UNIVERSITY OF MICHIGAN

Department of Nuclear Engineering
and Radiological Sciences

APPLIED MATHEMATICS FOR ENGINEERING PHYSICS

Brian Kiedrowski

March 20, 2025

Contact information: `bckiedro@umich.edu`.

Contents

Chapter 1

Introduction

This textbook is designed to review and add engineering context and relevance for the undergraduate math sequence at the University of Michigan with a specific focus on the discipline of nuclear engineering. Given the breadth of the nuclear field, the contents herein should also be applicable to other engineering disciplines as well. Specifically, this text contains mathematical techniques relevant to mechanics, fluids and heat transfer, electromagnetism, quantum mechanics, in addition to topics such as neutron diffusion and nuclear criticality.

For University of Michigan students in Nuclear Engineering and Radiological Sciences, the field makes heavy use of mathematics and computing and NERS 320 is meant to provide students a solid mathematical foundation relevant to the advanced undergraduate coursework and beyond, whether that be a career in industry, graduate school, etc. This course assumes that students have completed the standard mathematics sequence at the University of Michigan. In particular, this course relies heavily on the material in Calculus III, multivariate calculus, and Calculus IV, ordinary differential equations.

The course has five units: linear algebra, systems of ordinary differential equations, vector calculus, partial differential equations, and probability. Each unit is a course unto itself; as such, this text can only superficially cover these topics and is not a substitute for the aforementioned mathematics courses or other math courses that can cover these topics in greater depth.

1.1 Mathematical Models in Engineering

Mathematics is the language that scientists and engineers use to describe the workings of the universe. It offers a consistent framework and set of rules by which we can explain and manipulate the forces of nature for the benefit of humanity and the planet. A common goal of science and engineering is to take a complicated set of phenomena and construct a mathematical model that we hope is relevant and has high enough fidelity for its purpose. In constructing such as model, it is important

to realize that all such models are limited by either concerns of practicality of solving the equations or from ignorance about the underlying processes. For this it is the responsibility of the scientist or engineer to ensure that the mathematical model, the model parameters, and the techniques to solve the equations in the model are adequate for the task at hand.

Some examples of mathematical models we encounter in the nuclear engineering discipline are as follows. The first of these are the Bateman equations that describe the production and decay of a population of isotopes:

$$\frac{dN_i}{dt} = -\lambda_i N_i(t) + \sum_j f_{ij}\lambda_j N_j(t) + Q_i(N_1(t), N_2(t), \cdots, t). \tag{1-1}$$

This is a coupled set of equations and are important in the fields of health physics, the design of nuclear fission and fusion reactors, and the management of radioactive and nuclear materials. Note that the production term Q may be a function of the isotopes in the system itself and is an important consideration in nuclear fission reactors where the isotope mixture impacts the neutron radiation field that produces the isotopes.

Another important equation for NERS is the heat equation, which describes the flow of thermal energy through a system and obtains an unknown temperature field $T(x, y, z, t) = T(\boldsymbol{x}, t)$ that varies in space and time:

$$\rho c_p \frac{\partial T}{\partial t} - \nabla \cdot k(\boldsymbol{x}, T)\nabla T(\boldsymbol{x}, t) = q(\boldsymbol{x}, t). \tag{1-2}$$

This partial differential equation is important because nuclear reactions are often used as a heat source for various applications such as nuclear fission or fusion reactors, radioisotope thermoelectric generators, and as a consequence of energy deposition by fields of radiation.

A related and more complicated equation is the one that describes the flow of fluids with an unknown velocity field $\mathbf{u}(\boldsymbol{x}, t)$. These are the Navier-Stokes equations for compressible flow:

$$\frac{\partial}{\partial t}(\rho\mathbf{u}) + \nabla \cdot (\rho\mathbf{u} \otimes \mathbf{u}) = -\nabla p + \mu\nabla^2\mathbf{u} + \frac{1}{3}\mu\nabla(\nabla \cdot \mathbf{u}) + \rho\mathbf{g}. \tag{1-3}$$

This equation is inherently nonlinear and gives rise to turbulence, perhaps one of the most difficult phenomena to model in physics. The relevance of this equation is that it is the one that governs fluid motion, which is vital in heat removal for nuclear systems. It is also one of the equations that may be used for the flow of plasmas, which is vital to understanding systems involving nuclear fusion.

Nuclear fusion systems require extreme temperatures that require an understanding of the physics of plasmas or ionized gasses. In plasmas and fusion systems, the behavior of electromagnetic fields is of primary concern. Additionally, the equations of electromagnetism arises in the study of radiation detection. The relevant equations are Maxwell's equations, which can be written in either differential or integral form. The differential form of these are:

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0}, \text{ (Gauss' law)} \tag{1-4a}$$

$$\nabla \cdot \mathbf{B} = 0, \tag{1-4b}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \text{ (Faraday's law)} \tag{1-4c}$$

$$\nabla \times \mathbf{B} = \mu_0 \left(\mathbf{J} + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right), \text{ (Ampere's law).} \tag{1-4d}$$

Maxwell's equations involve a heavy amount of vector calculus with the unknowns being the components of the electric field $\mathbf{E}$ and magnetic field $\mathbf{B}$. Also, they are often combined with the fluids equations through the external force term $\mathbf{g}$ in plasma physics applications.

The last of this list is the equation describing the evolution of a radiation field. This is done with the linear Boltzmann equation that describes the position, direction, kinetic energy, and time (7 dimensions) of the radiation field:

$$\begin{aligned} &\frac{1}{v}\frac{\partial \psi}{\partial t} + \hat{\mathbf{\Omega}} \cdot \nabla \psi(\boldsymbol{x}, \hat{\mathbf{\Omega}}, E, t) + \Sigma_t(E)\psi(\boldsymbol{x}, \hat{\mathbf{\Omega}}, E, t) \\ &= \iint \Sigma_s(E' \to E, \hat{\mathbf{\Omega}}' \cdot \hat{\mathbf{\Omega}})\psi(\boldsymbol{x}, \hat{\mathbf{\Omega}}', E', t) d\Omega' dE' + Q(\boldsymbol{x}, \hat{\mathbf{\Omega}}, E, t). \end{aligned} \tag{1-5}$$

Of note is that this is an integro-differential equation in that the process of scattering (radiation changing direction and energy in collisions) is an integral over the incident directions and energies. This coupled with the dimensionality of the problem presents its own set of challenges that are largely unique to nuclear engineering.

By no means is this list comprehensive; rather, it is meant to illustrate the complexity and mathematical richness of the physics encountered in NERS and motivate the need for the mathematical content in this text.

1.2 Computational Physics

In the ideal word, we would encounter a set of equations describing physical phenomena (usually partial differential equations) and we would go about obtaining analytical solutions. Unfortunately, this is usually only possible in a limited set of cases that are usually too simplistic for practical engineering. Therefore, we invariably must take those equations in our mathematical model and apply techniques to approximate them as a simpler set of equations that we are capable of solving, usually only numerically. These equations are then often implemented in some computer software that we use to obtain numerical results.

This is the essence of using mathematics in engineering and will be a common theme of the course. Throughout this course, we will cover some of the basic considerations with approximating the complex models encountered in NERS with simpler ones. Before going into detail, it is important to cover two vital topics in scientific computing called verification and validation. In short, it is the responsibility of the engineer to ensure their mathematical models are being solved correctly and that they are applicable to the engineering problem. Failure to do suitable verification and validation can lead to erroneous results that can lead to flawed design decisions being

made with potentially life threatening impacts. For this reason, ensuring correctness and suitability of mathematical models is also an important theme of this course.

1.2.a Verification

An approximate set of equations are obtained from a more general mathematical model and these approximate equations are then solved using some computer code. For a given engineering design application, it is important to certify that the software is actually solving its equations correctly and as intended. Much of the responsibility for this falls onto the software developer; however, it is important for the end user engineer to ensure that the relevant portions of the software have been adequately verified.

Some examples of verification methods are as follows:

- **Unit testing** checks individual operations and functions within the software to ensure that each small unit of the code is doing the intended operation. While this strategy is important for catching many errors in the software, and vital to the development process, it is impractical to apply unit testing in a way that stresses all possible combinations of cases throughout a large simulation code. Usually this is done by the software developer and is noted by the end user in deciding the quality of the software.

- **Benchmark comparisons** check the ability of the code to calculate results of reference solutions. The "gold standard" for this are analytical benchmarks, which involve comparing simulation results against solutions of the underlying differential equation(s); these solutions tend to be few and only for very simple cases, but can be revealing of deficiencies in the software. The next best thing are numerical benchmarks that have been obtained using alternative approaches to solving the problem that themselves may be too complicated to be practical except for benchmarking. Often software developers will do some of this and, where relevant to the specific problem, should be noted. In cases where this has not been performed for the application, it may be advisable to search the literature for analytical or numerical benchmarks or devise some that have some of the mathematical features of the application, but are simple enough to permit a solution.

- **Code comparisons**, as the name implies, checks the solution from a particular software to another code that has hopefully been vetted. Ideally, the other code would use a different method to solve the same problem. When possible, it is good to occasionally double check the preferred design tool in an organization with another to provide confidence that the design calculations are being performed correctly. Usually, this is done exclusively by the end-user engineer and not the developers of the software.

Another consideration related to verification is ensuring the code is being used appropriately. Most methods involve approximating some differential equation in some

manner; the most common of these involves transforming the differential equations into a set of linear equations described by a spatial mesh. A finer mesh preserves more of the mathematics of the initial equations, but at the expense of computational resources (memory and computational time). Many engineering calculations involve, for example, a mesh sensitivity study to ensure that the spatial grid is sufficiently resolved to retain an acceptable amount of fidelity in the model.

1.2.b Validation

Validation is the process by determining whether the underlying mathematical models and the associated physical data (e.g., equation of state, physical properties, nuclear cross sections) are sufficiently accurate for the engineering application. This differs from verification in that it checks only that the mathematical equations are being solved correctly, not that the mathematical model itself is suitable. (It is possible, albeit very undesirable, for a code to solve equations incorrectly, but be "good enough" for the engineering design.)

The process of validation often involves comparing the results obtained from a code with numerical measurements of analogous quantities in scenarios ranging from small-scale laboratory experiments looking at one physical phenomenon, to prototypes that are simplified versions of the system, up to a full system in real-world operating conditions. From these comparisons, a quantitative assessment of how predictive the code is with reality can be made. Additionally, validation exercises are often used to quantify the bias. For example, if a particular calculation tends to predict a mechanical stress that is 5-10% lower than measurement for a given system, then this can be accounted for in the design process and mathematical models are calibrated accordingly. Many organizations will devote resources to performing experiments whose sole purpose is to test and calibrate computational models. Indeed, much of the historical nuclear engineering design work relied on comparing prototypes and then operating reactors to calculations and then calibrating results accordingly, and this is still ongoing today, albeit at a smaller scale.

As part of the engineer using software for design and analysis, it is important to be aware of what experiments have been performed relevant or similar systems that can be used to test the predictive capability of the computational design models and to perform those comparisons. Depending on the nature of the work, there may need to be conversations between the analysts and the experimenters to develop suitable experiments to test the models.

An example in the nuclear engineering field is the area of criticality safety, which concerns itself with the safe handling of fissionable materials in industrial processes to avoid the formation of a critical mass of nuclear material. For this work, software that simulates the neutron radiation throughout the systems of interest are employed to estimate the nuclear criticality or effective multiplication of the system. Unfortunately, the nuclear properties are not known well enough to give perfect agreement with reality, so criticality safety engineers perform validation studies by looking for benchmark experiments that are similar to the application, testing the radiation transport solver on those, and quantifying how well or poorly the code predicts the

experiments. These results are analyzed statistically and margins or safety factors are put into the design process to ensure subcriticality.

In short, validation of software is the responsibility of the end-user engineer performing the analysis. While software developers may perform some of their own validation on a wide set of systems, they cannot be familiar with and test every possible application. And even if they could, it is still up to the engineer to understand how well (or poorly) the mathematical models, software, and material data perform together on the application to build in appropriate safety factors.

1.2.c Example: Radioactive Decay

To illustrate the difference between the concepts of verification and validation, consider the simple case of radioactive decay. This follows the simple differential equation

$$\frac{dN}{dt} = -\lambda N(t), \quad N(0) = N_0. \tag{1-6}$$

While this has a simple solution of an exponential,

$$N(t) = N_0 e^{-\lambda t}, \tag{1-7}$$

numerical integration techniques can be applied to approximate the solution. (It is foolish to apply numerical techniques to such a simple case, but such techniques quickly become necessary as problems become too difficult to solve analytically.) The simplest numerical integration technique is called Forward Euler (see Sec. 3.5 for details) and involves taking a finite time step Δt. If implemented correctly, the approximate solution from forward Euler should approach the analytic solution as Δt becomes small. A verification exercise checks that this is indeed the case.

As a verification exercise, let $\lambda = 1$ and see how well the approximate numerical solution matches the analytic exponential solution in this case. This comparison is shown in Fig. 1.1. The dark solid line gives the analytic solution and dots (interpolated with lines for clarity) give points from the numerical solution for different time step sizes. Based on this result, one can say that the forward Euler method is correctly approximating the right equation and exhibits the expected limiting behavior.

Validation checks that the model is appropriate for the problem being analyzed and often involves using experimental data. For the purposes here, fictitious experimental data of radioactive detection counts are used. (This is generated for the plot by using the analytic solution and applying a random offset to each point to simulate the effect.) First, consider the isotope ^{18}F, which is used in positron emission tomography, a medical diagnostic. ^{18}F has a half life of 1.82 hr or a decay constant of $\lambda = 0.379$ hr^{-1}. The comparison of the activity computation versus the "experimental" data is given in Fig. 1.2. While there is some variation because of the noise in the data (that would arise from measurement uncertainties because of a finite counting time), the computation appears to adequately represent the experimental data trends.

Next, consider the radioisotope ^{135}I, which is a common product of nuclear fission and has a half life of about 6.6 hr or a decay constant of $\lambda = 0.105$ hr^{-1}. The

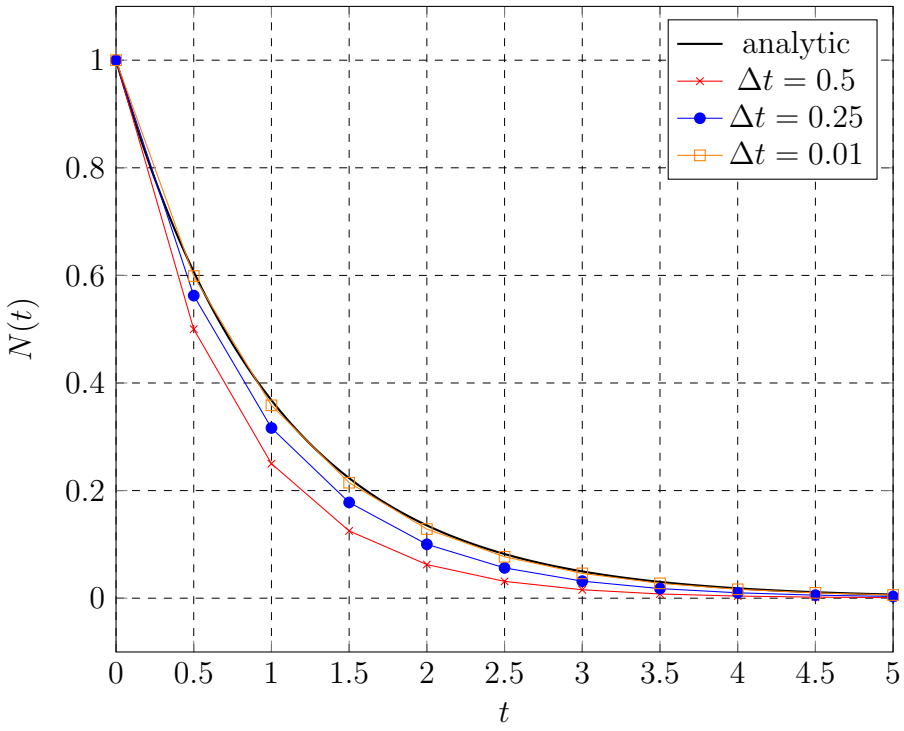

Figure 1.1: Verification exercise for simple radioactive decay.

calculation versus the (simulated) experimental data is shown in Fig. 1.3. Clearly, the computed activity does not agree with the experimental data at all. In fact, the measured activity increases initially and then decreases. What explains this is that ^{135}I decays to ^{135}Xe, which is itself radioactive and has a half life of 9.2 hours and decays to nearly stable ^{135}Cs. The single decay model being used by the computation does not account for the secondary decays of ^{135}Xe and cannot therefore adequately reproduce this experimental data. In terms of validation, it can be said that the mathematical model is not suitable for ^{135}I.

1.2.d Comment on Analytical Models

While it is rare for an analytical model to be good enough for practical design calculations, this course still covers many of those analytical techniques. The reason for this is while it may be impossible to solve the equations for the actual engineering application directly, most such applications follow some general principles and trends. The analytical methods can provide physical insight into, for example, the general shape of a particular temperature or radiation field, and the expected behavior of how making design changes will impact a system. While sometimes one encounters counterintuitive results, it is more often that when calculations do not follow mathematical intuition from simpler models, there is usually something wrong with the way the software is being used (e.g., errors in input) or there are misunderstandings on the part of the engineer about the physics of the system being analyzed.

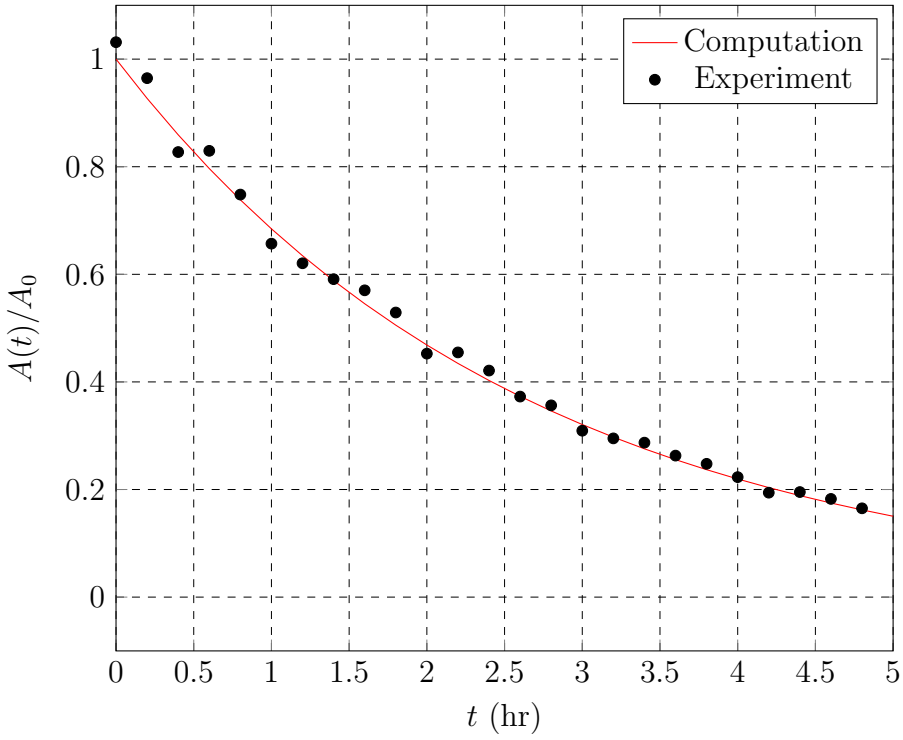

Figure 1.2: Validation of exponential decay model for ^{18}F.

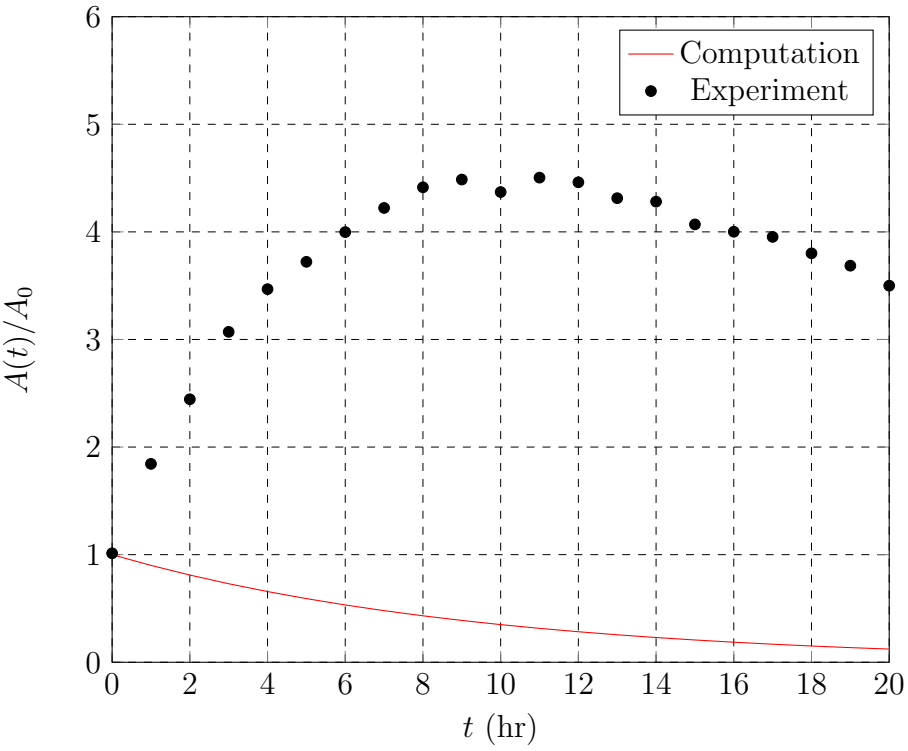

Figure 1.3: Validation of exponential decay model for ^{135}Xe.

Chapter 2

Linear Algebra

The techniques of linear algebra are widely used in the field of engineering. Many practical engineering problems often involve solving partial differential equations that can rarely be solved exactly using analytical techniques of calculus. In these cases, the partial differential equations are approximated as a large system of linear equations that can be solved to obtain results that are hopefully accurate enough to make design or operational decisions. In this manner, many of the simulation tools used regularly by engineers involve the techniques of linear algebra, and it is important that engineers using those tools have a basic understanding of the underlying mathematics so as to use them appropriately.

This chapter assumes the reader has had some exposure to linear algebra. It begins by giving a quick review of the fundamental objects of linear algebra: namely matrices and vectors and their fundamental operations.

The next topic is the change of basis, which is fundamental to understanding conversions between different coordinate systems. A key point for an engineer is that while the physics should yield the same end results, solving problems in one coordinate system versus another may prove to be significantly simpler.

Next, linear systems of algebraic equations are covered along with a few common algorithmic techniques used to solve them. These fall into two categories: direct methods and iterative methods. The former will yield the solution (should such a unique solution exist), but is often prohibitive for the large systems of equations encountered in routine engineering calculations. The special case of the direct solve for the tridiagonal matrix is also shown, which occurs commonly in approximate solutions of differential equations. The iterative solution methods, which under certain conditions that are discussed, provide successively more accurate estimates of the solution.

Finally, the chapter reviews eigenvalues and eigenvectors. First, this concept is fundamental to understanding a key quantity in nuclear engineering: the effective multiplication factor of a system containing fissionable material, which is an eigenvalue. Furthermore, eigenvalues and eigenvectors are incredibly useful for solving linear systems of ordinary differential equations, which is discussed in the following

chapter.

2.1 Matrices

A matrix is an two-dimensional object containing information. A matrix has dimensions $N \times M$, where N is the number of rows and M is the number of columns. A matrix can be represented as

$$\mathbf{A}_{N\times M} = \mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,M} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{N,2} & \cdots & a_{N,M} \end{bmatrix}. \tag{2-1}$$

The size subscript is given here for emphasis, but is almost always excluded.

2.1.a Matrix Transpose

The transpose operator takes a $N \times M$ matrix and constructs a $M \times N$ matrix by flipping the indices of the elements. The transpose is defined as:

$$\mathbf{A}^\top = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,M} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{N,2} & \cdots & a_{N,M} \end{bmatrix}^\top = \begin{bmatrix} a_{1,1} & a_{2,1} & \cdots & a_{M,1} \\ a_{1,2} & a_{2,2} & \cdots & a_{M,2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1,N} & a_{2,N} & \cdots & a_{M,N} \end{bmatrix}. \tag{2-2}$$

Note that $\top$ superscript. In other words, the elements of $\mathbf{A}^\top$ are $a_{j,i}$.

The transpose satisfies several properties related to matrix addition, multiplication, and inverses that is discussed in the subsequent sections. An important property is that the transpose operation applied twice produces the original matrix:

$$\left(\mathbf{A}^\top\right)^\top = \mathbf{A}. \tag{2-3}$$

The transpose is mostly used when the elements are strictly real. In fields such as quantum mechanics, complex numbers of the form $a + bi$ where $i = \sqrt{-1}$, the imaginary unit, are regularly encountered. In cases where the matrix elements are complex, it is common to take the *conjugate transpose* instead of just the transpose. The conjugate transpose is exactly like the transpose, in that it saws the rows and columns, but also takes the complex conjugate of each element in the resulting matrix $a + bi \to a - bi$. The conjugate transpose is denoted by a star superscript or $\mathbf{A}^*$.

For example, consider the following matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 2+i & 0 \\ 4 & -3i & i \end{bmatrix}.$$

Its conjugate transpose is

$$\mathbf{A}^* = \begin{bmatrix} 1 & 4 \\ 2-i & 3i \\ 0 & -i \end{bmatrix}.$$

2.1.b Special Types of Matrices

Some matrices have particular forms that occur regularly in scientific and engineering applications. Furthermore, when a matrix has a particular form, it often has useful properties. In this section, a few important classes of matrices are defined. (There are numerous others that are not mentioned here.)

A matrix is said to be a *square matrix* if the number of its rows and columns are equal, $N = M$. An example of a square matrix is

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}.$$

Specifically, this is a 3×3 matrix.

A matrix is said to be a *diagonal matrix* if it is both a square matrix and all $a_{ij} = 0, i \neq j$. In other words, all the off-diagonal elements are zero. Three examples of diagonal matrices are

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}, \quad \mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{0} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

The first of these is a 4×4 diagonal matrix. The second example is a 3×3 diagonal matrix where all elements on the diagonal are one. This is called an *identity matrix*, which has special properties for matrix multiplication. The third example is a 2×2 matrix where all the elements are zero. In addition to being a diagonal matrix, this is also a *zero matrix*, which has special properties for matrix addition and multiplication. Note that identity matrices are always diagonal matrices whereas zero matrices can be of any size, where the latter is when all elements $a_{ij} = 0$.

A matrix is said to be a *symmetric matrix* when $A = A^\top$ or $a_{ij} = a_{ji}, i \neq j$. This requires the matrix to also be a square matrix. Note that diagonal matrices are always symmetric matrices. An example of a symmetric matrix is

$$\mathbf{S} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 5 & 6 \\ 3 & 6 & 9 \end{bmatrix}.$$

For the case when the matrix elements are complex, we can define a *Hermitian matrix* as one that is equal to its own conjugate transpose. For example, the following matrix is Hermitian:

$$\mathbf{H} = \begin{bmatrix} 1 & 2i & 3-i \\ -2i & 2 & -1 \\ 3+i & -1 & 3 \end{bmatrix}.$$

Note that a real, symmetric matrix is also Hermitian since the complex conjugate of a real number is itself.

2.1.c Column and Row Vectors

Two special cases of matrices that are particularly important are called column and row vectors. A column vector is a matrix with a single column and N rows, i.e., a $N\times 1$ matrix and can be represented as

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix}. \tag{2-4}$$

Conversely, a row vector is a $1\times N$ matrix, which may be represented as

$$\mathbf{a}^\top = \begin{bmatrix} a_1 & a_2 & \cdots & a_N \end{bmatrix}. \tag{2-5}$$

Note the $\top$ superscript on the vector, which is the transpose operator discussed previously. The column vector is sometimes referred to as a (ordinary) vector and a row vector is sometimes called a co-vector. Both of these have geometrical interpretations that will be discussed.

2.1.d Matrix Addition, Subtraction, and Scaling

We can add two matrices of the same dimension. If $\mathbf{A}$ and $\mathbf{B}$ are both $N\times M$, then

$$\begin{aligned} \mathbf{A}+\mathbf{B} &= \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,M} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{N,2} & \cdots & a_{N,M} \end{bmatrix} + \begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,M} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ b_{N,1} & b_{N,2} & \cdots & b_{N,M} \end{bmatrix} \\ &= \begin{bmatrix} a_{1,1}+b_{1,1} & a_{1,2}+b_{1,2} & \cdots & a_{1,M}+b_{1,M} \\ a_{2,1}+b_{2,1} & a_{2,2}+b_{2,2} & \cdots & a_{2,M}+b_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1}+b_{N,1} & a_{N,2}+b_{N,2} & \cdots & a_{N,M}+b_{N,M} \end{bmatrix}. \end{aligned} \tag{2-6}$$

Matrix subtraction follows similarly:

$$\begin{aligned} \mathbf{A}-\mathbf{B} &= \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,M} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{N,2} & \cdots & a_{N,M} \end{bmatrix} - \begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,M} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ b_{N,1} & b_{N,2} & \cdots & b_{N,M} \end{bmatrix} \\ &= \begin{bmatrix} a_{1,1}-b_{1,1} & a_{1,2}-b_{1,2} & \cdots & a_{1,M}-b_{1,M} \\ a_{2,1}-b_{2,1} & a_{2,2}-b_{2,2} & \cdots & a_{2,M}-b_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1}-b_{N,1} & a_{N,2}-b_{N,2} & \cdots & a_{N,M}-b_{N,M} \end{bmatrix}. \end{aligned} \tag{2-7}$$

If the two matrices being added are not the same size, then neither addition nor subtraction are defined.

Matrix addition is commutative:

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}. \tag{2-8}$$

By extension, it is also associative:

$$\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}. \tag{2-9}$$

In other words, the order that matrices are added does not matter.

It is always possible to add the zero matrix of the appropriate size to a matrix and get the same result

$$\mathbf{A} + \mathbf{0} = \mathbf{A}. \tag{2-10}$$

This property is useful because many mathematical derivations involve adding $\mathbf{A} - \mathbf{A}$ to a side of an equation, which then allows for some simplification to occur.

The product of a number and a matrix simply scales all of the elements of the matrix:

$$\begin{aligned} r\mathbf{A} &= r \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,M} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{N,2} & \cdots & a_{N,M} \end{bmatrix} \\ &= \begin{bmatrix} ra_{1,1} & ra_{1,2} & \cdots & ra_{1,M} \\ ra_{2,1} & ra_{2,2} & \cdots & ra_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ ra_{N,1} & ra_{N,2} & \cdots & ra_{N,M} \end{bmatrix}. \end{aligned} \tag{2-11}$$

Scalar multiplication and addition can be combined to show that matrices satisfy the distributive property:

$$r(\mathbf{A} + \mathbf{B}) = r\mathbf{A} + r\mathbf{B}. \tag{2-12}$$

In other words, we are free to add and scale or scale and then add. Note that matrix subtraction is simply a compact way of stating addition of a matrix by another matrix that has been scaled by a factor of -1.

If the scalar 0 is multiplied by any matrix, the result will be the zero matrix of the same size:

$$0\mathbf{A} = \mathbf{0}. \tag{2-13}$$

The matrix transpose is distributive under addition and scaling:

$$(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top, \tag{2-14a}$$

$$(r\mathbf{A})^\top = r\mathbf{A}^\top. \tag{2-14b}$$

2.1.e Matrix Multiplication

The multiplication between two matrices is not as straightforward as the rules for addition, subtraction, and scaling. Suppose we have two matrices $\mathbf{A}$, which has dimension $N\times M$, and $\mathbf{B}$, which has dimension $L\times K$. We can only take the product of $\mathbf{AB}$ (note the order!) if number of columns M of matrix $\mathbf{A}$ is equal to the number of rows L of matrix $\mathbf{B}$, i.e., $M = L$. The result of the matrix multiplication produces an $N\times K$ matrix. If this is not the case, then matrix multiplication is not defined.

The rule for matrix multiplication is

$$\begin{aligned}\mathbf{C} = \mathbf{AB} &= \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,M} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{N,2} & \cdots & a_{N,M} \end{bmatrix} \begin{bmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,K} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ b_{M,1} & b_{M,2} & \cdots & b_{M,K} \end{bmatrix} \\ &= \begin{bmatrix} c_{1,1} & a_{1,2} & \cdots & a_{1,K} \\ c_{2,1} & a_{2,2} & \cdots & a_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ c_{N,1} & a_{N,2} & \cdots & a_{N,K} \end{bmatrix},\end{aligned}$$

$$c_{i,j} = \sum_{k=1}^{M} a_{i,k} b_{k,j}. \tag{2-15}$$

Unlike multiplication between numbers, the order that the matrices are multiplied matters. In many cases, the matrix multiplication will not be defined.

In the common case where the matrices are square (number of rows equals the number of columns) both $\mathbf{AB}$ and $\mathbf{BA}$ are defined, but **not** the same. In other words, matrix multiplication is *non-commutative* except for certain special cases:

$$\mathbf{AB} \neq \mathbf{BA}. \tag{2-16}$$

For this reason, it is important to state whether the multiplication occurs on the left or on the right. In this case matrix $\mathbf{A}$ left multiplies $\mathbf{B}$ and $\mathbf{B}$ right multiplies $\mathbf{A}$.

Matrix multiplication is associative. Suppose the product $\mathbf{ABC}$ is defined. The same result will be produced by either taking the product of $\mathbf{A}$ and $\mathbf{B}$ first and then left multiplying the result on $\mathbf{C}$ or taking the product of $\mathbf{B}$ and $\mathbf{C}$ first and then right multiplying the result on $\mathbf{A}$. In other words,

$$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC}). \tag{2-17}$$

Matrix multiplication is distributive so long as the multiplication is applied consistently on the left or right:

$$\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}, \tag{2-18a}$$

$$(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}. \tag{2-18b}$$

Multiplication of a matrix on the left and right by the appropriately sized (square) identity matrix yields the same matrix:

$$\mathbf{I}_{N\times N}\mathbf{A}_{N\times M} = \mathbf{A}_{N\times M}\mathbf{I}_{M\times M} = \mathbf{A}_{N\times M}. \tag{2-19}$$

Here $\mathbf{A}$ is a N×M matrix and subscripts are given to explicitly denote their size.

Likewise, multiplication of an N×M matrix by an appropriately-sized zero matrix on either size yields another zero matrix:

$$\mathbf{0}_{L\times N}\mathbf{A}_{N\times M} = \mathbf{0}_{L\times M}, \tag{2-20a}$$
$$\mathbf{A}_{N\times M}\mathbf{0}_{M\times L} = \mathbf{0}_{N\times L}. \tag{2-20b}$$

Here again the subscripts are included to emphasize the sizes for didactic purposes, but are normally excluded.

Additionally, the transpose operator is distributive under matrix multiplication:

$$(\mathbf{AB})^\top = \mathbf{A}^\top\mathbf{B}^\top \tag{2-21}$$

2.1.f Linear Mapping

Matrix multiplication has at least two geometric interpretations that are discussed in the notes. One of these is the notion of a linear map. (The other involves covectors and the discussion of that interpretation will be deferred until that section.) The notion of a linear map is important to understand coordinate transformations.

Formally, a linear map must satisfy *linearity*. This implies a matrix $\mathbf{T}$ multiplied by a vectors must satisfy the following:

$$\mathbf{T}(\mathbf{x}+\mathbf{y}) = \mathbf{T}\mathbf{x} + \mathbf{T}\mathbf{y}, \tag{2-22a}$$
$$\mathbf{T}(r\mathbf{x}) = r\mathbf{T}\mathbf{x}. \tag{2-22b}$$

The first equation states that we may add the vectors $\mathbf{x}$ and $\mathbf{y}$ and then multiply by $\mathbf{T}$ or multiply first and then add. The second states that we may scale the vector and multiply or multiply and then scale.

To understand the linear map geometrically, let us consider the identity matrix $\mathbf{I}$ as a square $N\times N$ matrix. The columns of $\mathbf{I}$ can be thought of as unit vectors along the principal axes. In 2-D these unit vectors can be added in both orders to draw a unit square, in 3-D a unit cube, and the geometrical analogs for higher dimensions. If we then multiply $\mathbf{I}$ by $\mathbf{T}$ on the left, $\mathbf{TI} = \mathbf{T}$. The columns of the result $\mathbf{T}$ can also be interpreted and added both ways to form a parallelogram. The linear map for matrix

$$\mathbf{T} = \begin{bmatrix} 2 & 1 \\ 1 & -1 \end{bmatrix}$$

maps the area within the unit square to a parallelogram described by that product. This is illustrated in Fig. 2.1 with the unprimed to the single primed basis vectors.

In general, suppose **A** is another square 2×2 matrix that describes a parallelogram, then the multiplication **TA** maps that parallelogram to another parallelogram. In other words, all points (or differential area elements) within the parallelogram **A** are each uniquely moved to corresponding points (or differential area elements) within the parallelogram described by **TA**.

