainbow shows several features one does not often see: first there is a double rainbow. Second if you look carefully into the arcs of each of the rainbow you can see that there are *supernumerary* rainbows inside, that is little replicas of rainbows. We would like to understand the phenomenology of rainbows: what sets the angle? why are the different colors at different angles? what is the intensity distribution of each color? Why does one sometimes see rainbows clearly and sometimes not see them so clearly?

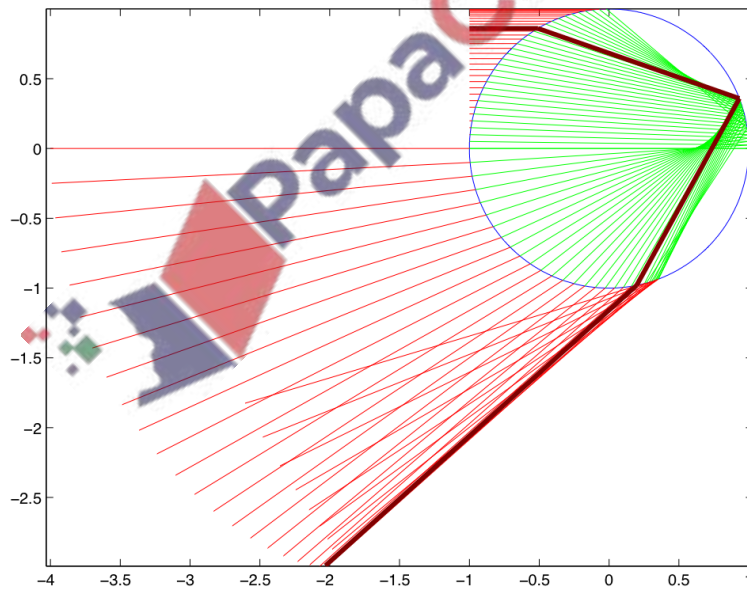
We will discuss these questions, and find out that to answer them we need to carry out oscillatory integrals. The types of oscillatory integrals that we will encounter come up in all sorts of applications (optics and beyond) but we will use the example of a rainbow to motivate our discussion.

Our account of the rainbow follows in places the excellent recent discussion of the rainbow by J.D. Jackson (Phys Reports 320, pgs 27-36 (1999)). This article is on the course web site.

The starting point in the theory of rainbows is due to Descartes, who pointed out that the intensity distribution around a rainbow is caused by the scattering of light off a spherical droplet. The figure shows the bundle of rays:



Figure of a double rainbow. (Taken off of a link from the UCAR website).



Scattering off of a spherical droplet.

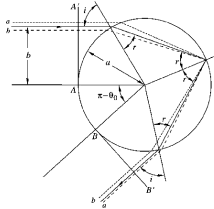


Fig. 1. Geometrical optics of a primary rainbow.

Notation for scattering off of a spherical droplet. This figure is taken from Fig. 1 of J.D. Jackson, Phys. Rep, 1999

What is apparent from the figure is that the scattering leads to a bunching of rays around the so called “ray of minimum deflection; we seek to compute the intensity distribution of the rays contributing to this bunching.

First we set up notation, shown in figure ??

The initial angle of the light relative to the tangent to the sphere is i ; the light is reflected initially to an angle r , where $n_1 \sin(i) = n_2 \sin(r)$,

n_1 and n_2 being the index of refraction of the two media, respectively. For air and water $n_2/n_1 = 4/3$. The total deflection angle of the light is given by

$\theta_D = 2(i - r) - (\pi - 2r)$; the light turns by angle $i - r$ on entering and leaving the sphere, and turns by $\pi - 2r$ while inside the sphere.

Let's first solve for the maximal angle of deflection. It is usually asserted that the angle of light reflected is at the maximum angle of deflection. Setting

$n = n_2/n_1$, we need to solve for $d\theta_D/di = 0$. Now the definition of θ_D implies that $d\theta_D = 2di - 4dr$,

but from Snell's law we have

$$\cos(i)di = n\cos(r)dr,$$

implying that the stationary angle satisfies

$$\frac{\cos(i)}{\cos(r)} = \frac{n}{2}.$$

Now $\cos(r) = \sqrt{1 - \sin(r)^2} = \sqrt{1 - \sin(i)^2/n^2}$, so that with a little algebra we have

$$\cos(i) = \sqrt{\frac{n^2-1}{3}}.$$

Taking $n = 4/3$, this gives the angle $\theta_D = 130^\circ$.

It is convenient for our analysis to introduce the 'impact parameter'. As depicted in figure ??, the impact distance $b = a \sin(i)$, and we will write the impact parameter as $x = b/a$. We then have that the deflection angle

$$\theta_D = \pi + 2\sin^{-1}(x) - 4\sin^{-1}(x/n). \quad (10.4)$$

Taking derivatives of θ_D with respect to x we obtain

$$\frac{d\theta_D}{dx} = \frac{2}{\sqrt{1-x^2}} - \frac{4}{\sqrt{n^2-x^2}} \quad (10.5)$$

$$\frac{d^2\theta_D}{dx^2} = \frac{2x}{(1-x^2)^{3/2}} - \frac{4x}{(n^2-x^2)^{3/2}}. \quad (10.6)$$

Written in these terms, $d\theta_D/dx = 0$ when $x_D = \sqrt{(4-n^2)/3}$ and evaluated at this x we have

$$\frac{d^2\theta_D}{dx^2} = \frac{9}{2} \frac{\sqrt{4-n^2}}{(n^2-1)^{3/2}}. \quad (10.7)$$

10.4.1 Colors

The colors of the rainbow arise because the index of refraction is wavelength dependent. For visible light the index of refraction varies from about 1.345 for violet (400nm) light to 1.330 for red (700nm) light. This is a change of $\Delta n = 0.015$. One can now ask how much the maximum deflection angle will change with changing index of refraction. One can show (Jackson/algebra) that with this change in the index of refraction there is a change in the deflection angle of $\Delta\theta = 1.89^\circ$.

10.4.2 Deflection angle of the rainbow and the Method of Stationary Phase

Now let us consider the intensity distribution of light through the rainbow. Let us call $\phi(x)$ the total amount of phase accumulated along a ray passing through the water droplet. Within the droplet, the total path length of the light ray is $4a \cos(r)$; hence the phase accumulated along this path is $4n_2 k a \cos(r)$. Outside the droplet we measure the phase accumulated between the entry and exit lines (see Fig 1 of Jackson, these lines

are AA' and BB'). Now the total length of this part of the ray is $2a(1 - \cos(i))$, so that the phase accumulated is $2kan_1(1 - \cos(i))$. Hence the total phase accumulated is

$$\phi(x) = 2ka(n_1(1 - \cos(i)) + 2n_2\cos(r)). \quad (10.8)$$

We can rewrite this in terms of the impact parameter x to be

$$\phi(x) = 2ka\left(1 - \sqrt{1 - x^2} + 2\sqrt{n^2 - x^2}\right)$$

Now, let us consider the total intensity of light that one observes far away from the scattering point.

This is just

$$\int_{-a}^a \exp\left[ik_{\parallel}z + i\mathbf{k}_{\perp} \cdot \mathbf{r}_{\perp} + i\phi(x)\right] dx \quad (10.9)$$

where here the vector $\mathbf{r} = (z, \mathbf{r}_{\perp})$ is the vector direction along which the observer is looking at the scattering, and $k_{\parallel}, \mathbf{k}_{\perp}$ are the wavevectors parallel and perpendicular to the propagation direction. The integral is over all incident rays, which we are parameterizing by the impact parameter x . We will assume that the z axis is oriented with an angle θ_* to the horizontal (measured in the same sense as the deflection angle itself). The vector \mathbf{k}_{\perp} is perpendicular to the propagation direction. If we write $\mathbf{r}_{\perp} = x$, where this is measured along the line BB' in the figure ??, then we have $\mathbf{r}_{\perp} \cdot \mathbf{k}_{\perp} = -k(\theta - \theta_*)x$, so that for a ray exiting the droplet at angle θ we have

$\mathbf{k}_{\perp} \cdot \mathbf{r} = -kb(\theta - \theta_*) = -ka(\theta - \theta_*)x = -ka\left((\theta - \theta_D) + (\theta_D - \theta_*)\right)$ where θ_D is the Descartes angle. The negative sign here is because \mathbf{k}_{\perp} is in the opposite direction to r_{\perp} . Hence we have that the intensity is given by

$$e^{ik_{\parallel}z} \int_{-a}^a \exp\left[i(-kax(\theta_D - \theta_*) - kax((\theta - \theta_D) + \phi(x)))\right] dx. \quad (10.10)$$

To determine the Descartes angle, let's suppose we are observing the system at $\theta_* = \theta_D$. Then we need to evaluate

$$e^{ik_{\parallel}z} \int_{-a}^a \exp\left[i\left(-kax(\theta - \theta_D) + \phi(x)\right)\right] dx. \quad (10.11)$$

10.4.3 Stationary Phase

How do we evaluate such integrals? Our integral is of the general form:

$\int_{-\infty}^{\infty} f(x)e^{ik\phi(x)} dx$? One option would be to follow our example of integrating by parts with Fourier transforms: We do this writing the integral as

$\int_{-\infty}^{\infty} \frac{d}{dx} \left(\frac{f(x)}{\phi'(x)} \right) e^{ik\phi(x)} dx = \frac{f(x)}{\phi'(x)} \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{f(x)}{\phi'(x)} e^{ik\phi(x)} \phi''(x) dx$. Now, if $\phi' = 0$ anywhere in the integration region, we have destroyed the wonderful convergence properties of this integral, as the function

f/ϕ' is no longer perfectly smooth—it diverges at the zeros of ϕ' . The consequence of this is that the integral will no longer decay exponentially for analytic f . How does it decay?

The method to determine this is called the *method of stationary phase*; it was invented by Stokes in a study of wave interference (Stokes's problem was not unlike the rainbow problem described above.). As you will see, the method looks a lot like Laplace's method: let us assume that $\phi'(c) = 0$, where

$-\infty < c < \infty$. Then we can expand the argument of the exponential to be

$$\int_{c-R}^{c+R} f(x) \exp \left[ik \left(\phi(c) + \frac{\phi''(c)}{2} (x-c)^2 \right) \right] dx.$$

Here we are doing the integration just in a small part around the place where ϕ' vanishes. Letting $x = c + \tau \sqrt{2/(\phi''(c)k)}$, so that the integral becomes

$$\int_{-R\sqrt{k\phi''(c)/2}}^{R\sqrt{k\phi''(c)/2}} f(c) e^{i\tau^2} \sqrt{\frac{2}{\phi''(c)k}} d\tau.$$

Clearly as $k \rightarrow \infty$ the integral becomes

$$2\sqrt{\frac{2}{\phi''(c)k}} \int_0^{\infty} e^{i\tau^2} d\tau.$$

(The factor of two is because the integral is even—note the integration range is half the integration range of the previous formula.) We can evaluate this integral by using a contour which moves out along the real τ axis, and comes in along the line at an angle

$\pi/4$ to the vertical. This demonstrates that

$$\int_0^{\infty} e^{i\tau^2} d\tau = -e^{i\pi/4} \int_0^{\infty} e^{-s^2} ds = -e^{i\pi/4} \sqrt{\pi}/2.$$

Hence we have shown that the integral is

$$-e^{ik\phi(c)} e^{i\pi/4} \sqrt{\frac{8\pi}{\phi''(c)k}} f(c), \text{ to leading order.}$$

What is the error? There are two contributions: first, we localized the integral; there will be corrections from the Taylor series of $f(x)$ about $x = c$, etc. Since the characteristic scale of the integral is $k^{-1/2}$, we expect that these corrections will be of this order. There is also an error from the fact that we excised the integral we did above from the full integral; the characteristic size of the remaining integrals \int_R^{∞}

and $\int_{-R}^{-\infty}$ are of order k^{-1} ; this is the same order as the local corrections described above. In particular, this estimate demonstrates that the leading order contribution to the integral is the behavior near the stationary point.

Note that we have assumed in this analysis that $f(c) \neq 0$. If

$f(c) = 0$, then it is no longer apparent whether the stationary point will give a relevant contribution to the integral. This can only be discovered by carrying out a full analysis.

Note that the one major difference between this method and the Laplace method is that here, when we excised the region of interest from the full integral, the contribution from the non-excised region decays *algebraically*, like k^{-1} . In contrast in the Laplace method this error decays exponentially. Therefore the application of the method of stationary phase is more subtle in general.

10.4.4 Back to the Rainbow

The analysis above demonstrates that the integral is dominated by its stationary point, and that this recovers the often-quoted argument that the rainbow is dominated by the ray of minimal deflection. Why? We need to evaluate equation (10.11), and according to our analysis the integral is dominated at the stationary phase point, which is

$$\frac{d}{dx} \left[-kax(\theta - \theta_D) + \phi(x) \right] = -ka(\theta - \theta_D) - ka \frac{d\theta}{dx} + \frac{d\phi}{dx} = 0.$$

But from formula (10.8) for ϕ , we have that

$$\frac{d\phi}{dx} = ka \left(\frac{x}{\sqrt{1-x^2}} - \frac{2x}{\sqrt{n^2-x^2}} \right) = kax \frac{d\theta}{dx}. \quad (10.12)$$

Hence, the stationary phase point occurs when $d\theta/dx = 0$, or at the Descartes angle!

On the other hand, we would really like to compute the intensity around this minimal deflection point—how does the intensity change? Naively we expect that the variation in the intensity will be qualitatively different when $\theta_* < \theta_D$ than when $\theta_* > \theta_D$. In the former case we are in no-man's land, to speak. None of the rays can arrive there. On the other hand in the latter case we expect to see interference between the different rays.