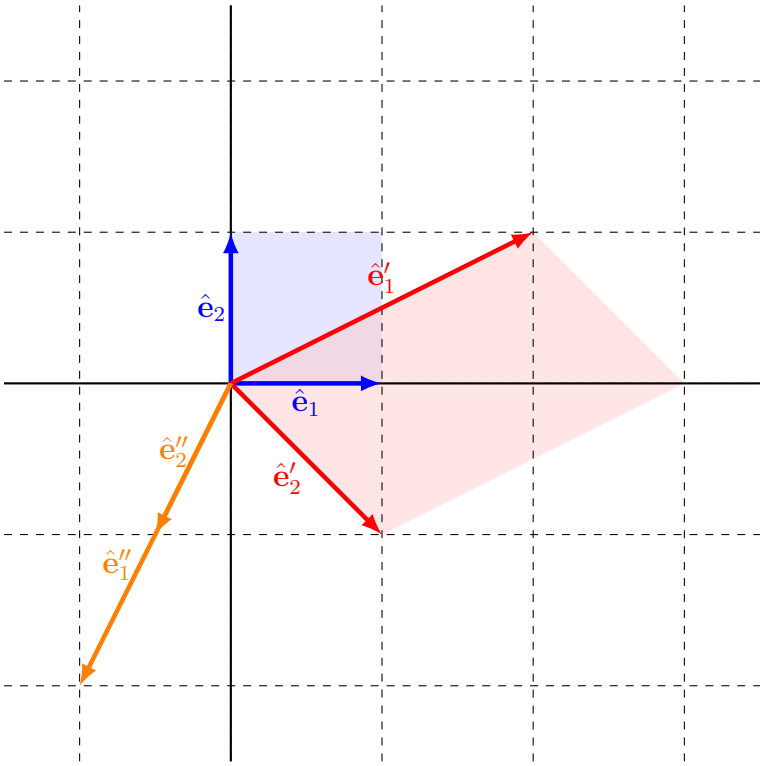

Figure 2.1: Illustration of a linear map of the unit square (unprimed basis vectors) to a parallelogram (single primed basis vectors) and to a line segment (double primed basis vectors).

This mapping can be readily extended to 3-D or any higher dimensions. In 3-D, for example, the product **TA** maps the parallelepiped described by all six combinations of sums of vectors described by the columns of **A** to another parallelepiped.

Thus far, we have looked at the case where a linear map takes an N dimensional parallelotope (an N dimensional analog of parallelogram) into another N dimensional parallelotope. It is also possible for an $N\times N$ matrix **T** to map to a lower dimensional space. Consider the matrix,

$$\mathbf{T} = \begin{bmatrix} -1 & -1/2 \\ -2 & -1 \end{bmatrix}.$$

Adding the columns of **A** in both orders forms parallelogram with zero width, which is a 1-D line segment. In other words, all points (or differential area elements) in the unit square have been mapped to points (or differential line elements) on the line. This is illustrated in Fig. 2.1 with the unprimed to the double primed basis vectors.

The matrix $\mathbf{A}$ for linear maps need not be square. In these cases, the linear map will change the dimensionality of the set of vectors being operated upon.

2.1.g Matrix Determinant

In this section we define an operation called the matrix determinant, which acts only on square $N\times N$ matrices denoted by

$$\det(\mathbf{A}) = \begin{vmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,N} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{N,2} & \cdots & a_{N,N} \end{vmatrix}. \tag{2-23}$$

Note the square braces have been replaced by vertical bars.

To understand the determinant, consider the linear map discussed in the previous section, which, takes one parallelotope and transforms it into another parallelotope. In general, this linear map does not preserve the volume. (Here volume is used to mean length, area, volume, etc. depending on the dimensionality of the space) The determinant is the ratio of the volume of the transformed parallelotope to the volume of the original parallelotope, or the volume scaling factor with a sign. This sign denotes whether or not there is a flip in orientation.

The determinant of a scalar or 1×1 matrix is simply the value of the scalar and just denotes that the length of the unit line is scaled by that amount.

The determinant of a 2×2 matrix is the area of the parallelogram defined by the vectors in the columns of $\mathbf{A}$. The formula for the determinant is given by

$$\begin{vmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{vmatrix} = a_{1,1}a_{2,2} - a_{1,2}a_{2,1}. \tag{2-24}$$

This formula is equivalent to the area of a parallelogram defined by the addition of the two column vectors of $\mathbf{A}$ in both orders.

A 3×3 matrix can be written as the sum of three 2×2 determinants:

$$\begin{vmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{vmatrix} = a_{1,1}\begin{vmatrix} a_{2,2} & a_{2,3} \\ a_{3,2} & a_{3,3} \end{vmatrix} - a_{1,2}\begin{vmatrix} a_{2,1} & a_{2,3} \\ a_{3,1} & a_{3,3} \end{vmatrix} + a_{1,3}\begin{vmatrix} a_{2,1} & a_{2,2} \\ a_{3,1} & a_{3,3} \end{vmatrix}. \tag{2-25}$$

Here each term includes a factor from the top row of the matrix and a 2×2 determinant containing the second and third rows while excluding the column of the factor. Also note that the second term has a minus sign as opposed to a plus sign.

From these example of a 3×3 determinant, we can extend the calculation of the determinant to an arbitrary number of dimensions. First, we define the quantity called the minor $M_{i,j}$ of matrix $\mathbf{A}$, which is the determinant of a submatrix $\mathbf{A}$ with the ith row and the jth column deleted. From the previous expression for a 3×3 determinant, we can write,

$$M_{1,1} = \begin{vmatrix} a_{2,2} & a_{2,3} \\ a_{3,2} & a_{3,3} \end{vmatrix}, \quad M_{1,2} = \begin{vmatrix} a_{2,1} & a_{2,3} \\ a_{3,1} & a_{3,3} \end{vmatrix}, \quad M_{1,3} = \begin{vmatrix} a_{2,1} & a_{2,2} \\ a_{3,1} & a_{3,3} \end{vmatrix}. \tag{2-26}$$

The 2×2 determinants can be evaluated using Eq. (2-24). The expression for the determinant of a square matrix of size greater than 2×2 is then

$$\det(\mathbf{A}) = \sum_{j=1}^{N} (-1)^{i+j} a_{i,j} M_{i,j}, \tag{2-27}$$

for any choice of row i (usually chosen to be one). This expression is called the Laplace expansion and must be evaluated recursively, as the minors require the computation of determinants, which may themselves require computation of more determinants, reducing in size until a set of determinants of 2×2 matrices can be evaluated.

Determinants follow a few useful identities. The determinant of an N×N matrix times a scaling factor r is the determinant of that matrix times r^N,

$$\det(r\mathbf{A}_{N\times N}) = r^N \det(\mathbf{A}_{N\times N}). \tag{2-28}$$

The determinant of a matrix is equal to the determinant of the transpose:

$$\det(\mathbf{A}^\top) = \det(\mathbf{A})\,. \tag{2-29}$$

The determinant of the product of two matrices is the product of the determinants:

$$\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B})\,. \tag{2-30}$$

2.1.h Matrix Inverse

If $\mathbf{A}$ is square, then we **might** be able to find an inverse $\mathbf{A}^{-1}$ that satisfies the following property:

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}. \tag{2-31}$$

If the matrix is not square, then there is no such inverse matrix.

To show a case where the inverse does not exist, consider the matrix

$$\mathbf{T} = \begin{bmatrix} 2 & 1 \\ 4 & 2 \end{bmatrix}.$$

This matrix maps the unit square onto a parallelogram of zero area, or a line segment. Now let's try to restore the unit square given by the columns of the identity matrix $\mathbf{I}$ by applying a transformation matrix $\mathbf{T}^{-1}$. For this to be true,

$$\begin{aligned} \mathbf{T}^{-1}\mathbf{T} &= \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 4 & 2 \end{bmatrix} \\ &= \begin{bmatrix} 2a+4b & a+2b \\ 2c+4d & c+2d \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \end{aligned} \tag{2-32}$$

The first row of the resultant matrix produces two equations:

$$2a + 4b = 1, \tag{2-33a}$$

$$a + 2b = 0. \tag{2-33b}$$

Solving the second equation yields $a = -2b$, plugging into the first gives $-4b + 4b = 0 = 1$. Since zero cannot obviously equal one, these equations are inconsistent or contradictory. It is therefore impossible to find a matrix that maps points (or differential lengths) along a line to unique points (or differential areas) within the unit square.

Recall that the resulting area of the "parallelogram" produced by applying matrix $\mathbf{T}$ has zero area. Therefore $\det(\mathbf{T}) = 0$, which means the volume scaling factor is zero, and multiplication of $\mathbf{T}$ by another matrix maps any parallelogram to a line (which has zero area). By extension, it can be said that $\mathbf{T}^{-1}$ exists if and only if $\det(\mathbf{T}) \neq 0$. Note that a matrix is said to be *singular* if its inverse does not exist.

There are a few cases for which computing the inverse by solving a system of equations directly is not too difficult. The first case is the 2×2 matrix. We write the equation $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$ in the following form:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} w & x \\ y & z \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{2-34}$$

where $\{a, b, c, d\}$ are known and $\{w, x, y, z\}$ are unknowns. Expanding this out gives the following four equations:

$$aw + by = 1, \tag{2-35a}$$
$$cw + dy = 0, \tag{2-35b}$$
$$ax + bz = 0, \tag{2-35c}$$
$$cx + dz = 1. \tag{2-35d}$$

These are two sets of two equations with two unknowns that can be solved algebraically. Taking the first set with unknowns w and y, we obtain,

$$w = \frac{d}{ad - bc}, \tag{2-36a}$$
$$y = \frac{-c}{ad - bc}, \tag{2-36b}$$

and for the second set with unknowns x and z give,

$$x = \frac{-c}{ad - bc}, \tag{2-36c}$$
$$z = \frac{a}{ad - bc}. \tag{2-36d}$$

Based on this result, we can write the following solution for the inverse of a 2×2 matrix:

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix}, \quad \mathbf{A}^{-1} = \frac{1}{a_{1,1}a_{2,2} - a_{1,2}a_{2,1}} \begin{bmatrix} a_{2,2} & -a_{1,2} \\ -a_{2,1} & a_{1,1} \end{bmatrix}. \tag{2-37}$$

In other words, the diagonal elements are flipped, the off-diagonal elements are given a minus sign, and then the result is divided by the determinant of the original matrix. The other simple case that can be obtained directly using the direct solution approach is the inverse of a diagonal matrix:

$$\mathbf{D} = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_N \end{bmatrix}, \quad \mathbf{D}^{-1} = \begin{bmatrix} \frac{1}{d_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{d_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{d_N} \end{bmatrix}. \tag{2-38}$$

That is, the inverse of a diagonal matrix has all of its elements being the reciprocal.

The inverse of an arbitrary square matrix $\mathbf{A}$ can be obtained the expression:

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}\mathbf{C}^{\top}, \tag{2-39}$$

provided $\det(\mathbf{A}) \neq 0$, where $\mathbf{C}$ is called the cofactor matrix (its transpose, as in the above expression, is called the adjugate matrix). The elements of the cofactor matrix are the minors $M_{i,j}$ defined for taking the determinant (see Sec. 2.1.g) times a factor of $(-1)^{i+j}$. The cofactor matrix can be expressed as

$$\mathbf{C} = \begin{bmatrix} +M_{1,1} & -M_{1,2} & +M_{1,3} & \cdots & (-1)^{1+N}M_{1,N} \\ -M_{2,1} & +M_{2,2} & -M_{2,3} & \cdots & (-1)^{2+N}M_{2,N} \\ +M_{3,1} & -M_{3,2} & +M_{3,3} & \cdots & (-1)^{3+N}N_{3,N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (-1)^{N+1}M_{N,1} & (-1)^{N+2}M_{N,2} & (-1)^{N+3}M_{N,3} & \cdots & +M_{N,N} \end{bmatrix}. \tag{2-40}$$

It is not too difficult to verify that the solutions for the inverses of 2×2 and diagonal matrices are special cases of this result.

While this equation provides a compact definition of the matrix inverse, it requires computation of the cofactor matrix, which involves the computation of numerous determinants. This is extremely tedious for even modestly-sized matrices and therefore this approach is not used much in practice. Rather, an algorithm computing the inverse of a matrix via a technique called Gaussian elimination is used to solve the equation $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$. This is discussed in Sec. 2.5.f.

The matrix inverse has several useful properties. First, the inverse of the inverse matrix is the original matrix:

$$(\mathbf{A}^{-1})^{-1} = \mathbf{A}. \tag{2-41}$$

Also, the following scaling property is satisfied:

$$(r\mathbf{A})^{-1} = \frac{1}{r}\mathbf{A}^{-1}. \tag{2-42}$$

The inverse of a product of matrices is the product of the inverse matrices in the reverse order:

$$(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}. \tag{2-43}$$

The inverse of the transpose of the matrix is the transpose of the inverse matrix:

$$\left(\mathbf{A}^{-1}\right)^{\top} = \left(\mathbf{A}^{\top}\right)^{-1}. \tag{2-44}$$

The determinant of the inverse of a matrix is one over the determinant of the matrix:

$$\det\left(\mathbf{A}^{-1}\right) = \frac{1}{\det(\mathbf{A})}. \tag{2-45}$$

There is also another important special case where all the rows and columns of a square matrix are orthogonal to each other, i.e., when the sum of the products of the respective elements of all the different rows or columns of a matrix are zero (see Sec. 2.2.d). If this is true, then the inverse is equal to the transpose divided by the determinant of the matrix:

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}\mathbf{A}^{\top}, \quad \text{if the rows and columns of } \mathbf{A} \text{ are orthogonal.} \tag{2-46}$$

A matrix is said to be an *orthogonal matrix* if the inverse of the matrix is equal to its transpose, which also requires the determinant to be one in addition to what was stated above. This case arises in transformations to and from orthogonal coordinate systems. This includes all rotations as well as transformations between Cartesian, cylindrical, and spherical coordinates. (In all of these listed cases, the systems are also orthonormal, so the determinant is equal to one.) Because of this, once the transformation matrix is known in one direction, the inverse transformation in the opposite direction is simple to calculate.

In the case where the elements are complex, a matrix is said to be *unitary* if the matrix times its conjugate transpose is equal to the identity matrix:

$$\mathbf{U}\mathbf{U}^{*} = \mathbf{U}^{*}\mathbf{U} = \mathbf{I}. \tag{2-47}$$

In other words, the conjugate transpose is its own inverse. This is a extension of the notion of an orthogonal matrix to complex elements. Note that a real orthogonal matrix is also a unitary matrix. However, a matrix with complex elements can be orthogonal but not unitary, since the transpose is not the same as the conjugate transpose. The importance of unitary matrices in quantum mechanics is that it ensures that under any transformation, the probability

2.1.i Matrix Trace

Another operation that sometimes arises is taking the trace of a square matrix. The trace is simply the sum of the diagonal elements:

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{N} a_{ii}. \tag{2-48}$$

An important property is that the trace of a matrix is equal to the sum of its eigenvalues, which will be discussed later in this chapter.

2.2 Vectors

A commonly used mathematical object used to describe physical quantities such as force, momentum, electric and magnetic fields, etc. is a vector. In 3-D Cartesian coordinates, vectors are can be written in the form:

$$\mathbf{a} = a_x\hat{\mathbf{i}} + a_y\hat{\mathbf{j}} + a_z\hat{\mathbf{k}}. \tag{2-49}$$

This object can be represented as an arrow in three dimensional space with a_x units in the x direction, a_y units in the y direction, and a_z units in the z direction. Here $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, and $\hat{\mathbf{k}}$ represent unit vectors along the respective x, y, and z axes.

In general 3-D coordinates, we can write a vector as

$$\mathbf{a} = a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + a_3\mathbf{e}_3, \tag{2-50}$$

where $\mathbf{e}_i$ is a vector describing the coordinate system.

In the context of matrices, the vector $\mathbf{a}$ is represented as a $N\times 1$ vector or column vector. For 3-D Cartesian coordinates:

$$\begin{bmatrix} a_x \\ a_y \\ a_z \end{bmatrix} = a_x \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + a_y \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + a_z \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{2-51}$$

2.2.a Addition and Scaling of Vectors

Since vectors are a special case of a matrix, they satisfy this same addition and scaling rules. The addition (and subtraction) of two vectors is

$$\mathbf{a} + \mathbf{b} = (a_x + b_x)\hat{\mathbf{i}} + (a_y + b_y)\hat{\mathbf{j}} + (a_z + b_z)\hat{\mathbf{k}} \tag{2-52}$$

and the multiplication of a scalar r and a vector is

$$r\mathbf{a} = ra_x\hat{\mathbf{i}} + ra_y\hat{\mathbf{j}} + ra_z\hat{\mathbf{k}}. \tag{2-53}$$

2.2.b Magnitude and Unit Vector

A vector has a length known as is magnitude. Sometimes this is referred to as the norm, Euclidian norm, or L2-norm. The expression for computing the magnitude in *Cartesian coordinates* is

$$|\mathbf{a}| = \sqrt{a_x^2 + a_y^2 + a_z^2}. \tag{2-54}$$

Here $|\cdot|$ is an operator for computing the magnitude of the vector.

A unit vector is a vector with a magnitude of one and often denoted with a hat. Given a vector $\mathbf{a}$, the corresponding unit vector may be computed as

$$\hat{\mathbf{a}} = \frac{\mathbf{a}}{|\mathbf{a}|}. \tag{2-55}$$

2.2.c Dot (Inner) Product

A closely related concept to the magnitude is the *dot product*, which is also called the *inner product* (the dot product is a special case of the more general mathematical concept of the inner product). The dot product maps two vectors and returns a scalar. The dot product between two vectors is given as

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos\theta, \tag{2-56}$$

where θ is the angle between the two vectors. Note that the definition implies commutativity of the dot product:

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}. \tag{2-57}$$

To illustrate, suppose $\mathbf{a}$ and $\mathbf{b}$ are 2-D vectors in Cartesian coordinates. The dot product between them can be written as

$$\begin{aligned} \mathbf{a} \cdot \mathbf{b} &= (a_x\hat{\mathbf{i}} + a_y\hat{\mathbf{j}}) \cdot (b_x\hat{\mathbf{i}} + b_y\hat{\mathbf{j}}) \\ &= a_x b_x(\hat{\mathbf{i}} \cdot \hat{\mathbf{i}}) + a_x b_y(\hat{\mathbf{i}} \cdot \hat{\mathbf{j}}) + a_y b_x(\hat{\mathbf{j}} \cdot \hat{\mathbf{i}}) + a_y b_y(\hat{\mathbf{j}} \cdot \hat{\mathbf{j}}) \\ &= a_x b_x(\hat{\mathbf{i}} \cdot \hat{\mathbf{i}}) + (a_x b_y + a_y b_x)(\hat{\mathbf{i}} \cdot \hat{\mathbf{j}}) + a_y b_y(\hat{\mathbf{j}} \cdot \hat{\mathbf{j}}). \end{aligned} \tag{2-58}$$

The dot products of the unit vectors can be understood through applying the right hand side of Eq. (2-56). Since $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ are unit vectors, their magnitudes are one. Furthermore, the angle between a vector and itself is $0°$, which implies the dot product of the a unit vector with itself is one, i.e., $\hat{\mathbf{i}} \cdot \hat{\mathbf{i}} = 1, \hat{\mathbf{j}} \cdot \hat{\mathbf{j}} = 1$. Further, $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ are aligned along the x and y axes respectively, which are $90°$ apart. Since $\cos 90° = 0$, then $\hat{\mathbf{i}} \cdot \hat{\mathbf{j}} = 0$.

By extension to 3-D (or any other number of dimensions), in Cartesian coordinates, the dot product between two vectors can be written as

$$\mathbf{a} \cdot \mathbf{b} = a_x b_x + a_y b_y + a_z b_z. \tag{2-59}$$

Note that this expression does **not** hold in other coordinate systems, since the unit vectors may not cancel in the same way.

From a linear algebra perspective, the dot product may be written as

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^\top \mathbf{b} = \begin{bmatrix} a_x & a_y & a_z \end{bmatrix} \begin{bmatrix} b_x \\ b_y \\ b_z \end{bmatrix}. \tag{2-60}$$

Here $\mathbf{a}^\top$ is the transpose of the vector $\mathbf{a}$.

An operation related to the dot product is finding the projection of one vector onto other. The projection of a vector $\mathbf{v}$ onto $\mathbf{u}$ is given as

$$\text{proj}_{\mathbf{u}}(\mathbf{v}) = \left(\frac{\mathbf{u} \cdot \mathbf{v}}{\mathbf{u} \cdot \mathbf{u}}\right) \mathbf{u}. \tag{2-61}$$

The term in parentheses is a scaling factor denoting the length that $\mathbf{v}$ shadows a line through $\mathbf{u}$. The projection operator is most commonly used to generate a set of orthogonal vectors, which will be discussed later.

2.2.d Orthogonality and Orthornormality

Two different vectors are said to be orthogonal if the vectors are both of non-zero magnitude and their dot product is zero. This implies $\cos\theta = 0$, which occurs when $\theta = 90^\circ = {}^{\pi}/_{2}$ or $\theta = 270^\circ = {}^{3\pi}/_{2}$. (More generally, two different mathematical objects, e.g., functions, are orthogonal if their inner product is zero, which requires a definition of the inner product.) Given a set of orthogonal vectors $\mathbf{u}_i$, the following holds:

$$\mathbf{u}_i \cdot \mathbf{u}_j = c_i \delta_{ij} = \begin{cases} c_i & i = j \\ 0 & i \neq j \end{cases}, \tag{2-62}$$

with constant $c_i \neq 0$. Here δ_{ij} is called the Kronecker delta, which is one when $i = j$ and zero otherwise. The first condition for $i = j$ of this expression is a bit obvious, since the dot product of a vector with itself will always be positive (except if it is the zero vector). It is the second condition for $i \neq j$ that is not immediately apparent for an arbitrary set of vectors and is the one that should be focused on in demonstrating orthogonality and understanding the importance of this result.

Two different vectors are orthonormal if in addition to being orthogonal, they are both unit vectors (have a magnitude of one). This implies that $c_i = 1$ for all i if the vectors are orthonormal.

The set of basis vectors $\mathbf{e}_i$ of the standard coordinate systems, Cartesian, cylindrical, and spherical, are orthonormal. Most work in science and engineering is in these coordinate systems, but occasionally it is more convenient to work in a coordinate system that is not.

The more general definition of orthogonality involving functions is useful as well and will be revisited in solutions of partial differential equations. The short version is that it is sometimes easier to solve for the integrals of the product of some orthogonal functions with the function of interest than it is to solve for that function directly.

2.2.e Cross Product

Another common application in engineering applications is the *cross product*, which is denoted by $\mathbf{u} \times \mathbf{v}$. The cross product gives another vector $\mathbf{c} = \mathbf{a} \times \mathbf{b}$ that is perpendicular to $\mathbf{a}$ and $\mathbf{b}$.

Since in 3-D space there are two vectors that are 180° apart for each other and perpendicular to any given pair of vectors, we require a convention to pick which one. The typical convention to select this vector is the *right-hand rule*. Note that this rule implies that the if we switch the order of $\mathbf{a}$ and $\mathbf{b}$, we would pick the other vector, therefore the cross product is anti-commutative, i.e.,

$$\mathbf{a} \times \mathbf{b} = -(\mathbf{b} \times \mathbf{a}). \tag{2-63}$$

Another important observation about the cross product is the it has the unusual property of being non-associative:

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) \neq (\mathbf{a} \times \mathbf{b}) \times \mathbf{c}. \tag{2-64}$$

Computing the cross product in Cartesian coordinates involves taking the 3×3 determinant. This proceeds as follows:

$$\begin{aligned}\mathbf{a}\times\mathbf{b} &= \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ a_x & a_y & a_z \\ b_x & b_y & b_z \end{vmatrix} \\ &= \hat{\mathbf{i}}\begin{vmatrix} a_y & a_z \\ b_y & b_z \end{vmatrix} - \hat{\mathbf{j}}\begin{vmatrix} a_x & a_z \\ b_x & b_z \end{vmatrix} + \hat{\mathbf{k}}\begin{vmatrix} a_x & a_y \\ b_x & b_y \end{vmatrix} \\ &= (a_y b_z - a_z b_y)\hat{\mathbf{i}} - (a_x b_z - a_z b_x)\hat{\mathbf{j}} + (a_x b_y - a_z b_y)\hat{\mathbf{k}}. \end{aligned} \tag{2-65}$$

The determinant expression does not hold in non-Cartesian coordinate systems.

The cross product can also be written similarly to the dot product as

$$\mathbf{a}\times\mathbf{b} = |\mathbf{a}||\mathbf{b}|\sin\theta\,\hat{\mathbf{n}}. \tag{2-66}$$

This expression for the cross product differs for the dot product of Eq. (2-56) in two ways. First, instead of the cosine of the angle between the vector, the equation uses the sine. Secondly, the expression includes some unit vector $\hat{\mathbf{n}}$, which is a unit vector perpendicular to $\mathbf{a}$ and $\mathbf{b}$.

Equation (2-66) provides a geometric interpretation of the cross product. By taking the magnitude of the cross product, $|\mathbf{a}\times\mathbf{b}|$, the right hand side becomes the area of the parallelogram formed by $\mathbf{a}$ and $\mathbf{b}$.

2.2.f Tensor (Outer) Product

Thus far, we have studies two types of multiplication. The dot product takes two vectors and produces a scalar. The cross product takes two vectors and produces another vector perpendicular to them. There is a third type of product that produces a matrix or bivector. This is called the *outer product*. The outer product of two vectors is denoted by

$$\mathbf{a}\otimes\mathbf{b} = \mathbf{a}\mathbf{b}^{\top}. \tag{2-67}$$

In the context of equations where all vectors are exclusively column vectors (common in mathematical physiscs), the product of two vectors is sometimes defined equivalently

$$\mathbf{a}\mathbf{b} \equiv \mathbf{a}\otimes\mathbf{b}. \tag{2-68}$$

This notation is called the dyadic form.

The left hand side is the product of two column vectors from linear algebra, giving a matrix. The outer product takes all combinations of component products of the two vectors and places them in the matrix. To illustrate, consider the unit vectors in 2-D space:

$$\hat{\mathbf{i}}\otimes\hat{\mathbf{i}} = \hat{\mathbf{i}}\hat{\mathbf{i}}^{\top} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}\begin{bmatrix} 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \tag{2-69a}$$

$$\hat{\mathbf{i}} \otimes \hat{\mathbf{j}} = \hat{\mathbf{i}}\hat{\mathbf{j}}^\top = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \tag{2-69b}$$

$$\hat{\mathbf{j}} \otimes \hat{\mathbf{i}} = \hat{\mathbf{j}}\hat{\mathbf{i}}^\top = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \tag{2-69c}$$

$$\hat{\mathbf{j}} \otimes \hat{\mathbf{j}} = \hat{\mathbf{j}}\hat{\mathbf{j}}^\top = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}. \tag{2-69d}$$

Therefore, the outer product of $\mathbf{a}$ and $\mathbf{b}$ is

$$\begin{aligned} \mathbf{a} \otimes \mathbf{b} &= (a_x\hat{\mathbf{i}} + a_y\hat{\mathbf{j}}) \otimes (b_x\hat{\mathbf{i}} + b_y\hat{\mathbf{j}}) \\ &= a_x b_x(\hat{\mathbf{i}} \otimes \hat{\mathbf{i}}) + a_x b_y(\hat{\mathbf{i}} \otimes \hat{\mathbf{j}}) + a_y b_x(\hat{\mathbf{j}} \otimes \hat{\mathbf{i}}) + a_y b_y(\hat{\mathbf{j}} \otimes \hat{\mathbf{j}}) \\ &= a_x b_x \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + a_x b_y \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + a_y b_x \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} + a_y b_y \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} a_x b_x & a_x b_y \\ a_y b_x & a_y b_y \end{bmatrix}. \end{aligned} \tag{2-70}$$

The generalization to 3-D follows the same procedure and yields a 3×3 matrix.

Similar to the cross product, the outer product is not commutative. Switching the order of the outer product results in the transpose of the matrix:

$$\mathbf{a} \otimes \mathbf{b} = (\mathbf{b} \otimes \mathbf{a})^\top. \tag{2-71}$$

The outer product (unlike the cross product) is associative:

$$\mathbf{a} \otimes (\mathbf{b} \otimes \mathbf{c}) = (\mathbf{a} \otimes \mathbf{b}) \otimes \mathbf{c}. \tag{2-72}$$

The dot product of the a vector with the outer product of two other vectors is a vector. This arises with relative frequency in applications of fluid dynamics. The following simplification is often used based on associative rules:

$$\mathbf{a} \cdot (\mathbf{b} \otimes \mathbf{c}) = (\mathbf{a} \cdot \mathbf{b})\mathbf{c}. \tag{2-73}$$

Recall the dot product is a scalar, so the overall expression is a vector.

2.2.g Scalar Triple Product

The dot and cross product can be mixed to calculate the volume of a three-dimensional box with arbitrarily aligned axes called a parallelepiped. If we have three vectors defining the parallelepiped, then the signed volume can be calculated using the scalar triple product:

$$V = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}). \tag{2-74}$$

The sign of V depends upon the orientation of the vectors, and the volume is simply the absolute value.

The interpretation of the scalar triple product as the volume is illustrated in Fig. 2.2. The vectors $\mathbf{u}$ and $\mathbf{v}$ reside in the x-y plane and form a parallelogram

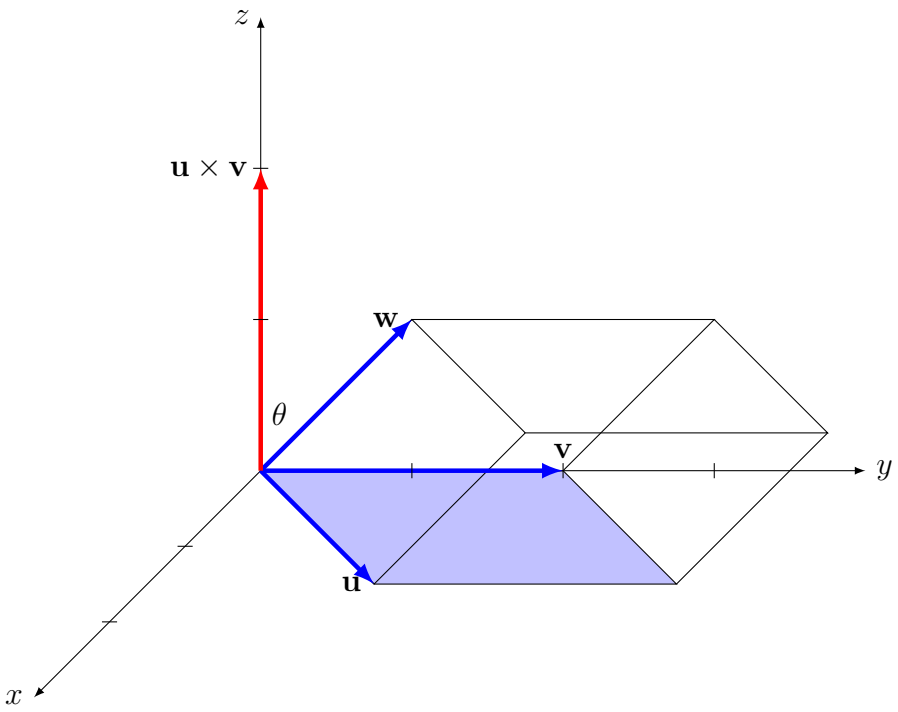

Figure 2.2: Illustration of the scalar triple product as a computation of the volume of a parallelepided.

(shaded in blue in the figure). The cross product between $\mathbf{u}$ and $\mathbf{v}$ (red vector) points in the z direction with a magnitude corresponding to the area of this parallelogram. The vector $\mathbf{w}$ is in the y-z plane at angle θ from the vector $\mathbf{u} \times \mathbf{v}$. Recall the dot product of these two can be expressed as is

$$\mathbf{w} \cdot (\mathbf{u} \times \mathbf{v}) = |\mathbf{u} \times \mathbf{v}||\mathbf{w}| \cos\theta. \tag{2-75}$$

Again, the magnitude of $\mathbf{u} \times \mathbf{v}$ is the area of the base of the parallelepidped. The quantity $|\mathbf{w}| \cos\theta$ is the magnitude of the projection onto the z-axis (or $\mathbf{u} \times \mathbf{v}$) and is the height with respect to the base. The product of this area and height gives the volume.

In Cartesian coordinates, the scalar triple product can be computed simply by taking the determinant of a 3×3 matrix where either the rows or columns correspond to the vector components. To see this we expand out the determinant for $\mathbf{b} \times \mathbf{c}$:

$$\begin{aligned}
\mathbf{b} \times \mathbf{c} &= \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ b_x & b_y & b_z \\ c_x & c_y & c_z \end{vmatrix} \\
&= \hat{\mathbf{i}} \begin{vmatrix} b_y & b_z \\ c_y & c_z \end{vmatrix} - \hat{\mathbf{j}} \begin{vmatrix} b_x & b_z \\ c_x & c_z \end{vmatrix} + \hat{\mathbf{k}} \begin{vmatrix} b_x & b_y \\ c_x & c_y \end{vmatrix}.
\end{aligned}$$

Taking the dot product with $\mathbf{a}$ leads to the unit basis vectors being replaced the the

components

$$\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c}) = a_x\begin{vmatrix} b_y & b_z \\ c_y & c_z \end{vmatrix} - a_y\begin{vmatrix} b_x & b_z \\ c_x & c_z \end{vmatrix} + a_z\begin{vmatrix} b_x & b_y \\ c_x & c_y \end{vmatrix}.$$

Therefore, rewriting the three minor determinants of 2×2 matrices as a single 3×3 determinant gives the result:

$$\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c}) = \begin{vmatrix} a_x & a_y & a_z \\ b_x & b_y & b_z \\ c_x & c_y & c_z \end{vmatrix}. \tag{2-76}$$

Again, it is important to note that the above expression is only valid in Cartesian coordinates.

It is not difficult to show from the determinant form, that the scalar triple product can be reordered using a circular shift, such that

$$\mathbf{a}\cdot(\mathbf{b}\times\mathbf{c}) = \mathbf{c}\cdot(\mathbf{a}\times\mathbf{b}) = \mathbf{b}\cdot(\mathbf{c}\times\mathbf{a}). \tag{2-77}$$

A circular shift is pushing each vector forward one position in the expression and moving the entry on the end to the beginning. The cross product of any of these can be reordered with a minus sign given the anticommutativity of the cross product.

2.2.h Vector Triple Product

Given the scalar triple product defined with the dot product of a cross product, it may be instructive to consider taking the cross product of another cross product. (Taking the cross product of a dot product does not make sense, or a dot product of a dot product does not make sense either if considering vectors.) This arises in physics when, for example, calculating centrifugal forces in a rotating coordinate system and occurs frequently in electrodynamics. The vector triple product produces another vector and is

$$\mathbf{v} = \mathbf{a}\times(\mathbf{b}\times\mathbf{c}). \tag{2-78}$$

The computation of this can be a bit unwieldy to work out, but there is a very important vector identity to recast this vector product in terms of much simpler dot products. This identity is referred to as the "BAC-CAB rule", which is a useful nemonic:

$$\mathbf{a}\times(\mathbf{b}\times\mathbf{c}) = \mathbf{b}\,(\mathbf{a}\cdot\mathbf{c}) - \mathbf{c}\,(\mathbf{a}\cdot\mathbf{b})\,. \tag{2-79}$$

Note the ordering has the vector out front of the scalar in the nemonic. This identity is frequently used to simplify complex vector equations and makes the numerical computation much simpler. Unlike the scalar triple product, which has a geometric interpretation of computing a volume, there is not an analogous physical interpretation.

2.3 Covectors

If ordinary vectors are matrix equivalents of column vectors, it may be natural to ask if there is an interpretation of row vectors. Indeed there is, and this is called a covector. (In many contexts this is referred to as a linear form or an algebraic 1-form.)

To illustrate the concept, let us consider the 2-D row vector

$$\boldsymbol{\alpha} = \begin{bmatrix} 1 & -1/2 \end{bmatrix}.$$

Let this row vector act upon a column vector as follows to yield an equation

$$\begin{bmatrix} 1 & -1/2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = x - \frac{y}{2} = r,$$

here r is some scalar value that is the numerical result for a given values of (x, y). This equation denotes the equation for the line $y = 2x - 2r$. If we plot a series of lines for various values of r, we have the geometric representation of the covector (see Fig. 2.3). This can be thought of as a series of contour lines with a provided orientation given by the small black arrows normal to those lines that point toward larger values of r. In 3-D the covector is represented graphically a stack of planes with a provided orientation. Covectors may also be represented in a similar manner in higher dimensions, but this is harder to visualize.

Similar to vectors, we can express covectors in terms of basis covectors:

$$\boldsymbol{\alpha} = \alpha_1 \boldsymbol{\epsilon}_1 + \alpha_2 \boldsymbol{\epsilon}_2 + \dots \tag{2-80}$$

In 2-D Cartesian coordinates the basis covectors $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2$ are $\mathbf{e}_1^\top = \hat{\mathbf{i}}^\top$ and $\mathbf{e}_2^\top = \hat{\mathbf{j}}^\top$. These represent the lines $x = r$ and $y = r$ respectively for given parameter r with orientation pointing toward the positive x and y directions.

2.3.a Covectors and Matrix Multiplication

Covector-vector multiplication gives us a second geometric interpretation of matrix multiplication (the first being a linear map).

A covector may operate on a vector to produce a number in the same manner as matrix multiplication or the dot product as follows:

$$\boldsymbol{\alpha}(\mathbf{v}) = \sum_i \alpha_i v_i. \tag{2-81}$$

More interesting is the geometric interpretation. Suppose now that our former 2-D covector

$$\boldsymbol{\alpha} = \begin{bmatrix} 1 & -1/2 \end{bmatrix}$$

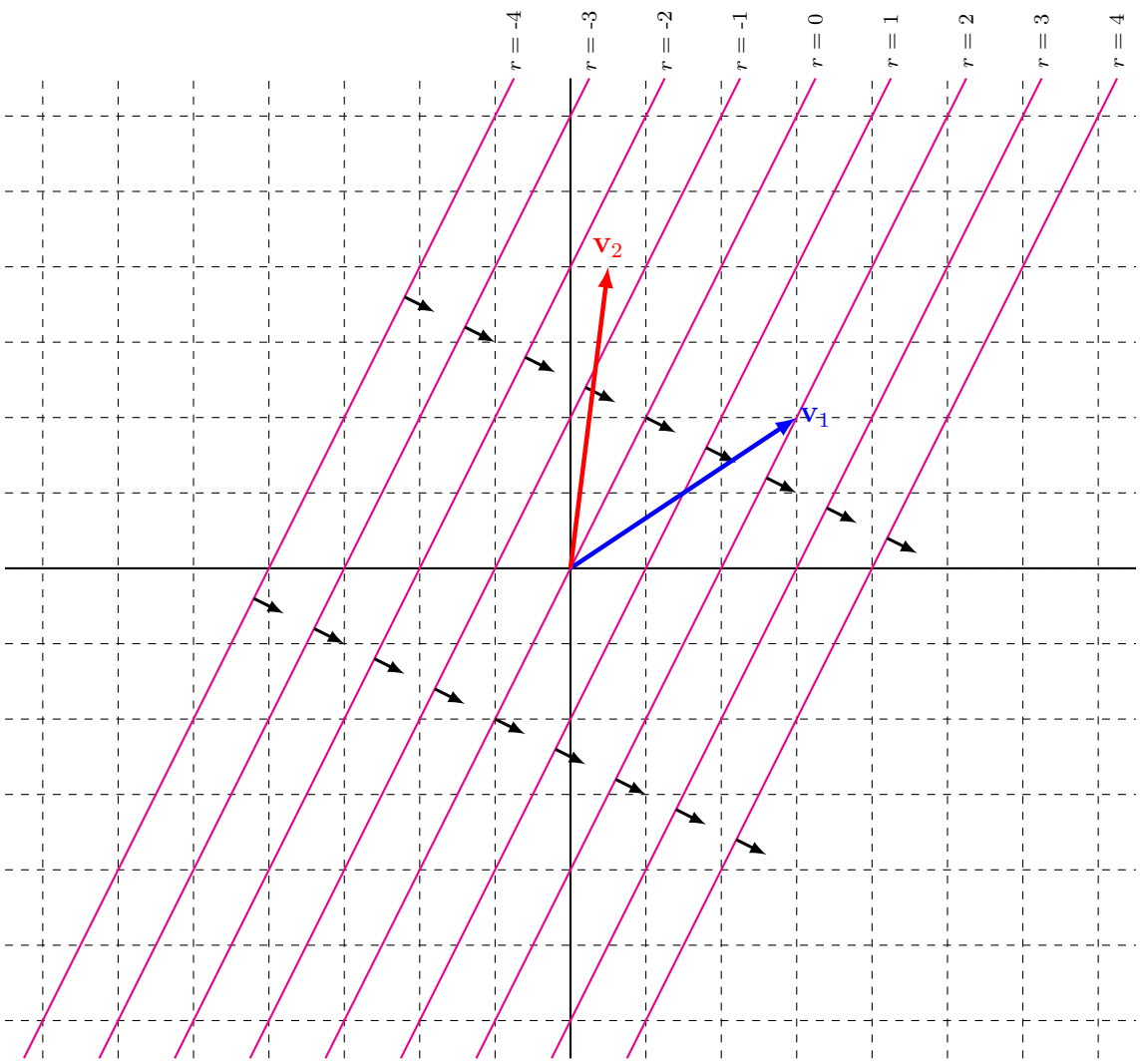

Figure 2.3: Illustration of vectors and a covector.

acts on a vector

$$\mathbf{v}_1 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}.$$

The result of $\boldsymbol{\alpha}(\mathbf{v}_1)$ is 2. This is plotted graphically in Fig. 2.3 in blue. We could think of the lines formed by a covector as contour lines on a physical map, which describe lines of constant elevation. The covector-vector multiplication can therefore be thought of as a change in elevation or a kind of signed distance measured with respect to the covector planes. In this specific case, the "distance" measured is 2 because the vector travels right up to the second covector line. Also, the length is positive the directionality with respect to the covector (going in the same direction

as the small black arrows). Considering another vector

$$\mathbf{v}_2 = \begin{bmatrix} 1/2 \\ 4 \end{bmatrix},$$

$\boldsymbol{\alpha}(\mathbf{v}_2)$ is $-3/2$, which is plotted in Fig. 2.3 in red. This means the vector travels a "distance" of $3/2$ with respect to the covector—the vector crosses one line and gets halfway to the next—but in the negative direction since the vector is crossing the lines in the opposite direction of their orientation. Note that the length is negative despite both components of $\mathbf{v}_2$ being positive; the sign is with respect to the orientation of the covector.

The second geometric interpretation of matrix multiplication $\mathbf{C} = \mathbf{AB}$ is that each element in the resultant matrix $c_{i,j}$ is a signed "distance" of a vector defined by column j of $\mathbf{B}$ measured with respect to the planes of the covector defined by row i of $\mathbf{A}$. More precisely "signed distance" can be thought of as a change in some quantity in a field, e.g., the work done in the presence of a force field.

2.4 Coordinate Systems

Engineering calculations require making a choice of coordinate system. Most often, the Cartesian coordinate system is sufficient. Other times, we must work in other common coordinate systems such as cylindrical or spherical coordinates. Occasionally, we will need to work in other nonstandard coordinate systems as well. Coordinate systems are described by a set of vectors called basis vectors, which can be thought of as the fundamental building blocks that can be used to construct any vector in that coordinate system. Moving from one coordinate system to another is referred to as a change of basis. The notion of a basis can be extended to not just include vectors, but can also include any object such as polynomials or matrices. This section will review the concepts of linear independence, span, a basis, and then transformations for converting between different coordinate systems.

2.4.a Linear Independence

Coordinate systems are defined by a set of linearly independent vectors called basis vectors. For this reason, we need to define the concept of linear independence. Suppose we have a set of vectors $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_N\}$. We say that some vector in the set $\mathbf{v}_k$ for $k = 1, \ldots, N$ is linearly independent of the others if and only if $\mathbf{v}_k$ cannot be expressed as a linear combination of the others. Conversely, a vector is linearly dependent (not linearly independent) if

$$\mathbf{v}_k = \sum_{\substack{i=1 \\ i \neq k}}^{N} a_i \mathbf{v}_i, \tag{2-82}$$

for some choice of nonzero coefficients a_i.

For example, consider the set of two-dimensional vectors

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}. \tag{2-83}$$

It is fairly evident that this set of vectors is linearly dependent. To show this, we can find a counterexample. We express $\mathbf{v}_3 = 3\mathbf{v}_1 - \mathbf{v}_2$ or

$$3\begin{bmatrix} 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 3 \\ 0 \end{bmatrix} + \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \tag{2-84}$$

Note that if we pick any two of the vectors from this set, those vectors are linearly independent. This is very simple to show since there is no single scalar that we can multiply by one of the vectors to get the other.

An important observation is that if the vectors in the set are N dimensional, then if there are more than N vectors in the set, the set cannot be linearly independent. The converse is not true. Just because there are N or fewer N-dimensional vectors in the set, that does not imply the vectors are linearly independent.

For another example, consider the set of vectors

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}. \tag{2-85}$$

We can show that $\mathbf{v}_1$ cannot be written as a linear combination of $\mathbf{v}_2$ and $\mathbf{v}_3$. Let us suppose that there exists nonzero coefficients a_1 and a_2 such that

$$\begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} \stackrel{?}{=} a_1 \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} + a_2 \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}.$$

This forms the linear system of equations

$$a_1 = 1, \quad a_2 - a_1 = 0, \quad a_2 = 2,$$

The first and third equations are inconsistent with the second, since $2 - 1 \neq 0$. Therefore, $\mathbf{v}_1$ is linearly independent of $\mathbf{v}_2$ and $\mathbf{v}_3$.

A test for whether a set of N vectors each having a length N is linearly independent is to check whether the determinant of a matrix having the columns be given by those vectors is nonzero. Using the previous example

$$\begin{vmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \\ 2 & 0 & 1 \end{vmatrix} = (1)(-1-0) - (1)(0-2) + (0)(-1-0) = -1 + 2 = 1.$$

By this test, the vectors are linearly independent.

A very similar, but non-trivial case where the vectors are linearly dependent is

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}. \tag{2-86}$$

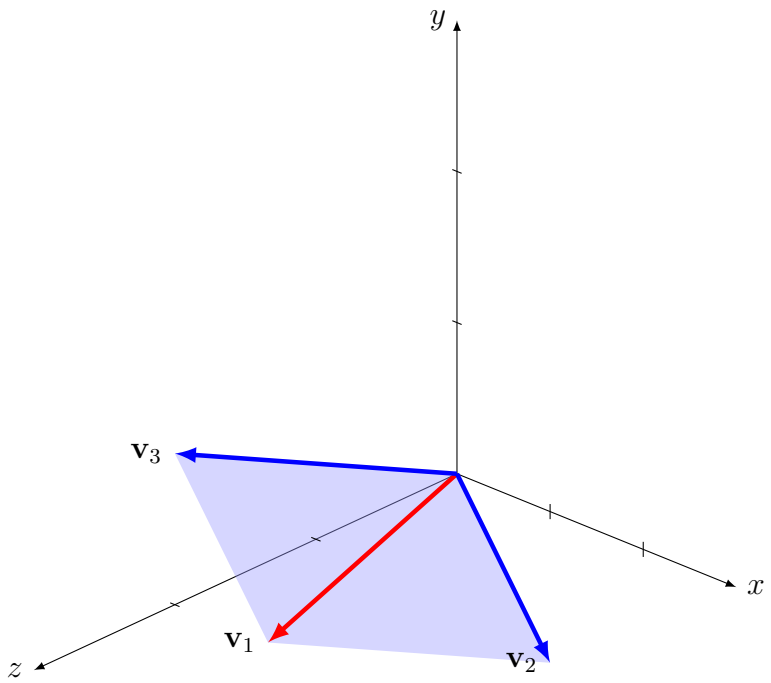

Figure 2.4: Example of linearly dependent vectors where $\mathbf{v}_1$ is on the plane formed by $\mathbf{v}_2$ and $\mathbf{v}_3$.

The only difference from the previous example is the vector $\mathbf{v}_3$. This can be simply shown by forming the linear system as before

$$a_1 = 1, \quad a_2 - a_1 = 0, \quad 2a_2 = 2.$$

This time $a_1 = 1$ and $a_2 = 1$, so the second equation is satisfied. Therefore, the system is linearly dependent. Alternatively, we can compute the determinant

$$\begin{vmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \\ 2 & 0 & 2 \end{vmatrix} = (1)(-2-0) - (1)(0-2) + (0)(-1-0) = -2+2 = 0.$$

Before proceeding, let us consider the geometric interpretation by way of looking at the example where the vectors are linearly dependent. The vectors $\mathbf{v}_2$ and $\mathbf{v}_3$ can be used to form a plane in 3-D space. The vector $\mathbf{v}_1$ is within this plane. This is illustrated in Fig. 2.4. Geometrically, three vectors that are linearly independent do not exist in the same plane.

2.4.b Basis and Span

A basis can be thought of as the fundamental linearly independent building blocks (e.g., vectors) that can be used to construct a particular class of objects by way of a linear combination. The span is the smallest set of these vectors that can be used to create the class.

The most common example encountered in engineering applications is the set of all real numbers in 3-D space, denoted by $\mathbb{R}^3$. This can be described by the unit basis vectors $\{\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}\}$. Equivalently all 3-D vectors can be defined by

$$\text{span}\left\{\begin{bmatrix}1\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\end{bmatrix}, \begin{bmatrix}0\\0\\1\end{bmatrix}\right\} = a_1\begin{bmatrix}1\\0\\0\end{bmatrix} + a_2\begin{bmatrix}0\\1\\0\end{bmatrix} + a_3\begin{bmatrix}0\\0\\1\end{bmatrix}. \tag{2-87}$$

Note that the unit basis vectors are a convenient choice for the span of $\mathbb{R}^3$. In reality, any three linearly independent 3-dimensional vectors would suffice.

Another example is the space of all vectors that live in a plane in 3-D space. Considering the vectors from Eq. (2-86), we can compute the equation for the plane that intersects the origin by taking the cross product of two of the vectors with the coefficients being the components of that vector:

$$2x + 2y - z = 0.$$

The span of all vectors on this plane consists of any two linearly independent vectors. For instance, the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ satisfy this criteria:

$$\text{span}\left\{\begin{bmatrix}1\\0\\2\end{bmatrix}, \begin{bmatrix}1\\-1\\0\end{bmatrix}\right\} = a_1\begin{bmatrix}1\\0\\2\end{bmatrix} + a_2\begin{bmatrix}1\\-1\\0\end{bmatrix}. \tag{2-88}$$

Any other pair of linearly independent vectors on the plane would suffice as arguments of the span.

The notion of linear independence and span is connected to geometry and vector calculus as well. For now, let us consider a curve C described by a function $f(x)$. The set of all points along that curve can be described by the pair $(a, f(a))$. We often consider the set of all vectors tangent to the curve at some point p in a local coordinate system with the origin at p. This is called the tangent space of curve C at point p denoted by T_pC.

Figure 2.5 gives an illustration. The curve C defined by $f(x)$ is in blue. Suppose the point p is at $x = a, y = f(a)$. We can draw a line in red that is tangent to the curve at $x = a$. Any vector tangent to the curve at $(a, f(a))$ along that line is in the tangent space T_pC. Writing this vector component wise can be done by looking at the shaded triangle. Suppose the x component is set to one, then the y component is then the slope, which is obtained from taking the first derivative of $f(x)$ and evaluating it at $x = a$. We can then write the span of T_pC as having a single 2-D vector

$$\text{span}\left\{\begin{bmatrix}1\\f'(a)\end{bmatrix}\right\} = c\begin{bmatrix}1\\f'(a)\end{bmatrix}, \tag{2-89}$$

where c is an arbitrary nonzero scalar and $f'(a)$ is the derivative evaluated at $x = a$.

Note that the notion of a tangent space generalizes to higher dimensions. For example, if we have some arbitrary curved 2-D surface in 3-D space, at any point p, we can define a plane that is tangent with respect to the surface at that point. The

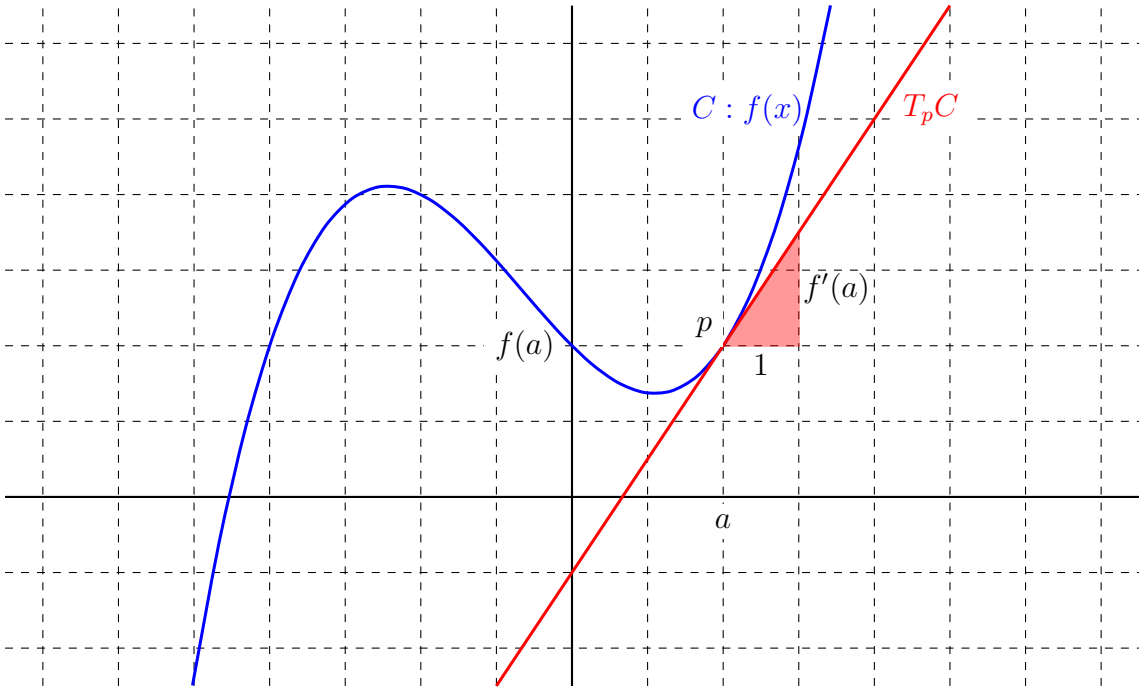

Figure 2.5: Illustration of the tangent space of a curve.

basis of the tangent space in this context contains two linearly independent vectors that reside on the plane. This notion will become important when we study vector calculus.

From these three examples, there is an observation that can be made regarding $\mathbb{R}^3$ a three-dimensional vectors: A span of three linearly independent vectors gives all possible vectors in three-dimensional space; a span of two linearly independent vectors gives the space of all vectors within a plane within $\mathbb{R}^3$; and a span of one linearly independent vector gives the space of all vectors along a line in $\mathbb{R}^3$. These conclusions can be extended to an arbitrary number of dimensions. For example, if we are in $\mathbb{R}^n$, then having a span of n linearly independent vectors gives all vectors in n-dimensional space.

These examples described the basis in terms of vectors. The notion of a basis is very general. For example, polynomials can be described using a basis of monomials x^k. For example, all quadratic polynomials are spanned by

$$\text{span}\left\{1, x, x^2\right\} = a_0 + a_1 x + a_2 x^2. \tag{2-90}$$

2.4.c Example: Quantum Spin and the SU(2) Group

An example that of a basis that is important in quantum mechanics and the description of spin in an electromagnetic field. This is called the *Special Unity Group* of degree 2 or SU(2). This describes all 2×2 matrices that are unitary (all matrices where the inverse is its conjugate transpose) and have a sum along the diagonal or trace of zero.

Similar to SU(2), we can describe the space of all 2×2 Hermitian (not necessarily unitary) matrices with zero trace as a linear combination of three 2×2 matrices called the Pauli matrices. Here the Pauli matrices are the basis objects forming the group. The span of these matrices is

$$\begin{aligned}
&\text{span}\left\{\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\right\} \\
&= a_1 \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} + a_2 \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} + a_3 \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \\
&= \begin{bmatrix} a_3 & a_1 - a_2 i \\ a_1 + a_2 i & -a_3 \end{bmatrix}. \qquad \text{(2-91)}
\end{aligned}$$

Here the coefficients a_k are real numbers and $i = \sqrt{-1}$, the imaginary unit. (We call the set of numbers that the coefficients may have the base field.) The resulting matrix is Hermitian because it is equal to its conjugate transpose. Its trace, or sum of the diagonals, is zero since $a_3 - a_3 = 0$.

The resulting matrix times its conjugate transpose is the sum of the squares of the coefficients times the identity matrix:

$$\begin{bmatrix} a_3 & a_1 - a_2 i \\ a_1 + a_2 i & -a_3 \end{bmatrix} \begin{bmatrix} a_3 & a_1 - a_2 i \\ a_1 + a_2 i & -a_3 \end{bmatrix}^* = (a_1^2 + a_2^2 + a_3^2) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \qquad \text{(2-92)}$$

The matrix is unitary when $a_1^2 + a_2^2 + a_3^2 = 1$. This implies the base field of SU(2) are components of all 3-D real vectors with unit magnitude, which can be geometrically thought of as all vectors that point from the origin to the surface of the unit sphere.

2.4.d Transformation of Vectors and Covectors

As we saw previously with linear maps, the act of performing matrix multiplication maps one set of vectors to another set. We define a matrix $\mathbf{T}$ as the *forward transformation matrix* that is a tool that can be used to convert basis vectors from one coordinate system to another. Recall that a coordinate system is defined by a set of linearly independent vectors called basis vectors that spans some space (e.g., all real numbers in 3-D space or $\mathbb{R}^3$). The forward transformation matrix $\mathbf{T}$ is constructed by creating a matrix where its columns are the basis vectors of the new coordinate system.

Note that if we multiply $\mathbf{T}$ onto a matrix containing the standard Cartesian basis vectors $\{\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}\}$, which is the identity matrix, then this maps these basis vectors to a new set of basis vectors.

A very simple example is a unit conversion. Suppose in our standard 3-D Cartesian coordinate system we define 1 unit of measurement in each direction to be a foot. Rather, we wish our measurement to be in inches, i.e., in the new coordinate system 1 unit corresponds to 1 inch. To do this, we would need to shorten each basis vector by a factor of 12, since there are 12 inches per foot. The matrix that would do this is

a simple scaling on the identity matrix such that all the diagonal elements are 1/12 and the off-diagonal elements are zero:

$$\mathbf{T} = \begin{bmatrix} {}^{1}\!/_{12} & 0 & 0 \\ 0 & {}^{1}\!/_{12} & 0 \\ 0 & 0 & {}^{1}\!/_{12} \end{bmatrix}.$$

Let us again emphasize that the matrix $\mathbf{T}$ is called the forward transformation matrix and its purpose is to change the *basis vectors* defining one coordinate system to a set defining another.

We define the forward transformation to be *covariant* with respect to the basis vectors. Said the other way around, the basis vectors change along with (or covary with) the transformation.

A change in the basis or coordinate system does not change the vectors in the sense that we would not expect a purely mathematical operation to change the actual distances, magnitude of the forces, intensities or electric or magnetic fields, etc. If the vector is to remain the same in this sense, then when we change the basis vectors via the forward transformation, we need to adjust the vector components in the precise manner to accomplish this.

Revisiting the example of the conversion from feet to inches. Suppose we have a vector that was 1/2 foot in the x direction, 1 foot in the y direction, and 0 feet in the z direction. Upon doing the change of basis that does the unit conversion by changing the meaning of 1 unit of measurement from a foot to an inch, we would need to convert the vector components from feet to inches. In the new coordinate system, the x component would go from 1/2 foot to 6 inches (or 6 units in the new coordinate system), the y component would go from 1 foot to 12 inches, and the z component would remain at zero. The matrix that accomplishes this via matrix multiplication is a diagonal matrix where all the elements are 12. Note that this is precisely the inverse of $\mathbf{T}$. Therefore, we define a new matrix called the *backward transformation matrix*; which, for our example is

$$\mathbf{T}^{-1} = \begin{bmatrix} 12 & 0 & 0 \\ 0 & 12 & 0 \\ 0 & 0 & 12 \end{bmatrix}.$$

The notion of the backwards transformation matrix for the vector components as the inverse of the forward transformation extends to more complicated coordinate transformations including rotations and skewing.

It bares repeating the key point that a vector is fundamentally unchanged by a change of basis or coordinate transformation. How we measure the vector (i.e., the components) changes to account for the change in basis. The vector components change using the backwards transformation and change against how the basis vectors change to preserve the overall meaning of the vector. We say that the vector components are *contravariant* with respect to the change of basis.

Like vectors, the covectors are fundamentally unchanged by a change of basis, only the covector components. To figure out how, we note that in Cartesian coordinates,

the basis covector operating on the basis vector gives

$$\boldsymbol{\epsilon}_j(\mathbf{e}_i) = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} . \tag{2-93}$$

Since a change of basis does not impact the vector or covector, merely how they are represented, then this relation must hold in all coordinate systems, i.e.,

$$\boldsymbol{\epsilon}'_j(\mathbf{e}'_i) = \boldsymbol{\epsilon}'_j(\mathbf{T}\mathbf{e}_i) = \delta_{ij}. \tag{2-94}$$

To show in 2-D, we can write out the following system:

$$\boldsymbol{\epsilon}'_1(\mathbf{e}'_1) = \begin{bmatrix} \epsilon'_{11} & \epsilon'_{21} \end{bmatrix} \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \epsilon'_{11} & \epsilon'_{21} \end{bmatrix} \begin{bmatrix} T_{11} \\ T_{21} \end{bmatrix} = 1, \tag{2-95a}$$

$$\boldsymbol{\epsilon}'_1(\mathbf{e}'_2) = \begin{bmatrix} \epsilon'_{11} & \epsilon'_{21} \end{bmatrix} \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} \epsilon'_{11} & \epsilon'_{21} \end{bmatrix} \begin{bmatrix} T_{12} \\ T_{22} \end{bmatrix} = 0, \tag{2-95b}$$

$$\boldsymbol{\epsilon}'_2(\mathbf{e}'_1) = \begin{bmatrix} \epsilon'_{12} & \epsilon'_{22} \end{bmatrix} \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \epsilon'_{12} & \epsilon'_{22} \end{bmatrix} \begin{bmatrix} T_{11} \\ T_{21} \end{bmatrix} = 0, \tag{2-95c}$$

$$\boldsymbol{\epsilon}'_2(\mathbf{e}'_1) = \begin{bmatrix} \epsilon'_{12} & \epsilon'_{22} \end{bmatrix} \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} \epsilon'_{12} & \epsilon'_{22} \end{bmatrix} \begin{bmatrix} T_{12} \\ T_{22} \end{bmatrix} = 1. \tag{2-95d}$$

The first and second equations can be combined, as can the third and fourth to write the system:

$$\begin{bmatrix} \epsilon'_{11} & \epsilon'_{21} \end{bmatrix} \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \end{bmatrix} = \begin{bmatrix} \epsilon_{11} & \epsilon_{21} \end{bmatrix}, \tag{2-96a}$$

$$\begin{bmatrix} \epsilon'_{12} & \epsilon'_{22} \end{bmatrix} \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} = \begin{bmatrix} 0 & 1 \end{bmatrix} = \begin{bmatrix} \epsilon_{12} & \epsilon_{22} \end{bmatrix}. \tag{2-96b}$$

The backward transformation $\mathbf{T}^{-1}$ may be multiplied on the right, to show the relationship

$$\boldsymbol{\epsilon}' = \boldsymbol{\epsilon}\mathbf{T}^{-1}. \tag{2-97}$$

This result generalizes to any dimensionality. In other words, to transform the basis covectors, we apply the backward transform, which is opposite to how the basis vectors transform. Therefore, the basis covectors transform contravariantly with respect to the transformation of the basis vectors.

Since, the covectors are not changed under a coordinate transformation and basis covectors transform with the inverse transform, then the covector components must transform with the forward transformation. The covector components transform covariantly with respect to the basis vectors.

To provide a concrete example, let us consider the covector and the two vectors in Sec. 2.3.a on covectors and matrix multiplication:

$$\boldsymbol{\alpha} = \begin{bmatrix} 1 & -1/2 \end{bmatrix}, \qquad \mathbf{v}_1 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}, \qquad \mathbf{v}_2 = \begin{bmatrix} 1/2 \\ 4 \end{bmatrix}.$$

Now we will apply the (forward) transformation matrix

$$\mathbf{T} = \begin{bmatrix} 1 & -1 \\ 3/2 & 0 \end{bmatrix}.$$

The backwards transformation matrix is its inverse:

$$\mathbf{T}^{-1} = \begin{bmatrix} 0 & 2/3 \\ -1 & 2/3 \end{bmatrix}.$$

The *components* of the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ transform contravariantly with respect to the basis vectors and therefore use the backwards transform:

$$\mathbf{v}_1' = \begin{bmatrix} 0 & 2/3 \\ -1 & 2/3 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 4/3 \\ -5/3 \end{bmatrix},$$
$$\mathbf{v}_2' = \begin{bmatrix} 0 & 2/3 \\ -1 & 2/3 \end{bmatrix} \begin{bmatrix} 1/2 \\ 4 \end{bmatrix} = \begin{bmatrix} 8/3 \\ 13/6 \end{bmatrix}.$$

The *components* of the covector $\boldsymbol{\alpha}$ transform covariantly with respect to the basis vectors and use the forward transformation matrix:

$$\boldsymbol{\alpha}' = \begin{bmatrix} 1 & -1/2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 3/2 & 0 \end{bmatrix} = \begin{bmatrix} 1/4 & -1 \end{bmatrix}.$$

Multiplying the covector and the two vectors in the transformed space gives:

$$\boldsymbol{\alpha}'(\mathbf{v}_1') = \begin{bmatrix} 1/4 & -1 \end{bmatrix} \begin{bmatrix} 4/3 \\ -5/3 \end{bmatrix} = 2,$$
$$\boldsymbol{\alpha}'(\mathbf{v}_2') = \begin{bmatrix} 1/4 & -1 \end{bmatrix} \begin{bmatrix} 8/3 \\ 13/6 \end{bmatrix} = -3/2.$$

This result is identical to the one from the untransformed coordinate system, which is expected since coordinate transformations do not fundamentally change the vectors and covectors, only their representation.

An illustration of the vectors and the covector in the transformed coordinate system is given in Fig. 2.6, similar to what was provided for the untransformed coordinate system (Fig. 2.3). The representation of the vectors in this coordinate system is different, having them both rotated and shortened. The covector is also rotated, but the contour lines are slightly more spread out than in the untransformed coordinate system. Per the result above, both the length of the vectors measured with respect to the covectors are the same in either coordinate system.

This result illustrates an important point in physics: the choice of coordinate system is completely arbitrary and the physical laws and the results we obtain applying them should be identical. For example, the distance traveled between two points should not depend on the coordinate system we choose. If we think of the covector acting on a vector as a measure such a distance, then we expect to get the same results either way and in any other coordinate system.

The key points from this section may be summarized as follows:

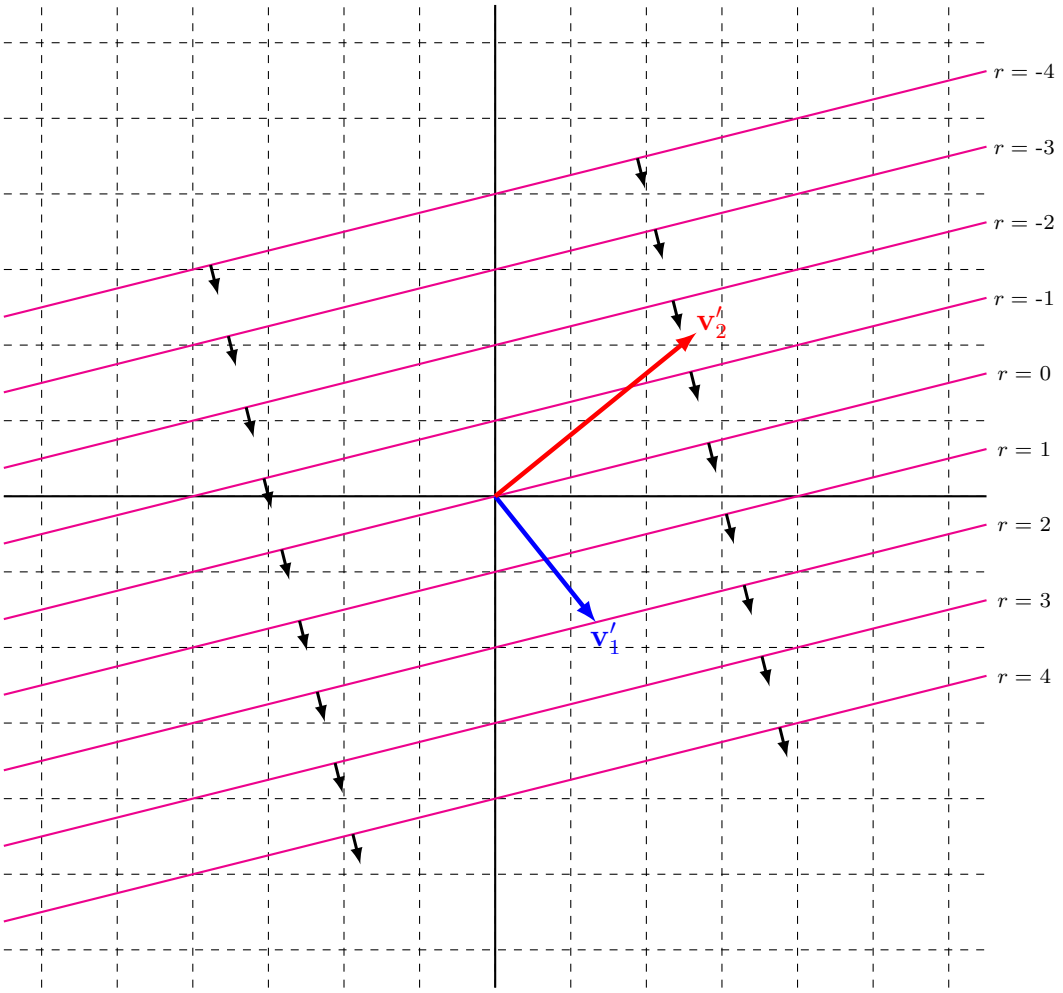

Figure 2.6: Illustration of the vectors and a covector in transformed coordinate system.

1. Neither vectors nor covectors are fundamentally changed by a coordinate transformation or change of basis, merely their representation through the components and basis vectors/covectors;

2. The *basis vectors* transform *covariantly* with respect to themselves and transform using the *forward transformation matrix*, while the *vector components* transform *contravariantly* with respect to the basis vectors and transform using the *backward transformation matrix*;

3. The *basis covectors* transform *contravariantly* with respect to the basis vectors and transform using the *backward transformation matrix*, while the *covector components* transform *covariantly* with respect to the basis vectors and transform using the *forward transformation matrix*.

2.4.e Distance, Angles and the Metric Tensor

The distance we calculate between two points should be the same regardless of our choice of basis. The same applies for the orientation between different objects, i.e., the angles. First, we will focus on the distance calculation, derive an object called the metric tensor, and finally discuss its application to finding angles between vectors.

Two points can be connected with a vector and its length squared can be determined by the dot product of a vector with itself. To illustrate in 2D, the length squared is

$$\begin{aligned} r^2 = \mathbf{a}\cdot\mathbf{a} &= (a_1\mathbf{e}_1 + a_2\mathbf{e}_2)\cdot(a_1\mathbf{e}_1 + a_2\mathbf{e}_2) \\ &= a_1^2(\mathbf{e}_1\cdot\mathbf{e}_1) + a_2^2(\mathbf{e}_2\cdot\mathbf{e}_2) + 2a_1a_2(\mathbf{e}_1\cdot\mathbf{e}_2). \end{aligned} \tag{2-98}$$

For the special case of the Cartesian coordinate system where $\mathbf{e}_1 = \hat{\mathbf{i}}$ and $\mathbf{e}_2 = \hat{\mathbf{j}}$, we end up with $r^2 = x^2 + y^2$, which is the Pythagorean theorem. (Recall $\hat{\mathbf{i}}\cdot\hat{\mathbf{i}} = 1, \hat{\mathbf{j}}\cdot\hat{\mathbf{j}} = 1, \hat{\mathbf{i}}\cdot\hat{\mathbf{j}} = 0$.) In any other coordinate system, the dot products need to be evaluated properly. As an example, recall the transformed coordinates earlier in this section where the transformed basis vectors are

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 3/2 \end{bmatrix}, \qquad \mathbf{e}_2 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}.$$

The dot products are

$$\begin{aligned} \mathbf{e}_1\cdot\mathbf{e}_1 &= 13/4, \\ \mathbf{e}_2\cdot\mathbf{e}_2 &= 1, \\ \mathbf{e}_1\cdot\mathbf{e}_2 &= -1. \end{aligned}$$

The distance formula for the transformed coordinate space is therefore

$$r^2 = \frac{13}{4}a_1^2 + a_2^2 - 2a_1a_2.$$

Now let's check to see if the length squared is the same in both coordinates. Recall the vector

$$\mathbf{v}_1 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}, \qquad \mathbf{v}_1' = \begin{bmatrix} 4/3 \\ -5/3 \end{bmatrix}.$$

Applying the dot product to both of these with the appropriate basis vectors gives

$$\begin{aligned} \mathbf{v}_1\cdot\mathbf{v}_1 &= 3^2 + 2^2 = 13, \\ \mathbf{v}_1'\cdot\mathbf{v}_1' &= \frac{13}{4}\left(\frac{4}{3}\right)^2 + \left(\frac{-5}{3}\right)^2 - 2\left(\frac{4}{3}\right)\left(\frac{-5}{3}\right) = 13, \end{aligned}$$

which is the expected result.