To study this, we recall our integral

$$e^{ik_{||}z} \int \exp i \left(-kax(\theta_D - \theta_*) - kax(\theta - \theta_D) + \phi(x) \right). \quad (10.13)$$

We can use equation 10.12 to expand the phase around the Descartes ray: Namely

$\phi' = kax\theta' = kax_D\theta' + ka(x-x_D)\theta'$, but on the other hand since θ_D is also a minimum of $\theta(x)$ we have $\theta(x) = \theta_D + \theta''(x_D)(x-x_D)^2/2$, so that

$\phi(x) \approx \phi_D + kax_D(\theta - \theta_D) + ka\theta''(x - x_D)^3/3 + \dots$, The complete phase within our integral is therefore

$$\Phi = -ka(\theta_D - \theta_*)x - kax(\theta - \theta_D) + kax_D(\theta - \theta_D) + ka\theta''(x - x_D)^3/3 + \dots \quad (10.14)$$

or if we use the fact that $\theta = \theta_D + \theta''/2(x - x_D)^2$ (θ_D being a minimum!) we have that

$$\Phi = -ka(\theta_D - \theta_*)x - ka\theta''(x - x_D)^3/6. \quad (10.15)$$

We therefore need to evaluate

$$e^{ik_{||}z} \int_{-a}^a \exp i \left(-kax(\theta_D - \theta_*) - ka\theta''(x - x_D)^3/6 \right). \quad (10.16)$$

Note that in contrast to our stationary phase methodology from above

\item When $\theta_D = \theta_*$, the leading term in the phase is actually a *cubic* instead of the generic quadratic we expected from our argument above. Thus the formula we derived there is not going to work for the intensity. Instead, setting $\theta_D = \theta_*$ we have the integral

$$\int_{-a}^a \exp \left[-ika(x - x_D)^3/6 \right] dx.$$

If we write $x - x_D = \left(\frac{6}{ka} \right)^{1/3} t$ then the integral is just

$$\left(\frac{6}{ka} \right)^{1/3} \int_{-\infty}^{\infty} e^{it^3} dt.$$

Hence we achieve the beautiful result that the intensity of light at the Descartes angle scales like $(ka)^{-1/3}$! \item If we have $\theta_* \neq \theta_D$, the stationary phase formula gives an imaginary answer! What to do? Hence the *point of stationary phase* is not along our integration path! The method therefore fails.

10.5 Saddle Points

How to proceed? Without a stationary phase point the integral is oscillating, and completely not obvious how to evaluate.

The idea we will use here is to use the one freedom we have that we have not yet exploited: name the deformation of the integration contour in the complex plane. This is an idea invented by Peter deBye. While we are deforming we might as well hope we can

get rid of the oscillations entirely, and this way we can just do a Laplace type integral and get the integrand to decay exponentially as in Laplace's method.

Namely, given

$$I(x) = \int_C h(t)e^{x\rho(t)} dt \quad (10.17)$$

the idea is to deform the contour to make the integral as much like Laplace's method (and as little like stationary phase) as possible. We will think of t as a complex variable and distort the contour into the complex t plane. Ideally we will find a contour for which $\Im\rho(t)$ is constant along the contour so there are no oscillations. $\Re\rho(t) < 0$ along the path of the integration, ie there is a maximum in the Real part along the contour. If we assume that $\rho(t)$ is analytic, then $\nabla^2\Re\rho = 0$ so that $\Re\rho$ has no maxima or minima. Thus the only way that the gradient of $\Re\rho$ can vanish along a contour is if it is a saddle point.

Now, the nature of functions in the complex plane implies something remarkable:

if we write

$$\rho(t) = \phi(t) + i\psi(t), \quad (10.18)$$

where both ϕ

and ψ are real valued functions, and we think of $t = u + iv$, then

$$\Re\nabla\rho = (\partial_u\phi, \partial_v\phi). \quad (10.19)$$

Now consider

$$\nabla\phi \cdot \nabla\psi = \partial_u\phi\partial_u\psi + \partial_v\phi\partial_v\psi = (\nabla\psi \cdot \nabla)\phi. \quad (10.20)$$

By the Cauchy Riemann equations, it is apparent that this quantity is zero. On the other hand, the quantity can be interpreted as the directional derivative of ψ along the direction of the gradient of ϕ . Hence, if we consider a curve in the complex plane that moves along the gradient of ϕ , the imaginary part is constant! Thus, we have achieved exactly what we had hoped for—in general there exist contours where the real part decreases along a contour (from a maximum point) and where the imaginary part of ρ is constant. Given that as we said above the extrema of $\Re\rho$ must be saddle points

To identify these contours, we need to look for saddle points of ρ , namely points where $\rho' = 0$. We then need to identify the contour which has $\Re\rho$ decreasing, and transform our integral to this contour. In doing the transformation it is possible we will need to circumvent poles or branch points and of course, we will have to pay for such indiscretions. It should be anticipated that along the new contour, the integral will reduce to a Laplace integral. Recall that Laplace integrals are dominated by the region around the maximum. Therefore, to leading order, we will only need to know the behavior of the integrand and the saddle point contour near the saddle point! Hence, the requirement that we know the saddle point contour is not as onerous as it might seem.

10.5.1 Elementary Examples

The first example here illustrates how powerful it can be to manipulate a contour. Then we will turn to an example of Carrier, and to our rainbow.

10.5.2 Example 1: Finite Fourier Transform

Before we considered the large k behavior of the fourier transform. Let us reconsider this problem for a finite fourier transform:

$$I(x) = \int_0^\pi f(t)e^{ixt} dt \tag{10.21}$$

To evaluate this consider the integral

$\int_C f(t)e^{ixt} dt$, over the rectangular contour with the following four legs: (1) $0 \rightarrow \pi$ along the x axis; (2) $\pi \rightarrow \pi + iT$ parallel to the y axis; (3) $\pi + iT \rightarrow iT$ parallel to the x axis; and (4) $iT \rightarrow 0$ along the y axis. The integral along leg (3) vanishes as we send $T \rightarrow \infty$, and clearly if f is analytic inside our domain there is no residue contribution, so we have that $I(x) = \text{Integral along leg 2} + \text{Integral along leg 4}$, or

$$I(x) = \int_\infty^0 f(is)e^{-xs} ds + \int_0^\infty f(\pi + is)e^{ix\pi} e^{-xs} ds. \tag{10.22}$$

In the limit that $x \rightarrow \infty$ we can just apply the main idea from Laplace's method (both integrals are dominated as $x \rightarrow 0$, to give us

$$I(x) = i(f(\pi)e^{ix\pi} - f(0))\frac{1}{x}. \tag{10.23}$$

10.5.3 Example 2: An example of Carrier

$$I(x, z) = \frac{1}{2\pi i} \int_{\gamma-i\infty}^{\gamma+i\infty} \frac{e^{x(sz-\sqrt{s})}}{s} ds. \tag{10.24}$$

Example 3: Back to the Rainbow Now let's go back to our rainbow. We have the integral

$$e^{ik_{||}z} \int_{-a}^a dx \exp i \left(-kax(\theta_D - \theta_*) - ka\theta''(x - x_D)^3/6 \right). \tag{10.25}$$

which is proportional to

$$\int_{-a}^a dx \exp i \left(-ka(\theta_D - \theta_*)(x - x_D) - ka\theta''(x - x_D)^3/6 \right). \quad (10.26)$$

Let us define $s = \left(\frac{ka\theta''}{2} \right)^{1/3} (x - x_D)$. Using this substitution in the integral, we arrive at

$$I(z) = \left(\frac{2}{ka\theta''} \right)^{1/3} \int_{-\infty}^{\infty} \exp \left[-i(zs + s^3/2) \right] ds, \quad (10.27)$$

where

$$z = \left(\frac{2(ka)^2}{\theta''} \right)^{1/3} (\theta_D - \theta_*). \quad (10.28)$$

The function

$$Ai(z) = \int_{-\infty}^{\infty} \exp \left[-i(zs + s^3/2) \right] ds, \quad (10.29)$$

is called the Airy function, after George Airy, who actually invented it in the context of the rainbow problem. Our job now is to figure out the behavior of $Ai(z)$ in the limit that z is both large and positive, and large and negative. Please note that the $z \rightarrow \infty$ limit is the relevant limit because in our rainbow $ka \gg 1$ —the droplet radius is much larger than the wavelength of the light.

The Airy Function: Deforming the Contour Now we would like to apply the Saddle point method to the Airy function, to discover its functional behavior for large z . Before doing this we need to briefly discuss how we are going to define our Contour integrals—ie what are the constraints we are going to place on our deformations of the contours?

Let us write

$$Ai(z) = \int_C \exp \left[-i(zs + s^3/2) \right] ds, \quad (10.30)$$

where here C is a contour in the complex s plane. Which contours are allowed? It is standard to define $s = it$, and then we find that

$$Ai(z) = \int_C \exp\left[zt - t^3/2\right] dt. \quad (10.31)$$

Clearly in the complex t plane our contour should go from roughly $-i\infty \rightarrow +i\infty$ –this is what is called for by our algebraic manipulations. How much can we move our contour around? For the integral to converge we need to require that the integrand approaches zero (or at least does not diverge!) as we move to the ends of the contour.

Let us therefore examine the integrand. This is $e^{zt-t^3/2}$. Clearly as $|t| \rightarrow \infty$, the t^3 term wins. If we write $t = Re^{i\theta}$ then this is just $t^3 = R^3 e^{3i\theta} = R^3(\cos(3\theta) + i\sin(3\theta))$. Hence, as $R \rightarrow \infty$ the integrand will behave as $e^{-R^3/2\cos(3\theta)}$. Requiring this to decay exponentially is tantamount to requiring that the contour obeys $\cos(3\theta) > 0$. There are three regions in the complex plane that allow this:

- \item $-\frac{\pi}{6} \leq \theta \leq \frac{\pi}{6}$.
- \item $\frac{3\pi}{6} \leq \theta \leq \frac{5\pi}{6}$.
- \item $-\frac{5\pi}{6} \leq \theta \leq -\frac{\pi}{2}$.

Thus for the integral to remain finite we must nudge our contour into moving from region (3) to region (2). This can be done with only a small change in the contour (indeed our definition of $Ai(z)$ above is along the extremes of this boundary).

Now we need to evaluate our integral. Where are the stationary points? Let us first assume that $z > 0$. Note that z can be either positive or negative depending on the sign of $\theta_* - \theta_D$. The assumption that $z > 0$ corresponds to $\theta_* < \theta_D$ –ie here we are in no-mans land, where the intensity should be zero.

Now, the function $\rho = zt - t^3/2$ obeys $\rho'(t) = 0$ at $t = \pm\sqrt{2z}$.

The stationary point which is closest to our integration path is at $t = -\sqrt{2z/3}$. Let us expand ρ about this stationary point, writing $t = -\sqrt{2z/3} + \tilde{t}$. This gives

$$\rho = -cz^{3/2} + \sqrt{3z/2}\tilde{t}^2.$$

Here the constant $c = (2/3)^{3/2}$. If we write $\tilde{t} = re^{i\theta}$, then this is

$$\rho = -cz^{3/2} + \sqrt{3z/2}r^2(\cos(2\theta) + i\sin(2\theta)).$$

Now we would like our contour to have the feature that the real part of ρ decays exponentially, and the imaginary part of ρ is constant. We can attain this if we choose $\theta = \pi/2$, namely our contour should go through the saddle point at a right angle to the access.

Deforming the integration contour to this path gives

$$\int e^\rho = e^{-(2z/3)^{3/2}} \int_{-\infty}^{\infty} d\tilde{t} e^{-\sqrt{3z/2}\tilde{t}^2} = e^{-(2z/3)^{3/2}} \sqrt{\frac{\pi}{\sqrt{3z/2}}}.$$

If we go back and use equation (49) 's definition for z , we can show that the intensity in the dark zone on the rainbow is given by the beautiful formula

$$I = \sqrt{\pi} \left[\frac{2}{3} \right]^{1/6} \frac{1}{\sqrt{ka}} \frac{1}{\theta^{1/3}(\theta_D - \theta_*)^{1/4}} \exp \left[-2/3ka(\theta_D - \theta_*) \right]. \quad (10.32)$$

The light side What about the light side of the rainbow? When $\theta_* > \theta_D$, z becomes negative and there are now two saddles at $t = \pm i\sqrt{2|z|/3}$. Our integration contour now requires that we go through both of them (NEED A FIGURE HERE). A calculation gives the answer

$$\int e^\rho = \sqrt{\frac{\pi}{\sqrt{|z|}}} \sin\left(\frac{2}{3}|z|^{3/2} + \frac{\pi}{4}\right).$$

We thus arrive at the beautiful formula for the intensity in the light region:

$$I = \sqrt{\pi} \left[\frac{2}{3} \right]^{1/6} \frac{1}{\sqrt{ka}} \frac{1}{\theta^{1/3}(\theta_D - \theta_*)^{1/4}} \sin \left[2/3ka(\theta_* - \theta_D) + \pi/4 \right]. \quad (10.33)$$

10.6 A terrible Integral

The following problem was assigned for homework one year. Consider the integral

$$I(x) = \int_0^{10} \frac{e^{-x(4t^2+5t)} \sin(13x(t+3t^3))}{1+8t^3} dt. \quad (10.34)$$

The question is to develop approximate expressions for the integral at large x and compare to numerical calculations of the integral.

At first glance, this is a saddle point problem; this motivates us to rewrite the problem in the form

$$I(x) = \Im \left(\int_0^\infty \frac{\exp(-x\phi(t))}{1+8t^3} dt \right),$$

where $\phi(t) = -4t^2 - 5t + 13i(t+3t^3)$. However, the integral is sufficiently complex that there are several possibilities for what might dominate the integral. These include: \item The saddle points. These will occur when $\phi' = 0$. From class we know that generically the saddle point contributions cause the integral to decay like $x^{-1/2}$. \item The possible residue contributions. Given that the poles exist at $t = 1/2(-1)^{1/3}$, these will contribute exponentially decaying contributions to the integral. \item The final possibility is that there could be contributions from the *endpoints* of the integration regions when the integral is extended to the complex plane. As discussed in class these generically give behaviors which decay like $1/x$.