Equation (2-98) for the length squared can be rewritten as

$$r^2 = \begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} \mathbf{e}_1\cdot\mathbf{e}_1 & \mathbf{e}_1\cdot\mathbf{e}_2 \\ \mathbf{e}_2\cdot\mathbf{e}_1 & \mathbf{e}_2\cdot\mathbf{e}_2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}. \tag{2-99}$$

For the specific example the matrix is

$$\mathbf{g} = \begin{bmatrix} {}^{13}/_{4} & -1 \\ -1 & 1 \end{bmatrix}.$$

This expression may be generalized to any number of dimensions:

$$r^2 = \begin{bmatrix} a_1 & a_2 & \cdots & a_N \end{bmatrix} \begin{bmatrix} \mathbf{e}_1 \cdot \mathbf{e}_1 & \mathbf{e}_1 \cdot \mathbf{e}_2 & \cdots & \mathbf{e}_1 \cdot \mathbf{e}_N \\ \mathbf{e}_2 \cdot \mathbf{e}_1 & \mathbf{e}_2 \cdot \mathbf{e}_2 & \cdots & \mathbf{e}_2 \cdot \mathbf{e}_N \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{e}_N \cdot \mathbf{e}_1 & \mathbf{e}_N \cdot \mathbf{e}_2 & \cdots & \mathbf{e}_N \cdot \mathbf{e}_N \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix}. \tag{2-100}$$

The matrix in this expression is called the metric tensor and is denoted by

$$\mathbf{g} = \begin{bmatrix} \mathbf{e}_1 \cdot \mathbf{e}_1 & \mathbf{e}_1 \cdot \mathbf{e}_2 & \cdots & \mathbf{e}_1 \cdot \mathbf{e}_N \\ \mathbf{e}_2 \cdot \mathbf{e}_1 & \mathbf{e}_2 \cdot \mathbf{e}_2 & \cdots & \mathbf{e}_2 \cdot \mathbf{e}_N \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{e}_N \cdot \mathbf{e}_1 & \mathbf{e}_N \cdot \mathbf{e}_2 & \cdots & \mathbf{e}_N \cdot \mathbf{e}_N \end{bmatrix}. \tag{2-101}$$

The metric tensor is used (as we have seen) for finding lengths of vectors and (as we will soon show) angles between vectors. Note that the inverse metric tensor $\mathbf{g}^{-1}$ can be used to find lengths and angles between covectors.

To find the angle between two vectors, we again use the dot product. Recall,

$$\cos\theta = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|}.$$

This can be written in terms of the metric tensor as vector-matrix-vector products as

$$\cos\theta = \frac{\mathbf{a}^\top \mathbf{g} \mathbf{b}}{(\mathbf{a}^\top \mathbf{g} \mathbf{a})^{1/2} \, (\mathbf{b}^\top \mathbf{g} \mathbf{b})^{1/2}}.$$

To explore this, let us find the angle between $\mathbf{v}_1$ and $\mathbf{v}_2$ as well as $\mathbf{v}_1'$ and $\mathbf{v}_2'$ from the example. Recall that

$$\mathbf{v}_2 = \begin{bmatrix} {}^{1}/_{2} \\ 4 \end{bmatrix}, \qquad \mathbf{v}_2' = \begin{bmatrix} {}^{8}/_{3} \\ {}^{13}/_{6} \end{bmatrix}.$$

In the untransformed coordinate system:

$$\theta = \cos^{-1}\left[\frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{(\mathbf{v}_1 \cdot \mathbf{v}_1)^{1/2} (\mathbf{v}_2 \cdot \mathbf{v}_2)^{1/2}}\right] = \cos^{-1}\left[\frac{{}^{19}/_{2}}{(13)^{1/2} \, ({}^{65}/_{4})^{1/2}}\right] \approx 49.2^\circ.$$

In the transformed coordinate system

$$\theta = \cos^{-1}\left[\frac{\mathbf{v}_1' \cdot \mathbf{v}_2'}{(\mathbf{v}_1' \cdot \mathbf{v}_1')^{1/2} (\mathbf{v}_2' \cdot \mathbf{v}_2')^{1/2}}\right].$$

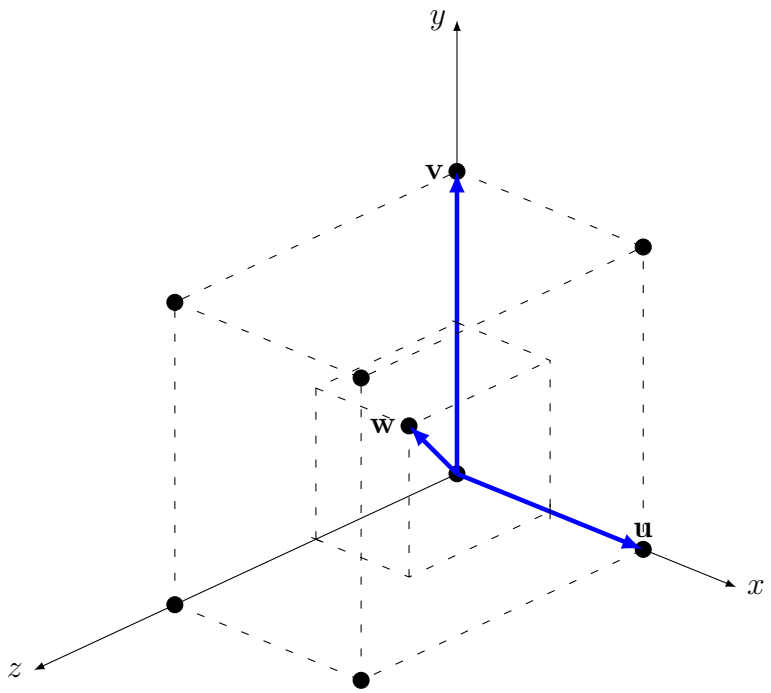

Figure 2.7: Body-Centered Cubic (BCC) Unit Cell and Basis Vectors.

We have already showed $\mathbf{v}_1' \cdot \mathbf{v}_1' = \mathbf{v}_1 \cdot \mathbf{v}_1 = 13$. It reasons that if the other two dot products are identical, then we will calculate the same angle. Now, using the metric tensor,

$$\mathbf{v}_1' \cdot \mathbf{v}_2' = \begin{bmatrix} 4/3 & -5/3 \end{bmatrix} \begin{bmatrix} 13/4 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 8/3 \\ 13/6 \end{bmatrix} = \frac{19}{2}.$$
$$\mathbf{v}_2' \cdot \mathbf{v}_2' = \begin{bmatrix} 8/3 & 13/6 \end{bmatrix} \begin{bmatrix} 13/4 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 8/3 \\ 13/6 \end{bmatrix} = \frac{65}{4},$$

we see that indeed the dot products are same, yielding the same angle. In comparing the plots of the vectors in the two coordinate systems in Figs. 2.3 and 2.6, the angles between the two vectors on both plots do not appear identical. If one were to naively calculate the angle in the transformed coordinate without considering the metric tensor, one would get 90.4°. This may seem odd, but one has to remember that transformations do not change vectors and covectors themselves, only their representation in a particular basis. So while the two appear different, when appropriate measures of lengths and angles for the coordinate systems basis vectors are used, one will get consistent results.

2.4.f Example: Body-Centered Cubic Lattice

Atoms in solids form grains consisting of regular patterns of atoms called a lattice. It is often convenient to write equations in terms of lattice positions, where the

integer values correspond to nominal atom sites within the lattice. In most cases, the coordinate system is nonorthogonal. Many calculations such as computing forces and energies involve computing distances between atoms, and the metric tensor is a useful computational tool to quickly calculate distances.

In this example, we consider the body-centered cubic (BCC) lattice as depicted in Fig. 2.7. The basis vectors connect an atom to its neighboring atoms. For this lattice, the basis vectors (see Fig. 2.7) are

$$\mathbf{u} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}. \tag{2-102}$$

The basis vectors $\mathbf{u}$ and $\mathbf{v}$ connect atoms on the same face and the basis vector $\mathbf{w}$ connects the face with the center. It is easy to verify by taking dot products that the basis vectors $\mathbf{u}$ and $\mathbf{v}$ are orthogonal to each other, but the basis vector $\mathbf{w}$ is not orthogonal to $\mathbf{u}$ or $\mathbf{v}$. Also it is important to note that the magnitude of $\mathbf{w}$ differs from the magnitudes of $\mathbf{u}$ and $\mathbf{v}$. Taken together, this implies that the distance between two positions given by lattice coordinates would not be given by anything resembling the standard Euclidian distance.

The metric tensor can be used to derive an expression to calculate the distance between any two points in the coordinate system. We can derive the metric tensor by taking dot products of the basis vectors:

$$\mathbf{g} = \begin{bmatrix} \mathbf{u}\cdot\mathbf{u} & \mathbf{u}\cdot\mathbf{v} & \mathbf{u}\cdot\mathbf{w} \\ \mathbf{v}\cdot\mathbf{u} & \mathbf{v}\cdot\mathbf{v} & \mathbf{v}\cdot\mathbf{w} \\ \mathbf{w}\cdot\mathbf{u} & \mathbf{w}\cdot\mathbf{v} & \mathbf{w}\cdot\mathbf{w} \end{bmatrix} = \begin{bmatrix} 4 & 0 & 2 \\ 0 & 4 & 2 \\ 2 & 2 & 3 \end{bmatrix}. \tag{2-103}$$

Suppose the distance between two arbitrary points is given by the vector

$$a\mathbf{u} + b\mathbf{v} + c\mathbf{w},$$

The formula for the distance squared (or, equivalently, the magnitude squared of the vector) is then

$$\begin{aligned} s^2 &= \begin{bmatrix} a & b & c \end{bmatrix} \begin{bmatrix} 4 & 0 & 2 \\ 0 & 4 & 2 \\ 2 & 2 & 3 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} \\ &= \begin{bmatrix} (4a+2c) & (4b+2c) & (2a+2b+3c) \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} \\ &= 4a^2 + 4b^2 + 3c^2 + 4ac + 4bc. \end{aligned} \tag{2-104}$$

The first collection of terms looks similar to the Euclidian distance, but with different scaling factors. The last two cross terms are different and arise from the nonorthogonality of the coordinate system.

2.4.g Rotation Matrix

A common operation involving a change of basis involves rotating the coordinate system. Sometimes reorienting the principal axes can simplify the equations that need to be solved.

In 2-D the forward transformation for the basis vectors of the rotation matrix is

$$\mathbf{R} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{2-105}$$

The backward transformation is

$$\mathbf{R}^{-1} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}. \tag{2-106}$$

Note that $\mathbf{R}^{-1} = \mathbf{R}^{\top}$. This is true of all valid rotation matrices in any number of dimensions.

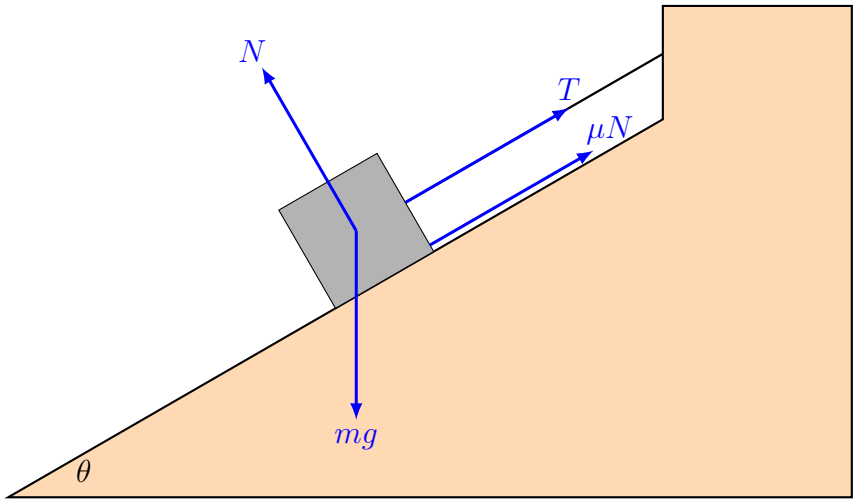

Figure 2.8: Simple statics problem to illustrate rotation matrix.

To illustrate where a rotation matrix may be useful, consider the statics problem illustrated in Fig. 2.8 where we want to solve for the tension T. The force vector components must sum to zero. The force vector is

$$\mathbf{F} = \begin{bmatrix} F_x \\ F_y \end{bmatrix} = \begin{bmatrix} T\cos\theta + \mu N\cos\theta - N\sin\theta \\ T\sin\theta + \mu N\sin\theta + N\cos\theta - mg \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

These equations can be simplified considerably if the coordinate system is rotated by angle θ. The components of the force vector transform contravariantly and are transformed using the backwards transformation matrix:

$$\mathbf{R}^{-1}\mathbf{F} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} F_x \\ F_y \end{bmatrix} = \begin{bmatrix} T + \mu N - mg\sin\theta \\ N - mg\cos\theta \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The normal force N can be easily eliminated to obtain an expression for the tension:

$$T = mg(\sin\theta - \mu\cos\theta).$$

Now, for a problem such as this, it was probably easier to solve the original equations directly (or to infer the transformed coordinate equations), but this can become considerably more difficult for more complicated problems.

In 3-D rotating about the x, y, and z axes can be easily deduced. These are:

$$\mathbf{R}_x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}, \tag{2-107a}$$

$$\mathbf{R}_y = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix}, \tag{2-107b}$$

$$\mathbf{R}_z = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{2-107c}$$

Where the backwards transformations are simply the transpose of the matrix. The 3-D general rotation matrix for rotating about an arbitrary axis is complicated to derive, and the derivation is not provided here. Suppose one wishes to rotate an angle θ about an axis given by vector $\mathbf{u}$, the rotation matrix is

$$\mathbf{R} = \begin{bmatrix} \cos\theta + u_x^2(1-\cos\theta) & u_x u_y(1-\cos\theta) - u_z\sin\theta & u_x u_z(1-\cos\theta) + u_y\sin\theta \\ u_x u_y(1-\cos\theta) + u_z\sin\theta & \cos\theta + u_y^2(1-\cos\theta) & u_y u_z(1-\cos\theta) - u_x\sin\theta \\ u_x u_z(1-\cos\theta) - u_y\sin\theta & u_y u_z(1-\cos\theta) + u_x\sin\theta & \cos\theta + u_z^2(1-\cos\theta) \end{bmatrix}. \tag{2-108}$$

By setting $\mathbf{u}$ to one of the orthogonal basis vectors, the general 3-D rotation matrix reduces to either of the special cases.

2.4.h Gram-Schmidt Orthogonalization Process

We sometimes find ourselves with a set of linearly independent vectors or basis that are not orthogonal to each other. The question is whether one can devise a scheme to come up with a set of orthogonal vectors. One method of doing this is called the *Gram-Schmidt Orthogonalization Process*.

Given a set of k vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ (the dimension of the vectors N must be greater than or equal to k), we can build the set of k orthogonal vectors $\mathbf{u}$ by

$$\begin{aligned} \mathbf{u}_1 &= \mathbf{v}_1, \\ \mathbf{u}_2 &= \mathbf{v}_2 - \text{proj}_{\mathbf{u}_1}(\mathbf{v}_2), \\ \mathbf{u}_3 &= \mathbf{v}_3 - \text{proj}_{\mathbf{u}_1}(\mathbf{v}_3) - \text{proj}_{\mathbf{u}_2}(\mathbf{v}_3), \\ &\vdots \end{aligned}$$

$$\mathbf{u}_k = \mathbf{v}_k - \sum_{i=1}^{k-1} \text{proj}_{\mathbf{u}_i}(\mathbf{v}_k). \tag{2-109}$$

The projection operator is defined in Eq. (2-61). Sometimes it is convenient to rescale the vector $\mathbf{u}_j$ in each step, especially when doing the calculations by hand. This is allowed since a constant scaling does not impact the orthogonality and the projection renormalizes each term anyway. In the same vein, the vectors $\mathbf{u}_j$ are orthogonal, but not naturally normalized; and it is often desired to construct an orthonormal basis:

$$\hat{\mathbf{e}}_j = \frac{\mathbf{u}_j}{|\mathbf{u}_j|}. \tag{2-110}$$

As an example, consider the space of vectors in $\mathbb{R}^4$ with the following nonorthogonal span:

$$\text{span}\left\{ \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ -2 \\ 1 \end{bmatrix} \right\}.$$

We wish to transform to an orthonormal basis using the Gram-Schmidt process.

First, we have simply that

$$\mathbf{u}_1 = \mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}. \tag{2-111}$$

The second orthogonal vector is

$$\mathbf{u}_2 = \mathbf{v}_2 - \left(\frac{\mathbf{v}_2 \cdot \mathbf{u}_1}{\mathbf{u}_1 \cdot \mathbf{u}_1} \right) \mathbf{u}_1 = \begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \frac{1}{3} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} -2 \\ 3 \\ 1 \\ 1 \end{bmatrix}. \tag{2-112}$$

The final form is written with a factor of $\frac{1}{3}$ because scaling does not change the orthogonality. The third orthogonal vector is

$$\begin{aligned} \mathbf{u}_3 &= \mathbf{v}_3 - \left(\frac{\mathbf{v}_3 \cdot \mathbf{u}_1}{\mathbf{u}_1 \cdot \mathbf{u}_1} \right) \mathbf{u}_1 - \left(\frac{\mathbf{v}_3 \cdot \mathbf{u}_2}{\mathbf{u}_2 \cdot \mathbf{u}_2} \right) \mathbf{u}_2 \\ &= \begin{bmatrix} 2 \\ 1 \\ 0 \\ 0 \end{bmatrix} - \frac{2}{3} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} - \frac{1}{15} \begin{bmatrix} 2 \\ -3 \\ -1 \\ -1 \end{bmatrix} = \frac{3}{5} \begin{bmatrix} 2 \\ 2 \\ -1 \\ -1 \end{bmatrix}. \end{aligned} \tag{2-113}$$

The final orthogonal vector is

$$\mathbf{u}_4 = \mathbf{v}_4 - \left(\frac{\mathbf{v}_4 \cdot \mathbf{u}_1}{\mathbf{u}_1 \cdot \mathbf{u}_1} \right) \mathbf{u}_1 - \left(\frac{\mathbf{v}_4 \cdot \mathbf{u}_2}{\mathbf{u}_2 \cdot \mathbf{u}_2} \right) \mathbf{u}_2 - \left(\frac{\mathbf{v}_4 \cdot \mathbf{u}_3}{\mathbf{u}_3 \cdot \mathbf{u}_3} \right) \mathbf{u}_3$$

$$= \begin{bmatrix} 0 \\ 0 \\ -2 \\ 1 \end{bmatrix} + \frac{1}{3}\begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} - \frac{1}{15}\begin{bmatrix} 2 \\ -3 \\ -1 \\ -1 \end{bmatrix} - \frac{1}{10}\begin{bmatrix} 2 \\ 2 \\ -1 \\ -1 \end{bmatrix} = \frac{3}{2}\begin{bmatrix} 0 \\ 0 \\ -1 \\ 1 \end{bmatrix}. \tag{2-114}$$

Therefore, an orthogonal basis spanning $\mathbb{R}^4$ based on the original set of vectors is

$$\text{span}\left\{\begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -2 \\ 3 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \\ -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ -1 \\ 1 \end{bmatrix}\right\}.$$

To find an orthonormal basis, we divide each vector by its respective magnitude:

$$\hat{\mathbf{e}}_1 = \frac{1}{\sqrt{3}}\begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \quad \hat{\mathbf{e}}_2 = \frac{1}{\sqrt{15}}\begin{bmatrix} -2 \\ 3 \\ 1 \\ 1 \end{bmatrix}, \quad \hat{\mathbf{e}}_3 = \frac{1}{\sqrt{10}}\begin{bmatrix} 2 \\ 2 \\ -1 \\ -1 \end{bmatrix}, \quad \hat{\mathbf{e}}_4 = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 \\ 0 \\ -1 \\ 1 \end{bmatrix}. \tag{2-115}$$

2.5 Systems of Linear Equations

Many phenomena in science and engineering can be described using systems of linear equations. These include: engineering statics, electrical circuits, and thermodynamics of power conversion cycles. Furthermore, many physical phenomena are explained by partial differential equations, which cannot be solved in general, but can be represented approximately as a system of linear equations.

All linear systems can be written in the form

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{2-116}$$

where $\mathbf{A}$ is a known matrix of coefficients, $\mathbf{x}$ is a column vector of unknowns, and $\mathbf{b}$ is a known column vector of constant or inhomogeneous terms.

The expanded form for this set of equations is as follows:

$$\begin{aligned} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,M}x_M &= b_1, \\ a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,M}x_M &= b_2, \\ &\vdots \\ a_{N,1}x_1 + a_{N,2}x_2 + \cdots + a_{N,M}x_M &= b_N, \end{aligned} \tag{2-117}$$

and written equivalently as

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,M} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{N,2} & \cdots & a_{N,M} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix}. \tag{2-118}$$

Here we have a system of N equations with M unknowns. If $N = M$, then the system *might* have a unique solution vector $\mathbf{x}$. It is also possible even if $N = M$ for some the equations to be inconsistent, meaning there is no solution or, alternatively, one or more of the equations could be redundant (linearly dependent) and therefore $\mathbf{x}$ has infinitely many solutions that satisfy these equations. The same can be said if $N > M$, more equations than unknowns. If there are fewer equations than unknowns, then there is insufficient information to uniquely determine $\mathbf{x}$ so there may, at best, be an infinite number of solutions; however, the equations could be inconsistent as well meaning that there are two or more linearly dependent equations that have different results, in which case there is no solution to the linear system.

Recall that matrix multiplication has an interpretation of a linear map or coordinate transformation. A geometric interpretation for $\mathbf{Ax} = \mathbf{b}$ is as follows: for a given $N \times M$ transformation matrix $\mathbf{A}$, find the $M \times 1$ vector, or range of such vectors, $\mathbf{x}$ that, upon applying the transformation to them, results in the $N \times 1$ vector $\mathbf{b}$.

Next we will focus on methods of solving these equations (should a solution exist). The algorithm is called Gauss-Jordan elimination or simply Gaussian elimination. This procedure consists of two major steps. The first, is forward elimination and the second is backward substitution. The geometric interpretation to this process is to transform the basis or coordinate system using matrix operations on $\mathbf{A}$ and $\mathbf{b}$ into an orthonormal basis, or as close to one as possible in cases where there is not a unique solution.

2.5.a Forward Elimination

The goal with forward elimination is to apply a set of linear operations to transform the coefficient matrix into a particular form called *row echelon form*. First, we write our linear system in an augmented form:

$$\left[\begin{array}{ccccc|c} a_{1,1} & a_{1,2} & \cdots & a_{1,N-1} & a_{1,N} & b_1 \\ a_{2,1} & a_{2,2} & \cdots & a_{2,N-1} & a_{2,N} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ a_{N-1,1} & a_{N-1,2} & \cdots & a_{N-1,N-1} & a_{N-1,N} & b_{N-1} \\ a_{N,1} & a_{N,2} & \cdots & a_{N,N-1} & a_{N,N} & b_N \end{array}\right].$$

This vertical bar in the augmented form separates out the left- and right-hand sides of the equation. This specific form is where the number of equations matches the number of unknowns.

The goal is to manipulate the matrix to be in row echelon form using elementary row operations. This form requires that the matrix be organized in a descending staircase pattern such that there are zeroes below the pattern. An example of row

echelon form for a square matrix is:

$$\left[\begin{array}{ccccc|c} c_{1,1} & c_{1,2} & \cdots & c_{1,N-1} & c_{1,N} & d_1 \\ 0 & c_{2,2} & \cdots & c_{2,N-1} & c_{2,N} & d_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & c_{N-1,N-1} & c_{N-1,N} & d_{N-1} \\ 0 & 0 & \cdots & 0 & c_{N,N} & d_N \end{array}\right].$$

In this case, we will see that the matrix has a unique solution provided that all the diagonal elements are nonzero. Should any of them be zero, then there will be infinitely many solutions.

Another example with actual numbers that is also in row-echelon form for a 4×5 system is

$$\left[\begin{array}{ccccc|c} 1 & 1 & 0 & 2 & 1 & 1 \\ 0 & 0 & 1 & 1 & 2 & 0 \\ 0 & 0 & 0 & 1 & -1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

This case has two features that distinguish it from the first. One is that the second row does not have a diagonal matrix in the second column entry. The other is that the fourth column is all zeroes. This is still in row echelon form because the pattern is still a descending staircase and backward substitution is possible. It will turn out this case does not have a unique solution, but rather infinitely many solutions over a space of them that can be defined geometrically (in this specific case a plane in 5-dimensional space).

A similar example where no solution exists is as follows:

$$\left[\begin{array}{cccc|c} 1 & 1 & -1 & 0 & 3 \\ 0 & 1 & 1 & 1 & -2 \\ 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right].$$

Here, the last row is all zeroes on the left-hand side (all coefficients zero), but the right-hand side is nonzero. This would imply zero equals a nonzero number, which is impossible. This implies that the original system included information that was contradictory.

Note that some authors define row echelon form with the diagonal as one. This is helpful with writing efficient numerical algorithms, but is entirely optional for advancing to backward substitution.

To accomplish the task of transforming an augmented matrix into row-echelon form, we are permitted to do three operations. Each of these elementary row operations may be performed by multiplying the system $\mathbf{Ax} = \mathbf{b}$ on the left by an equivalent transformation matrix $\mathbf{T}$, i.e., $\mathbf{TAx} = \mathbf{Tb}$. These operations are:

1. Swap the position of any two rows; this is simply rewriting the equations in a different order. In geometric terms, this is the same as performing a rotation

that swaps the coordinate axes. An example of a 4×4 transformation matrix that swaps the second and fourth rows is

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}. \tag{2-119a}$$

This matrix is similar to the identity matrix, except that the location of the ones have been swapped. Here the 1 in the second row is in the fourth column and the 1 in the fourth row is in the second column. This yields the appropriate swap.

2. Scale any equation by a constant, non-zero factor; this is multiplying both sides of the equation by a constant. An example of a different 4×4 matrix that scales the second row by a factor of 2 is

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{2-119b}$$

This is similar to the identity matrix except that there is a 2 as opposed to a 1 in the diagonal entry of the second row. This scales the coordinate system along the second coordinate axis.

3. Replace an equation with a linear combination of that equation and another equation. For example, if we want to replace the third equation by the sum of that equation and 4 times the first equation, the transformation is

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 4 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{2-119c}$$

Again, this is similar to the identity matrix except that in the third row, there is a 4 in the first column. When multiplying the third row by the columns of $\mathbf{A}$ this takes 4 times the first element of the column and adds it to the third element to produce the result. The geometric interpretation of this is it rotates and scales one of the coordinate axes.

Note that a common mistake is attempting to directly swap columns, which is not allowed. Column swapping cannot also be done by multiplying a matrix on the left, but rather would require one to multiplied on the right. Since $\mathbf{A}$ multiplies the solution vector $\mathbf{x}$ on the left, this right multiplication would have to be to the right of $\mathbf{x}$ and would therefore act on $\mathbf{Ax}$ and the right-hand side vector $\mathbf{b}$ and not on $\mathbf{A}$ itself.

By chaining together elementary transformation matrices $\mathbf{T}_i$, the matrix $\mathbf{A}$ can be transformed into row echelon form and then reduced-row echelon form. Furthermore, if $\mathbf{A}^{-1}$ exists, then it is the result of the products of all the transformations $\mathbf{T}_i$ applied in the appropriate order given by an algorithm that will be discussed shortly. In practice, we do not perform the matrix multiplication for $\mathbf{T}_i$, as it would be too inefficient either to perform by hand or on a computer. Rather, these are given to show the geometric connection between the elementary row operations and changes of basis.

Using these three operations, we can develop an algorithm for forward elimination, which will be followed by one for backward substitution. The algorithm applied to a square matrix is given in Fig. 2.9. In short, the algorithm loops over the columns and tries to turn all elements in that column below a pivot row `k` into zeros by replacing an equation on a row with a linear combination of equations. The condition on line 3 handles the case where the pivot element is zero; when this occurs, we need to attempt to perform a swap with another equation. If we cannot find such an equation, we know the system does not have a unique solution, but we move onto the next column keeping the pivot point `k` fixed and start there. If we can get the pivot element to be nonzero, we then proceed and make the elements in the current column below that point zero by replacing an equation with a linear combination of that equation and the equation of the pivot row.

Note that this a mechanical algorithm that will always work, but is not necessarily the most efficient way. When doing forward elimination by hand, it is often beneficial to make heuristic simplifications to make the process more efficient.

```
 1. let k = 1
 2. loop j over the columns of the matrix A in augmented matrix:
 3.   if a(k,j) = 0:
 4.     loop i down the rows from k+1 until a(i,j) != 0 found
 5.     if such an a(i,j) found:
 6.       swap rows k and i
 7.     else:
 8.       skip to the next column (do not increment k)
 9.   optional: scale row k by 1/a(k,j)
10.   loop i down the rows from k+1 until the last row:
11.     if a(i,j) != 0:
12.       add a(i,j)/a(k,j) times row k to row i
13.   k += 1
14. return modified augmented matrix
```

Figure 2.9: Algorithm for forward elimination.

Example 1

As an example, consider the following linear system:

$$2y - 1z = -1, \tag{2-120a}$$
$$x - 2y = 0, \tag{2-120b}$$
$$3x + y - 2z = 1. \tag{2-120c}$$

This may be represented in matrix-vector form as

$$\begin{bmatrix} 0 & 2 & -1 \\ 1 & -2 & 0 \\ 3 & 1 & -2 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \tag{2-121a}$$

or in augmented form as

$$\left[\begin{array}{ccc|c} 0 & 2 & -1 & -1 \\ 1 & -2 & 0 & 0 \\ 3 & 1 & -2 & 1 \end{array}\right]. \tag{2-121b}$$

To begin the process, start with the first column and go down the rows from ① to ③ and attempt to make the diagonal terms nonzero and the lower triangle zero. Since the (1,1) element is zero, we must flip this row (equation) with another. Either row ② or ③ will do, but to be consistent, we will flip rows ① and ②. The result becomes:

$$\left[\begin{array}{ccc|c} 0 & 2 & -1 & -1 \\ 1 & -2 & 0 & 0 \\ 3 & 1 & -2 & 1 \end{array}\right] : ① \leftrightarrow ② : \left[\begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & 2 & -1 & -1 \\ 3 & 1 & -2 & 1 \end{array}\right]. \tag{2-121c}$$

This is merely reordering the equations. Now that we have a non-zero element to work with, we must turn all the elements below the diagonal in the first column to zeroes. The (2,1) element is already zero (as a consequence of the flip), so there remains nothing to be done for row ②. Moving onto row ③, we can eliminate the (3,1) element by subtracting 3 times row ① from row ③. This step results in

$$\left[\begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & 2 & -1 & -1 \\ 3 & 1 & -2 & 1 \end{array}\right] : ③ \rightarrow ③ - 3 \times ① : \left[\begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & 2 & -1 & -1 \\ 0 & 7 & -2 & 1 \end{array}\right] \tag{2-121d}$$

Now that all elements of the first column below the diagonal are zero, we move onto the second column and do the same using the second row, ②. Next, make the element in the diagonal of the second row 1 by multiplying by $^1/_2$:

$$\left[\begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & 2 & -1 & -1 \\ 3 & 1 & -2 & 1 \end{array}\right] : ② \rightarrow {}^1/_2 \times ② : \left[\begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & 1 & {}^{-1}/_2 & {}^{-1}/_2 \\ 0 & 7 & -2 & 1 \end{array}\right] \tag{2-121e}$$

This step is not strictly necessary, but will make the subsequent step easier. The (3,2) element is 7 and should be turned to zero. This can be done by subtracting 7 times row (2) to row (3):

$$\left[\begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & 1 & -1/2 & -1/2 \\ 0 & 7 & -2 & 1 \end{array}\right] : (3) \to (3) - 7 \times (2) : \left[\begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & 1 & -1/2 & -1/2 \\ 0 & 0 & 3/2 & 9/2 \end{array}\right]. \quad \text{(2-121f)}$$

Now (optionally) multiplying row (3) by $2/3$ to remove the fractions and give a one on the diagonal:

$$\left[\begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & 1 & -1/2 & -1/2 \\ 0 & 0 & 3/2 & 9/2 \end{array}\right] : (3) \to 2/3 \times (3) : \left[\begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & 1 & -1/2 & -1/2 \\ 0 & 0 & 1 & 3 \end{array}\right]. \quad \text{(2-121g)}$$

At this point forward elimination is complete as the system of equations is in row echelon form. All elements below the diagonal are zero and to continue we will proceed with backward substitution in Sec. 2.5.c. Before we discuss this, let us do another couple examples.

Example 2

Another example is as follows:

$$x_1 - x_2 + 2x_4 = 0 \quad \text{(2-122a)}$$
$$x_3 - x_4 = 0 \quad \text{(2-122b)}$$
$$x_1 + 2x_3 + x_4 = 1 \quad \text{(2-122c)}$$
$$x_2 + x_4 = 1. \quad \text{(2-122d)}$$

This system may be converted into matrix-vector form

$$\left[\begin{array}{cccc} 1 & -1 & 0 & 2 \\ 0 & 0 & 1 & -1 \\ 1 & 0 & 2 & 1 \\ 0 & 1 & 0 & 1 \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \end{array}\right] = \left[\begin{array}{c} 0 \\ 0 \\ 1 \\ 1 \end{array}\right] \quad \text{(2-123a)}$$

or as an augmented matrix

$$\left[\begin{array}{cccc|c} 1 & -1 & 0 & 2 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 1 & 0 & 2 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{array}\right]. \quad \text{(2-123b)}$$

To begin, we start with the first column and turn all elements below the diagonal zero. Thankfully the (2,1) and (4,1) elements are already zero, so we only have to do

operations on row ③. The (3,1) element can be made zero by subtracting row ① from ③:

$$\left[\begin{array}{cccc|c} 1 & -1 & 0 & 2 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 1 & 0 & 2 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{array}\right] : ③ \rightarrow ③ - ① : \left[\begin{array}{cccc|c} 1 & -1 & 0 & 2 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 1 & 2 & -1 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{array}\right]. \tag{2-123c}$$

Now that the elements in the first column below the diagonal are all zero, proceed to the second column. Since the diagonal element (2,2) is zero, we search down the column and swap this row with the row with next available non-zero element, which happens to be (3,2):

$$\left[\begin{array}{cccc|c} 1 & -1 & 0 & 2 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 1 & 2 & -1 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{array}\right] : ② \leftrightarrow ③ : \left[\begin{array}{cccc|c} 1 & -1 & 0 & 2 & 0 \\ 0 & 1 & 2 & -1 & 1 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & 1 & 1 \end{array}\right]. \tag{2-123d}$$

Now that there is a one on the diagonal of the second column, proceed down the column and make all elements below the diagonal zero. The element (3,2) is zero by virtue of the swap. The (4,2) element can be made zero by subtracting row ② from row ④:

$$\left[\begin{array}{cccc|c} 1 & -1 & 0 & 2 & 0 \\ 0 & 1 & 2 & -1 & 1 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & 1 & 1 \end{array}\right] : ④ \rightarrow ④ - ② : \left[\begin{array}{cccc|c} 1 & -1 & 0 & 2 & 0 \\ 0 & 1 & 2 & -1 & 1 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & -2 & 2 & 0 \end{array}\right]. \tag{2-123e}$$

Proceeding to the third column, there is a nonzero element in (4,3) which can be made zero by adding two times ③ to row ④:

$$\left[\begin{array}{cccc|c} 1 & -1 & 0 & 2 & 0 \\ 0 & 1 & 2 & -1 & 1 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & -2 & 2 & 0 \end{array}\right] : ④ \rightarrow ④ + 2 \times ③ : \left[\begin{array}{cccc|c} 1 & -1 & 0 & 2 & 0 \\ 0 & 1 & 2 & -1 & 1 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right]. \tag{2-123f}$$

Notice row ④ is now all zeroes. This implies that the last row contains no new information and that one of the equations in the system is redundant or a linear combination of other equations. As we will see, this will imply that there is no unique solution. Rather, as we will discuss shortly, there is a range of vectors that satisfy this linear system. This is as far as one can take forward elimination for this example. Before proceeding with backward substitution, we will discuss the concept of the rank of a matrix, which gives information on whether the system of equations has a unique solution or, if it does not, the nature of the solutions that may satisfy the equations.