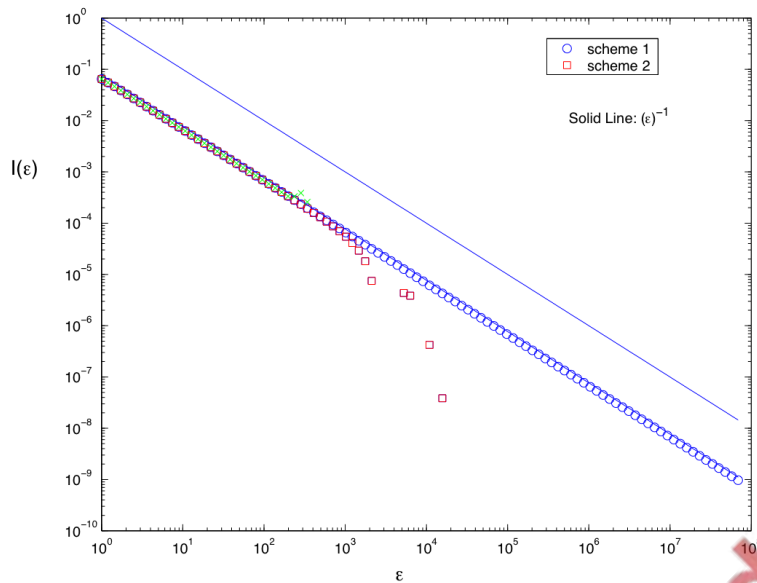


Figure 10.1. Numerical integration of the integral.

On the other hand, the algebra is complicated, and it is very tedious to go through and calculate all of the contributions, especially when we might recognize *a priori* that all that really matters is the *dominant* contribution. For this reason, let's solve the problem backwards by doing the numerics first. In the present day, I think this is just as valid a way of solving problems in today's world as any other.

The figure below plots a numerical evaluation of the integral as a function of x . Four different data sets are shown: the circles, squares and x's are three methods of integration; the solid like is the law $(x)^{-1}$.

The numerical integration shown in the above figure was performed first with Matlab's numerical integration routine 'quad'; and second by brute force: Given the integral $\int_a^b f(x) dx$, I simply broke the interval into N intervals and then approximated the integral as $\sum_{i=0}^{N-1} f(x_i)\Delta x$. By increasing the number of points in the interval N one can then see that the integral converges to a well defined law over a wider and wider range. If one discretizes the entire interval the convergence rate is terribly slow: the squares show what happens if the entire integral is discretized into 10^4 points. Instead I discretized the region between $0 < x < 10/\epsilon$, reasoning that the integrand decays away to nothing beyond this scale. These are the circles, which agree with the squares up to some point

On the other hand, neither Matlab nor Mathematica's numerical integration routines fared so well—both produced either nonsense (matlab) or error messages (mathematica) when the integration routine got too large. The reason is that both of these programs are supposed to intelligently disperse the integration points to get the most accurate answer possible—the algorithms for doing this apparently miss that all of the action is near the origin!

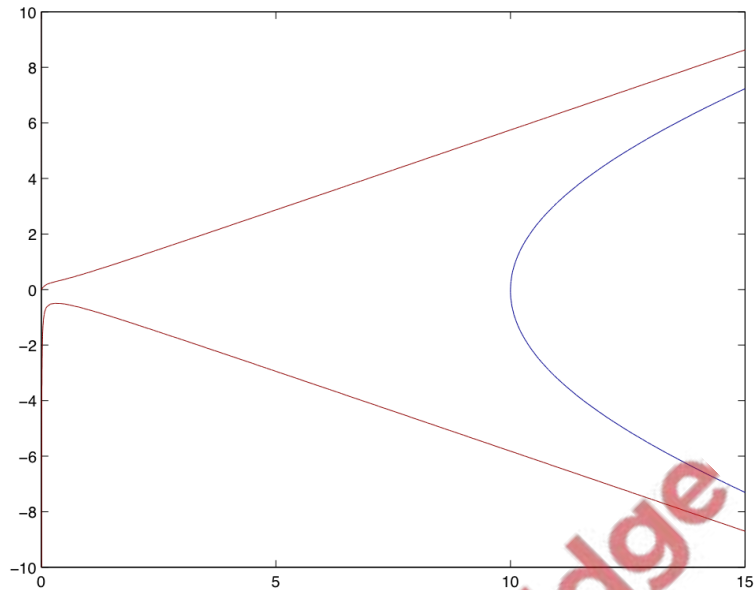


Figure 10.2. Contours of constant phase going through the endpoints of the integral (at $t = 0, 10$). Note that near the origin the constant phase contour is linear..

Thus we have shown that the integral decays like $1/x$. Recall that (a) saddle points tend to decay exponentially when they are off of the real axis; (b) Laplace contributions decay like

$1/\sqrt{x}$ when they are on the real axis in the interior of the domain; (c) poles should also decay exponentially, as should branch points. Thus the only possibility for the $1/x$ is that it must be an end point contribution to the integral.

It is very likely that the endpoint in question is the one at $t = 0$, since at $t = 10$ the integrand is very small. (If you doubt this, you could test it numerically by varying the upper limit of integration and showing that the integral does not change very much.)

With this information, we are almost done: we need to investigate the constant phase contours emanating from the origin. Now since

$\phi(t) = -4t^2 - 5t + 13i(t + 3t^3)$, near $t = 0$ we have that $\phi \approx t(5 + 13i)$. Thus we have that near the origin, the constant phase contour moves along the line

$$t = s \frac{5+13i}{5^2+13^2}.$$

Just as a check, figure 2 shows constant phase contours for this problem, and you can see that near the origin, the contour is linear. I have also included in figure 2 the constant phase contours that pass through $t = 10$: the integration path we must take starts along the real axis from $t=0:10$ (our original integration path), moves out along one constant phase path and in along the other.

Now, we are almost done: only the constant phase path through the origin will contribute to leading order, since the residue contribution is exponentially small, and the contribution at

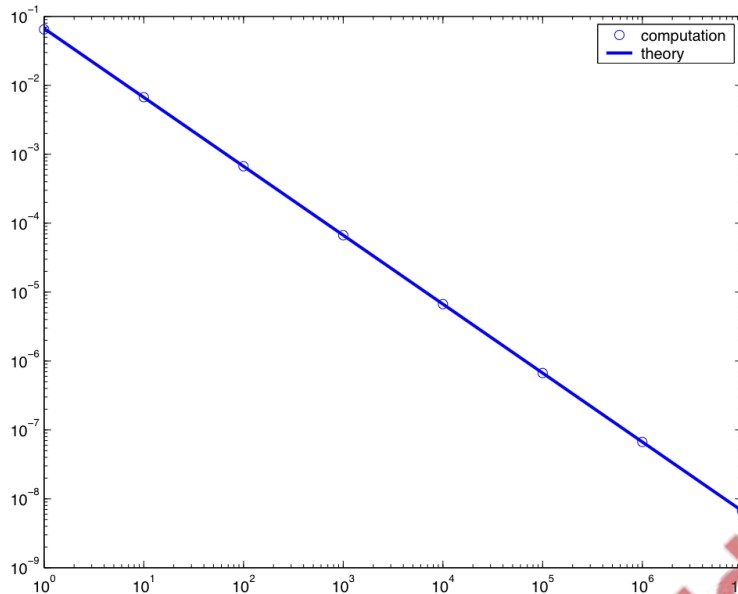


Figure 10.3. Comparison of the formula $I(x) = 15/194x^{-1}$ with the numerical simulations.

$t = 10$ is also exponentially small. As for the saddle points, neither is along an integration path, and so we don't have to worry about these. We therefore have

$$I(x)\Im \approx \int ds \frac{5+13i}{5^2+13^2} \exp -xs = \frac{13}{5^2+13^2} \frac{1}{x} = \frac{13}{194} \frac{1}{x}.$$

Figure 10.3 compares the numerical computations to this exceedingly simple formula.

Thus, we understand the integral asymptotically. Our understanding is nonrigorous, but convincing, combining numerics and asymptotics to understand what is going on.

10.7 Some Applications of Stationary Phase Ideas

Here we discuss a few applications of the method of stationary phase. First, in many different problems one often decomposes the solution to some partial differential equation into Fourier modes, of the form

$$\phi(x, t) = \int_{-\infty}^{\infty} dk \hat{\phi}(k, t) e^{ikx},$$

where further analysis shows that

$$\hat{\phi}(k, t) = \hat{\phi}(k, 0) e^{-i\omega(k)t}.$$

The quantity $\omega(k)$, the frequency at which a disturbance at wavenumber k oscillates, is called a *dispersion relation*. Now we have c The question we would like to address is the value of this integral at large time. To address this we write $x = Ut$, for some velocity U , so that the integral becomes:

$$\phi(x, t) = \int_{-\infty}^{\infty} dk \hat{\phi}(k, t) e^{it(kU - \omega(k))}.$$

The method of stationary phase then states that the integral is dominated at the value of k so that

$U = \frac{d}{dk}\omega(k)$. Thus, at long times the integral is dominated by signals that are moving at the velocity U . This velocity is called the *group velocity*, distinguished from the phase velocity $U_{phase} = \omega(k)/k$, the velocity at which the phase varies.

To continue the analysis, if we let $\theta(x, t) = t(kU - \omega(k))$, then if k_0 is the wavevector of stationary phase, then we can expand $\theta = (k_0U - \omega(k_0)) + \omega''(k_0)/2(k - k_0)^2 + \dots$. Our integral therefore becomes

$$\phi(x, t) = e^{it(k_0U - \omega(k_0))} \int_{-\infty}^{\infty} dk e^{-i\omega''(k_0)/2(k - k_0)^2} = \sqrt{\frac{2\pi}{t|\omega''(k_0)|}} e^{i(k_0U - \omega(k_0)) - i\pi/4 \text{sgn}(\omega''(k_0))}. \quad (10.35)$$

10.7.1 The Front of a Wavetrain

The aforementioned analysis assumed that $\omega''(k_0) \neq 0$. This condition is often violated in practice. For example consider *gravity* waves in a channel of finite depth h_0 . The dispersion relation for such waves is

$$\omega(k) = \sqrt{gk \tanh(kh_0)}. \quad (10.36)$$

A typical wavetrain is thus

$$\phi(x, t) = \int_{-\infty}^{\infty} F(k) e^{ikx - i\omega(k)t}. \quad (10.37)$$

The integral is dominated by the stationary phase points

$$\omega'(k) = \frac{x}{t}. \quad (10.38)$$

Now, it is easy to verify that the group velocity $\omega'(k)$ has a maximum value $= \sqrt{gh_0}$ as $k \rightarrow 0$. Hence, the *leading edge* of the disturbance will travel at this velocity.

The issue with this is that since the dispersion relation has a maximum, then at this maximum $\omega''(k_0) = 0$. Thus, our analysis must break down. Naively, one would expect that above this maximum velocity, the water in front of the disturbance will be unperturbed, so there should be a transition from a wave pattern to a flat one. To find out how this transition happens, we pick $x/t = \sqrt{gh_0}$, and evaluate the integral here. The phase function $\phi(x, t) = kx - \omega(k)t = (k(x - \sqrt{gh_0}t) - \omega'''(k_0)(k)^3/6t)$. Hence, the integral reduces to

$$\phi(x, t) = \int_{-\infty}^{\infty} F(0) e^{i(k(x - \sqrt{gh_0}t))} e^{k^3/6t i \omega'''} \quad (10.39)$$

It turns out that this formula comes up enough to have a name—it is the so-called Airy function, defined as

$$\text{Ai}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i(ux+u^3/3)} du. \quad (10.40)$$

Hence, if we redefine variables $u^3/3 = k^3/6t$

so that $u = \omega''' t^{1/3} k / 2^{1/3}$, we have

$$\phi(x, t) = \frac{2^{1/3}}{(t\omega''')^{1/3}} \text{Ai}\left(\frac{x - \sqrt{gh_0}t}{(t\omega''')^{1/3}}\right). \quad (10.41)$$

We will discuss the form of the Airy function very shortly.

10.7.2 Free particle Schrodinger Equation

As an example, let us consider solutions to the Schrodinger equation of a free particle. If $\psi(x)$ is the wavefunction, then Schrodinger's equation is

$$i\partial_t \psi + \frac{\hbar}{2m} \partial_x^2 \psi = 0.$$

If we decompose the solution in Fourier modes, writing

$$\psi = \int a(k, t) e^{ikx},$$

we find that $a(k, t)$ obeys

$$i\dot{a} = \frac{k^2 \hbar}{2m} a,$$

so that

$$\psi = \int a(k, t) e^{i(kx - \frac{k^2 \hbar}{2m} t)}.$$

Writing

$x = Ut$ as above, the stationary phase implies that the integral will be dominated when $U = 2k\hbar/2m$, so that

$$\psi \approx e^{it(x/t)^2 \hbar / (2m)} a\left(\frac{x}{t}\right) \sqrt{\frac{\pi}{t\hbar/m}} e^{i\pi/4},$$

where we used the result from the last class.

10.7.3 Waves behind a barge

Let us consider the waves produced by a boat moving down a canal. Waves in a canal satisfy a dispersion relation $\omega = \omega(k)$; let us also imagine that a wave with a wavenumber k is damped with a damping rate $\gamma(k)$. The wave amplitude at a location

x produced by the boat when it was at location x' , a time

Δt later is then given by

$$\int \hat{f}(k) e^{ik(x-x') - i\omega(k)\Delta t - \gamma(k)\Delta t} dk.$$

If the boat is moving with a steady velocity U , the time lapse is related to the initial location of the boat by $\Delta t = -x'/U$. If we want to total amplitude of the disturbance in the water we need to add up these contributions over all locations where the boat has been x' : if we suppose the current location of the boat is $x = 0$ at time $t = 0$ we have

$$\int e^{ikx} dk \int_{-\infty}^0 e^{(i(k-\omega(k)/U)+\gamma(k)/U)x'} dx' = U \int \frac{e^{ikx}}{-i(kU-\omega(k))+\gamma(k)}.$$

Now the integral can be performed by contour integration. There is a pole when $kU - \omega(k) + i\gamma(k) = 0$; If we solve for k where

$$U = \frac{\omega(k)}{k} \text{ (call this } k_* \text{), then writing}$$

$k = k^* + iq$, the pole exists when $q = -\gamma(k)/d\omega(k^*)/dk$. This is in the negative half plane. Hence, when $x > 0$, the disturbance of the boat is zero; when $x < 0$ we pick up the pole contribution, so the disturbance is

$\frac{2\pi U}{\omega'(k^*)} e^{ik^*x} e^{\gamma(k^*)/\omega'(k^*)x}$. This decays exponentially behind the boat, at a rate depending on the dissipation.