Example 3

The final example in this section is the following linear system:

$$x_1 - x_2 + x_5 = 1, \tag{2-124a}$$
$$x_1 + 2x_3 - x_4 + x_5 = 0, \tag{2-124b}$$
$$-2x_1 + x_3 + x_4 + x_5 = 0, \tag{2-124c}$$
$$x_2 + 2x_3 - x_4 = -1, \tag{2-124d}$$
$$-2x_2 + x_3 + x_4 + 3x_5 = 2. \tag{2-124e}$$

Writing the coefficient matrix in augmented form gives

$$\left[\begin{array}{ccccc|c} 1 & -1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 2 & -1 & 1 & 0 \\ -2 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & -2 & 1 & 1 & 3 & 2 \end{array}\right]. \tag{2-124f}$$

As before, we eliminate the non-zero elements in the first column. The (2,1) element can be eliminated by subtracting row (2) from row (1)

$$\left[\begin{array}{ccccc|c} 1 & -1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 2 & -1 & 1 & 0 \\ -2 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & -2 & 1 & 1 & 3 & 2 \end{array}\right] : (2) \to (2) - (1) : \left[\begin{array}{ccccc|c} 1 & -1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ -2 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & -2 & 1 & 1 & 3 & 2 \end{array}\right]. \tag{2-124g}$$

The (3,1) element can be eliminated by adding twice row (1) to row (3):

$$\left[\begin{array}{ccccc|c} 1 & -1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ -2 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & -2 & 1 & 1 & 3 & 2 \end{array}\right] : (3) \to (3) + 2 \times (1) : \left[\begin{array}{ccccc|c} 1 & -1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & -2 & 1 & 1 & 3 & 2 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & -2 & 1 & 1 & 3 & 2 \end{array}\right]. \tag{2-124h}$$

The elements of the first column are now all zero below the diagonal, so we move onto the second column. Making the (3,2) element zero is done by adding twice row (2) to row (3):

$$\left[\begin{array}{ccccc|c} 1 & -1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & -2 & 1 & 1 & 3 & 2 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & -2 & 1 & 1 & 3 & 2 \end{array}\right] : (3) \to (3) + 2 \times (2) : \left[\begin{array}{ccccc|c} 1 & -1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & 0 & 5 & -1 & 3 & 0 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & -2 & 1 & 1 & 3 & 2 \end{array}\right]. \tag{2-124i}$$

Eliminating the (4,2) element can be done by subtracting row ② from row ④:

$$\left[\begin{array}{ccccc|c} 1 & -1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & 0 & 5 & -1 & 3 & 0 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & -2 & 1 & 1 & 3 & 2 \end{array}\right] : ④ \to ④ - ② : \left[\begin{array}{ccccc|c} 1 & -1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & 0 & 5 & -1 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -2 & 1 & 1 & 3 & 2 \end{array}\right]. \tag{2-124j}$$

Note that row ④ are now all zeros, indicating that the original equation is a linear combination of at least two other equations. Continuing to eliminate (5,2) by adding twice row ② to row ⑤:

$$\left[\begin{array}{ccccc|c} 1 & -1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & 0 & 5 & -1 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -2 & 1 & 1 & 3 & 2 \end{array}\right] : ④ \to ④ + 2 \times ② : \left[\begin{array}{ccccc|c} 1 & -1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & 0 & 5 & -1 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 5 & -1 & 3 & 0 \end{array}\right]. \tag{2-124k}$$

It should now be apparent that row ⑤ is identical to row ③, but for the sake of following an programmed algorithm, move onto the third column and scale row ③ by $^1/_5$ to make the diagonal element one:

$$\left[\begin{array}{ccccc|c} 1 & -1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & 0 & 5 & -1 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 5 & -1 & 3 & 0 \end{array}\right] : ③ \to {}^1/_5 \times ③ : \left[\begin{array}{ccccc|c} 1 & -1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & 0 & 1 & -{}^1/_5 & {}^3/_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 5 & -1 & 3 & 0 \end{array}\right]. \tag{2-124l}$$

Finally, eliminate (5,3) by subtracting 5 times row ③ from row ⑤:

$$\left[\begin{array}{ccccc|c} 1 & -1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & 0 & 1 & -{}^1/_5 & {}^3/_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 5 & -1 & 3 & 0 \end{array}\right] : ⑤ \to ⑤ - 5 \times ③ : \left[\begin{array}{ccccc|c} 1 & -1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & 0 & 1 & -{}^1/_5 & {}^3/_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]. \tag{2-124m}$$

This matrix is in now in row-echelon form. Note that we now have two rows of zeros meaning that two of the equations were entirely redundant. In the next sections, we will see that this implies there is no unique solution as well.

2.5.b Matrix Rank

Before proceeding attempt backward substitution, we need to check something called the rank of the matrix. To find the rank of a matrix, first we must find its row-echelon form. Once this is found, the rank is equal to the number of nonzero rows.

If the rank of the matrix equals the number of rows, then we can assert that there is a unique solution to the system of equations, and it makes sense to proceed with backward substitution. If the rank is less than the number of rows, this implies that there will not be a unique solution. It may be that the equations are inconsistent; in which case there will be no solution. Or it may also be that there are an infinite number of solutions.

For example, if the we find the rank of a 5×5 matrix is 4 and all the equations are consistent, then geometrically, the solution consists of any set of points along a line in 5-D space. Likewise, if the rank of our 5×5 matrix is 3, then the solution corresponds to any point along a plane in that 5-D space.

Example 1

To illustrate, consider the first example from Sec. 2.5.a:

$$\begin{aligned} 2y - 1z &= -1, \\ x - 2y &= 0, \\ 3x + y - 2z &= 1. \end{aligned}$$

which has the row-echelon form of:

$$\left[\begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & 1 & {}^{-1}/_{2} & {}^{-1}/_{2} \\ 0 & 0 & 1 & 3 \end{array}\right].$$

There are three non-zero rows, so the system has a rank of three. Because there are three unknowns, we expect there to be a set of solutions that belong to a $3 - 3 = 0$ dimensional space, or a point in 3-D space. This means that there is one unique set of values that satisfies this system of equations. We say this system is fully determined.

Example 2

Now, let us consider the second example from Sec. 2.5.a:

$$\begin{aligned} x_1 - x_2 + 2x_4 &= 0 \\ x_3 - x_4 &= 0 \\ x_1 + 2x_3 + x_4 &= 1 \\ x_2 + x_4 &= 1. \end{aligned}$$

which has a row-echelon form of

$$\left[\begin{array}{cccc|c} 1 & -1 & 0 & 2 & 0 \\ 0 & 1 & 2 & -1 & 1 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

This has three non-zero rows, and therefore has a rank of three. Since there are four unknowns, we expect the space of solutions to live on a $4 - 3 = 1$ dimensional space, or a line. It is said that this system is underdetermined in the sense that there is no unique solution, however, we can classify the solution space as a set of vectors that point to anywhere along a line in 4-D space.

It is also possible for a system of equations to be inconsistent, suppose that the last row of the previous example did not have a zero to the right of the vertical bar. This would state that $0 \neq 0$, which is, of course, impossible. Therefore, while the rank is still three, there is no set of numbers that can satisfy all the equations consistently and therefore there is no solution at all.

Example 3

The third example from Sec. 2.5.a is a system of five equations,

$$\begin{aligned} x_1 - x_2 + x_5 &= 1, \\ x_1 + 2x_3 - x_4 + x_5 &= 0, \\ -2x_1 + x_3 + x_4 + x_5 &= 0, \\ x_2 + 2x_3 - x_4 &= -1, \\ -2x_2 + x_3 + x_4 + 3x_5 &= 2. \end{aligned}$$

that has the row-echelon form of

$$\left[\begin{array}{ccccc|c} 1 & -1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & 0 & 1 & -1/5 & 3/5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

This has three non-zero rows and is also a rank of three. Since there are five unknowns, the space of solutions lives on a $5 - 3 = 2$ dimensional space, or a line in 5-D space, assuming the equations are consistent.

2.5.c Backward Substitution

If the rank of the matrix (number of non-zero rows) of the row-echelon form of matrix **A** is equal to the number of its rows, then there exists a unique solution and we can proceed with doing backward substitution.

Before writing an algorithm, it may be worth seeing an example. A possible system of equations in row-echelon form is as follows:

$$\begin{aligned} 2x_1 + x_2 - x_3 &= -1, \\ -x_2 + 3x_3 &= 1, \\ -2x_3 &= 2. \end{aligned}$$

Starting with the third equation we can see

$$x_3 = \frac{2}{-2} = -1.$$

Now we can use the second equation to get

$$x_2 = \frac{1 - 3 \cdot (-1)}{-1} = -4.$$

Finally, from the first equation we can write

$$x_1 = \frac{-1 - (-1) \cdot (-1) - (1) \cdot (-4)}{2} = 1.$$

This suggests a general equation for backward substitution

$$x_i = \frac{b_i - \sum_{j=i+1}^{N} a_{i,j} x_j}{a_{i,i}}. \tag{2-125}$$

An algorithm for this is given in Fig. 2.10.

```
1. loop i backward over the rows:
2.   let s = 0
3.   loop j forward over the columns starting from i+1:
4.     s += a(i,j)*x(j)
5.   x(i) = ( b(i) - s )/a(i,i)
6. return x
```

Figure 2.10: Algorithm for backward substitution.

Example 1

Continuing with the first example from Sec. 2.5.a, the row echelon form obtained

$$\left[\begin{array}{ccc|c} 1 & -2 & 0 & 0 \\ 0 & 1 & {}^{-1}/_{2} & {}^{-1}/_{2} \\ 0 & 0 & 1 & 3 \end{array}\right]. \tag{2-126a}$$

results in the following system of equations:

$$x - 2y = 0, \tag{2-126b}$$

$$y - \frac{1}{2}z = -\frac{1}{2}, \tag{2-126c}$$

$$z = 3. \tag{2-126d}$$

Starting with the third equation in row (3), we can trivially solve for z,

$$z = 3. \tag{2-126e}$$

Substituting this value of z into the second equation in row (2):

$$y = \frac{-\frac{1}{2} + \frac{1}{2}(3)}{1} = 1. \tag{2-126f}$$

Finally, substituting y and z into the first equation in row (1) gives the value of x:

$$x = \frac{0 + 2(1)}{1} = 2. \tag{2-126g}$$

Therefore, the solution is

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}. \tag{2-126h}$$

This result yields a single point in 3-D space, which was expected since the rank of the system is 3 and matches the number of unknowns.

Example 2

Moving on to the second example, we obtained the row echelon form of

$$\left[\begin{array}{cccc|c} 1 & -1 & 0 & 2 & 0 \\ 0 & 1 & 2 & -1 & 1 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right] \tag{2-127a}$$

Since the last row is simply $0 = 0$ and provides no new information, we can simply delete that row:

$$\left[\begin{array}{cccc|c} 1 & -1 & 0 & 2 & 0 \\ 0 & 1 & 2 & -1 & 1 \\ 0 & 0 & 1 & -1 & 0 \end{array}\right], \tag{2-127b}$$

which is described by the equations

$$x_1 - x_2 + 2x_4 = 0, \tag{2-127c}$$
$$x_2 + 2x_3 - x_4 = 1, \tag{2-127d}$$
$$x_3 - x_4 = 0. \tag{2-127e}$$

Since there are only three equations and four unknowns, we cannot have a unique solution. Nonetheless, we may proceed and obtain a set of solutions by applying

backward substitution to the first three columns of the linear system. Using the third equation, we can obtain:

$$x_3 = x_4. \tag{2-127f}$$

Using the second equation,

$$x_2 + 2x_4 - x_4 = 1,$$

and solving for x_2 gives an expression in terms of x_4:

$$x_2 = 1 - x_4. \tag{2-127g}$$

Finally, using the first equation,

$$x_1 - (1 - x_4) + 2x_4 = 0,$$

and solving for x_1 results in the equation

$$x_1 = 1 - 3x_4. \tag{2-127h}$$

Rearranging these three equations results in the system:

$$x_1 + 3x_4 = 1, \tag{2-127i}$$
$$x_2 + x_4 = 1, \tag{2-127j}$$
$$x_3 - x_4 = 0, \tag{2-127k}$$

with the following augmented form

$$\left[\begin{array}{cccc|c} 1 & 0 & 0 & 3 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 & 0 \end{array}\right]. \tag{2-127l}$$

Note that this is called the *reduced-row echelon form* of the linear system.

We can also allow x_4 to be some free parameter, as in

$$x_4 = \alpha. \tag{2-127m}$$

Substituting this in and writing as a vector system yields the solution:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \alpha \begin{bmatrix} -3 \\ -1 \\ 1 \\ 1 \end{bmatrix}. \tag{2-127n}$$

This describes a line in four dimensional space that contain valid solutions to the linear system of equations.

Example 3

Finally, for the third example, forward elimination yields

$$\left[\begin{array}{ccccc|c} 1 & -1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & 0 & 1 & -1/5 & 3/5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

The final two rows are all zero, providing no information and can be removed,

$$\left[\begin{array}{ccccc|c} 1 & -1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 2 & -1 & 0 & -1 \\ 0 & 0 & 1 & -1/5 & 3/5 & 0 \end{array}\right].$$

We can write an equation for x_3 and solve in terms of x_4 and x_5:

$$x_3 = \frac{1}{5}x_4 - \frac{3}{5}x_5. \tag{2-128a}$$

Solving for x_2 in terms of x_4 and x_5 yields

$$x_2 = -1 + \frac{3}{5}x_4 + \frac{6}{5}x_5. \tag{2-128b}$$

Finally, x_1 in terms of x_4 and x_5 is

$$x_1 = \frac{3}{5}x_4 + \frac{1}{5}x_5. \tag{2-128c}$$

This gives the system

$$x_1 - \frac{3}{5}x_4 - \frac{1}{5}x_5 = 0, \tag{2-128d}$$

$$x_2 - \frac{3}{5}x_4 - \frac{6}{5}x_5 = -1, \tag{2-128e}$$

$$x_3 - \frac{1}{5}x_4 + \frac{3}{5}x_5 = 0. \tag{2-128f}$$

Arranging this into the reduced-row echelon form gives

$$\left[\begin{array}{ccccc|c} 1 & 0 & 0 & -3/5 & -1/5 & 0 \\ 0 & 1 & 0 & -3/5 & -6/5 & -1 \\ 0 & 0 & 1 & -1/5 & 3/5 & 0 \end{array}\right]. \tag{2-128g}$$

There are two free parameters, so define

$$x_4 = 5\alpha, \tag{2-128h}$$

$$x_5 = 5\beta, \tag{2-128i}$$

where α and β are any real number. Putting the equations into vector form, the solution then is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \alpha \begin{bmatrix} 3 \\ 3 \\ 1 \\ 5 \\ 0 \end{bmatrix} + \beta \begin{bmatrix} 1 \\ 6 \\ -3 \\ 0 \\ 5 \end{bmatrix}. \tag{2-128j}$$

This vector equation describes a plane in 5-D space.

2.5.d Discussion of Solutions

To understand the solution to the first problem geometrically, the original matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 2 & -1 \\ 1 & -2 & 0 \\ 3 & 1 & -2 \end{bmatrix}$$

can be thought of as a linear transformation that maps a vector $\mathbf{x}$ onto another vector $\mathbf{b}$. In this case, the right-hand side vector is

$$\mathbf{b} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}.$$

The solution to this mapping is unique and given by Eq. (2-126h) as

$$\mathbf{x} = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}.$$

Multiplying $\mathbf{A}$ by $\mathbf{x}$ yields the vector $\mathbf{b}$. Furthermore, there is only a single mapping of $\mathbf{x}$ onto $\mathbf{b}$, so the linear map given by $\mathbf{A}$ is said to be one-to-one. This also implies that $\mathbf{A}$ is invertible. The solution space is therefore a unique point in 3-D space.

With the second problem, the original matrix

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 & 2 \\ 0 & 0 & 1 & -1 \\ 1 & 0 & 2 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

maps a vector $\mathbf{x}$ onto another vector

$$\mathbf{b} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

and we wish to know which vectors $\mathbf{x}$ would map to this solution $\mathbf{b}$, which are given by the solution in Eq. (2-128j). Examples of vectors that map to $\mathbf{b}$ lie along a line in 4-D space parameterized by a constant α that may take on any real number. Examples of possible solutions are using $\alpha = 0, -1, 2$ respectively,

$$\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 4 \\ 2 \\ -1 \\ -1 \end{bmatrix}, \quad \begin{bmatrix} -5 \\ -1 \\ 2 \\ 2 \end{bmatrix}.$$

All of these vectors, when multiplied by $\mathbf{A}$ yield the same solution vector $\mathbf{b}$. Correspondingly, the transformation in not one-to-one in that there is no unique mapping of $\mathbf{b}$ to $\mathbf{x}$. This implies that $\mathbf{A}$ is not invertible, i.e., $\mathbf{A}^{-1}$ does not exist.

2.5.e Example: Electrical Circuit

Electric circuits are important in applications in nuclear engineering. Nuclear power plants convert heat generated from nuclear reactions into electricity that must be then supplied to the grid. A nuclear power plant itself has numerous electric circuits as part of its operation involving sensors and other equipment that must be designed and analyzed. Radiation detectors also involve electronics that are analyzed using electrical circuits. Furthermore, most experiments involve some form of electrical circuitry.

Electric circuits involving batteries (power sources) and resistors can be described by systems of linear algebraic equations, and the methods of writing these systems of equations is provided in this section. Electric circuits involving capacitors and inductors, conversely, result in linear systems of ordinary differential equations, which is the subject of the next chapter.

To analyze electrical circuits involving batteries and resistors, we apply Kirchoff's rules, which states that the electrical potential (voltage) V is equal to the product of the net electrical current I and the resistance R. The applied voltage and resistances are generally known, whereas the currents I are not.

Consider the following example in Fig. 2.11. We proceed to write Kirchoff's rule for each loop within the system numbered by the index on the subscript of the current in that loop I_i. For the first loop, we write:

$$I_1 R_1 + (I_1 - I_2)R_2 + (I_1 - I_3)R_3 = V. \tag{2-129a}$$

Each term applies to a resistor in the loop, which are $j = 1, 2, 3$. The current is the net electric current with respect to the loop being analyzed. The first term is $I_1 - 0$ since there is no adjacent loop. For the second term, we must subtract off the current in the second loop. And likewise for the third. Finally, the right-hand side has the voltage applied to the circuit V. The second loop has the following equation:

$$(I_2 - I_1)R_2 + (I_2 - I_3)R_4 + I_2 R_5 = 0. \tag{2-129b}$$

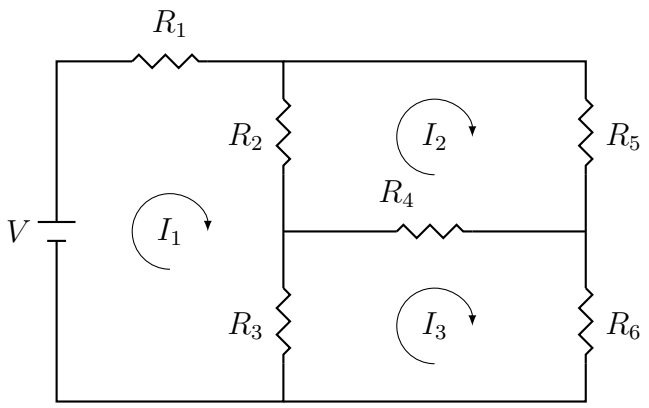

Figure 2.11: Example of an electrical circuit diagram.

As with the first loop, there is a term for each resistor with the net current multiplied by each. The right-hand side is now zero since there is no applied voltage in this section of the loop. The third loop has the equation

$$(I_3 - I_1)R_3 + (I_3 - I_2)R_4 + I_3R_6 = 0. \tag{2-129c}$$

Taking these equations and rearranging to be in terms of the unknown electric currents gives the following linear system:

$$(R_1 + R_2 + R_3)I_1 - R_2I_2 - R_3I_3 = V, \tag{2-129d}$$
$$-R_2I_1 + (R_2 + R_4 + R_5)I_2 - R_4I_3 = 0, \tag{2-129e}$$
$$-R_3I_1 - R_4I_2 + (R_3 + R_4 + R_6)I_3 = 0. \tag{2-129f}$$

The resulting matrix-vector form is

$$\begin{bmatrix} (R_1 + R_2 + R_3) & -R_2 & -R_3 \\ -R_2 & (R_2 + R_4 + R_5) & -R_4 \\ -R_3 & -R_4 & (R_3 + R_4 + R_6) \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \\ I_3 \end{bmatrix} = \begin{bmatrix} V \\ 0 \\ 0 \end{bmatrix}. \tag{2-129g}$$

This system of equations can then be solved using Gaussian elimination.

2.5.f Matrix Inversion

In Sec. 2.1.h we introduced the inverse of a matrix, but did not discuss how it may be computed except for providing some formulas for 2×2 and diagonal matrices. In this section, we will discuss how to use Gaussian elimination to solve for the matrix inverse. Note that formally, we may solve the linear system equations as

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}. \tag{2-130}$$

Therefore, if we know $\mathbf{A}^{-1}$, we could easily solve for $\mathbf{x}$ for any inhomogeneous column vector $\mathbf{b}$. This being said, it is rare, at least for large systems of equations, that

we actually require solving for $\mathbf{A}^{-1}$. Rather Gaussian elimination can be used more efficiently to solve $\mathbf{Ax} = \mathbf{b}$ as we discussed, or, should $\mathbf{A}$ have a specific structure, there may be specialized techniques that can be employed.

Should we actually need to compute $\mathbf{A}^{-1}$, we start by writing the equation of the form

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}. \tag{2-131}$$

Here $\mathbf{A}$ is known and its inverse $\mathbf{A}^{-1}$ is not. As with Gaussian elimination, we apply elementary row operations that can be formally described by $\mathbf{T}_i$ on left side in the order according to the algorithm. This removes the matrix $\mathbf{A}$ on the left-hand side leaving the solution for $\mathbf{A}^{-1}$ on the right. Practically, this is done by constructing the augmented matrix of $\mathbf{A}$ with the identity matrix $\mathbf{I}$

$$\left[\begin{array}{ccccc|ccccc} a_{1,1} & a_{1,2} & \cdots & a_{1,N-1} & a_{1,N} & 1 & 0 & \cdots & 0 & 0 \\ a_{2,1} & a_{2,2} & \cdots & a_{2,N-1} & a_{2,N} & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{N-1,1} & a_{N-1,2} & \cdots & a_{N-1,N-1} & a_{N-1,N} & 0 & 0 & \cdots & 1 & 0 \\ a_{N,1} & a_{N,2} & \cdots & a_{N,N-1} & a_{N,N} & 0 & 0 & \cdots & 0 & 1 \end{array}\right]$$

in a similar manner to the system $\mathbf{Ax} = \mathbf{b}$. We then use the same three allowed operations, swapping, scaling, and replacing an equation with a linear combination of that equation with another, to arrive at the following augmented matrix:

$$\left[\begin{array}{ccccc|ccccc} 1 & 0 & \cdots & 0 & 0 & \widetilde{a}_{1,1} & \widetilde{a}_{1,2} & \cdots & \widetilde{a}_{1,N-1} & \widetilde{a}_{1,N} \\ 0 & 1 & \cdots & 0 & 0 & \widetilde{a}_{2,1} & \widetilde{a}_{2,2} & \cdots & \widetilde{a}_{2,N-1} & \widetilde{a}_{2,N} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & \widetilde{a}_{N-1,1} & \widetilde{a}_{N-1,2} & \cdots & \widetilde{a}_{N-1,N-1} & \widetilde{a}_{N-1,N} \\ 0 & 0 & \cdots & 0 & 1 & \widetilde{a}_{N,1} & \widetilde{a}_{N,2} & \cdots & \widetilde{a}_{N,N-1} & \widetilde{a}_{N,N} \end{array}\right].$$

Here $\widetilde{a}_{i,j}$ are the coefficients of $\mathbf{A}^{-1}$.

As stated previously, algorithm for finding the matrix inverse first uses forward elimination in the same manner as with the $\mathbf{Ax} = \mathbf{b}$ solve before. The difference arises in that backward substitution is replaced with a second "backward" elimination step that mirrors the forward elimination step, except everything is done in reverse. If at any point during the forward elimination process, a row of all zeroes is detected, then the algorithm can stop as the matrix is singular and $\mathbf{A}^{-1}$ does not exist.

As an example, consider the matrix:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & -1 & 2 \\ -1 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & 2 & 1 & -1 \end{bmatrix}. \tag{2-132a}$$

This matrix can be augmented with the identity matrix as follows:

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 2 & 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & -1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 2 & 1 & -1 & 0 & 0 & 0 & 1 \end{array}\right]. \tag{2-132b}$$

Our goal is to turn the left side into the identity matrix where the right side becomes the matrix inverse. To do this algorithmically, we perform forward elimination on the left side. For the first column to make (2,1) and (3,1) zero, replace row ② with the sum of ② and ① and ③ with the sum of ③ and ①:

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 2 & 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & -1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 2 & 1 & -1 & 0 & 0 & 0 & 1 \end{array}\right] : ② \to ② + ① :$$

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 2 & 1 & -1 & 0 & 0 & 0 & 1 \end{array}\right], \tag{2-132c}$$

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 2 & 1 & -1 & 0 & 0 & 0 & 1 \end{array}\right] : ③ \to ③ + ① :$$

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 & 0 & 1 & 0 \\ 0 & 2 & 1 & -1 & 0 & 0 & 0 & 1 \end{array}\right]. \tag{2-132d}$$

Moving onto the second column, make the (4,2) element zero by replacing row ④ with ④ minus twice ②:

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 & 0 & 1 & 0 \\ 0 & 2 & 1 & -1 & 0 & 0 & 0 & 1 \end{array}\right] : ④ \to ④ - 2 \times ② :$$

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 & 0 & 1 & 0 \\ 0 & 0 & 3 & -3 & -2 & -2 & 0 & 1 \end{array}\right]. \tag{2-132e}$$

Moving onto the third column, there is a zero in the (3,3) element, so we move down and swap rows ③ and ④:

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 & 0 & 1 & 0 \\ 0 & 0 & 3 & -3 & -2 & -2 & 0 & 1 \end{array}\right] : ③ \leftrightarrow ④ :$$

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 3 & -3 & -2 & -2 & 0 & 1 \\ 0 & 0 & 0 & 2 & 1 & 0 & 1 & 0 \end{array}\right]. \tag{2-132f}$$

This finishes the forward elimination phase. Next, we do the process in reverse, starting at the last column and working up and to the left. First, scale row ④ by $^1/_2$ and row ③ by $^1/_3$:

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 3 & -3 & -2 & -2 & 0 & 1 \\ 0 & 0 & 0 & 2 & 1 & 0 & 1 & 0 \end{array}\right] : ④ \to {}^1/_2 \times ④ :$$

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 3 & -3 & -2 & -2 & 0 & 1 \\ 0 & 0 & 0 & 1 & {}^1/_2 & 0 & {}^1/_2 & 0 \end{array}\right], \tag{2-132g}$$

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 3 & -3 & -2 & -2 & 0 & 1 \\ 0 & 0 & 0 & 1 & {}^1/_2 & 0 & {}^1/_2 & 0 \end{array}\right] : ③ \to {}^1/_3 \times ③ :$$

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & {}^{-2}/_3 & {}^{-2}/_3 & 0 & {}^1/_3 \\ 0 & 0 & 0 & 1 & {}^1/_2 & 0 & {}^1/_2 & 0 \end{array}\right]. \tag{2-132h}$$

Now moving up the fourth column, replace row ③ with the sum of ③ and ④, row ② with the difference of ② and ④, and ① with ① minus twice ④:

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & {}^{-2}/_3 & {}^{-2}/_3 & 0 & {}^1/_3 \\ 0 & 0 & 0 & 1 & {}^1/_2 & 0 & {}^1/_2 & 0 \end{array}\right] : ③ \to ③ + ④ :$$

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & {}^{-1}/_6 & {}^{-2}/_3 & {}^1/_2 & {}^1/_3 \\ 0 & 0 & 0 & 1 & {}^1/_2 & 0 & {}^1/_2 & 0 \end{array}\right], \tag{2-132i}$$

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & {}^{-1}/_6 & {}^{-2}/_3 & {}^1/_2 & {}^1/_3 \\ 0 & 0 & 0 & 1 & {}^1/_2 & 0 & {}^1/_2 & 0 \end{array}\right] : ② \to ② - ④ :$$

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & {}^1/_2 & 1 & {}^{-1}/_2 & 0 \\ 0 & 0 & 1 & 0 & {}^{-1}/_6 & {}^{-2}/_3 & {}^1/_2 & {}^1/_3 \\ 0 & 0 & 0 & 1 & {}^1/_2 & 0 & {}^1/_2 & 0 \end{array}\right], \tag{2-132j}$$

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & {}^1/_2 & 1 & {}^{-1}/_2 & 0 \\ 0 & 0 & 1 & 0 & {}^{-1}/_6 & {}^{-2}/_3 & {}^1/_2 & {}^1/_3 \\ 0 & 0 & 0 & 1 & {}^1/_2 & 0 & {}^1/_2 & 0 \end{array}\right] : ① \to ① - 2 \times ④ :$$

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 & 1/2 & 1 & -1/2 & 0 \\ 0 & 0 & 1 & 0 & -1/6 & -2/3 & 1/2 & 1/3 \\ 0 & 0 & 0 & 1 & 1/2 & 0 & 1/2 & 0 \end{array}\right]. \tag{2-132k}$$

Moving onto the third column, we can replace row (2) with the sum of (2) and (3) and then also replace row (1) with the sum of (1) and (3):

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 & 1/2 & 1 & -1/2 & 0 \\ 0 & 0 & 1 & 0 & -1/6 & -2/3 & 1/2 & 1/3 \\ 0 & 0 & 0 & 1 & 1/2 & 0 & 1/2 & 0 \end{array}\right] : (2) \to (2) + (3) :$$

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 & -1/6 & -2/3 & 1/2 & 1/3 \\ 0 & 0 & 0 & 1 & 1/2 & 0 & 1/2 & 0 \end{array}\right], \tag{2-132l}$$

$$\left[\begin{array}{cccc|cccc} 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 & -1/6 & -2/3 & 1/2 & 1/3 \\ 0 & 0 & 0 & 1 & 1/2 & 0 & 1/2 & 0 \end{array}\right] : (1) \to (1) + (3) :$$

$$\left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & -1/6 & -2/3 & -1/2 & 1/3 \\ 0 & 1 & 0 & 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 & -1/6 & -2/3 & 1/2 & 1/3 \\ 0 & 0 & 0 & 1 & 1/2 & 0 & 1/2 & 0 \end{array}\right]. \tag{2-132m}$$

Observe that the left side is the identity matrix, so therefore the right side is

$$\mathbf{A}^{-1} = \begin{bmatrix} -1/6 & -2/3 & -1/2 & 1/3 \\ 1/3 & 1/3 & 0 & 1/3 \\ -1/6 & -2/3 & 1/2 & 1/3 \\ 1/2 & 0 & 1/2 & 0 \end{bmatrix} = \frac{1}{6}\begin{bmatrix} -1 & -4 & -3 & 2 \\ 2 & 2 & 0 & 2 \\ -1 & -4 & 3 & 2 \\ 3 & 0 & 3 & 0 \end{bmatrix} \tag{2-132n}$$

2.5.g Tridiagonal Systems

An important class of matrices are those that are tridiagonal. A tridiagonal matrix is zero everywhere except for the diagonal and the elements immediately above and below the diagonal, called the superdiagonal and subdiagonal respectively. This matrix is of the form:

$$\mathbf{A} = \begin{bmatrix} d_1 & u_1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ \ell_2 & d_2 & u_2 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & \ell_3 & d_3 & u_3 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \ell_{N-2} & d_{N-2} & u_{N-2} & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & \ell_{N-1} & d_{N-1} & u_{N-1} \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \ell_N & d_N \end{bmatrix}. \tag{2-133}$$

Here d_i, u_i, and ℓ_i are the diagonal, superdiagonal, and subdiagonal elements of the ith row.

Many linear systems in science and engineering have symmetry and the resulting matrix is symmetric. In this case, there are a series of relatively inexpensive rotations that can be performed to convert the linear system to a tridiagonal one. The other common case is for approximately solving a second-order ordinary differential equation. Suppose we have the equation:

$$\frac{d^2f}{dx^2} + p(x)f(x) = q(x), \quad 0 \leq x \leq a, \quad f(0) = f_\ell, \quad f(a) = f_r, \tag{2-134}$$

where $f(x)$ is the unknown function and $p(x)$ and $q(x)$ are prescribed functions and f_ℓ and f_r are known boundary conditions on the left and right sides of the problem. Unless $p(x)$ takes on a special form, e.g., a constant or monomial in x, there may not be an analytic solution for $f(x)$. We can, however, approximate the solution at discrete points x_i separated by adjacent grid points by distance Δ using approximations to the derivative. We can rewrite this equation at x_i approximately as

$$f(x_1) = f_\ell, \tag{2-135a}$$

$$\frac{f(x_{i-1}) - 2f(x_i) + f(x_{i+1})}{\Delta^2} + p(x_i)f(x_i) = q(x_i)$$

$$f(x_{i-1}) + \left[p(x_i)\Delta^2 - 2\right] f(x_i) + f(x_{i+1}) = q(x_i)\Delta^2, \quad i = 2, \ldots, N-1, \tag{2-135b}$$

$$f(x_N) = f_r. \tag{2-135c}$$

This system of equations can be written as

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 1 & a_2 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 1 & a_3 & 1 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & a_{N-2} & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 & a_{N-1} & 1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_{N-2} \\ f_{N-1} \\ f_N \end{bmatrix} = \begin{bmatrix} f_\ell \\ q_2 \\ q_3 \\ \vdots \\ q_{N-2} \\ q_{N-1} \\ f_r \end{bmatrix}. \tag{2-136}$$

where $a_i = p(x_i)\Delta^2 - 2$, $f_i = f(x_i)$, and $q_i = q(x_i)$.

In a general system, Gaussian elimination can be very inefficient and time consuming, especially for a large system of equations. The number of steps in a full Gaussian elimination solve for a square matrix scales as N^3, where N is the number of columns. It turns out for a tridiagonal system, we can simplify this significantly and the time requirement scales as N.

The method of solving a tridiagonal system can be derived using a simplification

of Gaussian elimination. Consider the augmented tridiagonal system:

$$\left[\begin{array}{cccccccccc} d_1 & u_1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & r_1 \\ \ell_2 & d_2 & u_2 & 0 & \cdots & 0 & 0 & 0 & 0 & r_2 \\ 0 & \ell_3 & d_3 & u_3 & \cdots & 0 & 0 & 0 & 0 & r_3 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \ell_{N-2} & d_{N-2} & u_{N-2} & 0 & r_{N-2} \\ 0 & 0 & 0 & 0 & \cdots & 0 & \ell_{N-1} & d_{N-1} & u_{N-1} & r_{N-1} \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \ell_N & d_N & r_N \end{array}\right].$$

We will use forward elimination to make the subdiagonal terms zero by iterating down the rows. While there is nothing to do for the first row yet, let's define, for notational consistency,

$$\widetilde{d}_1 = d_1, \tag{2-137a}$$

$$\widetilde{r}_1 = r_1. \tag{2-137b}$$

For the second row, we must get the term with ℓ_2 to zero. To do this, we subtract $ell_2/\widetilde{d}_1$ times row 1 from row 2. This eliminates ℓ_2 from the second row,

$$\widetilde{\ell}_2 = \ell_2 - \frac{\ell_2}{\widetilde{d}_1}\widetilde{d}_1 = \ell_2 - \ell_2 = 0, \tag{2-137c}$$

and d_2 and r_2 are modified as follows:

$$\widetilde{d}_2 = d_2 - \frac{\ell_2}{\widetilde{d}_1}u_1, \tag{2-137d}$$

$$\widetilde{r}_2 = r_2 - \frac{\ell_2}{\widetilde{d}_1}\widetilde{r}_1. \tag{2-137e}$$

Note that because the element above u_2 is zero, that entry is unmodified. The augmented matrix now appears as

$$\left[\begin{array}{cccccccccc} \widetilde{d}_1 & u_1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \widetilde{r}_1 \\ 0 & \widetilde{d}_2 & u_2 & 0 & \cdots & 0 & 0 & 0 & 0 & \widetilde{r}_2 \\ 0 & \ell_3 & d_3 & u_3 & \cdots & 0 & 0 & 0 & 0 & r_3 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \ell_{N-2} & d_{N-2} & u_{N-2} & 0 & r_{N-2} \\ 0 & 0 & 0 & 0 & \cdots & 0 & \ell_{N-1} & d_{N-1} & u_{N-1} & r_{N-1} \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \ell_N & d_N & r_N \end{array}\right].$$

We then proceed to do this to the third row, then the fourth, and until we reach row N. The general expression for forward elimination of a tridiagonal matrix is therefore:

$$\widetilde{d}_1 = d_1, \tag{2-138a}$$

$$\widetilde{r}_1 = r_1, \tag{2-138b}$$

$$\widetilde{d}_i = d_i - \frac{\ell_i}{\widetilde{d}_{i-1}} u_{i-1}, \quad i = 2, \ldots, N, \tag{2-138c}$$

$$\widetilde{r}_i = r_i - \frac{\ell_i}{\widetilde{d}_{i-1}} \widetilde{r}_{i-1}, \quad i = 2, \ldots, N. \tag{2-138d}$$

After forward elimination, the augmented matrix is

$$\left[\begin{array}{cccccccccc|c} \widetilde{d}_1 & u_1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \widetilde{r}_1 \\ 0 & \widetilde{d}_2 & u_2 & 0 & \cdots & 0 & 0 & 0 & 0 & \widetilde{r}_2 \\ 0 & 0 & \widetilde{d}_3 & u_3 & \cdots & 0 & 0 & 0 & 0 & \widetilde{r}_3 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & \widetilde{d}_{N-2} & u_{N-2} & 0 & \widetilde{r}_{N-2} \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & \widetilde{d}_{N-1} & u_{N-1} & \widetilde{r}_{N-1} \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \widetilde{d}_N & \widetilde{r}_N \end{array}\right].$$

Now we use backward substitution to solve for the solution vector $\mathbf{x}$:

$$x_N = \frac{\widetilde{r}_N}{\widetilde{d}_N}, \tag{2-139a}$$

$$x_i = \frac{\widetilde{r}_i - u_i x_{i+1}}{\widetilde{d}_i}, \quad i = N-1, \ldots, 1. \tag{2-139b}$$

```
1. initialize vectors td, tr, and x of length N
2. td(1) = d(1), rd(1) = r(1)
3. loop i down rows from 2 to N:
4.   td(i) = d(i) -  u(i-1) * l(i)/td(i-1)
5.   tr(i) = r(i) - tr(i-1) * l(i)/td(i-1)
6. x(N) = tr(N) / td(N)
7. loop i up rows from N-1 to 1:
8.   x(i) = ( tr(i) - u(i)*x(i+1) )/td(i)
9. return x
```

Figure 2.12: Algorithm for solution of a tridiagonal linear system.