10.7.4 Waves behind a boat

Now lets consider the more interesting problem of waves behind a boat on open water, so that the waves can move in two dimensions. We first need to set up a coordinate system. Suppose the boat is moving in the \hat{x} direction with velocity U . Suppose that we are interested in the wave amplitude at a position \mathbf{r}

which is far from the boat and at an angle θ to the travelling direction. If we now repeat the derivation above we find that the wave amplitude behind the boat is

$\int \frac{e^{i\mathbf{k}\cdot\mathbf{r}}}{i(\mathbf{k}\cdot\hat{x}-\omega(k))+\gamma(k)} d\mathbf{k} = \int \frac{e^{i\mathbf{k}\cdot\mathbf{r}}}{i(k\cos(\phi)-\omega(k))+\gamma(k)} k dk d\phi$. If we write $\mathbf{k}\cdot\mathbf{r} = krcos(\theta - \phi)$, we can then performing the contour integral over k to find

$$\int e^{ik(\phi)rcos(\theta-\phi)} d\phi,$$

where $k(\phi)$ satisfies

$$\frac{\omega(k(\phi))}{k} = U\cos(\phi).$$

If we now use the dispersion relation for gravity waves,

$$\omega = \sqrt{kg}, \text{ we find that}$$

$$k(\phi) = \frac{g}{U^2\cos^2\phi}.$$

Thus, our integral becomes

$$\int e^{i\frac{gr}{U^2} \frac{\cos(\theta-\phi)}{\cos^2(\phi)}} d\phi.$$

This integral can be evaluated at the point of stationary phase! The stationary point occurs when

$$\tan(\theta - \phi) = 2\tan(\phi). \text{ Using}$$

$$\tan(\theta - \phi) = (\tan(\theta) - \tan(\phi))/(1 + \tan(\theta)\tan(\phi)), \text{ this is just}$$

$$\tan\theta = \frac{\tan\phi}{1+2\tan^2\phi}.$$

There are only a certain range of θ where this equation can be satisfied. This sets the size of the wake! The maximum angle happens when $\tan(\phi) = 1/\sqrt{2}$, which implies

$$\tan(\theta) = 1/(2\sqrt{2}). \text{ This gives}$$

$$\phi = \text{atan}(1/2\sqrt{2}) = 19.5^\circ ! \text{ Thus, the size of the wake is}$$

39° . This remarkable result is due to lord Kelvin, and is called *Kelvin's wedge*.

10.7.5 Dispersive Electromagnetic Waves

In free space, electromagnetic waves move at constant velocity

$$c = \frac{1}{\sqrt{\epsilon\mu}}, \quad (10.42)$$

where ϵ is the dielectric constant and μ the magnetic permeability of vacuum. In a real material however both ϵ and μ can depend on frequency, and this causes dispersion of the waves. In the early days after Einstein proposed his theory of relativity, there was much work investigating whether the hypothesis that nothing could move faster than the speed of light would hold up when taking these dispersive effects into account. In fact many of the mathematical methods we are discussing were invented in response to this question.

The relation for electromagnetic radiation is, in general,

$$k^2 = \omega^2 \mu \epsilon. \quad (10.43)$$

The simplest model for how ϵ

can depend on ω is called the Lorenz model, which is as follows: consider the equation of motion for an electron in an external electric field

$$m\ddot{x} + \gamma\dot{x} + kx = -eE, \quad (10.44)$$

where γ is the damping constant, k is a spring constant and m is the mass, and e the charge. Then if the electric field is periodic $E = E_0 e^{i\omega t}$ the displacement of the electron $x(t) = X(\omega) e^{i\omega t}$, where

$$X(\omega) = \frac{-eE/m}{\omega_0^2 - \omega^2 - i\gamma/m\omega}, \quad (10.45)$$

where $\omega_0^2 = k/m$. Using the fact that the electric polarizability is given by $P = -ne x(t)$ where n is the density of electron oscillators, and that $\epsilon E = \epsilon_0 E + P$ we therefore have

$$\epsilon(\omega) = \epsilon_0 \left(1 + \frac{\omega_p^2}{\omega_0^2 - \omega^2 - i\Gamma\omega} \right). \quad (10.46)$$

Here $\Gamma = \gamma/m$ and $\omega_p^2 = ne^2/(\epsilon_0 m)$ is the plasma frequency.

The question of the general structure of an electromagnetic pulse input into a metal is therefore quite interesting. As always we need to evaluate

$$\int f(k) e^{i(kx - \omega(k)t)}. \quad (10.47)$$

The nontrivial structure of the dispersion relation leads to interesting answers, which we will explore in the homework.

10.7.6 Absolute and Convective Instability

Consider the following model for the development of an instability

$$\partial_t A = \epsilon A + D\partial_{xx}A. \quad (10.48)$$

If we use the ansatz $A(x, t) = e^{ikx}e^{\omega t}$, then it is straightforward to show that

$$\omega = \epsilon - Dk^2. \quad (10.49)$$

This implies that an initial perturbation will grow in time if the wavenumber $k < \sqrt{\epsilon/D}$; if $k > \sqrt{\epsilon/D}$

the perturbation will decay. There are many examples of natural phenomena in which instabilities occur; the mathematical framework for describing the development of the instability is often very similar to the model presented here.

Given an initial condition $A(x, t = 0) = A_0(x)$, the solution at time t is given by the following integral

$$A(x, t) = \int_{-\infty}^{\infty} A_k e^{ikx} e^{\omega(k)t}. \quad (10.50)$$

This integral can be performed by saddle point method: the saddle point occurs at

$$ix + \omega'(k)t = 0, \quad (10.51)$$

or

$$k = i \frac{x}{2Dt}. \quad (10.52)$$

The appropriate contour is therefore parallel to the real axis $k = ik_I + \Delta k$ –one can check that the phase of the integrand is constant along this contour. Hence we obtain

$$A(x, t) = A_{k=0} e^{\epsilon t} e^{-x^2/(4Dt)} \frac{\sqrt{\pi}}{Dt} = A_{k=0} \exp\left(\frac{4Det^2 - x^2}{4Dt}\right) \quad (10.53)$$

as the long time behavior. The behavior is therefore the precise superposition of growth and diffusion!

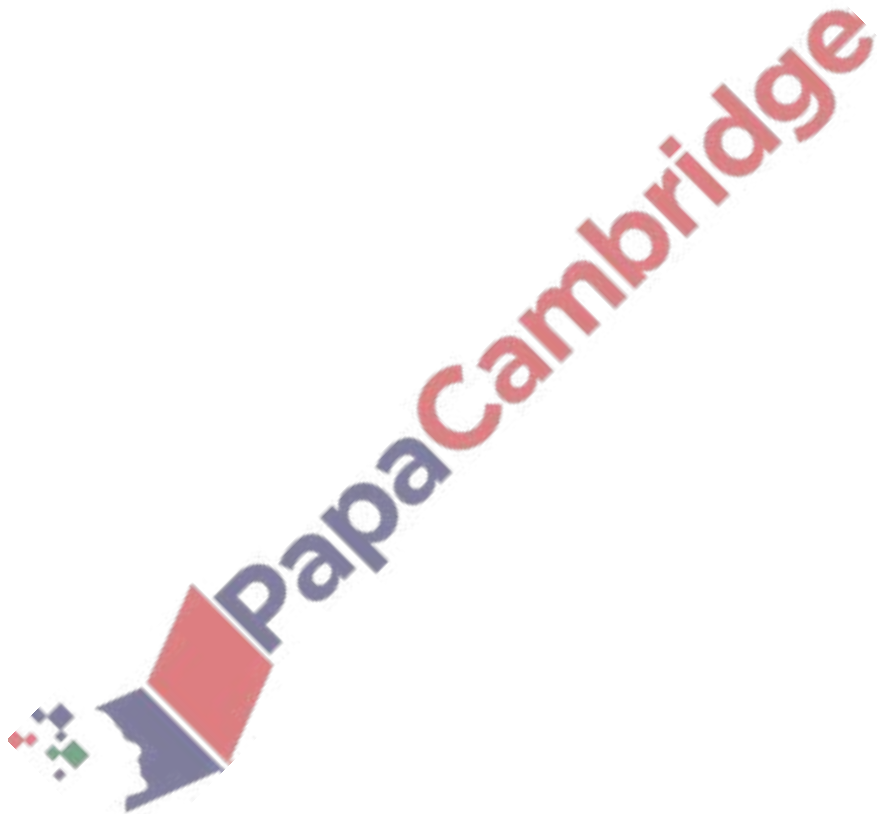
Note that the solution has the interesting behavior that it grows within the region $4Det^2 > x^2$ or $x < 2\sqrt{Det}$ and it decays exponentially outside of this region. The edge of the region separating instability from stability moves at a constant velocity.

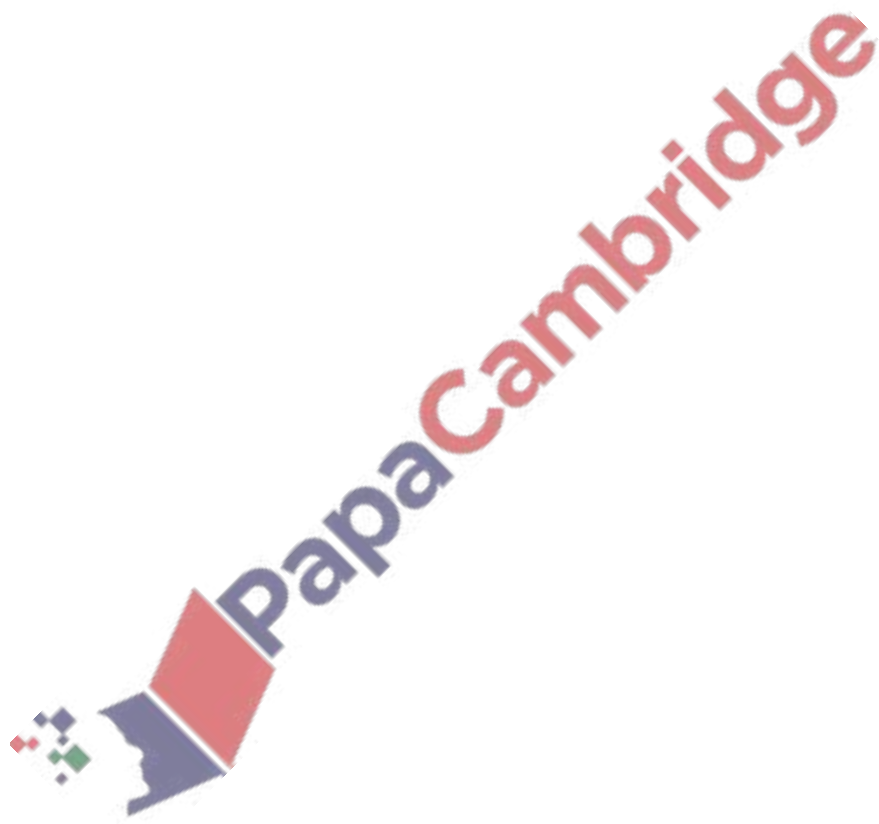
Now, suppose we modify our model to include a convection of the perturbation. That is

$$\partial_t A + U\partial_x A = \epsilon A + D\partial_{xx}A. \quad (10.54)$$

This contains the additional bit of physics that the function A

is being advected away at velocity U . Clearly, if the advection rate is fast enough, the perturbation will grow but it will be advected out of our field of view as it grows. If on the other hand the velocity is very small we should recover the situation above where the perturbation grows in the laboratory frame. We can expect on purely intuitive grounds that the transition velocity will be $U = 2\sqrt{\epsilon D}$, since the latter is the characteristic velocity at which the unstable front moves out of the field of view. When $U < 2\sqrt{\epsilon D}$, the system is said to be *absolutely unstable*; when $U > 2\sqrt{\epsilon D}$ the system is convectively unstable.





11 Nonlinear Partial Differential Equations

We will now turn to examining how the method of thinking we have advocated in this course can be applied to nonlinear partial differential equations. We will limit ourselves to equations with one space and one time dimension, in order to get across the essential concepts and allow easy numerical implementation and testing.

Recall from our previous remarks on partial differential equations that it is useful to classify equations in essentially four types:

1. Diffusion. The canonical diffusion equation is $\partial_t u = \partial_{xx} Du$, where D is the diffusion constant. Such equations lead to spreading like \sqrt{Dt} . We will generalize this concept to mean all 'diffusion like' equations. Instead of trying to define this (vague) notion in detail, we'll learn what belongs to this class by example. For instance, clearly when we let $D = D(u, x)$ (ie the diffusion constant depend on the value of the quantity that is diffusing), then this is diffusion like; similarly the fourth order equation $\partial_t u = -\partial_{xxxx} u$ is also diffusion like. You can verify that the Fourier mode $A(t)e^{ikx}$ decays in time with $A(t) = e^{-k^4}$.
2. Advection. Here the canonical equation is $\partial_t u + V\partial_x u = 0$. The solution to this equation (if V is a constant) is $u(x - Vt)$. If $V = V(x, u, t)$ then solutions can be more complicated. Advection equations come up often in science, largely through wave equations, which have the form $\partial_{tt} u = \partial_{xx} u$.
3. Dispersion Here the canonical equation is $\partial_t u = \kappa\partial_{xxx} u$.
4. Laplace equation $\partial_{tt} u = -\partial_{xx} u$. I have written this here in a strange form—using a time variable. I've done this to emphasize that despite the superficial similarity to the wave equation (note the only difference is the negative sign on the right hand side!) the nature of the solutions could not be more different.
5. To this list we might also include the general class of ordinary differential equations, e.g. $\frac{du}{dt} = f(u, t)$. This is not of course a PDE, but if we think of u being a function of space and time then the ODE could describe the properties at every point in space.

11.0.7 Solving Nonlinear Pde's using Matlab

Matlab has a nifty function called *pdepe* that solves nonlinear partial differential equations, with one space dimension. Here we will outline how to use this function for solving the diffusion equation—subsequently we will see how to modify the program to solve more complicated equations (it is easy!)

pdepe solves equations of the following general form

$$c(x, t, u, \partial_x u) \partial_t u = \partial_x F(u, \partial_x u, x, t) + s(u, \partial_x u, x, t). \quad (11.1)$$

Here F is the flux, and s is a source. The boundary conditions must be of the form

$$p(x, t, u) + q(x, t) F(u, \partial_x u, x, t) = 0. \quad (11.2)$$

For diffusion equation we set $c = 1$, $F = \partial_x u$ and $s = 0$. The program is written in such a way that u can be a vector, so one can solve multiple equations simultaneously. Although for 'professional applications' I have often found it necessary to rewrite the program myself, it will serve our purposes perfectly: as a testing ground for how to easily discover the behavior of nonlinear partial differential equations, so we can see how it works.

The following is a fragment of code that solves the diffusion equation. (I simply modified this from the excellent example programs that are given in MATLAB's helpdesk, under *pdepe*).

```
1 function [x,t,sol]=pdex1
2
3 x = linspace(-20,20,200);
4 t = linspace(0,2,10);
5
6 sol = pdepe(0,@pdex1pde,@pdex1lic,@pdex1bc,x,t);
7 % the first argument of pdepe, m, defines the symmetry
8 % of the problem. For slab symmetry we use m = 0.
9
10
11 function [c,F,s] = pdex1pde(x,t,u,DuDx)
12 % defines c, F, s in equation 10.1 in terms of x, t, u, and
13 % the partial of u with respect to x: DuDx.
14 c = 1;
15 F = DuDx;
16 s = 0;
17
18 function u0 = pdex1lic(x)
19 % defines the initial condition vector
20 u0 = 1.*exp(-x^2);
21
22 function [pl,ql,pr,qr] = pdex1bc(xl,ul,xr,ur,t)
23 % defines the boundary conditions according to
24 % p and q in equation 10.2
25 % xl = x(1) and xr = x(end), ul = u(xl) and ur = u(xr),
26 % and the same for pl, pr, ql and qr.
27 pl = ul; % sets the b.c. u(x = -20) = 0
28 ql = 0;
29 pr = ur; % sets the b.c. u(x = 20) = 0
30 qr=0;
```

We will use this program in what follows.

11.1 The diffusion equation, and nonlinear diffusion

Recall our previous discussion of the diffusion equation: we found the green's function by using dimensional analysis. The argument was as follows: If we look at $\partial_t n = D\partial_x^2 n$, we see that roughly $\partial_t \sim D\partial_x^2$

There is a sense in which this equality is meaningless. What I mean by it is that if you have a function n which obeys a diffusion equation, taking a single time derivative of the function gives a number of about the same size as when you take two spatial derivatives. What this means is that the characteristic length scale over which n varies is of order \sqrt{Dt} . Now since the initial distribution δ is perfectly localized, and therefore has no length scale, we expect that at time t , $G(x - x', t)$ will have characteristic width \sqrt{Dt} .

Thus, we guess a (so-called) similarity solution:

$$G(x - x', t) = A(t)F\left(\frac{x - x'}{\sqrt{Dt}}\right). \quad (11.3)$$

The time dependence of $A(t)$ is determined by mass conservation. Since

$$\int_{-\infty}^{\infty} \rho dx = \int_{-\infty}^{\infty} A(t)F\left(\frac{x - x'}{\sqrt{Dt}}\right) dx = A(t)\sqrt{Dt} \int_{-\infty}^{\infty} dy F(y) \quad (11.4)$$

must be constant in time, we see that $A(t) = 1/\sqrt{Dt}$. Here ρ is the density and we have changed variables from x to $y = x/\sqrt{Dt}$.

Now let's just plug in $G(x, t) = 1/\sqrt{Dt}F((x - x')/\sqrt{Dt})$ into the diffusion equation. This gives the following ordinary differential equation for $F(y)$:

$$\frac{1}{Dt^{3/2}}D\left(-\frac{1}{2}F - \frac{1}{2}yF'\right) = \frac{1}{Dt^{3/2}}DF'' \quad (11.5)$$

Cancelling out the time factors, and integrating this equation once gives

$$F' = -\frac{1}{2}Fy. \quad (11.6)$$

This equation can be immediately integrated to give

$$F(y) = F_0 e^{-y^2/4}, \quad (11.7)$$

or

$$G(x - x', t) = \frac{F_0}{\sqrt{Dt}} \exp\left(-\frac{(x - x')^2}{4Dt}\right), \quad (11.8)$$

where the constant $F_0 = 1/\sqrt{\pi}$ is determined by requiring that $\int G = 1$.

Now, what is particularly amazing about this solution is that it actually works for an arbitrary initial condition (that is localized in space) as long as we wait sufficiently long.

The reason is that any information encoded in an arbitrary $n(x, t = 0)$ will *diffuse away*, and eventually the solution will spread according to the Green's function, multiplied by a constant factor representing the total mass of the solution. Hence if $\int n_0(x')dx' = M$, then at large times

$$n(x, t) = \frac{M}{\sqrt{\pi Dt}} \exp -\frac{x^2}{4Dt}. \quad (11.9)$$

Let us demonstrate this: Figure 11.1A. shows the solution of the diffusion equation starting from the initial condition $n_0(x) = 1$ if $|x| < 3$ and $u_0(x) = 0$ otherwise.

Figure 11.1B. compares our analytic solution with the simulation. Finally, Figure 11.1C shows the time dependence of the solution. We compare $n(x = 0, t)$ with the prediction from the diffusion equation $n(x, 0) = \frac{M}{\sqrt{D\pi t}}$, where $D = 1$ and $M = 3$.

After $t \sim 10$ the time dependence of the formula agrees quantitatively with the simulation. Can we estimate when this should have occurred? The initial condition had an initial width of 3. The time for information to propagate across the solution is of order $3^3 = 9$. Hence this agrees with our expectations.

11.1.1 A nonlinear diffusion equation

Let's now consider a slightly more complicated variation on this theme. Consider the equation

$$\partial_t u = \partial_x \left(u^2 \partial_x u \right). \quad (11.10)$$

with the initial conditions $u(x, 0) = u_0(x)$.

What happens at long times? This equation corresponds to a situation where the diffusion constant is linearly proportional to u^2 .

Can we make the same argument as before? Let's go through the logic. On one hand, the entire premise of the above discussion was based on the idea of a Green's function—where the equation is linear, and thus we can use superposition of solutions to construct the general solution. On the other hand at the end of the discussion above we observed that the Green's function itself worked only in the limit of large times—the reason for this was that when the characteristic width of the solution is much larger than that of the initial condition, these lengths should no longer play a role in the solution.

Can we apply the same logic here? On what parameters can the solution at large times depend? As above, there must be some characteristic width of the solution $L(t)$. Moreover, mass conservation must be obeyed. In principle the solution at any time should also depend on any length scales implicit in the initial conditions—lets call these ℓ .

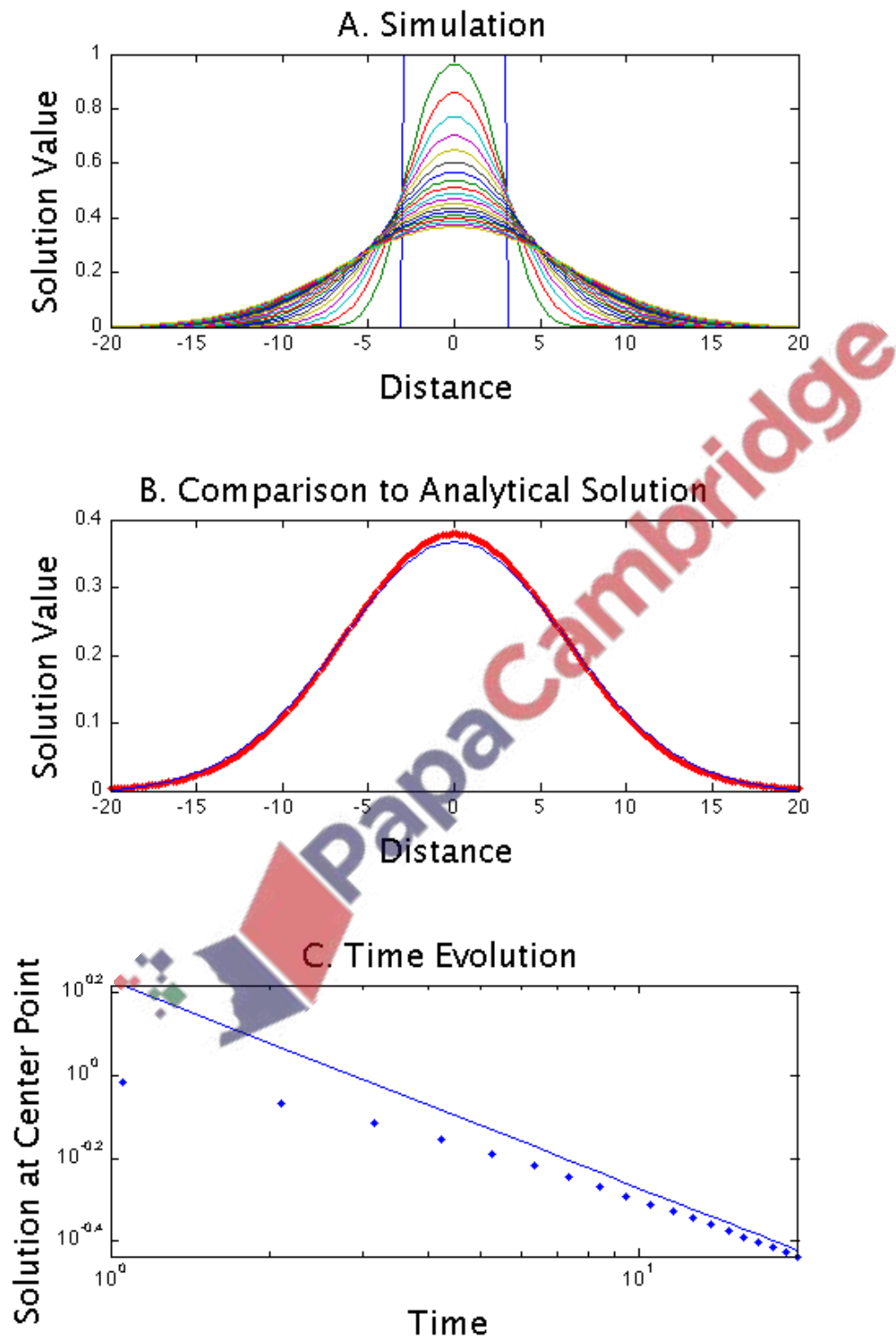


Figure 11.1. Top: Simulation of the diffusion equation, starting out from $n_0(x) = 1$ for $|x| \leq 3$. Middle: Comparison of simulation (blue dots) at $t = 20$ with the Green's function solution (red line). Bottom: Time dependence of $u(0, t)$, comparing the simulation (dots) with the derived formula. Note the quantitative agreement after $t \approx 10$.

Program 18 MATLAB code used to create figure 11.1

```
1 function [x,t,sol]=CH9a_diffusion
2
3 x = linspace(-20,20,200);
4 t = linspace(0,20,20);
5
6 sol = pdepe(0,@pdexlpde,@pdexlic,@pdexlbc,x,t);
7 % the first argument of pdepe, m, defines the symmetry
8 % of the problem. For slab symmetry we use m = 0.
9
10 *** analytical solution**
11 ana = (3/sqrt(pi*20))*exp(-(x.^2)/(4*20));
12 % code for plotting
13 quickplot(x,t,sol,ana);
14 subplot(3,1,3),hold on, loglog(t,3./sqrt(pi.*t),'-b')
15
16
17
18 function [c,F,s] = pdexlpde(x,t,u,DuDx)
19 % defines c, F, s in equation 10.1 in terms of x, t, u, and
20 % the partial of u with respect to x: DuDx.
21 c = 1;
22 F = DuDx;
23 s = 0;
24
25 function u0 = pdexlic(x)
26 % defines the intial condition vector
27 for i = 1:length(x)
28     if abs(x(i)) >= 3
29         u0(i) = 0;
30     else
31         u0(i) = 1;
32     end
33 end
34
35
36 function [pl,ql,pr,qr] = pdexlbc(xl,ul,xr,ur,t)
37 % defines the boundary conditions according to
38 % p and q in equation 10.2
39 % xl = x(1) and xr = x(end), ul = u(xl) and ur = u(xr),
40 % and the same for pl, pr, ql and qr.
41 pl = ul;           % sets the b.c. u(x = -20) = 0
42 ql = 0;
43 pr = ur;           % sets the b.c. u(x = 20) = 0
44 qr=0;
```


Hence we know that

$$u(x, t) = \frac{1}{L(t)} F\left(\frac{x}{L}, \frac{\ell}{L}\right). \quad (11.11)$$

In the limit that the characteristic scale of the solution $L(t)$ is much larger than ℓ we thus reduce to the previous case.