The pseudocode is given in Fig. 2.12. Here `td(i)` and `tr(i)` are $\widetilde{d}_i$ and $\widetilde{r}_i$ respectively.

2.6 Iterative Methods

Gaussian elimination is an effective way of solving linear systems, however, it can be very inefficient if there are no simplifications that can be made because of some

structure of the matrix. In general, for a square matrix, the scaling is the cube of the number of rows/columns, or N^3. Many modern engineering calculations involve systems of equations that are in the hundreds of thousands, millions, or more; and in these cases, Gaussian elimination may become impractical. Furthermore, computers use floating point arithmetic and operations induce errors because of numerical roundoff. In the Gaussian elimination algorithm, these errors tend to compound, meaning that solutions may become inaccurate, especially if the matrix or solution vector are of very different magnitudes.

Iterative methods have been developed and are useful for solving the large systems of equations found in many engineering calculations. The two that will be discussed in these notes are the Jacobi iteration and Gauss-Seidel iteration schemes.

The advantage of iterative schemes is that they tend to be more computationally efficient than Gaussian elimination and do not suffer to the same degree from accumulating errors from roundoff involved in floating point arithmetic—although those errors are still present, as they always are from any numerical calculation.

A disadvantage is that we do not solve the system of linear equations, rather each iteration gives a successively better approximation (at least until errors from floating point arithmetic become important). Therefore, it is necessary to define a *convergence criterion* that should be satisfied before the iteration stops. (It is also useful to define a maximum number of iterations in case issues related to errors from floating point arithmetic exceed the convergence criterion, as this guarantees the program stops.) A common choice a convergence criterion checks that the magnitude of the difference between the solution vectors in iterations (k) and $(k+1)$ divided by the magnitude of the updated solution vector is less than a prescribed tolerance:

$$\frac{\left|\mathbf{f}^{(k+1)} - \mathbf{f}^{(k)}\right|}{\left|\mathbf{f}^{(k+1)}\right|} < \epsilon. \tag{2-140}$$

Another common choice checks to ensure the magnitude of the largest error is less than some tolerance:

$$\frac{\left|\max\left\{\mathbf{f}^{(k+1)} - \mathbf{f}^{(k)}\right\}\right|}{\left|\max\left\{\mathbf{f}^{(k+1)}\right\}\right|} < \epsilon. \tag{2-141}$$

Note that some implementations do not divide the convergence measure by the magnitude, i.e. they use absolute versus relative metrics. The advantage of relative convergence criteria is they are mostly independent of the magnitude of the solution vector and therefore more robust measures.

2.6.a Diagonal Dominance and Convergence

Before going into the specifics of the iterative methods, an important consideration is whether or not the iteration scheme will actually converge. It is possible that it may not. (It could stall out or even diverge.) While it is difficult to enumerate all possible cases whether a system of equations with a given iteration scheme will or will not converge, it is possible to state that for the two iteration schemes to be discussed here, convergence is guaranteed if the matrix is *diagonally dominant*.

A matrix is diagonally dominant if for all rows, the magnitude of the diagonal element in a row is greater than the sum of the magnitudes of the off-diagonal elements in that row, i.e.,

$$|a_{i,i}| > \sum_{j=1, j\neq i}^{N} |a_{i,j}|, \text{ for all rows } i. \tag{2-142}$$

If a matrix is *not* diagonally dominant, the iterative algorithm may still converge for a given matrix; however, this convergence cannot be assured. Another useful property is that if a matrix is diagonally dominant, then that matrix is also invertible. (This does not imply the converse: an invertible matrix is not necessarily diagonally dominant.)

2.6.b Jacobi Iteration

The Jacobi iteration method is simplest of the iterative methods. First, we break a matrix A into the sum of a diagonal matrix D, an upper triangular matrix $\mathbf{U}$, and a lower triangular matrix $\mathbf{L}$:

$$\mathbf{A} = \mathbf{D} + \mathbf{U} + \mathbf{L}; \tag{2-143}$$

these are:

$$\mathbf{D} = \begin{bmatrix} a_{1,1} & 0 & 0 & 0 & \cdots \\ 0 & a_{2,2} & 0 & 0 & \cdots \\ 0 & 0 & a_{3,3} & 0 & \cdots \\ 0 & 0 & 0 & a_{4,4} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{2-144a}$$

$$\mathbf{U} = \begin{bmatrix} 0 & a_{1,2} & a_{1,3} & a_{1,4} & \cdots \\ 0 & 0 & a_{2,3} & a_{2,4} & \cdots \\ 0 & 0 & 0 & a_{3,4} & \cdots \\ 0 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{2-144b}$$

$$\mathbf{L} = \begin{bmatrix} 0 & 0 & 0 & 0 & \cdots \\ a_{2,1} & 0 & 0 & 0 & \cdots \\ a_{3,1} & a_{3,2} & 0 & 0 & \cdots \\ a_{4,1} & a_{4,2} & a_{4,3} & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{2-144c}$$

Given these definitions, one can rearrange $\mathbf{Ax} = \mathbf{b}$ as

$$\mathbf{Dx} = -(\mathbf{L} + \mathbf{U})\mathbf{x} + \mathbf{b}. \tag{2-145}$$

Next, we introduce iteration indices on $\mathbf{x}$ as follows:

$$\mathbf{Dx}^{(k+1)} = -(\mathbf{L} + \mathbf{U})\mathbf{x}^{(k)} + \mathbf{b}. \tag{2-146}$$

Here iteration indices are denoted with superscripts in parentheses. It is necessary to provide a guess for $\mathbf{x}^{(0)}$ and usually guessing the zero vector will suffice, unless there is information about an approximate form of $\mathbf{x}$ available. For the duration of the iteration, the right-hand side is held constant, and an updated value of the solution vector $\mathbf{x}$ is obtained by solving the resulting linear system. This new value of $\mathbf{x}$ is then used on the right-hand side and the process continues until some convergence criterion is satisfied. Explicitly, the iteration is

$$\begin{aligned}
\mathbf{D}\mathbf{x}^{(1)} &= \mathbf{b}, \\
\mathbf{D}\mathbf{x}^{(2)} &= -(\mathbf{L}+\mathbf{U})\mathbf{x}^{(1)} + \mathbf{b}, \\
\mathbf{D}\mathbf{x}^{(3)} &= -(\mathbf{L}+\mathbf{U})\mathbf{x}^{(2)} + \mathbf{b}, \\
&\vdots
\end{aligned}$$

Each iteration is very simple since the matrix $\mathbf{D}$ is diagonal. If the right-hand side for the kth iteration is $\mathbf{r}^{(k)}$ then

$$x_i^{(k+1)} = \frac{r_i^{(k)}}{a_{i,i}}. \tag{2-147}$$

To illustrate the idea, consider the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$:

$$\begin{bmatrix} 3 & 1 & 1 & 0 & 0 & 0 \\ 0 & 5 & 0 & 1 & 1 & 2 \\ 2 & 0 & 4 & 0 & 0 & 1 \\ 0 & 1 & 0 & 3 & 1 & 0 \\ 1 & 1 & 1 & 1 & 6 & 1 \\ 0 & 0 & 1 & 0 & 2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} 1 \\ -8 \\ 6 \\ 2 \\ 0 \\ -6 \end{bmatrix}. \tag{2-148a}$$

First, we can check that the system is indeed invertible by checking if $\det(\mathbf{A}) = 0$, and indeed it is. Next, we can verify diagonal dominance: for row (1), the diagonal is 3 and the sum of the magnitudes of off-diagonal terms is 2; for row (2), the diagonal is54 and the off-diagonal magnitude is 4; for row (3), these are 4 and 4; row (4), 3 and 2; row (5), 6 and 5; and finally row (6) is 4 and 3. Since all diagonal magnitudes of each row exceed the magnitude of the off-diagonal elements, we can guarantee convergence.

We can then define the matrices:

$$\mathbf{D} = \begin{bmatrix} 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 \end{bmatrix}, \tag{2-148b}$$

and

$$\mathbf{L}+\mathbf{U}=\begin{bmatrix}0&1&1&0&0&0\\0&0&0&1&1&2\\2&0&0&0&0&1\\0&1&0&0&1&0\\1&1&1&1&0&1\\0&0&1&0&2&0\end{bmatrix} \tag{2-148c}$$

If we guess $\mathbf{x}^{(0)}=\mathbf{0}$, the zero vector, then the right-hand side for the first iteration becomes:

$$\begin{bmatrix}3&0&0&0&0&0\\0&5&0&0&0&0\\0&0&4&0&0&0\\0&0&0&3&0&0\\0&0&0&0&6&0\\0&0&0&0&0&4\end{bmatrix}\begin{bmatrix}x_1^{(1)}\\x_2^{(1)}\\x_3^{(1)}\\x_4^{(1)}\\x_5^{(1)}\\x_6^{(1)}\end{bmatrix}=\begin{bmatrix}1\\-8\\6\\2\\0\\-6\end{bmatrix}. \tag{2-148d}$$

Since the matrix is diagonal we can easily find $\mathbf{x}^{(1)}$:

$$\begin{bmatrix}x_1^{(1)}\\x_2^{(1)}\\x_3^{(1)}\\x_4^{(1)}\\x_5^{(1)}\\x_6^{(1)}\end{bmatrix}=\begin{bmatrix}1/3\\-8/5\\3/2\\2/3\\0\\-3/2\end{bmatrix}. \tag{2-148e}$$

Proceeding to find the right-hand side for the second iteration:

$$-(\mathbf{L}+\mathbf{U})\mathbf{x}^{(1)}=\begin{bmatrix}0&-1&-1&0&0&0\\0&0&0&-1&-1&-2\\-2&0&0&0&0&-1\\0&-1&0&0&-1&0\\-1&-1&-1&-1&0&-1\\0&0&-1&0&-2&0\end{bmatrix}\begin{bmatrix}1/3\\-8/5\\3/2\\2/3\\0\\-3/2\end{bmatrix}=\begin{bmatrix}1/10\\7/3\\5/6\\8/5\\3/5\\-3/2\end{bmatrix}. \tag{2-148f}$$

Adding on the solution vector $\mathbf{b}$ gives the right-hand side and the system

$$\begin{bmatrix}3&0&0&0&0&0\\0&5&0&0&0&0\\0&0&4&0&0&0\\0&0&0&3&0&0\\0&0&0&0&6&0\\0&0&0&0&0&4\end{bmatrix}\begin{bmatrix}x_1^{(2)}\\x_2^{(2)}\\x_3^{(2)}\\x_4^{(2)}\\x_5^{(2)}\\x_6^{(2)}\end{bmatrix}=\begin{bmatrix}11/10\\-17/3\\41/6\\18/5\\3/5\\-15/2\end{bmatrix}. \tag{2-148g}$$

Solving the system again gives the solution for the second iteration, $\mathbf{x}^{(2)}$:

$$\begin{bmatrix} x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \\ x_4^{(2)} \\ x_5^{(2)} \\ x_6^{(2)} \end{bmatrix} \approx \begin{bmatrix} 0.3667 \\ -1.1333 \\ 1.7083 \\ 1.2000 \\ 0.1000 \\ -1.8750 \end{bmatrix}. \tag{2-148h}$$

Repeating the process gives $\mathbf{x}^{(3)}$:

$$\begin{bmatrix} x_1^{(3)} \\ x_2^{(3)} \\ x_3^{(3)} \\ x_4^{(3)} \\ x_5^{(3)} \\ x_6^{(3)} \end{bmatrix} \approx \begin{bmatrix} 0.1417 \\ -1.1100 \\ 1.0111 \\ -0.0444 \\ 0.1000 \\ -1.9771 \end{bmatrix}, \tag{2-148i}$$

and again to get $\mathbf{x}^{(4)}$:

$$\begin{bmatrix} x_1^{(4)} \\ x_2^{(4)} \\ x_3^{(4)} \\ x_4^{(4)} \\ x_5^{(4)} \\ x_6^{(4)} \end{bmatrix} \approx \begin{bmatrix} 0.1082 \\ -1.0025 \\ 1.9234 \\ 1.0515 \\ 0.0248 \\ -1.9241 \end{bmatrix}. \tag{2-148j}$$

If we do this several more times, we can observe the result of tenth iteration,

$$\begin{bmatrix} x_1^{(10)} \\ x_2^{(10)} \\ x_3^{(10)} \\ x_4^{(10)} \\ x_5^{(10)} \\ x_6^{(10)} \end{bmatrix} \approx \begin{bmatrix} 0.0071 \\ -0.9957 \\ 2.0019 \\ 1.0058 \\ 0.0054 \\ -1.9933 \end{bmatrix}. \tag{2-148k}$$

This process continues until we decide that our result is "close enough" to the exact answer, which in this case is

$$\mathbf{x} = \begin{bmatrix} 0 \\ -1 \\ 2 \\ 1 \\ 0 \\ -2 \end{bmatrix}. \tag{2-148l}$$

If we write the Jacobi algorithm on a computer and iterate until the error $\epsilon < 10^{-15}$, we see that it requires 118 iterations.

While the iterations are, in principle, straightforward, there are a few issues. First, the rate of convergence for Jacobi may be slow. In other words, the algorithm may require numerous iterations and this is almost always significantly greater than for Gauss-Seidel iteration. Second, performing the matrix multiplication involved in $(\mathbf{L}+\mathbf{U})\mathbf{x}^{(k)}$ can be computationally expensive unless there is some structure or sparsity to the matrix. For the case where $\mathbf{A}$ is dense (few sections of zeroes), it could be that the Jacobi iteration is actually *slower* than Gaussian elimination. Thankfully, there exists a more efficient algorithm called Gauss-Seidel that will be discussed now.

2.6.c Gauss-Seidel Iteration

Gauss-Seidel differs from Jacobi in how the matrices are split to the left- and right-hand sides. For Gauss-Seidel, the iteration scheme is as follows:

$$(\mathbf{L}+\mathbf{D})\mathbf{x}^{(k+1)} = -\mathbf{U}\mathbf{x}^{(k)} + \mathbf{b}. \tag{2-149}$$

The matrix on the left-hand side is now a lower-triangular matrix. For a fixed right-hand side, the linear system may be solved with backward substitution alone, except now this is done descending the rows as opposed to the typical ascending.

At first glance, it may seem that worse than Jacobi iteration because now each iteration is more computationally involved, but it is actually better. The most important feature is that the number of iterations required to get a specific level of precision is fewer. Informally, the reason for this is that each iteration, by using the lower triangular form, is propagating more information per iteration than Jacobi. An additional benefit is that since $\mathbf{U}$ is all zeroes on its lower triangle, the matrix multiplication for $\mathbf{U}\mathbf{x}^{(k)}$ can be performed with few operations than the analogous operation in Jacobi. While it is true that each now requires a backward substitution, this increased cost is almost always more than offset by the rate of information gained each iteration as well as the simpler matrix multiply on the right-hand side.

Revisiting our example from the Jacobi iteration, we again wish to solve the linear system

$$\begin{bmatrix} 3 & 1 & 1 & 0 & 0 & 0 \\ 0 & 5 & 0 & 1 & 1 & 2 \\ 2 & 0 & 4 & 0 & 0 & 1 \\ 0 & 1 & 0 & 3 & 1 & 0 \\ 1 & 1 & 1 & 1 & 6 & 1 \\ 0 & 0 & 1 & 0 & 2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} 1 \\ -8 \\ 6 \\ 2 \\ 0 \\ -6 \end{bmatrix}, \tag{2-150a}$$

but this time using Gauss-Seidel iteration. For this we define:

$$\mathbf{L}+\mathbf{D}=\begin{bmatrix}3&0&0&0&0&0\\0&5&0&0&0&0\\2&0&4&0&0&0\\0&1&0&3&0&0\\1&1&1&1&6&0\\0&0&1&0&2&4\end{bmatrix} \tag{2-150b}$$

and

$$\mathbf{U}=\begin{bmatrix}0&1&1&0&0&0\\0&0&0&1&1&2\\0&0&0&0&0&1\\0&0&0&0&1&0\\0&0&0&0&0&1\\0&0&0&0&0&0\end{bmatrix}. \tag{2-150c}$$

As before, our initial guess is $\mathbf{x}^{(0)}=\mathbf{0}$, the zero vector. The first iteration satisfies the linear system:

$$\begin{bmatrix}3&0&0&0&0&0\\0&5&0&0&0&0\\2&0&4&0&0&0\\0&1&0&3&0&0\\1&1&1&1&6&0\\0&0&1&0&2&4\end{bmatrix}\begin{bmatrix}x_1^{(1)}\\x_2^{(1)}\\x_3^{(1)}\\x_4^{(1)}\\x_5^{(1)}\\x_6^{(1)}\end{bmatrix}=\begin{bmatrix}1\\-8\\6\\2\\0\\-6\end{bmatrix}. \tag{2-150d}$$

Since the matrix is triangular, we can solve this system with backwards substitution going down the rows:

$$x_1^{(1)}=\frac{1}{3}; \tag{2-150e}$$

$$x_2^{(1)}=-\frac{8}{5}; \tag{2-150f}$$

$$2\left(\frac{1}{3}\right)+4x_3^{(1)}=6,$$

$$x_3^{(1)}=\frac{4}{3}; \tag{2-150g}$$

$$\left(-\frac{8}{5}\right)+3x_4^{(1)}=2,$$

$$x_4^{(1)}=\frac{6}{5}; \tag{2-150h}$$

$$\left(\frac{1}{3}\right)+\left(-\frac{8}{5}\right)+\left(\frac{4}{3}\right)+\left(\frac{6}{5}\right)+6x_5^{(1)}=0,$$

$$x_5^{(1)}=-\frac{19}{90}; \tag{2-150i}$$

$$\left(\frac{4}{3}\right) + 2\left(-\frac{19}{90}\right) + 4x_6^{(1)} = -6,$$
$$x_6^{(1)} = -\frac{221}{180}. \tag{2-150j}$$

Plugging in these results into the right-hand side for the second iteration gives:

$$\begin{bmatrix} 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 & 0 \\ 2 & 0 & 4 & 0 & 0 & 0 \\ 0 & 1 & 0 & 3 & 0 & 0 \\ 1 & 1 & 1 & 1 & 6 & 0 \\ 0 & 0 & 1 & 0 & 2 & 4 \end{bmatrix} \begin{bmatrix} x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \\ x_4^{(2)} \\ x_5^{(2)} \\ x_6^{(2)} \end{bmatrix} = \begin{bmatrix} 0 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 & -2 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} {}^{1}/_{3} \\ {}^{-8}/_{5} \\ {}^{4}/_{3} \\ {}^{6}/_{5} \\ {}^{-19}/_{90} \\ {}^{-221}/_{180} \end{bmatrix} + \begin{bmatrix} 1 \\ -8 \\ 6 \\ 2 \\ 0 \\ -6 \end{bmatrix}$$
$$\approx \begin{bmatrix} 1.2667 \\ -5.5333 \\ 2.2111 \\ 1.7278 \\ 1.7278 \\ -6 \end{bmatrix}. \tag{2-150k}$$

Doing backward substitution again gives the result of the second iteration:

$$\mathbf{x}^{(2)} = \begin{bmatrix} 0.4222 \\ -1.1067 \\ 1.7208 \\ 1.1059 \\ -0.0691 \\ -1.8957 \end{bmatrix}. \tag{2-150l}$$

Repeating this for the third iteration:

$$\mathbf{x}^{(3)} = \begin{bmatrix} 0.1286 \\ -1.0491 \\ 1.9096 \\ 1.0394 \\ -0.0221 \\ -1.9663 \end{bmatrix}. \tag{2-150m}$$

And again for the fourth iteration:

$$\mathbf{x}^{(4)} = \begin{bmatrix} 0.0465 \\ -1.0169 \\ 1.9683 \\ 1.0130 \\ -0.0074 \\ -1.9884 \end{bmatrix}. \tag{2-150n}$$

If we keep iterating, we get after the tenth iteration:

$$\mathbf{x}^{(10)} = \begin{bmatrix} 7.952 \times 10^{-5} \\ -1.0000 \\ 1.9999 \\ 1.0000 \\ -1.2742 \times 10^{-5} \\ -2.0000 \end{bmatrix}. \tag{2-150o}$$

This is getting very close to the reference solution (only four digits of precision are displayed, so the error is less than that), which is, again,

$$\mathbf{x} = \begin{bmatrix} 0 \\ -1 \\ 2 \\ 1 \\ 0 \\ -2 \end{bmatrix}. \tag{2-150p}$$

If we continue iterating, the iteration convergence to a tolerance of $\epsilon < 10^{-15}$ in 35 iterations. This is significantly faster than what was observed for the Jacobi iteration, which requires 118 iterations to get to the same level of convergence. This is typical for these two iteration schemes and, generally speaking, Gauss-Seidel (or some more advanced method) is used in practice for this reason.

Gauss-Seidel is the last iteration method we will discuss in these notes, but it is worth mentioning an enhancement on Gauss-Seidel is called *successive over-relaxation*. The general idea is to write the system iteration as

$$(\omega\mathbf{L} + \mathbf{D})\mathbf{x}^{(k+1)} = -(\omega\mathbf{U} + (\omega - 1)\mathbf{D})\mathbf{x}^{(k)} + \omega\mathbf{b}. \tag{2-151}$$

Here ω is an over-relaxation parameter that is usually $1 < \omega < 2$. Values of $\omega \geq 2$ can lead to the iteration diverging. A good choice of ω will cause the linear system to converge at a significantly faster rate than standard Gauss-Seidel. Unfortunately, it is difficult to predict ahead of time for a given linear system which values of ω will lead to improvements in the convergence rates, let alone finding the optimal one, and usually this involves a bit of hand tuning for the application of interest.

2.7 Eigenvalues and Eigenvectors

The final major topic of this chapter are a special type of linear system that leads to eigenvalues and eigenvectors. These eigenvalues and eigenvectors have numerous applications in nuclear engineering such as: defining the concept of nuclear criticality, finding the energy levels and wavefunctions in quantum mechanical systems, radioactive decay, understanding vibrations in nuclear systems, describing oscillatory behavior in fusion plasmas, and many more. It is, as we will see in a future chapter, important for obtaining solutions to systems of linear ordinary differential

equations. We will see going forward that the space of solutions for such a problem can be described by a linear combination of eigenvectors. In the case where the largest eigenvalue is real, this along with its eigenvector provides information about its equilibrium behavior.

The eigenvalue problem is defined as

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}, \tag{2-152}$$

where λ is a scalar constant. In the context of linear algebra, $\mathbf{A}$ is a square matrix; however, in physics this can be any combination of linear operators and is often shorthand for describing a partial differential equation with the matrix $\mathbf{A}$ being an approximate description of the continuous operators from calculus.

To understand Eq. (2-152), recall that a matrix acting on a vector can be viewed as a linear map that takes a vector (or set of vectors) to another set of vectors. The eigenvalue problem describes the case where when $\mathbf{A}$ is applied to a specific vector $\mathbf{x}$ that we call the eigenvector, we get that same vector scaled by a multiplicative constant λ, the eigenvalue corresponding to that eigenvector. In other words, the eigenvectors describe the axes that are invariant under the application of $\mathbf{A}$ to within a multiplicative scaling factor λ, or the eigenvalue.

To illustrate this concept, consider the matrix $\mathbf{A}$,

$$\mathbf{A} = \begin{bmatrix} 4 & -1 \\ 2 & 1 \end{bmatrix}.$$

Using techniques we will discuss, we can find that the eigenvalues are $\lambda = 3, 2$ with the respective eigenvectors $[1\ 1]^\top$ and $[1\ 2]^\top$ to within a multiplicative scaling constant. Let us consider two three vectors. Let $\mathbf{u}$ be the unit basis vector in the x direction and the vectors $\mathbf{v}$ and $\mathbf{v}_2$ be the eigenvectors corresponding to the eigenvalue $\lambda = 3$ and 2 respectively. Let us multiply $\mathbf{A}$ onto each of the vectors and plot the results in Fig. 2.13. Notice that $\mathbf{A}$ applies a rotation and stretching to $\mathbf{u}$ whereas $\mathbf{A}$ only stretches $\mathbf{v}_1$ by a factor of $\lambda = 3$, and $\mathbf{v}_2$ by a factor of $\lambda = 2$.

One thing to note is that eigenvectors are unique up to an arbitrary multiplicative scaling constant. So in the example, $[1/2\ 1/2]^\top$ and $[-3\ -6]^\top$ are also eigenvectors. Also, many matrices we encounter in practical applications have entries that are real and symmetric or are Hermitian. In these cases, the eigenvalues are real and the eigenvectors are orthogonal, which can be then used to form an orthogonal basis. (As seen in the example, this is not true of a general matrix.)

Next, we will discuss computing the eigenvalues and eigenvectors using analytical techniques. Then, we will show how to decompose a matrix into a product of matrices containing the eigenvalues and eigenvectors. This decomposition permits numerous applications including the evaluation of a function of a matrix and the solution of differential equations. This section will briefly touch upon an iterative numerical method (there are unfortunately no direct methods like Gaussian elimination for computing eigenvalues) for computing a single dominant (largest in magnitude) eigenvalue and eigenvector pair. Numerical methods for computing the entire set of eigenvalues and eigenvectors are very complicated and involve multiple steps; these

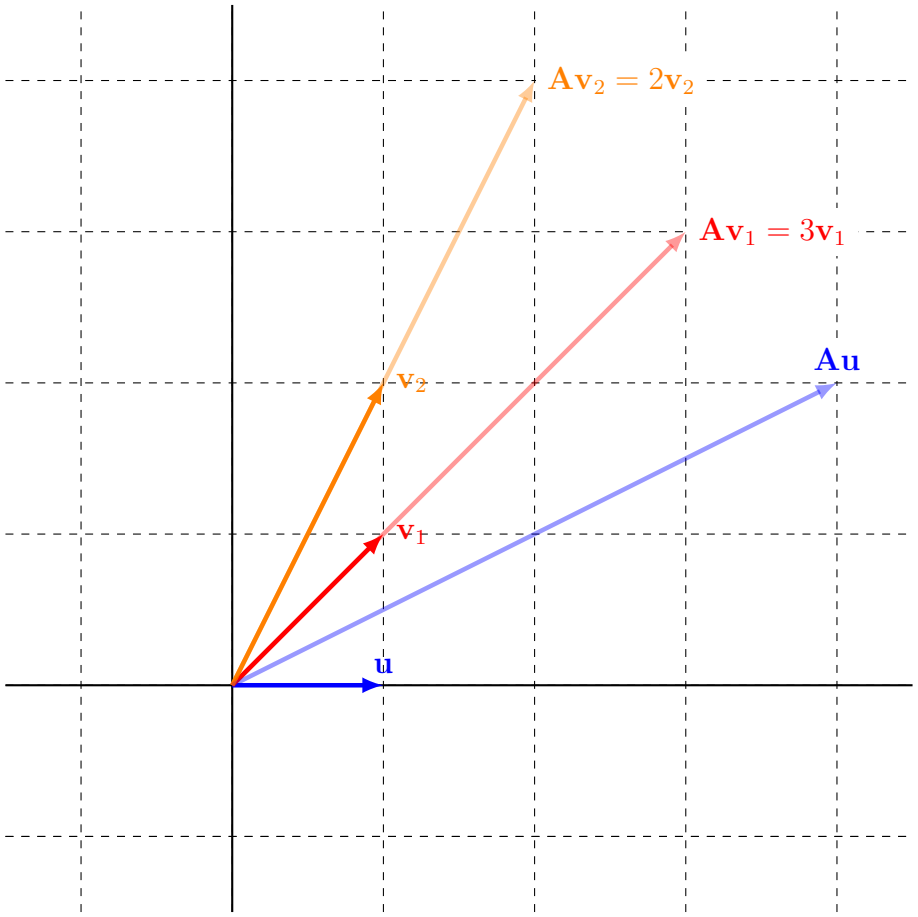

Figure 2.13: Illustration of a the application of a matrix on a non-eigenvector $\mathbf{u}$ and an eigenvector $\mathbf{v}$.

are therefore outside the scope of this text and left for a more advanced course on numerical linear algebra.

2.7.a Calculating Eigenvalues

To compute the eigenvalues, we solve the system:

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}. \tag{2-153}$$

We know that for $\mathbf{x}$ to be non-zero, then the determinant of $\mathbf{A} - \lambda\mathbf{I}$ must be zero. Therefore, we compute:

$$\begin{vmatrix} a_{1,1} - \lambda & a_{1,2} & \cdots & a_{1,N-1} & a_{1,N} \\ a_{2,1} & a_{2,2} - \lambda & \cdots & a_{2,N-1} & a_{2,N} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{N-1,1} & a_{N-1,2} & \cdots & a_{N-1,N-1} - \lambda & a_{N-1,N} \\ a_{N,1} & a_{N,2} & \cdots & a_{N,N-1} & a_{N,N} - \lambda \end{vmatrix} = 0. \tag{2-154}$$

This yields, a polynomial up to degree N in unknown λ, which has N roots or eigenvalues. For $N > 4$, usually the roots λ will have to be found numerically. (Rarely is numerical root finding used for large values of N, as there are more robust iterative methods available.)

An important special case is when $\mathbf{A}$ is a triangular matrix, which shows up frequently enough (e.g., in radioactive decay problems) to merit its own discussion. When $\mathbf{A}$ can be written in upper triangular form,

$$\begin{vmatrix} a_{1,1}-\lambda & 0 & \cdots & 0 & 0 \\ a_{2,1} & a_{2,2}-\lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{N-1,1} & a_{N-1,2} & \cdots & a_{N-1,N-1}-\lambda & 0 \\ a_{N,1} & a_{N,2} & \cdots & a_{N,N-1} & a_{N,N}-\lambda \end{vmatrix} = 0, \tag{2-155}$$

we can easily evaluate the determinant. As we decompose the matrix from an $N\times N$ determinant, to a $(N-1)\times(N-1)$, to a $(N-2)\times(N-2)$, and so on, we can observe that in all cases all but the first element of the first row are zero. This yields a product

$$\det(\mathbf{A}-\lambda\mathbf{I}) = (a_{1,1}-\lambda)(a_{2,2}-\lambda)\cdots(a_{N,N}-\lambda) = 0, \tag{2-156}$$

which means the roots λ are the diagonal elements of the matrix. (As we will see, these corresponds to the radioactive decay constants.)

To show this in practice, let us do a few examples.

Example 1

First, consider the matrix

$$\mathbf{A} = \begin{bmatrix} 5 & 4 & -2 \\ 4 & 5 & 2 \\ 0 & 2 & -2 \end{bmatrix}. \tag{2-157a}$$

To find the eigenvectors of $\mathbf{A}$, we take the determinant of $\mathbf{A}-\lambda\mathbf{I} = 0$ and solve for λ:

$$\begin{aligned} \det(\mathbf{A}-\lambda\mathbf{I}) &= \begin{vmatrix} 5-\lambda & 4 & -2 \\ 4 & 5-\lambda & 2 \\ 0 & 2 & -2-\lambda \end{vmatrix} = 0, \\ &= (5-\lambda)\left[(5-\lambda)(-2-\lambda)-4\right] \\ &\quad - 4\left[4(-2-\lambda)-0\right] \\ &\quad + (-2)\left(8-0\right) = 0, \\ &= \lambda^3 - 8\lambda^2 - 15\lambda + 54 = 0. \end{aligned} \tag{2-157b}$$

Solving for the roots λ of this cubic polynomial gives the result:

$$\lambda = 9, 2, -3. \tag{2-157c}$$

Example 2

Next, consider the matrix

$$\mathbf{A} = \begin{bmatrix} 3 & -1 & 2 \\ 3 & -1 & 6 \\ -2 & 2 & -2 \end{bmatrix}. \tag{2-158a}$$

To find the eigenvectors of $\mathbf{A}$, we again take the determinant of $\mathbf{A} - \lambda\mathbf{I} = 0$ and solve for λ:

$$\begin{aligned} \det(\mathbf{A} - \lambda\mathbf{I}) &= \begin{vmatrix} 3-\lambda & -1 & 2 \\ 3 & -1-\lambda & 6 \\ -2 & 2 & -2-\lambda \end{vmatrix} = 0, \\ &= (3-\lambda)\left[(-1-\lambda)(-2-\lambda) - 12\right] \\ &\quad - (-1)\left[3(-2-\lambda) - (-12)\right] \\ &\quad + 2\left[6 - (-2)(-1-\lambda)\right] = 0, \\ &= \lambda^3 - 12\lambda - 16 = 0. \end{aligned} \tag{2-158b}$$

Solving for the roots λ gives:

$$\lambda = -4, 2, 2. \tag{2-158c}$$

Note that here we have repeated eigenvalues.

Example 3

Finally, consider a third matrix

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 0 \\ 0 & 1 & 1 \\ -1 & 1 & 2 \end{bmatrix}. \tag{2-159a}$$

To find the eigenvectors of $\mathbf{A}$, we again take the determinant of $\mathbf{A} - \lambda\mathbf{I} = 0$ and solve for λ:

$$\begin{aligned} \det(\mathbf{A} - \lambda\mathbf{I}) &= \begin{vmatrix} 2-\lambda & -1 & 0 \\ 0 & 1-\lambda & 1 \\ -1 & 1 & 2-\lambda \end{vmatrix} = 0, \\ &= (2-\lambda)\left[((1-\lambda)(-2-\lambda) - 1\right] + 1, \\ &= \lambda^3 - 5\lambda^2 + 7\lambda - 3 = 0. \end{aligned} \tag{2-159b}$$

Solving for the roots λ gives:

$$\lambda = 3, 1, 1. \tag{2-159c}$$

As with the previous example, this has a repeated root.

2.7.b Calculating Eigenvectors

To find the eigenvectors, we solve the linear system

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0} \tag{2-160}$$

for each value of λ. Since the determinant of the matrix $(\mathbf{A} - \lambda\mathbf{I})$ is zero, the system does not have a unique solution and will have a rank that is less than the number of rows. All solutions may therefore be scaled by an arbitrary multiplicative constant. To illustrate the idea, let us revisit our examples.