Let us therefore work out the solution under the assumption that

$$u(x, t) = \frac{1}{L} F\left(\frac{x}{L}\right). \quad (11.12)$$

Plugging this ansatz into the nonlinear diffusion equation gives

$$-\frac{\dot{L}}{L^2} \left(F + \eta F_\eta \right) = \frac{1}{L^5} \left(F^2 F_\eta \right)_\eta, \quad (11.13)$$

where $\eta = \frac{x}{L}$. If we now balance the time dependences we have that

$$\frac{\dot{L}}{L^2} = \frac{1}{L^5}, \quad (11.14)$$

which implies that $L = (4t)^{1/4}$. The function $F(\eta)$ obeys the ordinary differential equation

$$-\eta F = F^2 F_\eta, \quad (11.15)$$

where we have integrated once using

$$F + \eta F_\eta = (\eta F)_\eta, \quad (11.16)$$

and also set the constant of integration to zero since $F \rightarrow 0$ as $\eta \rightarrow \infty$. We can integrate this equation again to give

$$F = \sqrt{A - \eta^2}. \quad (11.17)$$

An interesting feature of this solution is that $F(\eta = \pm\sqrt{A}) = 0$. Therefore in contrast to diffusion we expect the solution to be nonzero in only a finite region of space.

The free constant here is determined by the mass of the solution: we want

$$\int_{-\sqrt{A}}^{\sqrt{A}} F(\eta) d\eta = M. \quad (11.18)$$

If we substitute

$$\eta = \sqrt{A} \cos(\theta), \quad (11.19)$$

the integral becomes

$$\int_{-\pi}^0 \sin^2(\theta) A d\theta = A\pi/2, \quad (11.20)$$

so that $A = 2M/\pi$. Combined this gives us the solution

$$u(x, t) = \frac{1}{(4t)^{1/4}} \sqrt{\frac{2M}{\pi} - \left(\frac{x}{(4t)^{1/4}}\right)^2}. \quad (11.21)$$

Let us now compare this solution with numerical simulations: Figure 11.2A shows the solutions to our equation starting out with the same initial condition that we used in the previous section. You will see that the spreading is much slower than before, and as predicted the edge of the solution is sharp: u is nonzero only in a finite region around the origin.

Figure 11.2C. shows the time dependence of $u(0, t)$. This is compared with our prediction, that

$$u(0, t) = \sqrt{\frac{2M}{\pi}} \frac{1}{(4t)^{1/4}}. \quad (11.22)$$

The agreement between the theory and simulation becomes quite good as t becomes large, and is quite close beyond $t \sim 40$. Where does this threshold time come from? The initial condition has width 3, so we would expect agreement beyond when $L = (4t)^{1/4} = 3$ or $t = 3^4/4 = 20$, consistent with our measurements.

What about the profiles? Figure 11.2B. compares the simulation and theory for $t = 100$

The agreement is essentially perfect!

11.1.2 Radial Nonlinear Diffusion Equation

Here we consider the nonlinear diffusion equation, and also the ordinary diffusion equation, in multidimensions, assuming spherical symmetry.

11.2 A reaction diffusion equation

Now we turn to a different example. Consider the equation

$$\partial_t u = \partial_{xx} u + u^4. \quad (11.23)$$

This is the combination of diffusion, with some sort of reaction: here the reaction is encoded by the u^4 term which says that the rate of producing u is proportional to u^4 .

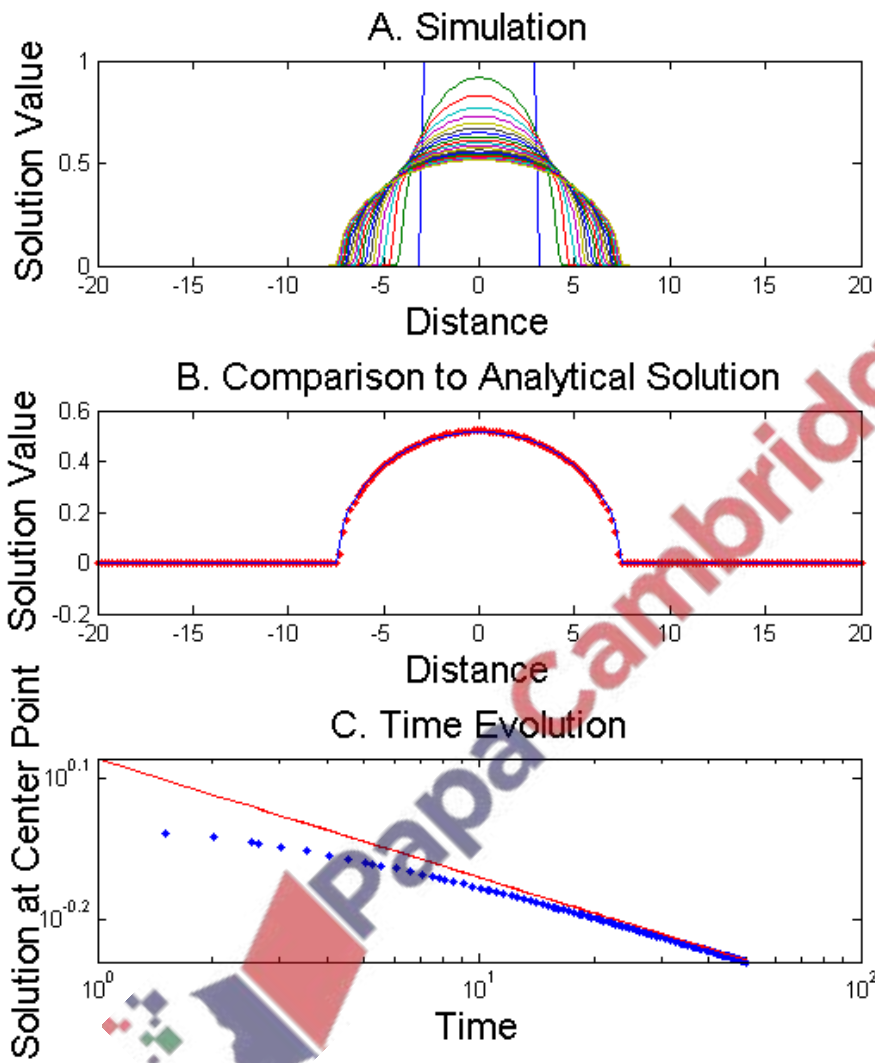


Figure 11.2. Top: Simulation of the diffusion equation, starting out from $u_0(x) = 1$ for $|x| \leq 3$. Middle: Comparison of simulation (blue dots) at $t = 50$ with the analytic solution (red line). Bottom: Time dependence of $u(0, t)$, comparing the simulation (dots) with the derived formula. Note the quantitative agreement after $t \approx 20$.

Program 19 MATLAB code used to create figure 11.2

```
1 function [x,t,sol]=CH9a.nonlineardiffusionu2
2
3 x = linspace(-20,20,200);
4 t = linspace(0,50,20);
5
6 sol = pdepe(0,@pdexlpde,@pdexlic,@pdexlbc,x,t);
7 % the first argument of pdepe, m, defines the symmetry
8 % of the problem. For slab symmetry we use m = 0.
9
10 ** analytical solution**
11 M = 6;
12 L = 3;
13 ana = (1/((4*50)^(1/4)))*sqrt(2*M/pi-(abs(x)/((4*50)^(1/4))).^2);
14 % code for plotting
15 quickplot(x,t,sol,ana)
16 subplot(3,1,3),hold on, loglog(t, sqrt(2*M/pi)/((4*t)^(1/4)), '-r')
17
18
19 function [c,F,s] = pdexlpde(x,t,u,DuDx)
20 % defines c, F, s in equation 10.1 in terms of x, t, u, and
21 % the partial of u with respect to x: DuDx.
22 c = 1;
23 F = (u.^2).*DuDx; % note the nonlinear u^2 term
24 s = 0;
25
26 function u0 = pdexlic(x)
27 % defines the intial condition vector
28 for i = 1:length(x)
29     if abs(x(i)) > 3
30         u0(i) = 0;
31     else
32         u0(i) = 1;
33     end
34 end
35
36 function [pl,ql,pr,qr] = pdexlbc(xl,ul,xr,ur,t)
37 % defines the boundary conditions according to
38 % p and q in equation 10.2
39 % xl = x(1) and xr = x(end), ul = u(xl) and ur = u(xr),
40 % and the same for pl, pr, ql and qr.
41 pl = ul; % sets the b.c. u(x = -20) = 0
42 ql = 0;
43 pr = ur; % sets the b.c. u(x = 20) = 0
44 qr=0;
```

This could represent a chemical reaction, where four u molecules combined to catalyze the production of more u molecules.

What happens to the solutions of this equation? First let us recall the behavior of solutions to the ordinary differential equation, which has a similar structure to our PDE $\frac{du}{dt} = u^4$. The solution here is $u = \left(\frac{3}{t^* - t}\right)^{1/3}$. Here the blow up time t^* is encoded by the initial condition so that $u(t = 0) = (3/t^*)^{1/3}$.

But what happens in our equation? Let's analyze (11.23) by asserting that the solution $u(x, t)$ has a characteristic size $A(t)$, and a characteristic length scale $L(t)$. Then the equation obeys the scaling relation

$$\frac{dA}{dt} \sim -\frac{A}{L^2} + A^4. \tag{11.24}$$

Here the three terms refer to the $\partial_t u$, $\partial_{xx} u$ and u^4 , respectively. We have written a negative sign in front of the diffusive term, because diffusion will always work to smear things out and decrease the amplitude.

As has now become our routine, we have three possible dominant balances: The first and second term represents diffusion, the first and third term represents reaction, and the second and third represents a steady balance between reaction and diffusion. What happens?

Roughly speaking, we should expect the following: if the first term on the right hand side is much larger than the second, diffusion will dominate and the amplitude will decrease in time. If the second term on the right hand side dominates, the reaction will dominate and the solution will diverge in time. The transition between these two regimes clearly happens when $A/L^2 \sim A^4$, or when $A^3 \sim L^{-2}$. Let's turn to simulations to see if this picture is correct. Figure ?? shows a simulation of equation (11.23) starting from our initial condition $u(x) = 1$ for $|x| < 3$. You see that the solution grows in time. Indeed, the time dependence shown in figure ?? shows that the solution apparently diverges in finite time, as would be expected if the dominant balance were the reaction term. Note that since the initial condition has a length scale $L = 3$, we should expect this solution to diverge since $A^3 = 1 \gg 1/L^2 = 1/9$.

The numerical solution reveals a subtlety that we must return to: as the solution is blowing up, the spatial scale of the region where the solution is large also seems to be changing!

Before examining this in more detail let's consider one other simulation: Figure ?? shows the solution to the reaction diffusion equation but now we decrease the value of the initial condition by an order of magnitude: take $u(x, 0) = 0.1$ for $|x| < 3$. This solution does not blow up, but appears to diffuse outwards! Figure ?? shows the time dependence of this solution, and indeed the solution appears to decrease like $1/\sqrt{t}$. Now we use our criterion, $A^3 = 10^{-3} \ll 1/L^2 = 1/9$, which confirms the dominance of diffusion, as can be readily seen in Figures ?? & ??.

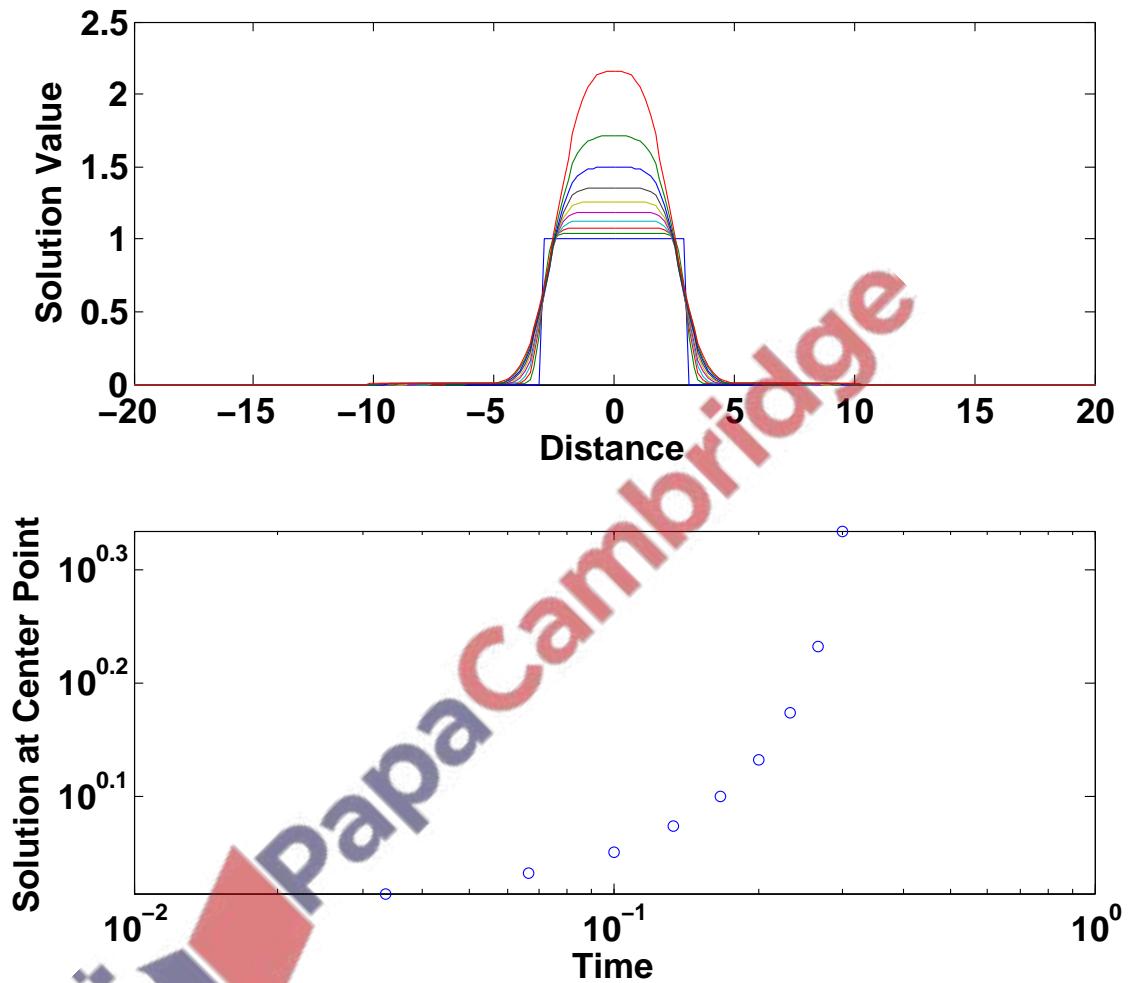


Figure 11.3. In the top figure, we see the solution evolve in time. The width of the region that is blowing up appears to change as time progresses. To get a general idea of the behavior away from this shrinking length, we also look at the behavior of the center of the domain as a function of time, in the second plot. The center point is clearly blowing up in finite time, encoded by the domain width.

Program 20 MATLAB code used to create figure 11.3

```
1 function [x,t,sol] = c10g
2
3 x = linspace(-20,20,200);
4 t = linspace(0,.3,10);
5
6 sol = pdepe(0,@c10gpde,@c10gic,@c10gbc,x,t);
7
8 subplot(2,1,1), plot(x,sol)
9 subplot(2,1,2), loglog(t,sol(:,100))
10
11 function [c,f,s] = c10gpde(x,t,u,DuDx)
12 c = 1;
13 f = DuDx;
14 s = u.^4;
15
16 function u0 = c10gic(x)
17 u0 = zeros(1:length(x));
18 for i = 1:length(x)
19     if abs(x(i)) >= 3
20         u0(i) = 0;
21     else
22         u0(i) = 1;
23     end
24 end
25 function [pl,ql,pr,qr] = c10gbc(xl,ul,xr,ur,t)
26 pl = ul;
27 ql = 0;
28 pr = ur;
29 qr = 0;
```

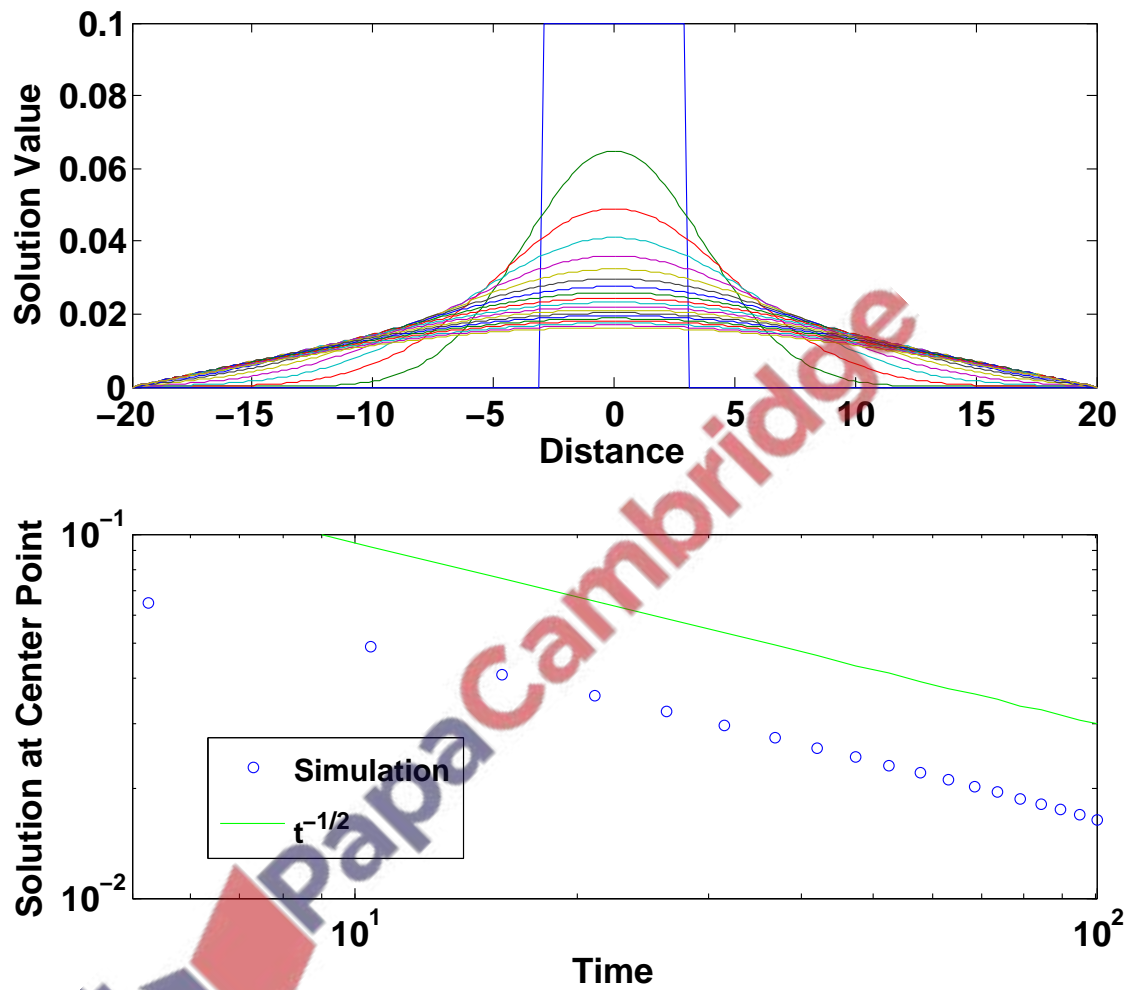


Figure 11.4. Here we see the simulation result for an initial condition an order of magnitude smaller than the previous simulation. The character of the behavior of the solution has changed dramatically; we no longer see it blowing up, rather diffusion has dominated the process! In the second figure, the behavior of the simulation center point is compared with the derived scaling of $1/\sqrt{t}$. The comparison holds quite well over a decade

Program 21 MATLAB code used to create figure 11.4

```
1 function [x,t,sol] = c10rd2
2
3 x = linspace(-20,20,200);
4 t = linspace(0,100,20);
5
6 sol = pdepe(0,@c10rd2pde,@c10rd2ic,@c10rd2bc,x,t);
7
8 subplot(2,1,1), plot(x,sol)
9 subplot(2,1,2), loglog(t,sol(:,100))
10 hold on
11 subplot(2,1,2), loglog(t,0.3./sqrt(t),'g-')
12
13 function [c,f,s] = c10rd2pde(x,t,u,DuDx)
14 c = 1;
15 f = DuDx;
16 s = u.^4;
17
18 function u0 = c10rd2ic(x)
19 u0 = zeros(1:length(x));
20 for i = 1:length(x)
21     if abs(x(i)) >= 3
22         u0(i) = 0;
23     else
24         u0(i) = 0.1;
25     end
26 end
27 function [pl,q1,pr,qr] = c10rd2bc(xl,ul,xr,ur,t)
28 pl = ul;
29 q1 = 0;
30 pr = ur;
31 qr = 0;
```

Are there really two different dominant balances for this equation? Let's evaluate the two unsteady balances that arise in the numerical solution to the reaction diffusion equation we're studying. If diffusion dominates, then our theory leads us to expect $u \sim (A/\sqrt{t})F(x/\sqrt{t})$. This solution implies that $\partial_{xx}u = A/t^{3/2}$, whereas a dominant reaction term implies $u^4 = A^4/t^2$. Hence as $t \rightarrow \infty$, we have $A/t^{3/2} \gg A^4/t^2$, conditioned upon $t > A^8$. Thus the diffusive term is logically self consistent, as we have seen.

What about the reaction dominated regime? Here we have one weak part of our argument: we argued above that reaction dominates diffusion as long as $A^3 > L^{-2}$. In using this criterion we assumed that L was given by the initial condition. But, our simulation seemed to indicate that L changed, and indeed decreased, in time, as can be seen in the numerical solution profiles in Figure ??

Why does this occur? Let's first consider the dominant balance where diffusion is negligible, so that the equation is just $\dot{u} = u^4$. Following our discussion of the beginning of this section, the solution is

$$u(x, t) = \frac{u_0(x)}{(1 - tu_0(x)^3 t/3)^{1/3}}. \quad (11.25)$$

Is diffusion negligible in this solution? We've already seen two different behaviors in our simulation by modifying the initial value; we can understand this behavior if we compute $\partial_{xx}u \sim u_0(x)^{-9}(u_0(x)^{-3} - t/3)^{-7/3}$, we can compare this against the scaling of the u^4 term, $u^4 \sim (3u_0(x)^{-3} - t)^{-4/3}$, noting that both solutions exhibit blow-up like behavior with different exponents. We expect that for initial values larger than one, we will see comparable contributions between these terms, showing our dominant balance is inconsistent!

So what happens? Apparently the solution will sharpen until diffusion also becomes important. At that point, the reaction term, diffusion term and the time dependent term are all exactly balanced with each other. By realizing that the only dimensionless argument for a functional form must be x/L , we explore this balance of three terms by writing the similarity form

$$u(x, t) = A(t)F\left(\frac{x}{L(t)}\right) \quad (11.26)$$

and use this ansatz in our reaction diffusion equation. This gives

$$\dot{A}F - \frac{\dot{L}}{L}A\eta F_\eta = \frac{A}{L^2}F_{\eta\eta} + A^4F^4. \quad (11.27)$$

We demand that all of the terms are the same order of magnitude. Hence we have that $\dot{A} = A^4$, which implies $A = (t^* - t)^{-1/3}$; also, $A^4 \sim A/L^2$, which implies that

I understand the argument that all three terms should exactly balance each other, but why does that motivate the particular form of the ansatz given below?

$L^2 \sim A^{-3}$ or implicitly $L \sim (t^* - t)^{1/2}$. Thus we have that the function F obeys the ordinary differential equation

$$-\left(\frac{1}{3}F + \frac{1}{2}\eta F_\eta\right) = F_{\eta\eta} + F^4. \quad (11.28)$$

A solution of this type would still diverge, but would have diffusion contribute nonetheless to the profile of the solution. We'll proceed to seek a reasonable solution.

What do we mean by reasonable? The solution that we wrote in equation 11.26 cannot apply across the entire region of our solution—e.g. we do not expect this solution to satisfy whatever boundary conditions we impose at the two spatial boundaries of our simulation region! This solution should apply only locally in our simulation, near the place where u is diverging.

But, we must impose boundary conditions on our solution—what are they? We claim that the correct boundary conditions to impose are the following: our solution is of the form

$$u(x, t) = \frac{1}{L^{2/3}} F\left(\frac{x}{L(t)}\right). \quad (11.29)$$

We need that as $\eta = x/L \rightarrow \infty$, since the solution cannot vary quickly in time. The reason we require this boundary condition is that clearly near the singularity the solution is changing extraordinarily rapidly—but far from its singular region, the solution does not change this fast. For this criterion to hold, we need to balance the pre-factor, implying the scaling $F(\eta) \sim \eta^{-2/3}$ as $\eta \rightarrow \infty$. If this scaling holds then we have

$$u(x, t) \sim L^{-2/3} \left(\frac{x}{L}\right)^{-2/3} \sim x^{-2/3}. \quad (11.30)$$

In order to respect symmetry about the blow-up point, we additionally require $F'(0) = 0$.

Are there solutions to equation 11.28 that obey these conditions? Figure ?? shows a set of solution to the equation 11.28 with $F(0) = 0.3$ and $F'(0) = 0$. It is seen that numerical solutions solved using continuation do exist that decay as $\eta^{-2/3}$.

11.3 An advection diffusion equation

Now let us consider a problem with nonlinear advection. Consider

$$\partial_t u + uu_x = \nu u_{xx}. \quad (11.31)$$

This equation is called *Burger's equation*. The second term on the right hand side is the advection term. As before we consider a simulation with the initial condition $u_0(x) = 1$

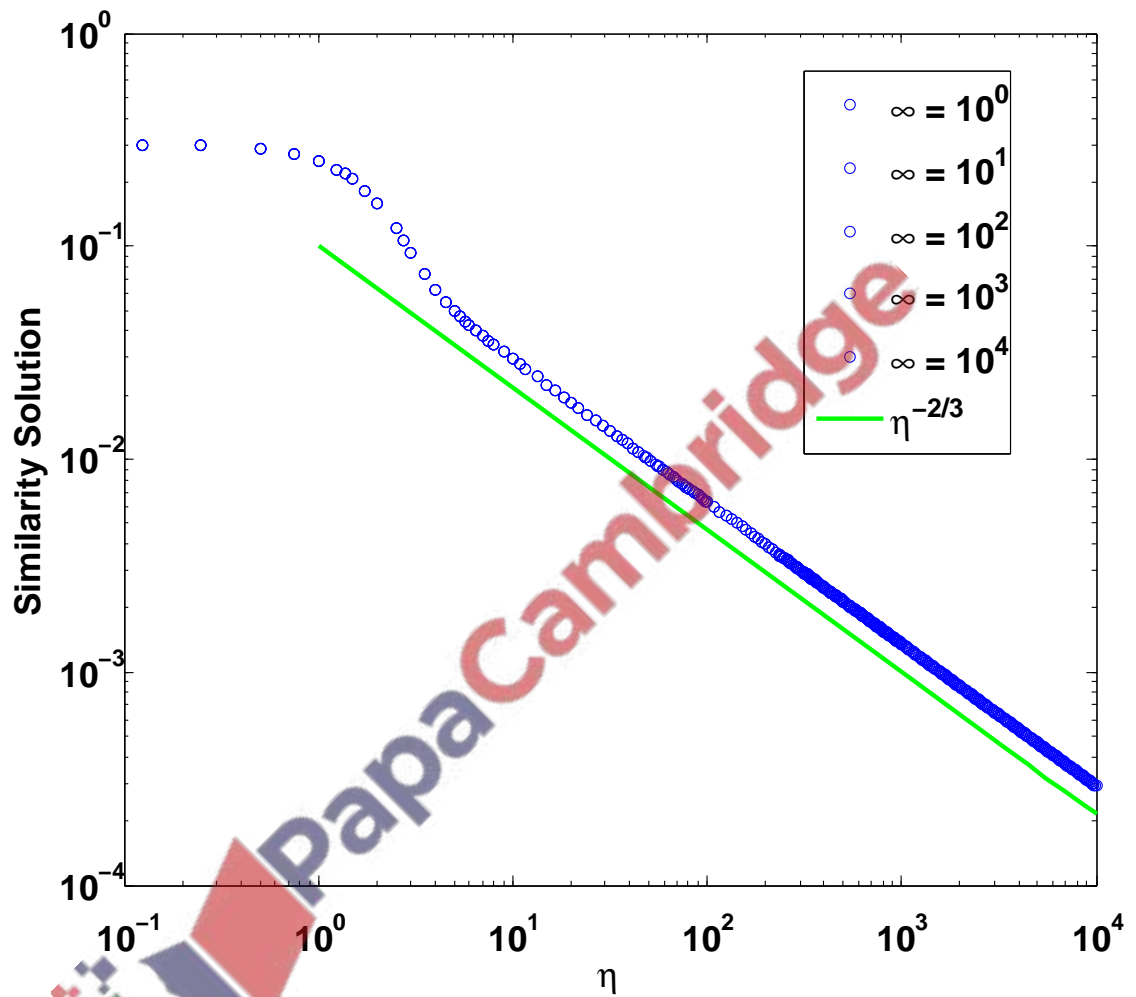


Figure 11.5. Using a continuation method, we solve the boundary value problem associated with our ODE subject to the boundary conditions described in the text. It is clear from the plot that the solution has a long regime that respects the scaling $\eta^{-2/3}$, which we derived in the text.