Example 1

The first example matrix

$$\mathbf{A} = \begin{bmatrix} 5 & 4 & -2 \\ 4 & 5 & 2 \\ 0 & 2 & -2 \end{bmatrix}. \tag{2-161a}$$

has three eigenvalues: $\lambda_1 = 9, \lambda_2 = 2, \lambda_3 = -3$, which we index in descending order. Evaluating $(\mathbf{A} - \lambda\mathbf{I})$ for $\lambda_1 = 9$ gives the matrix

$$\mathbf{A} - 9\mathbf{I} = \begin{bmatrix} -4 & 4 & -2 \\ 4 & -4 & 2 \\ 0 & 2 & -11 \end{bmatrix}. \tag{2-161b}$$

Now we use Gaussian elimination to put the matrix into a reduced-row echelon form. First, replace row (2) by the sum of rows (2) and (1):

$$\begin{bmatrix} -4 & 4 & -2 \\ 4 & -4 & 2 \\ 0 & 2 & -11 \end{bmatrix} : (2) \to (2) + (1) : \begin{bmatrix} -4 & 4 & -2 \\ 0 & 0 & 0 \\ 0 & 2 & -11 \end{bmatrix} \tag{2-161c}$$

Next, swap rows (2) and (3):

$$\begin{bmatrix} -4 & 4 & -2 \\ 0 & 0 & 0 \\ 0 & 2 & -11 \end{bmatrix} : (2) \leftrightarrow (3) : \begin{bmatrix} -4 & 4 & -2 \\ 0 & 2 & -11 \\ 0 & 0 & 0 \end{bmatrix}. \tag{2-161d}$$

Now to eliminate the (1,2) element, replace row (1) with the difference of rows (1) and twice (2):

$$\begin{bmatrix} -4 & 4 & -2 \\ 0 & 2 & -11 \\ 0 & 0 & 0 \end{bmatrix} : (1) \to (1) - 2 \times (2) : \begin{bmatrix} -4 & 0 & 20 \\ 0 & 2 & -11 \\ 0 & 0 & 0 \end{bmatrix}. \tag{2-161e}$$

Since we have a row of all zeros, we have one free parameter. Define

$$x_3 = \alpha. \tag{2-161f}$$

Then, using the second row:

$$\begin{aligned} 2x_2 - 11\alpha &= 0, \\ x_2 &= \frac{11}{2}\alpha. \end{aligned} \tag{2-161g}$$

Finally, using the first row:

$$\begin{aligned} -4x_1 + 20\alpha &= 0, \\ x_1 &= 5\alpha. \end{aligned} \tag{2-161h}$$

Writing out the solution vector:

$$\mathbf{x} = \alpha \begin{bmatrix} 5 \\ {}^{11}/_{2} \\ 1 \end{bmatrix} \rightarrow \alpha \begin{bmatrix} 10 \\ 11 \\ 2 \end{bmatrix}. \tag{2-161i}$$

Since α is an arbitrary constant, we may redefine α to scale the results to get the solution vector in terms of whole numbers. This implies the eigenvector corresponding to $\lambda = 9$ is

$$\mathbf{v}_1 = \begin{bmatrix} 10 \\ 11 \\ 2 \end{bmatrix}. \tag{2-161j}$$

Moving onto $\lambda_2 = 2$, the matrix

$$\mathbf{A} - 2\mathbf{I} = \begin{bmatrix} 3 & 4 & -2 \\ 4 & 3 & 2 \\ 0 & 2 & -4 \end{bmatrix}. \tag{2-161k}$$

After performing Gaussian elimination we obtain the matrix

$$\begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \end{bmatrix}. \tag{2-161l}$$

As before, define

$$\alpha = x_3; \tag{2-161m}$$

then,

$$\begin{aligned} x_2 - 2\alpha &= 0, \\ x_2 &= 2\alpha, \end{aligned} \tag{2-161n}$$

and

$$x_1 + 2\alpha = 0,$$

$$x_1 = -2\alpha. \tag{2-161o}$$

Therefore, the solution vector is

$$\mathbf{x} = \alpha \begin{bmatrix} -2 \\ 2 \\ 1 \end{bmatrix} \tag{2-161p}$$

and the eigenvector corresponding to $\lambda_2 = 2$ is

$$\mathbf{v}_2 = \begin{bmatrix} -2 \\ 2 \\ 1 \end{bmatrix}. \tag{2-161q}$$

Repeating the same procedure for $\lambda = -3$, we find the corresponding eigenvector

$$\mathbf{v}_3 = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}. \tag{2-161r}$$

Example 2

For the second example, recall the matrix

$$\mathbf{A} = \begin{bmatrix} 3 & -1 & 2 \\ 3 & -1 & 6 \\ -2 & 2 & -2 \end{bmatrix} \tag{2-162a}$$

has eigenvalues $\lambda_1 = -4, \lambda_2 = 2, \lambda_3 = 2$ with one of these repeated. For $\lambda_1 = -4$, the process is identical to the previous example and we obtain the eigenvector

$$\mathbf{v}_1 = \begin{bmatrix} -1 \\ -3 \\ 2 \end{bmatrix}. \tag{2-162b}$$

For the $\lambda_2 = \lambda_3 = 2$ eigenvector, we obtain the matrix

$$\mathbf{A} - 2\mathbf{I} = \begin{bmatrix} 1 & -1 & 2 \\ 3 & -3 & 6 \\ -2 & 2 & -4 \end{bmatrix}. \tag{2-162c}$$

We proceed with Gaussian elimination to obtain

$$\begin{bmatrix} 1 & -1 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{2-162d}$$

Since we have a single nonzero row, the rank is 1 and we have 3 - 1 = 2 free parameters. Define

$$\alpha = x_2, \tag{2-162e}$$

$$\beta = x_3. \tag{2-162f}$$

Then,

$$x_1 = \alpha - 2\beta. \tag{2-162g}$$

This results in the solution vector

$$\mathbf{x} = \alpha \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + \beta \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}. \tag{2-162h}$$

Each of these column vectors corresponds to an eigenvector, so therefore

$$\mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \tag{2-162i}$$

$$\mathbf{v}_3 = \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}. \tag{2-162j}$$

Example 3

Continuing to the third example, the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 0 \\ 0 & 1 & 1 \\ -1 & 1 & 2 \end{bmatrix} \tag{2-163a}$$

has eigenvalues $\lambda_1 = 3, \lambda_2 = 1, \lambda_3 = 1$. The eigenvector for $\lambda_1 = 3$ is

$$\mathbf{v}_1 = \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix}. \tag{2-163b}$$

For the eigenvector $\lambda_2 = \lambda_3 = 1$, we obtain the following matrix in reduced-row echelon form:

$$\begin{bmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}. \tag{2-163c}$$

From this, we know there will be one free nonzero parameter. However,

$$x_3 = 0, \tag{2-163d}$$

$$x_1 = x_2. \tag{2-163e}$$

Since x_3 is uniquely determined at zero, we cannot assign the free parameter to it, so we are left with assigning the free parameter to one of the others. This implies the solution vector is

$$\mathbf{x} = \alpha \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}. \tag{2-163f}$$

Therefore, we have only one linearly independent eigenvector corresponding to the repeated eigenvalue 1, which is

$$\mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}. \tag{2-163g}$$

The upshot shown by these three examples is that if we have N distinct eigenvalues, we will have N distinct eigenvectors. If we have N eigenvalues with some of them repeating, we may have N distinct eigenvectors, but we could have fewer.

2.7.c Example: Nuclear Criticality

The criticality of a nuclear system describes its ability to sustain a nuclear chain reaction. The criticality is described by the effective multiplication factor, which is an eigenvalue. The corresponding eigenvector (or eigenfunction) describes the neutron distribution throughout the system. Analysis or nuclear reactors depends upon being able to determine the criticality of a system. Furthermore, even outside the context of nuclear reactors, many industrial processes in the nuclear industry involve fissile materials and it is crucial to the safety of personnel that these processes never be in a configuration that would lead to a self sustaining chain reaction.

A simplified model of a nuclear system will treat it as an infinite homogeneous system (not a terrible first cut at analyzing a large reactor) where the dependent variable is the distribution of neutron kinetic energy $\phi(E)$ or the energy spectrum. The equation has the following form:

$$\sigma_t(E)\phi(E) - \int_0^\infty \sigma_s(E' \to E)\phi(E')dE' = \frac{\chi(E)}{k}\int_0^\infty \nu\sigma_f(E')\phi(E')dE'. \tag{2-164}$$

The first term on the left-hand side in the equation is the total collision rate of neutrons, the second term on the left-hand side describes the scattering process that leads to neutrons slowing down and thermalizing, and the term on the right-hand side describes the emission of fission neutrons. Note that the fission term has a factor of $\frac{1}{k}$, which describes the criticality of the system with k being the effective multiplication factor.

This equation is an integral equation (similar to a differential equation) and is difficult to solve except in the most simple conditions. As is the case when encountering a problem that is too difficult to solve, we develop an approximate form. This form is obtained by assuming the neutron kinetic energies can be described by a series of

discrete *energy groups*. We can do through a series of manipulations (you will see this in a reactor physics course) to obtain an approximate linear system for a group g:

$$\sigma_{tg}\phi_g - \sum_{g'=1}^{G} \sigma_{sg,g'}\phi_{g'} = \frac{\chi_g}{k}\sum_{g'=1}^{G} \nu\sigma_{fg'}\phi_{g'}. \tag{2-165}$$

Here: ϕ_g is the unknown neutron scalar flux (path-length density) in energy group g, σ_{tg} is the cross section for total neutron interactions in group g, $\sigma_{sg,g'}$ is the rate that neutrons scatter from group g' into group g, χ_g is the probability that a fission neutron is born in group g, and σ_{fg} is the neutron fission production cross section in group g.

We can put this in matrix vector form by defining the neutron net removal matrix:

$$\mathbf{T} = \begin{bmatrix} \sigma_{t1} - \sigma_{s1,1} & -\sigma_{s1,2} & -\sigma_{s1,3} & \cdots & -\sigma_{s1,G-1} & -\sigma_{s1,G} \\ -\sigma_{s2,1} & \sigma_{t2} - \sigma_{s2,2} & -\sigma_{s2,3} & \cdots & -\sigma_{s2,G-1} & -\sigma_{s2,G} \\ \cdots & \cdots & \cdots & \ddots & \cdots & \cdots \\ -\sigma_{sG-1,1} & \sigma_{sG-1,2} & -\sigma_{sG-1,3} & \cdots & \sigma_{tG-1} - \sigma_{sG-1,G-1} & -\sigma_{sG-1,G} \\ -\sigma_{sG,1} & \sigma_{sG,2} & -\sigma_{sG,3} & \cdots & -\sigma_{sG,G-1} & \sigma_{tG} - \sigma_{sG,G} \end{bmatrix}; \tag{2-166}$$

and the neutron production matrix as a product of a column and row vector:

$$\mathbf{F} = \begin{bmatrix} \chi_1 \\ \chi_2 \\ \chi_3 \\ \vdots \\ \chi_{G-1} \\ \chi_G \end{bmatrix} \begin{bmatrix} \nu\sigma_{f1} & \nu\sigma_{f2} & \nu\sigma_{f3} & \cdots & \nu\sigma_{fG-1} & \nu\sigma_{fG} \end{bmatrix}. \tag{2-167}$$

In matrix-vector form, this equation is then

$$\mathbf{T}\boldsymbol{\phi} = \frac{1}{k}\mathbf{F}\boldsymbol{\phi} \tag{2-168}$$

where $\boldsymbol{\phi}$ is a column vector of group neutron scalar fluxes.

It is often the case that $\mathbf{F}$ is non-invertable because $\chi_g \approx 0$ for low energy groups. Therefore, we can recast this problem as

$$k\boldsymbol{\phi} = \mathbf{T}^{-1}\mathbf{F}\boldsymbol{\phi}. \tag{2-169}$$

Since $\mathbf{T}^{-1}\mathbf{F}$ is an operator, this is an eigenvalue problem. This is often written in terms of the fission source where we define

$$\mathbf{f} = \mathbf{F}\boldsymbol{\phi}, \tag{2-170}$$

which gives the rate that fission neutrons are born for each energy group. Multiplying the neutron balance equation by $\mathbf{F}$ we can obtain the form

$$\mathbf{F}\mathbf{T}^{-1}\mathbf{f} = k\mathbf{f}. \tag{2-171}$$

The operator $\mathbf{FT}^{-1}$ has a physical significance. The meaning of $\mathbf{T}^{-1}$ is to transport neutrons from a source through the lifetime while treating fission as a loss (since the net removal operator does not involve fission). The meaning of $\mathbf{FT}^{-1}$ is to transport neutrons from a source through a single fission generation. Therefore, the equation can be viewed as taking neutrons from a fission source, transporting them through one fission generation to form a new fission source, with k being a scaling factor on the magnitude of that source. This gives an interpretation of k as the number of neutrons in a fission generation divided by the number of neutrons in the previous fission generation.

2.7.d Matrix Eigendecomposition

One of the most important applications of eigenvalues and eigenvectors is that many matrices can be expressed in terms of them. One example that we will encounter in the next chapter is in the solution of systems of linear ordinary differential equations. If an $N \times N$ matrix has N linearly independent eigenvectors, then we say the matrix is *diagonalizable* and can be written as

$$\mathbf{A} = \mathbf{VDV}^{-1}. \tag{2-172}$$

Here $\mathbf{D}$ is a diagonal matrix containing the eigenvalues,

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_{N-1} & 0 \\ 0 & 0 & \cdots & 0 & \lambda_N \end{bmatrix}, \tag{2-173}$$

and $\mathbf{V}$ is a matrix where its columns are the corresponding eigenvectors,

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_1 & \cdots & \mathbf{v}_{N-1} & \mathbf{v}_N \\ \end{bmatrix}. \tag{2-174}$$

The ordering is arbitrary, but conventionally, the matrices are ordered such that $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_N|$.

Most matrices encountered in science and engineering applications are diagonalizable. Occasionally, there are cases where a matrix does not have N linearly independent eigenvectors. These matrices are called *defective matrices*. It is possible to generalize the eigendecompositon in this case. The matrix $\mathbf{V}$ has the eigenvectors as columns plus additional vectors using something called the Jordan Normal Form. This is not encountered often, so is not discussed further here.

Revisiting our example, the matrix

$$\mathbf{A} = \begin{bmatrix} 5 & 4 & -2 \\ 4 & 5 & 2 \\ 0 & 2 & -2 \end{bmatrix}$$

has three unique eigenvalues $\lambda = 9, -3, 2$. Note the conventional ordering from largest to smallest magnitude. The eigendecomposition has the following matrices:

$$\mathbf{D} = \begin{bmatrix} 9 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} 10 & -2 & 1 \\ 11 & 2 & -1 \\ 2 & 1 & 2 \end{bmatrix}.$$

In the second example,

$$\mathbf{A} = \begin{bmatrix} 3 & -1 & 2 \\ 3 & -1 & 6 \\ -2 & 2 & -2 \end{bmatrix}$$

the eigenvalues are $\lambda = -4, 2, 2$. We have a repeated eigenvalue, but there are still three linearly independent eigenvectors. Therefore this matrix is diagonalizable:

$$\mathbf{D} = \begin{bmatrix} -4 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} -1 & 1 & -2 \\ 2 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix}.$$

For the third example,

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 0 \\ 0 & 1 & 1 \\ -1 & 1 & 2 \end{bmatrix},$$

there is only two linearly independent eigenvectors. This means this matrix is defective and not diagonalizable.

2.7.e Functions of Matrices

One important application of the eigendecomposition is taking a function of a matrix $f(\mathbf{A})$. This arises, for example, when solving systems of ordinary linear differential equations. We can evaluate the function of any diagonalizable matrix provided that the function is analytic, in that it has a power series expansion:

$$f(\mathbf{A}) = c_0 + c_1\mathbf{A} + c_2\mathbf{A}^2 + \ldots = \sum_{n=0}^{\infty} c_n\mathbf{A}^n. \tag{2-175}$$

If $\mathbf{A}$ is diagonalizable, then we can expand it as an eigendecomposition:

$$f(\mathbf{A}) = \sum_{n=0}^{\infty} c_n \left(\mathbf{V}\mathbf{D}\mathbf{V}^{-1}\right)^n. \tag{2-176}$$

Next, for an arbitrary power n, we note

$$\begin{aligned} \mathbf{A}^n &= \left(\mathbf{V}\mathbf{D}\mathbf{V}^{-1}\right)^n = \left(\mathbf{V}\mathbf{D}\mathbf{V}^{-1}\right)\left(\mathbf{V}\mathbf{D}\mathbf{V}^{-1}\right)\cdots\left(\mathbf{V}\mathbf{D}\mathbf{V}^{-1}\right)\left(\mathbf{V}\mathbf{D}\mathbf{V}^{-1}\right) \\ &= \mathbf{V}\mathbf{D}\mathbf{D}\cdots\mathbf{D}\mathbf{D}\mathbf{V}^{-1} = \mathbf{V}\mathbf{D}^n\mathbf{V}^{-1}. \end{aligned} \tag{2-177}$$

Notice that for internal term has an adjacent $\mathbf{V}$ and $\mathbf{V}^{-1}$ so they all cancel, leaving only a single term on the end. Returning to the power series expansion:

$$f(\mathbf{A}) = \sum_{n=0}^{\infty} c_n \mathbf{V}\mathbf{D}^n\mathbf{V}^{-1} = \mathbf{V}\left(\sum_{n=0}^{\infty} c_n \mathbf{D}^n\right)\mathbf{V}^{-1} = \mathbf{V}f(\mathbf{D})\mathbf{V}^{-1}. \tag{2-178}$$

This means we need to evaluate the function of a diagonal matrix. Using the definition of a power series and noting that the sums and products of diagonal matrices are just diagonal matrices (they function as effectively independent variables), we have

$$\begin{aligned}
\sum_{n=0}^{\infty} c_n \mathbf{D}^n &= c_0\mathbf{I} + c_1\begin{bmatrix} \lambda_1 & 0 & \cdots \\ 0 & \lambda_2 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} + c_2\begin{bmatrix} \lambda_1 & 0 & \cdots \\ 0 & \lambda_2 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}^2 + \cdots \\
&= c_0\mathbf{I} + c_1\begin{bmatrix} \lambda_1 & 0 & \cdots \\ 0 & \lambda_2 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} + c_2\begin{bmatrix} \lambda_1^2 & 0 & \cdots \\ 0 & \lambda_2^2 & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} + \cdots \\
&= \begin{bmatrix} \displaystyle\sum_{n=0}^{\infty} c_n\lambda_1^n & 0 & \cdots \\ 0 & \displaystyle\sum_{n=0}^{\infty} c_n\lambda_2^n & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} = \begin{bmatrix} f(\lambda_1) & 0 & \cdots \\ 0 & f(\lambda_2) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}.
\end{aligned} \tag{2-179}$$

Therefore, the procedure for taking $f(\mathbf{A})$ is to first find the eigenvalues and eigenvectors of $\mathbf{A}$. The matrices $\mathbf{V}$ and $\mathbf{V}^{-1}$ are computed as before. The matrix $f(\mathbf{D})$ is a diagonal matrix where all of the elements are the function evaluations of the eigenvalues.

Perhaps the most important example is matrix exponential, which arises in the solution of systems of linear ordinary differential equations. We have

$$\exp(\mathbf{A}) = \mathbf{V}\exp(\mathbf{D})\mathbf{V}^{-1}, \tag{2-180}$$

where $\mathbf{D}$ contains the exponentials of the eigenvalues of $\mathbf{A}$. It is important to keep in mind that the matrix exponential is *not* in general the exponential of the individual elements (diagonal matrices excepted). Rather, one needs to apply the eigendecomposition first.

2.7.f Power Iteration

Solving for all of the eigenvalues and eigenvectors for a given matrix $\mathbf{A}$ on a computer is a complicated algorithmic task involving numerous steps.

In many applications, we are primarily interested in the fundamental or largest (in magnitude) eigenvalue and its corresponding eigenvector, which we refer to as the dominant eigenpair. This eigenpair gives information about the long-time or

asymptotic behavior of a particular system. In nuclear reactor analysis, the dominant eigenvalue is k, the effective multiplication factor and the corresponding eigenvector has information about the steady-state distribution of neutrons during continuous operation. Fortunately, there is an iterative algorithm that is fairly simple and can be used for obtaining this called the *Power Iteration Method*.

The general idea is that we make an initial guess of the eigenvector and make successive matrix multiplications of the matrix $\mathbf{A}$ upon it. Eventually, the resulting vector should, under certain circumstances, converge. The power iteration method is guaranteed to converge if the dominant eigenvalue is unique and all of the eigenvalues are real and positive, such as when $\mathbf{A}$ is a real-symmetric matrix. (In practice, this works extremely well for the application of nuclear criticality even though this has not rigorously been proven.) Problems with convergence may arise when eigenvalues are complex or when the second-largest eigenvalue is equal to or close to the largest.

The power iteration starts with a nonzero initial guess for a normalized eigenvector $\mathbf{v}^{(0)}$, and finds an updated guess by

$$\mathbf{u}^{(k)} = \mathbf{A}\mathbf{v}^{(k-1)}, \tag{2-181}$$

$$\mathbf{v}^{(k)} = \frac{\mathbf{u}^{(k)}}{|\mathbf{u}^{(k)}|}. \tag{2-182}$$

Here the superscript k in parentheses is an iteration index. Note that the eigenvector needs to be normalized every iteration, else the magnitude of the eigenvector will grow without bound or decay to zero (unless the fundamental eigenvalue is 1).

The eigenvalue for the kth iteration is computed using something called the Rayleigh quotient. This can be derived by taking the dot product of $\mathbf{v}^{(k)}$ with the equation $\mathbf{A}\mathbf{v}^{(k)} = \lambda^{(k)}\mathbf{v}^{(k)}$ and solving for λ:

$$\lambda^{(k)} = \frac{(\mathbf{A}\mathbf{v}^{(k)}) \cdot \mathbf{v}^{(k)}}{\mathbf{v}^{(k)} \cdot \mathbf{v}^{(k)}}. \tag{2-183}$$

We then check the following convergence criterion:

$$|\mathbf{A}\mathbf{v}^{(k)} - \lambda^{(k)}\mathbf{v}^{(k)}| < \epsilon, \tag{2-184}$$

where ϵ is a user-defined tolerance. If it is not met, we iterate again and again until the equation is satisfied to the specified tolerance; however, care needs to be taken to restrict the maximum number of iterations because there are situations (e.g., complex eigenvalues) where the algorithm will never converge.

The power iteration method is quite robust for most linear systems that arise from approximating differential equations found in scientific and engineering applications. The primary issue with the method in practical applications is that it may exhibit slow convergence. Assuming the eigenvectors form an orthogonal basis that completely spans the space, the initial guess can be expanded as a linear combination of eigenvectors:

$$\mathbf{v}^{(0)} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \ldots + c_N\mathbf{v}_N. \tag{2-185}$$

Here the subscripts denote the eigenvectors $\mathbf{v}_i$ corresponding to the eigenvalues λ_i such that $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_N|$. To get the kth iterate of the eigenvector, we apply multiply the initial guess by the matrix $\mathbf{A}$ k times:

$$\mathbf{v}^{(k)} = \mathbf{A}^k \mathbf{v}^{(0)} = c_1 \mathbf{A}^k \mathbf{v}_1 + c_2 \mathbf{A}^k \mathbf{v}_2 + \ldots + c_N \mathbf{A}^k \mathbf{v}_N. \tag{2-186}$$

Here $\mathbf{A}^k$ is the matrix to the kth power. Noting that the $\mathbf{v}_k$ are eigenvectors upon which $\mathbf{A}$ is being applied, we can then write:

$$\mathbf{v}^{(k)} = c_1 \lambda_1^k \mathbf{v}_1 + c_2 \lambda_2^k \mathbf{v}_2 + \ldots + c_N \lambda_1^N \mathbf{v}_N. \tag{2-187}$$

Factoring out $c_1 \lambda_1^k$ gives

$$\mathbf{v}^{(k)} = c_1 \lambda_1^k \left[\mathbf{v}_1 + \frac{c_2}{c_1} \left(\frac{\lambda_2}{\lambda_1} \right)^k \mathbf{v}_2 + \frac{c_3}{c_1} \left(\frac{\lambda_3}{\lambda_1} \right)^k \mathbf{v}_3 + \ldots + \frac{c_N}{c_1} \left(\frac{\lambda_N}{\lambda_1} \right)^k \mathbf{v}_N \right]. \tag{2-188}$$

By the ordering of the eigenvalues, the ratio of the eigenvalues is less then one in magnitude and λ_2/λ_1 is the largest (in magnitude). This means when taken to the kth power, the λ_2/λ_1 decays slower than the others. Therefore, the asymptotic rate of convergence is described the ratio of the second largest (in magnitude) eigenvalue to the largest:

$$\rho = \left| \frac{\lambda_2}{\lambda_1} \right|, \tag{2-189}$$

sometimes referred to as the dominance ratio in the nuclear criticality discipline. When $\rho \approx 1$, this implies that the convergence rate is slow. In the application of nuclear reactor analysis, this is fairly typical of large, commercial reactors. Therefore, more advanced methods are used to accelerate convergence, but these are beyond the scope of this text.

To illustrate the power iteration method, let us revisit one of our examples:

$$\mathbf{A} = \begin{bmatrix} 5 & 4 & -2 \\ 4 & 5 & 2 \\ 0 & 2 & -2 \end{bmatrix},$$

where we found that the dominant eigenvalue and (normalized) eigenvector are

$$\lambda_1 = 9, \quad \mathbf{v}_1 = \frac{1}{15} \begin{bmatrix} 10 \\ 11 \\ 2 \end{bmatrix} \approx \begin{bmatrix} 0.66667 \\ 0.73333 \\ 0.13333 \end{bmatrix}.$$

Let us guess the eigenvector

$$\mathbf{v}^{(0)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

The first application of $\mathbf{A}$ gives the vectors

$$\mathbf{u}^{(1)} = \begin{bmatrix} 5 & 4 & -2 \\ 4 & 5 & 2 \\ 0 & 2 & -2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 11 \\ 11 \\ 0 \end{bmatrix}, \quad \mathbf{v}^{(1)} = \frac{\mathbf{u}^{(1)}}{|\mathbf{u}^{(1)}|} \approx \begin{bmatrix} 0.536875 \\ 0.843661 \\ 0 \end{bmatrix}.$$

The predicted eigenvalue is

$$\lambda^{(1)} = \frac{(\mathbf{A}\mathbf{v}^{(1)}) \cdot \mathbf{v}^{(1)}}{\mathbf{v}^{(1)} \cdot \mathbf{v}^{(1)}} \approx 8.62353.$$

The error estimate on the equation is

$$|\mathbf{A}\mathbf{v}^{(1)} - \lambda^{(1)}\mathbf{v}^{(1)}| \approx 2.39104.$$

As expected the agreement with the equation is not good and the eigenvalue is still significantly different than the correct result of 9.

Multiplying $\mathbf{A}$ onto $\mathbf{v}^{(1)}$ and renormalizing gives

$$\mathbf{v}^{(2)} \approx \begin{bmatrix} 0.677071 \\ 0.711353 \\ 0.188551 \end{bmatrix}, \quad \lambda^{(2)} \approx 8.88541, \quad \text{Error} \approx 0.725475.$$

Getting closer, but still not there. A few more iterations gives:

$$\begin{aligned}
\mathbf{v}^{(3)} &\approx \begin{bmatrix} 0.656610 & 0.745055 & 0.117286 \end{bmatrix}^\top, \quad \lambda^{(3)} \approx 9.01291, \quad \text{Error} \approx 0.261228. \\
\mathbf{v}^{(4)} &\approx \begin{bmatrix} 0.668615 & 0.730455 & 0.139246 \end{bmatrix}^\top, \quad \lambda^{(4)} \approx 8.99207, \quad \text{Error} \approx 0.082992. \\
\mathbf{v}^{(5)} &\approx \begin{bmatrix} 0.665714 & 0.734530 & 0.131490 \end{bmatrix}^\top, \quad \lambda^{(5)} \approx 9.00214, \quad \text{Error} \approx 0.028480. \\
\mathbf{v}^{(6)} &\approx \begin{bmatrix} 0.666918 & 0.732988 & 0.133976 \end{bmatrix}^\top, \quad \lambda^{(5)} \approx 8.99920, \quad \text{Error} \approx 0.009308. \\
&\vdots \\
\mathbf{v}^{(10)} &\approx \begin{bmatrix} 0.666670 & 0.733329 & 0.133341 \end{bmatrix}^\top, \quad \lambda^{(10)} \approx 8.99999, \quad \text{Error} \approx 1.15591 \times 10^{-4}.
\end{aligned}$$

It appears that within ten iterations, the eigenvalue and eigenvector are fairly close and converging rapidly. Looking at the theoretical asymptotic convergence rate we have $\rho = \left|\frac{\lambda_2}{\lambda_1}\right| = \left|\frac{-3}{9}\right| = \frac{1}{3}$, which implies that the power iteration should quickly in this case.

2.8 Singular Values

Any matrix $\mathbf{A}$, even if it is not square, has a set of numbers called singular values, which are related to eigenvalues. Recall that the eigenvectors of a matrix give the directions that are invariant with respect to multiplication $\mathbf{A}$ up to a multiplicative scaling constant given by the corresponding eigenvalues. The singular vectors, on the other hand, give the directions of maximal action when multiplied by $\mathbf{A}$ where the respective singular values give a scaling constant.

There has been a growing importance of singular values and the singular value decomposition (which is analogous to the eigendecomposition) over the last few decades, especially in the area of data science. One important application is finding the principal components, or the set of vectors that describe most of the information in a system. These singular vectors can be used to discern what quantities are most important for describing a system. They can also compactly approximate a matrix having information that describes a complicated system, allowing for data compression. In addition, we can use the singular value decomposition to compute a matrix called the pseudoinverse, which is a generalization of the normal inverse that exists for any matrix. This pseudoinverse can be used to solve optimization problems to find best fit for incomplete or even inconsistent data, which invariably arises in real-world measurements.

The equations that describe the singular values are

$$\mathbf{A}\mathbf{v} = \sigma\mathbf{u}, \tag{2-190a}$$

$$\mathbf{A}^*\mathbf{u} = \sigma\mathbf{v}. \tag{2-190b}$$

Here $\mathbf{A}$ is any (even non square) matrix, σ is a singular value, and $\mathbf{u}$ and $\mathbf{v}$ are respective left and right singular (unit) vectors, which are orthonormal. As with eigenvalues, there is a largest singular value σ_1. From this relationship, we can take the magnitude of the multiplicative action of $\mathbf{A}$ onto the corresponding right singular value $\mathbf{v}_1$ and get

$$|\mathbf{A}\mathbf{v}_1| = \sigma_1. \tag{2-191}$$

The application of $\mathbf{A}$ onto $\mathbf{v}_1$ is the maximum possible stretching of any unit vector. In other words,

$$|\mathbf{A}\mathbf{v}_1| \geq |\mathbf{A}\mathbf{x}|, \quad |\mathbf{x}| = 1. \tag{2-192}$$

To illustrate this geometrically, let us consider the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ 1 & -2 \end{bmatrix}.$$

Using techniques to be discussed, we will find that the singular values are $\sigma = 3, 1$; the corresponding right singular vectors are

$$\mathbf{v}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}; \tag{2-193}$$

and the respective left singular vectors are

$$\mathbf{u}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ -1 \end{bmatrix}, \quad \mathbf{u}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{2-194}$$

Observe that these two vectors within each set are orthonormal: their dot product $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$ and $\mathbf{u}_1 \cdot \mathbf{u}_2 = 0$ and their magnitudes are one. We can therefore draw

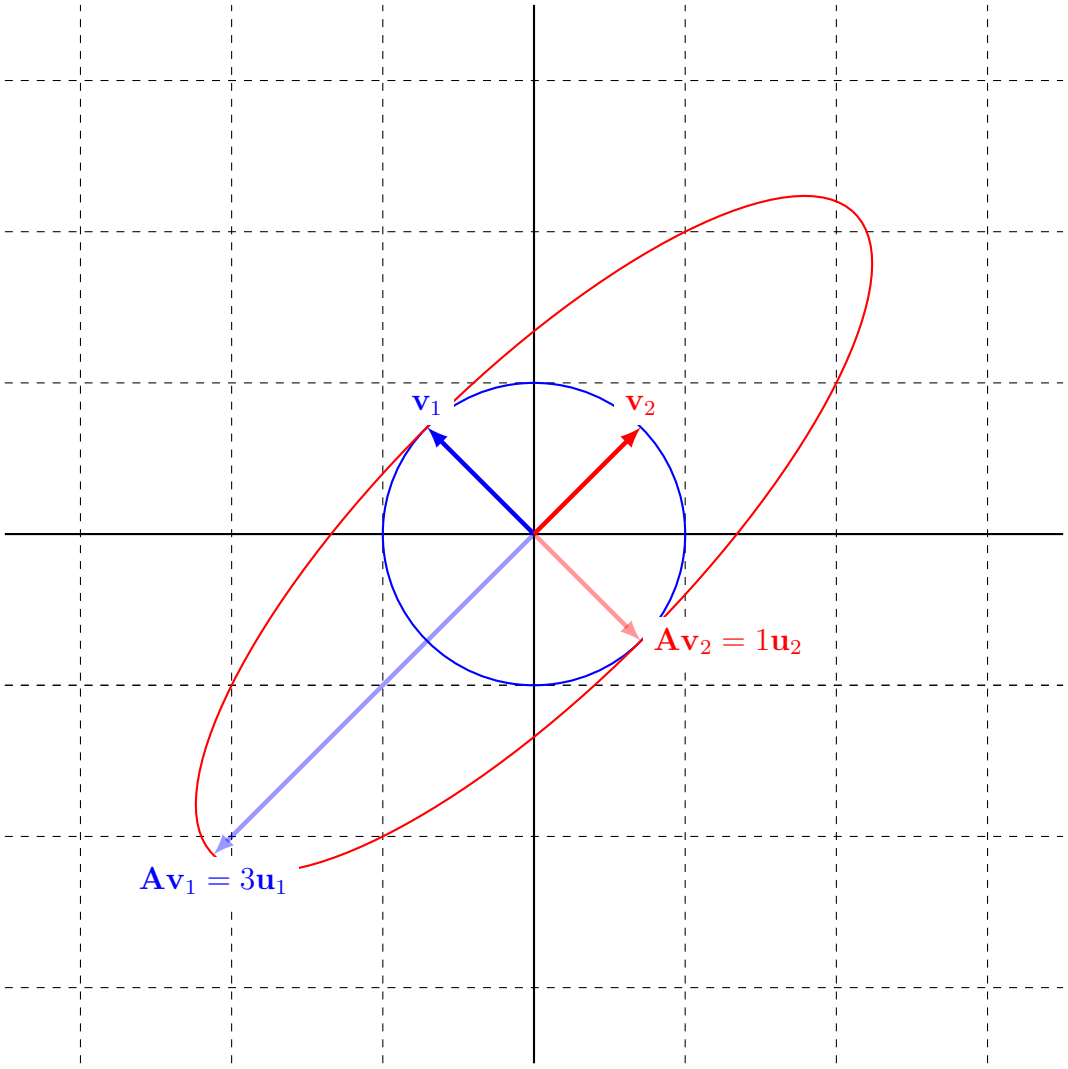

Figure 2.14: Illustration of a the application of a matrix on its right singular vectors.

the right singular values as two vectors on the unit circle. The application of $\mathbf{A}$ onto these vectors $\mathbf{v}$ yields the vectors $\mathbf{u}$ scaled by the respective singular values σ. These resulting vectors can now be thought of as the semi-axes of an ellipse. This is all illustrated in Fig. 2.14.

In general, the application of $\mathbf{A}$ maps its right singular vectors $\mathbf{v}$ from a N-dimensional sphere to an M-dimensional ellipsoid where $N \geq M$ with the left singular values $\mathbf{u}$ being the set of orthogonal directions of maximal stretching.

2.8.a Singular Value Decomposition

Any real $N \times M$ matrix $\mathbf{A}$ can be written as the product of three matrices using the singular value decomposition:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top. \tag{2-195}$$

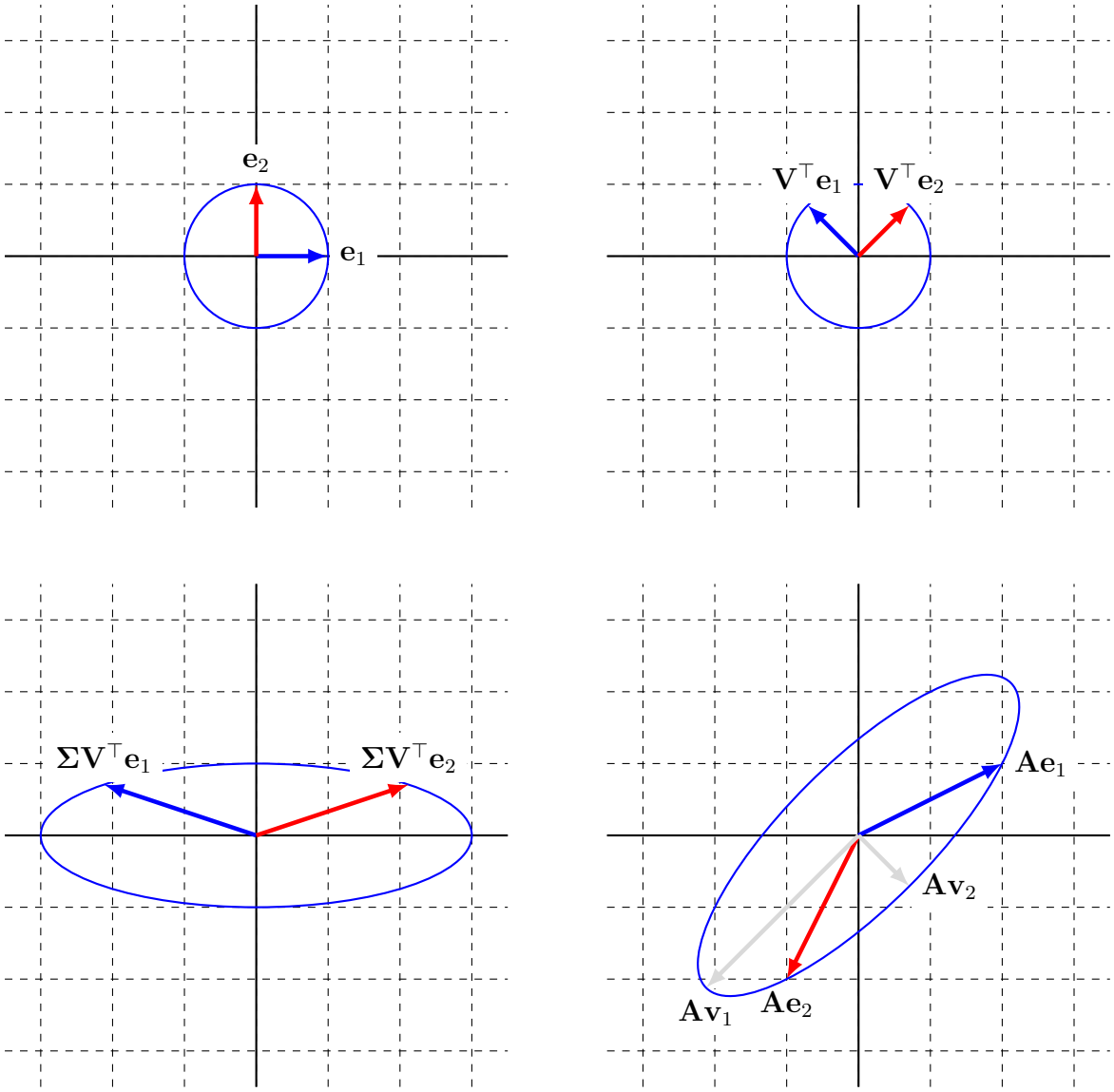

Figure 2.15: Application of each operation from right to left of the singular value decomposition onto the unit basis vectors

Here $\mathbf{\Sigma}$ is a $N\times M$ matrix where the diagonal elements $i = j$ contain the singular values with the remaining elements being zero, $\mathbf{V}$ is an $M\times M$ matrix containing the right singular values, and $\mathbf{U}$ is a $N\times N$ matrix containing the left singular values. An important property of $\mathbf{U}$ and $\mathbf{V}$ is that they are both unitary matrices, which means they equal their responses conjugate transposes (or just the transpose if all elements are real).