Program 22 MATLAB code used to create figure 11.5

```
1 function c10ss1
2
3 infinity = 1;
4 maxinfinity = 5;
5
6 solinit = bvpinit(linspace(0,infinity,5),[0 -0.3]);
7 sol = bvp4c(@ss1ode,@ss1bc,solinit);
8 eta = sol.x;
9 f = sol.y;
10
11 figure
12 loglog(eta,f(1,:), 'bo');
13
14 hold on
15 drawnow
16 shg
17
18 for nb = infinity+1:maxinfinity
19
20     solinit = bvpextend(sol,10^(nb-1));
21     sol = bvp4c(@ss1ode,@ss1bc,solinit);
22     eta = sol.x;
23     f = sol.y;
24     loglog(eta,f(1,:), 'bo');
25     drawnow
26
27 end
28 hold off
29 function dfdeta = ss1ode(eta,f)
30     dfdeta = [ f(2)
31               -(eta.*f(2)/2 + f(1)/3 + f(1).^4) ];
32 end
33
34 function res = ss1bc(f0,finf)
35     res = [f0(1)-0.3
36           f0(2)];
37 end
38 end % c10ss1
```

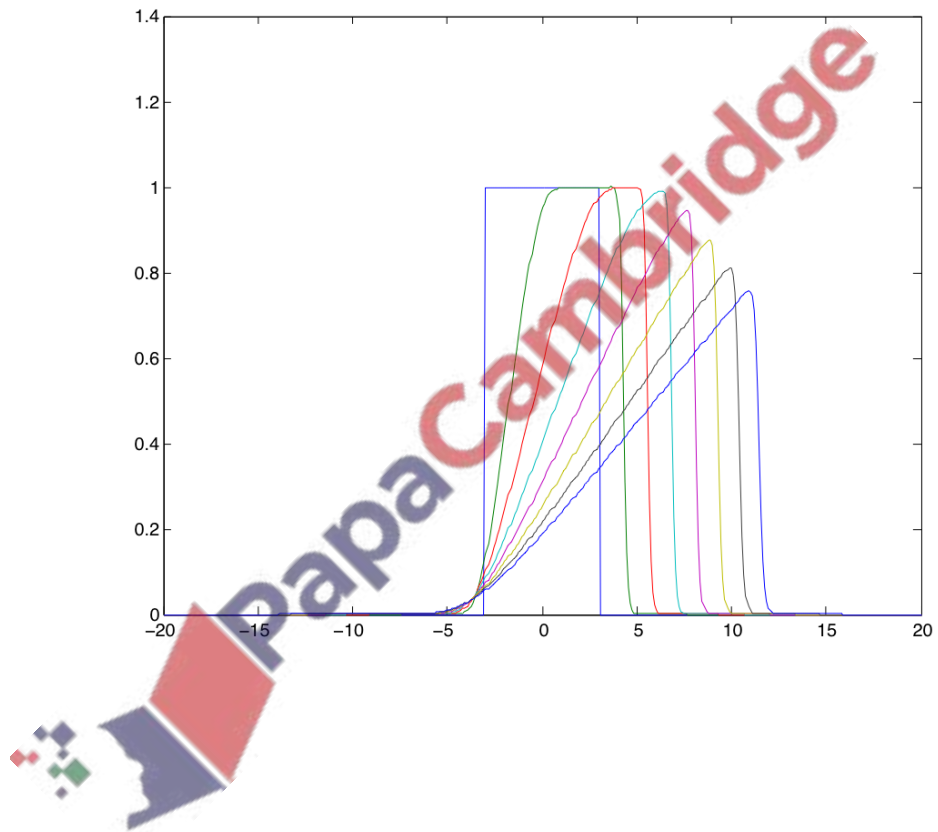


Figure 11.6. Solution to Burger's equation.

when $|x| < 3$. We also take $\nu = 0.1$. The solution is shown in figure 11.6. Three features are noteworthy: The solution propagates to the right; the solution does not smear out at all on the right hand side but appears to maintain its characteristic width; and finally on the left hand side the solution does smear out.

It is sometimes useful to rewrite the Burgers equation in this form:

$$\partial_t u = \partial_x \left[-\frac{u^2}{2} + \nu \partial_x u \right] \tag{11.32}$$

First lets look at some properties of Burgers equation.

1. Total u is conserved. In other words $\frac{d}{dt} \int_{-\infty}^{\infty} u dx = 0$ and $\int_{-\infty}^{\infty} u dx = \text{constant}$ if $\int_{-\infty}^{\infty} u dx = \text{finite}$, for $t = 0$.

Proof: Integrate Burgers equations from $x = -\infty$ to $x = \infty$. $\int_{-\infty}^{\infty} \partial_t u dx = \int_{-\infty}^{\infty} \partial_x \left[-\frac{u^2}{2} + \nu \partial_x u \right] dx = \left[-\frac{u^2}{2} + \nu \partial_x u \right]_{-\infty}^{\infty} = 0$

We used the fact that $u \rightarrow 0$ as $x \rightarrow \pm\infty$ so that $\int_{-\infty}^{\infty} u dx$ is finite at $t = 0$.

For the second part: $\frac{d}{dt} \int_{-\infty}^{\infty} u dx = 0$ implies that $\int_{-\infty}^{\infty} u dx = \text{constant}$, where the constant is decided by intial conditions.

2. Total u^2 decays monotonically if $\int_{-\infty}^{\infty} u^2 dx$ is finite at $t = 0$. This means u does not blow up.

Proof: Integrate u times the Burgers equation from $x = -\infty$ to ∞ . $\int_{-\infty}^{\infty} u \partial_t u dx = \int_{-\infty}^{\infty} -u^2 \partial_x u + \nu u \partial_{xx} u dx$

Using integration by parts and the fact that $u \rightarrow 0$ as $x \rightarrow \pm\infty$, $\frac{1}{2} \int_{-\infty}^{\infty} \partial_t u^2 dx = \int_{-\infty}^{\infty} -\frac{1}{3} \partial_x u^3 + \nu u \partial_{xx} u |_{-\infty}^{\infty} - \nu \int_{-\infty}^{\infty} (\partial_x u)^2 dx = -\frac{1}{3} u^3 |_{-\infty}^{\infty} - \nu \int_{-\infty}^{\infty} (\partial_x u)^2 dx$
 $\frac{d}{dt} \int_{-\infty}^{\infty} u^2 dx = -2\nu \int_{-\infty}^{\infty} (\partial_x u)^2 dx$.

Since this is always negative, $\int_{-\infty}^{\infty} u^2 dx$ must always be decaying.

3. At any finite t , $u(x, t)$ is smooth.

Lets try to understand the shape of $u(x, t)$: It appears from the numerical solution, there are traveling waves. Let us first consider whether there are solutions that move with constant velocity: we take the ansatz $u(x, t) = F(x - Vt)$. Plugging this into Burger's equation gives

$$-VF' + FF' = \nu F'' \tag{11.33}$$

If we now integrate this equation between $x = -\infty$ and $x = \infty$, and require that as $|x| \rightarrow \infty$, $u \rightarrow \text{constant}$, we see that

$$-(Vu_+ - Vu_-) + (u_+^2/2 - u_-^2/2) = \nu u' = 0 \tag{11.34}$$

Or, the velocity V must satisfy

$$V = \frac{1}{2} \frac{u_+^2 - u_-^2}{u_+ - u_-} = \frac{1}{2} (u_+ + u_-) \tag{11.35}$$

Thus the velocity is the average of the value of u between the front and the back! Additionally one can integrate the differential equation to get the exact solution

$$u = u_- + \frac{u_+ - u_-}{1 + \exp\left[\frac{u_+ - u_-}{2\nu}(x - Vt)\right]}. \quad (11.36)$$

Hence the characteristic width of the solution is $\ell = 2\nu/(u_+ - u_-)$.

This solution shows that solutions with $u > 0$ will propagate to the right (since then the average u must be positive), and moreover that the characteristic width of the solution that is propagating scales linearly with ν . For our simulation above we expect that roughly $\ell \sim 0.1$ and this is in accord with the figure shown. On the other hand it is equally clear that this solution does not explain what we see in the figure: there, the front propagates to the right, but its amplitude decreases in time. According to the solution we just constructed we should expect that if the amplitude decreases, the velocity should also decrease and the characteristic width should increase. Can we construct a solution with these properties?

The fundamental reason that the computed solution differs from a simple travelling wave is volume conservation: the equation implies that $\int_{-\infty}^{\infty} u(x, t) = \text{constant}$. This therefore motivates us to look for a solution of the form

$$\frac{1}{L(t)} F\left(\frac{x - x_0(t)}{L(t)}\right). \quad (11.37)$$

If we plug this into the PDE we find that all of the terms balance as long as $L = \sqrt{\nu t}$ and $x_0 \sim \sqrt{\nu t}$. The function F obeys the ordinary differential equation

$$-(F + \eta F_\eta) - F_\eta + FF_\eta = \nu F_{\eta\eta}. \quad (11.38)$$

This equation can be integrated once to yield

$$\nu F_\eta = -F(1 + \eta) + \frac{F^2}{2}, \quad (11.39)$$

where we have zeroed the constant of integration because we require that $F \rightarrow 0$ as $\eta \rightarrow \infty$.

Note that this equation explicitly depends on η : hence we do not have translation invariance. If $F(\eta)$ is a solution then it is not true that $F(\eta + c)$ is a solution.

In integrating this solution, we start at $\eta = 0$ and specify a single initial condition $F(0)$. Figure 11.7 shows these solutions for a range of initial conditions—note that each initial condition has a different total mass.

Thus the procedure for understanding the solution of initial value problem is as follows: first find the $F(0)$ such that the mass of the solution in question corresponds to the mass of the initial value problem. Then this is the mass of your solution at long times!

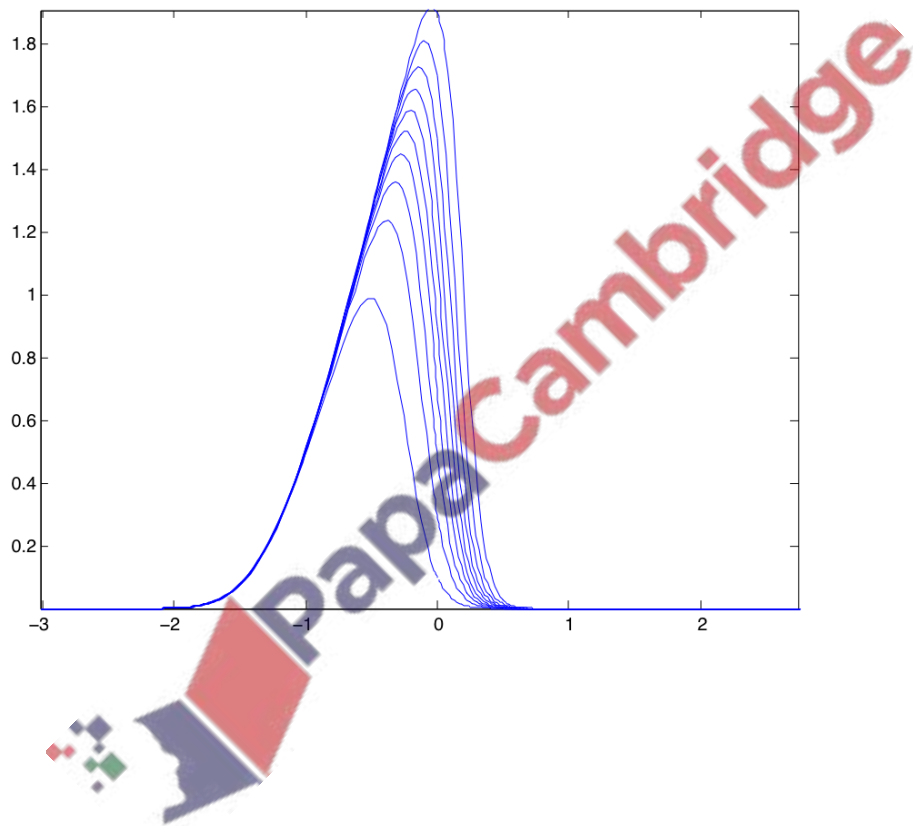


Figure 11.7. similarity solutions for burger's equation for a range of initial conditions.

11.4 Burger's Equation

To be filled in.

11.5 Pattern Formation

The equations thus far have led to either the amplification or dissipation of the initial condition. The equations did not have, in themselves, characteristic length scales that could lead to the formation of a pattern. Now we consider an equation that has this property.

Consider

$$\partial_t u = -\partial_{xx} u - \partial_{xxxx} u + u_x^2. \quad (11.40)$$