To illustrate the singular value decomposition, let us consider the action of $\mathbf{A} = \mathbf{U\Sigma V}^\top$ onto the unit basis vectors (i.e., the 2×2 identity matrix), as depicted in Fig. 2.15. The top-left panel shows that the two unit basis vectors point along the x and y axes and lie on the unit circle. Note the description here applies to any unit vector, and the unit Cartesian basis vectors are chosen as an example.

Going from right to left, we first apply $\mathbf{V}^\top$ onto the unit basis vectors. Since $\mathbf{V}$ is unitary with all columns being unit vectors, the application of $\mathbf{V}$ or $\mathbf{V}^\top$ only reorients the vectors and does not stretch them. This reorientation is illustrated in the top-right panel of Fig. 2.15.

The application of $\mathbf{\Sigma}$ onto $\mathbf{V}^\top\mathbf{I}$ stretches the vectors along each principle coordinate direction by their respective singular value. In this case, the x components are multiplied by a factor of $\sigma_1 = 3$ and the y components are left as is because $\sigma_2 = 1$. This is shown in the lower-left panel of Fig. 2.15. The vectors after applying $\mathbf{\Sigma}$ now live on an ellipse rather than a circle.

Finally, the application of $\mathbf{U}$ onto $\mathbf{\Sigma}\mathbf{V}^\top\mathbf{I}$ completes the action of $\mathbf{A}$. This operation reorients the vectors, but does not change their magnitude because $\mathbf{U}$ is unitary, having columns that are orthonormal vectors. This also rotates the ellipse. This is drawn on the lower-right panel of Fig. 2.15; also depicted are the application of $\mathbf{A}$ onto the right singular vectors. The application of $\mathbf{A}$ onto any unit vector will yield another vector that points to some location on an ellipse. As can be seen in the figure, the application of $\mathbf{A}$ onto the right singular value $\mathbf{v}_1$, corresponding to the largest singular value σ_1, maps that vector onto the semi-major (or longest) axis of the ellipse, which is the maximum possible stretching. Note that the the application of $\mathbf{A}$ onto $\mathbf{v}_2$, which is the right singular vector corresponding to the smallest value, maps it onto the semi-minor axis, which is the smallest possible scaling.

The process for the singular value decomposition is as follows:

1. Given $N\times M$ matrix $\mathbf{A}$, compute $\mathbf{A}^\top\mathbf{A}$, which is an $M\times M$ real-symmetric matrix, and find its eigenvalues and eigenvectors. Since $\mathbf{A}^\top\mathbf{A}$ is real and symmetric, so we know its eigenvalues are real and nonnegative, there will be r nonzero eigenvalues where r is the rank of $\mathbf{A}$, and these r eigenvectors should be orthogonal.

2. Sort the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$. Construct the matrix $\mathbf{V}$ by setting its columns equal to the eigenvectors of $\mathbf{A}^\top\mathbf{A}$ in the order of the eigenvalues.

3. Compute the singular values by taking the square root of the eigenvalues,

$$\sigma_i = \sqrt{\lambda_i}. \tag{2-196}$$

   Construct the $N\times M$ matrix $\mathbf{\Sigma}$ by setting the diagonal elements ($i = j$) to the singular values σ_i in the same order as the eigenvectors. Note that a eigenvalue or singular value of zero occurs when the rank of the matrix $\mathbf{A}$ is smaller than N.

4. Next, construct the first r columns of the $N\times N$ matrix $\mathbf{U}$. This can be done by using the definition of the singular value decomposition given in Eq. (2-197), noting that $\mathbf{V}$ is unitary and equal to its transpose (for real elements, or more generally its conjugate transpose), we can multiply the equation by $\mathbf{V}$ on the right to obtain

$$\mathbf{AV} = \mathbf{U\Sigma}. \tag{2-197}$$

Since $\mathbf{\Sigma}$ is a diagonal matrix, we can write $\mathbf{U\Sigma}$ as a matrix where each column is an (unknown) column of $\mathbf{U}$, call it $\mathbf{u}_i$, times the diagonal element σ_i (a scalar singular value). Therefore, we can solve for each column as

$$\mathbf{u}_i = \frac{1}{\sigma_i}\mathbf{A}\mathbf{v}_i. \tag{2-198}$$

5. If $r = N$, then we can go straight to the next step and form $\mathbf{U}$. Otherwise, if $r < N$, denoted by singular values $\sigma_i = 0$, then we need to find the remaining $N - r$ columns of $\mathbf{U}$. These columns must be orthonormal. These can be found using some orthogonalization procedure such as Gram-Schmidt process (see Sec. 2.4.h). To use this procedure, one needs to generate $N-r$ vectors that are linearly independent of all the other vectors in $\mathbf{U}$. Any linearly independent vector will do, so good candidates are often the unit basis vectors. Once these orthonormal vectors are computed, they should be inserted into the remaining columns of $\mathbf{U}$. Note that the $N - r$ orthonormal vectors are not unique, which implies the singular value decomposition is not unique when the rank of $\mathbf{A}$ is less than the number of rows N.

6. Finally, we form $\mathbf{U}$ by using, from left to right, first the r columns computed in step 4 in the same order as the singular values and then, if $r < N$, the remaining $N - r$ orthonormal vectors computed in step 5.

Now we will do a few examples to illustrate this process.

Example 1

Consider again the matrix we used earlier in this section:

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ 1 & -2 \end{bmatrix}. \tag{2-199}$$

Step 1 in the process involves computing $\mathbf{A}^\top\mathbf{A}$ and finding the eigenvalues and eigenvectors. First,

$$\mathbf{A}^\top\mathbf{A} = \begin{bmatrix} 2 & 1 \\ -1 & -2 \end{bmatrix}\begin{bmatrix} 2 & -1 \\ 1 & -2 \end{bmatrix} = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix}. \tag{2-200}$$

Observe that this a real symmetric matrix, which is guaranteed to have real and nonnegative eigenvalues. Compute the eigenvalues by evaluating

$$\begin{vmatrix} 5-\lambda & -4 \\ -4 & 5-\lambda \end{vmatrix} = (5-\lambda)^2 - 16 = \lambda^2 - 10\lambda + 9 = 0, \quad \lambda = 9, 1. \tag{2-201}$$

Inserting each eigenvalue into the matrix $\mathbf{A}^\top\mathbf{A}$ and finding where it equals zero yields the eigenvectors. Doing this and applying elementary row operations gives

$$\lambda_1 = 9: \quad \begin{bmatrix} -4 & -4 \\ -4 & -4 \end{bmatrix} \to \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{v}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ 1 \end{bmatrix}; \tag{2-202a}$$

$$\lambda_2 = 1: \quad \begin{bmatrix} 4 & -4 \\ -4 & 4 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{v}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{2-202b}$$

Note that the eigenvectors are normalized to be unit vectors and are orthogonal.

Step 2 involves forming the matrix $\mathbf{V}$ taking the eigenvectors and putting them as the columns in descending order of the magnitude of the eigenvalue:

$$\mathbf{V} = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}. \tag{2-203}$$

Step 3 is to form the matrix $\mathbf{\Sigma}$ by taking the square root of the eigenvalues to compute the singular values and putting them along the diagonal and zeroes elsewhere:

$$\mathbf{\Sigma} = \begin{bmatrix} \sqrt{9} & 0 \\ 0 & \sqrt{1} \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}. \tag{2-204}$$

We observe that we have two nonzero singular values. Therefore, the rank of the matrix $\mathbf{A}$ is $r = 2$. Step 4 involves computing the first $r = 2$ columns of matrix $\mathbf{U}$ by Eq. (2-198). These are

$$\mathbf{u}_1 = \frac{1}{3}\frac{1}{\sqrt{2}} \begin{bmatrix} 2 & -1 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \tag{2-205a}$$

$$\mathbf{u}_2 = \frac{1}{1}\frac{1}{\sqrt{2}} \begin{bmatrix} 2 & -1 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{2-205b}$$

Since the rank $r = 2$, which is equal to the number of rows (and columns) of the matrix, the matrix is said to be full rank. We therefore, have all the required orthonormal vectors to form $\mathbf{U}$ and can skip step 5 and proceed to step 6. Forming the matrix $\mathbf{U}$ with the vectors computing in step 4 gives the resulting matrix:

$$\mathbf{U} = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix}. \tag{2-206}$$

Example 2

The previous example uses a full-rank matrix, so we could skip step 5 having an orthonormalization process. Let us now consider a case where the matrix is not full rank:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix}. \tag{2-207}$$

It is easy to observe that the third and fourth rows are, respectively, the sum and difference of the first and second rows.

Step 1 involves finding $\mathbf{A}^\top\mathbf{A}$ and computing the eigenvalues and the eigenvectors. First,

$$\mathbf{A}^\top\mathbf{A} = \begin{bmatrix} 3 & 0 & -3 & 0 \\ 0 & 3 & 0 & 3 \\ -3 & 0 & 3 & 0 \\ 0 & 3 & 0 & 3 \end{bmatrix}. \tag{2-208}$$

Evaluate the determinant to find the eigenvalues:

$$\begin{vmatrix} 3-\lambda & 0 & -3 & 0 \\ 0 & 3-\lambda & 0 & 3 \\ -3 & 0 & 3-\lambda & 0 \\ 0 & 3 & 0 & 3-\lambda \end{vmatrix} = \lambda^4 - 12\lambda^3 + 36\lambda^2 = \lambda^2(6-\lambda)^2, \quad \lambda = 6, 6, 0, 0. \tag{2-209}$$

The eigenvectors, in the order of $\lambda_1 = 6, \lambda_2 = 6, \lambda_3 = 0, \lambda_4 = 0$ are

$$\mathbf{v}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}, \mathbf{v}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \mathbf{v}_3 = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 \\ -1 \\ 0 \\ 1 \end{bmatrix}, \mathbf{v}_4 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}. \tag{2-210}$$

Note that despite there being repeated eigenvalues, we still have four linearly independent eigenvectors.

In step 2, we construct the matrix $\mathbf{V}$ from the eigenvectors, which are already in the order of descending eigenvalue:

$$\mathbf{V} = \frac{1}{\sqrt{2}}\begin{bmatrix} 0 & -1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}. \tag{2-211}$$

Note that one could swap the first and second columns or third and fourth columns, since they belong to a repeated eigenvalue.

Step 3 has us computing the matrix of singular values by taking the square root of the eigenvalues and placing them along the diagonal:

$$\mathbf{\Sigma} = \begin{bmatrix} \sqrt{6} & 0 & 0 & 0 \\ 0 & \sqrt{6} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{2-212}$$

Step 4 is computing the first r columns of $\mathbf{U}$. Here $r = 2$, which corresponds to the number of nonzero singular values. We have

$$\mathbf{u}_1 = \frac{1}{\sqrt{6}}\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} = \frac{1}{\sqrt{3}}\begin{bmatrix} 0 \\ 1 \\ 1 \\ -1 \end{bmatrix}, \tag{2-213a}$$

$$\mathbf{u}_2 = \frac{1}{\sqrt{6}}\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix}\begin{bmatrix} -1 \\ 0 \\ 1 \\ 0 \end{bmatrix} = \frac{1}{\sqrt{3}}\begin{bmatrix} -1 \\ 0 \\ -1 \\ -1 \end{bmatrix}. \tag{2-213b}$$

Step 5 requires finding two additional orthonormal vectors with respect to $\mathbf{u}_1$ and $\mathbf{u}_2$. This may be done using the Gram-Schmidt orthogonalization procedure. To do this, we require any two additional vectors that are linearly independent with the others. Two acceptable choices are the unit basis vectors

$$\mathbf{w}_3 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{w}_4 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}. \tag{2-214}$$

It is easy to check that these cannot be written as a linear combination of each other or with $\mathbf{u}_1$ and $\mathbf{u}_2$. The unnormalized vector for $\mathbf{u}_3$ is obtained by taking

$$\tilde{\mathbf{u}}_3 = \mathbf{w}_3 - \left(\frac{\mathbf{w}_3 \cdot \mathbf{u}_1}{\mathbf{u}_1 \cdot \mathbf{u}_1}\right)\mathbf{u}_1 - \left(\frac{\mathbf{w}_3 \cdot \mathbf{u}_2}{\mathbf{u}_2 \cdot \mathbf{u}_2}\right)\mathbf{u}_2, \quad \mathbf{u}_3 = \frac{1}{\sqrt{6}}\begin{bmatrix} 2 \\ 0 \\ -1 \\ -1 \end{bmatrix}. \tag{2-215a}$$

Here $\tilde{\mathbf{u}}_3$ is the unnormalized version of $\mathbf{u}_3$. The actual vector $\mathbf{u}_3$ is computed by taking $\tilde{\mathbf{u}}_3/|\tilde{\mathbf{u}}_3|$. Finally, the vector $\mathbf{u}_4$ is computed by

$$\tilde{\mathbf{u}}_4 = \mathbf{w}_4 - \left(\frac{\mathbf{w}_4 \cdot \mathbf{u}_1}{\mathbf{u}_1 \cdot \mathbf{u}_1}\right)\mathbf{u}_1 - \left(\frac{\mathbf{w}_4 \cdot \mathbf{u}_2}{\mathbf{u}_2 \cdot \mathbf{u}_2}\right)\mathbf{u}_2 - \left(\frac{\mathbf{w}_4 \cdot \mathbf{u}_3}{\mathbf{u}_3 \cdot \mathbf{u}_3}\right)\mathbf{u}_3,$$

$$\mathbf{u}_4 = \frac{1}{\sqrt{6}}\begin{bmatrix} 0 \\ 2 \\ -1 \\ 1 \end{bmatrix}. \tag{2-215b}$$

It is important to note that the choices for $\mathbf{w}_3$ and $\mathbf{w}_4$ are arbitrary so long as they satisfy the criteria of being linearly independent with themselves and $\mathbf{u}_1$ and $\mathbf{u}_2$. This means that $\mathbf{u}_3$ and $\mathbf{u}_4$ are not unique, and there are infinitely many valid choices that could be used to form the matrix $\mathbf{U}$ and reproduce $\mathbf{A}$.

Step 6 forms the matrix $\mathbf{U}$ from the column vectors from steps 4 and 5:

$$\mathbf{U} = \frac{1}{\sqrt{6}}\begin{bmatrix} 0 & -\sqrt{2} & 2 & 0 \\ \sqrt{2} & 0 & 0 & 2 \\ \sqrt{2} & -\sqrt{2} & -1 & -1 \\ -\sqrt{2} & -\sqrt{2} & -1 & 1 \end{bmatrix}. \tag{2-216}$$

Example 3

The previous two examples used square matrices. The singular value decomposition, however, is defined for any matrix. Let us compute this for a 2×3 matrix:

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}. \tag{2-217}$$

In step 1, we compute $\mathbf{A}^\top\mathbf{A}$,

$$\mathbf{A}^\top\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 2 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 5 & 1 & -2 \\ 1 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix}, \tag{2-218}$$

and compute the eigenvalues,

$$\lambda_1 = 6, \quad \lambda_2 = 1, \quad \lambda_3 = 0, \tag{2-219a}$$

and the corresponding eigenvectors,

$$\mathbf{v}_1 = \frac{1}{\sqrt{30}} \begin{bmatrix} 5 \\ 1 \\ -2 \end{bmatrix}, \quad \mathbf{v}_2 = \frac{1}{\sqrt{5}} \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}, \quad \mathbf{v}_3 = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}. \tag{2-219b}$$

Note that there are two nonzero eigenvalues, signifying that $\mathbf{A}$ has a rank of 2. This is the maximum one could expect from a 2×3 matrix.

Step 2 forms the matrix $\mathbf{V}$ from the eigenvectors. This is

$$\mathbf{V} = \frac{1}{\sqrt{30}} \begin{bmatrix} 5 & 0 & \sqrt{5} \\ 1 & 2\sqrt{6} & -\sqrt{5} \\ -2 & \sqrt{6} & 2\sqrt{5} \end{bmatrix}. \tag{2-220}$$

Observe that the matrix $\mathbf{V}$ is a square matrix dimensioned by the number of columns of $\mathbf{A}$.

Step 3 computes the matrix $\mathbf{\Sigma}$ of singular values. This matrix is the same dimension as $\mathbf{A}$ or 2×3. In this case, we populate the only two diagonal elements with the nonzero singular values, which are computed by taking the square root of the eigenvalues. This is then

$$\mathbf{\Sigma} = \begin{bmatrix} \sqrt{6} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{2-221}$$

In step 4, we compute the first $r = 2$ columns of the matrix $\mathbf{U}$. This is a square matrix with the dimension corresponding to the number of rows of $\mathbf{A}$, so 2×2. Since the rank equals the number of columns of $\mathbf{U}$, we can determine all the left singular vectors $\mathbf{u}$ in this step. We have

$$\mathbf{u}_1 = \frac{1}{\sqrt{6}}\frac{1}{\sqrt{30}} \begin{bmatrix} 5 & 1 & -2 \\ 1 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 5 \\ 1 \\ -2 \end{bmatrix} = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \tag{2-222a}$$

$$\mathbf{u}_2 = \frac{1}{\sqrt{5}} \begin{bmatrix} 5 & 1 & -2 \\ 1 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} = \frac{1}{\sqrt{5}} \begin{bmatrix} -1 \\ 2 \end{bmatrix}. \tag{2-222b}$$

Since we have all the needed left singular values, we skip step 5 and move onto step 6 and form the $\mathbf{U}$ matrix from the vectors computed in step 4:

$$\mathbf{U} = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}. \tag{2-223}$$

2.8.b Matrix Pseudoinverse

A given matrix $\mathbf{A}$ only has an inverse if it is full rank, i.e., its rank is equal to the number of rows and columns. Just as the singular values exist for any matrix $\mathbf{A}$, there is a generalization of the inverse called the *pseudoinverse* that exists for any matrix $\mathbf{A}$, which is denoted by $\mathbf{A}^+$. The singular value decomposition can always be used to compute $\mathbf{A}^+$.

The pseudoinverse is useful when finding optimal solutions given by linear systems that have incomplete information (such that there are infinitely many solutions) or are inconsistent (such that there are no unique solutions). This case often arises in practice because the amount of data we can collect is often far less than the number of model parameters, so we cannot uniquely determine the solution, but can only find one that best fits the data. Furthermore, when making experimental measurements, there is always some associated error or uncertainty, which means that the linear system could be inconsistent and, again, one has to settle for a solution that best matches the data. This notion is what connects the singular value decomposition, via the computation of the psuedoinverse, to optimization problems.

The pseudoinverse discussed here is sometimes called the Moore-Penrose inverse and is unique for a given matrix $\mathbf{A}$. There are four criteria that the psuedoinverse must satisfy

$$\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}, \tag{2-224a}$$
$$\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+, \tag{2-224b}$$
$$\left(\mathbf{A}\mathbf{A}^+\right)^* = \mathbf{A}\mathbf{A}^+, \tag{2-224c}$$
$$\left(\mathbf{A}^+\mathbf{A}\right)^* = \mathbf{A}^+\mathbf{A}. \tag{2-224d}$$

Note that one could generalize an inverse to satisfy a subset of these four criteria, but then $\mathbf{A}^+$ would not be unique. Recall that the "star" superscript denotes the conjugate transpose, which, for real matrices is just the transpose. Another important property is that if $\mathbf{A}$ is full rank, then $\mathbf{A}^+ = \mathbf{A}^{-1}$. In a sense, the inverse is a special case of the pseudoinverse. One other fact about the pseudoinverse is that if $\mathbf{A}$ is $N \times M$ then $\mathbf{A}^+$ is $M \times N$.

There are special cases for which the pseudoinverse can be computed. One such important case is a rectangular $N \times M$ diagonal matrix $\mathbf{D}$, which has potentially nonzero values on the diagonal elements ($i = j$) and zeros everywhere else. (This is the case for the matrix $\mathbf{\Sigma}$ in the singular value decomposition). One can show (through) much effort, that the pseudoinverse of rectangular diagonal matrix is the transpose of that matrix with all of the nonzero diagonal elements inverted. For example:

$$\mathbf{D}_1 = \begin{bmatrix} 2 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix}, \quad \mathbf{D}_1^+ = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & -1 \\ 0 & 0 \end{bmatrix};$$

$$\mathbf{D}_2 = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{2}{3} \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{D}_2^+ = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & \frac{3}{2} & 0 & 0 \end{bmatrix}.$$

This can be shown by writing the general form for tall or wide rectangular diagonal matrices and showing that they satisfy the four criteria given above. (This is not done here because the derivation is rather tedious.)

The general method for computing a pseudoinverse uses the singular value decomposition and is given by

$$\mathbf{A}^+ = \mathbf{V}\boldsymbol{\Sigma}^+\mathbf{U}^\top. \tag{2-225}$$

Since $\boldsymbol{\Sigma}$ is a rectangular diagonal matrix, it can be computed in the manner described previously.

To illustrate this, let us compute the pseudoinverse for the examples in Sec. 2.8.a. The matrix for the first example is

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ 1 & -2 \end{bmatrix}.$$

Using the previously obtained results, the pseudoinverse is then

$$\mathbf{A}^+ = \frac{1}{2}\begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} \frac{1}{3} & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} -1 & -1 \\ 1 & -1 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 2 & -1 \\ 1 & -2 \end{bmatrix}. \tag{2-226}$$

Note that the pseudoinverse is identical to the actual inverse in this case since the matrix $\mathbf{A}$ is full rank.

In the second example, the matrix is

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix}. \tag{2-227}$$

This matrix again is not full rank because the third and fourth rows are linear combinations of the first and second rows. The pseudoinverse is

$$\begin{aligned} \mathbf{A}^+ &= \frac{1}{\sqrt{2}}\frac{1}{\sqrt{6}}\begin{bmatrix} 0 & -1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} \frac{1}{\sqrt{6}} & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{6}} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} 0 & \sqrt{2} & \sqrt{2} & -\sqrt{2} \\ -\sqrt{2} & 0 & -\sqrt{2} & -\sqrt{2} \\ 2 & 0 & -1 & -1 \\ 0 & 2 & -1 & 1 \end{bmatrix} \\ &= \frac{1}{6}\begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & -1 \\ -1 & 1 & -1 & -1 \\ 0 & 1 & 1 & -1 \end{bmatrix}. \end{aligned} \tag{2-228}$$

The matrix in the third example is

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}.$$

The pseudoinverse is

$$\begin{aligned} \mathbf{A}^+ &= \frac{1}{\sqrt{30}}\frac{1}{\sqrt{5}} \begin{bmatrix} 5 & 0 & \sqrt{5} \\ 1 & 2\sqrt{6} & -\sqrt{5} \\ -2 & \sqrt{6} & 2\sqrt{5} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ -1 & 2 \end{bmatrix} \\ &= \frac{1}{6} \begin{bmatrix} 2 & 1 \\ -2 & 5 \\ -2 & 2 \end{bmatrix}. \end{aligned} \tag{2-229}$$

2.8.c Example: Optimal Solution from Conflicting Data

When taking measurements, for example, either in a controlled laboratory experiment or from sensor data of an industrial process, the data is never strictly consistent because of inherent uncertainties, random noise, and errors in collecting the data. A common task is to determine the "best" set of parameters that fits the measured data. There are many methods for doing this, and one approach applies the pseudoinverse. Suppose whatever process can be described by a system of linear equations with known coefficients (or at least those predicted from some model and taken as known) given in a matrix $\mathbf{A}$, measurement results given in a column vector $\mathbf{b}$, and a set of unknown parameters in solution vector $\mathbf{x}$. The matrix $\mathbf{A}$ need not be square and could simply contain multiple instances of the same equation or linear combinations thereof.

In this case, there may be no inverse matrix and because of the contradictory information, there is no solution $\mathbf{x}$ that satisfies the system of linear equations $\mathbf{Ax} = \mathbf{b}$. Rather, we can find the "best" solution $\tilde{\mathbf{x}}$ from the equation

$$\mathbf{A}^+\mathbf{b} = \tilde{\mathbf{x}}. \tag{2-230}$$

Here $\tilde{\mathbf{x}}$ is the best approximate for the solution vector $\mathbf{x}$ in that this choice minimizes

$$|\mathbf{b} - \mathbf{Ax}|,$$

the magnitude of the difference between the measurements and $\mathbf{A}$ times the unknown model parameters $\mathbf{x}$.

For a numerical example, consider the mathematical model with three unknown variables and four measurements of different linear combinations of two of them

$$\begin{aligned} x + y &= 2, \\ y + z &= 1, \\ x - z &= \frac{3}{2}, \end{aligned}$$

$$x - z = \frac{3}{4}.$$

Inspecting this system, the first two equations are independent. The third and fourth equations are identical except for the right-hand side being different (e.g. because of measurement uncertainties) and the left-hand side is the difference between the first and second equations. The matrix $\mathbf{A}$ for this system is

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}. \tag{2-231}$$

Since this matrix is not square, we know that $\mathbf{A}^{-1}$ cannot exist. Furthermore, since the equations are obviously inconsistent, there is no solution vector that satisfies the linear system. While the inverse may not exist, the pseudoinverse does. This may be computed by taking the singular value decomposition and then computing the pseudoinverse from $\mathbf{A}^+ = \mathbf{V}\boldsymbol{\Sigma}^+\mathbf{U}^\top$. The result of the singular value decomposition is

$$\mathbf{V} = \frac{1}{\sqrt{6}} \begin{bmatrix} -\sqrt{3} & 1 & \sqrt{2} \\ 0 & 2 & -\sqrt{2} \\ \sqrt{3} & 1 & \sqrt{2} \end{bmatrix}, \tag{2-232a}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & \sqrt{5} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{2-232b}$$

$$\mathbf{U} = \frac{1}{\sqrt{30}} \begin{bmatrix} -\sqrt{3} & \sqrt{15} & -\sqrt{10} & -\sqrt{2} \\ \sqrt{3} & \sqrt{15} & \sqrt{10} & \sqrt{2} \\ -2\sqrt{3} & 0 & 0 & 3\sqrt{2} \\ -2\sqrt{3} & 0 & \sqrt{10} & -2\sqrt{2} \end{bmatrix}. \tag{2-232c}$$

The pseudoinverse is then

$$\mathbf{A}^+ = \frac{1}{15} \begin{bmatrix} 4 & 1 & 3 & -3 \\ 5 & 5 & 0 & 0 \\ 1 & 4 & -3 & -3 \end{bmatrix}. \tag{2-233}$$

The solution vector that minimizes the difference between the measurement and $\mathbf{Ax}$ is

$$\tilde{\mathbf{x}} = \frac{1}{15} \begin{bmatrix} 4 & 1 & 3 & -3 \\ 5 & 5 & 0 & 0 \\ 1 & 4 & -3 & -3 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \\ \frac{3}{2} \\ \frac{3}{4} \end{bmatrix} = \begin{bmatrix} \frac{3}{4} \\ 1 \\ -\frac{1}{20} \end{bmatrix}. \tag{2-234}$$

Chapter 3

Ordinary Differential Equations

In the previous chapter, we studied algebraic systems of linear equations. As we saw, some problems in science and engineering naturally yield linear systems, whereas many others, such as those involving partial differential equations, can be approximated by large systems of linear equations. There is an important class of problems that are described by coupled systems of ordinary differential equations (ODEs). This chapter is split into two parts: the first covers first-order ordinary differential equations and the second covers second-order ordinary differential equations.

In the first part, the first-order linear ordinary differential equation are be reviewed. Within this, we focus first on analytical techniques with a specific focus on the method of diagonalization. Next the discussion moves onto numerical methods for solving non-linear systems of first-order ordinary differential equations including the forward, backward, and improved Euler methods for solving linear systems of equations as well as the Newton-Raphson iteration for handling nonlinear equations.

The second part of the text reviews systems of second-order ordinary differential equations with a focus on initial and boundary value problems. In the first part, coupled systems of initial value problems is covered. The second part develops the common boundary and interface conditions and shows their application to several problems relevant to nuclear engineering.

3.1 Linear First-Order ODEs

The linear first-order ordinary differential equation can be written in the following form:

$$\frac{dy}{dt} + p(t)y(t) = q(t), \quad y(0) = y_0, \tag{3-1}$$

where $y(t)$ is the unknown function, t is some parameter (usually time), $p(t)$ and $q(t)$ are known functions and $y(0) = y_0$ describes the known initial condition.

3.1.a Integrating Factor Method

The solution method employs a technique called the integrating factor. Let this be given as

$$I(t) = \exp\left[\int_0^t p(t')dt'\right]. \tag{3-2}$$

If we multiply both sides of the equation by the integrating factor

$$I(t)\left[\frac{dy}{dt} + p(t)y(t)\right] = I(t)q(t),$$

we can rewrite the left-hand side as

$$\frac{d}{dt}\left(I(t)y(t)\right) = I(t)q(t). \tag{3-3}$$

The equivalence between these two equations can be shown by taking the product rule:

$$\begin{aligned}
\frac{d}{dt}\left(\exp\left[\int_0^t p(t')dt'\right]y(t)\right) &= \exp\left[\int_0^t p(t')dt'\right]\frac{dy}{dt} + y(t)\frac{d}{dt}\left(\exp\left[\int_0^t p(t')dt'\right]\right), \\
&= I(t)\frac{dy}{dt} + y(t)\exp\left[\int_0^t p(t')dt'\right]\frac{d}{dt}\left[\int_0^t p(t')dt'\right], \\
&= I(t)\frac{dy}{dt} + y(t)I(t)p(t), \\
&= I(t)\left[\frac{dy}{dt} + p(t)y(t)\right].
\end{aligned}$$

Before proceeding to derive a general formula, let us do a couple specific examples to illustrate.

Example 1

First consider the radioactive decay problem with decay constant $p(t) = \lambda$, constant source $q(t) = Q$, and $y(0) = y_0$:

$$\frac{dy}{dt} + \lambda y(t) = Q, \quad y(0) = y_0. \tag{3-4a}$$

The integrating factor is

$$I(t) = \exp\left[\int_0^t \lambda dt'\right] = e^{\lambda t}. \tag{3-4b}$$

Multiplying the equation by $I(t) = e^{\lambda t}$ and factoring gives

$$\frac{d}{dt}\left[e^{\lambda t}y(t)\right] = Qe^{\lambda t}. \tag{3-4c}$$

To solve this differential equation, integrate both sides from 0 to t, replacing the t with a dummy variable t' in the integrand:

$$\int_0^t \frac{d}{dt'}\left[e^{\lambda t'}y(t')\right]dt' = \int_0^t Qe^{\lambda t'}dt'. \tag{3-4d}$$

From the second fundamental theorem of calculus, the integral of the derivative results in the function being differentiated evaluated at the endpoints, and the right-hand side can be integrated directly:

$$e^{\lambda t}y(t) - e^0 y(0) = \frac{Q}{\lambda}\left(e^{\lambda t} - 1\right). \tag{3-4e}$$

Next, we solve for $y(t)$ to obtain the final result:

$$y(t) = y(0)e^{-\lambda t} + \frac{Q}{\lambda}\left(1 - e^{\lambda t}\right). \tag{3-4f}$$

The physics of this problem is such that (i) the initial population of radionuclides decays away exponentially and (ii) there is a buildup of radionuclides from the source that eventually equilibrates with its decay rate to reach a population of Q/λ.

Example 2

As a second example, consider the differential equation:

$$\frac{dy}{dt} + 2\alpha t y(t) = \beta t, \quad y(0) = 0 \tag{3-5a}$$

where α and β are known constants. The integrating factor is

$$I(t) = \exp\left[\int_0^t 2\alpha t' dt'\right] = e^{\alpha t^2}. \tag{3-5b}$$

Multiplying by the integrating factor and integrating both sides from 0 to t result in

$$e^{\alpha t^2}y(t) = \beta\int_0^t t' e^{\alpha(t')^2}dt'. \tag{3-5c}$$

Note that $y(0) = 0$ has been applied. The integral of the right-hand side can be performed using the substitution $u = \alpha(t')^2$:

$$e^{\alpha t^2}y(t) = \frac{\beta}{2\alpha}\int_0^{\alpha t^2} e^u du. \tag{3-5d}$$

Carrying out the integral and solving for $y(t)$ gives the result

$$y(t) = \frac{\beta}{2\alpha}\left(1 - e^{-\alpha t^2}\right). \tag{3-5e}$$

3.1.b General Solution Form

Now, with those examples, we can develop a general solution to the first-order linear ODE. Returning to Eq. (3-3) integrate both sides from 0 to t:

$$\int_0^t \frac{d}{dt'}\left(\exp\left[\int_0^{t'} p(t'')dt''\right] y(t')\right) dt' = \int_0^t \exp\left[\int_0^{t'} p(t'')dt''\right] q(t')dt'. \tag{3-6a}$$

Note that all of the t in the integral are switched to dummy variables t' and, keeping with the same trend, the t' become t''. The integral on the left-hand side can be carried out using the second fundamental theorem of calculus, where the derivative of an integral is the argument of the derivative evaluated at the end points:

$$\exp\left[\int_0^t p(t')dt'\right] y(t) - \exp\left[\int_0^0 p(t')dt'\right] y(0) = \int_0^t \exp\left[\int_0^{t'} p(t'')dt''\right] q(t')dt'. \tag{3-6b}$$

The integral of zero to zero of any function is zero and $e^0 = 1$. Moving the second term to the right and side and solving for $y(t)$ gives:

$$\begin{aligned} y(t) = y(0)\exp\left[-\int_0^t p(t')dt'\right] \\ + \exp\left[-\int_0^t p(t')dt'\right]\int_0^t \exp\left[\int_0^{t'} p(t'')dt''\right] q(t')dt'. \end{aligned} \tag{3-6c}$$

Now consider the term

$$\exp\left[-\int_0^t p(t')dt'\right];$$

the t' in the integral is a dummy variable and we are free to use any symbol, so instead let us write t' as t'':

$$\exp\left[-\int_0^t p(t')dt'\right] \to \exp\left[-\int_0^t p(t'')dt''\right].$$

Since this expression is not a function of t' we can bring it into the integral. The second term on the right-hand side becomes:

$$\begin{aligned} &\int_0^t \exp\left[-\int_0^t p(t'')dt''\right]\exp\left[\int_0^{t'} p(t'')dt''\right] q(t')dt' \\ &= \int_0^t \exp\left[\int_0^{t'} p(t'')dt'' - \int_0^t p(t'')dt''\right] q(t')dt', \\ &= \int_0^t \exp\left[\int_0^{t'} p(t'')dt'' - \left(\int_0^{t'} p(t'')dt'' + \int_{t'}^t p(t'')dt''\right)\right] q(t')dt', \end{aligned}$$

$$= \int_0^t \exp\left[-\int_{t'}^t p(t'')dt''\right] q(t')dt'. \tag{3-6d}$$

Substituting this into our equation, we obtain an expression for the general solution for the first-order linear ordinary differential equation:

$$y(t) = y(0)\exp\left[-\int_0^t p(t')dt'\right] + \int_0^t \exp\left[-\int_{t'}^t p(t'')dt''\right] q(t')dt'. \tag{3-7}$$

This general expression is a bit unwieldy, but let us try to parse each term. First, $p(t)$ can be thought of as a time-dependent growth or decay factor governed by the physics of the underlying system, e.g., this equation could be used to model the number of neutrons in a nuclear reactor where the rate coefficient depends on the criticality state of the reactor. The term on the left

$$y(0)\exp\left[-\int_0^t p(t')dt'\right]$$

takes the initial population and exponentially grows or decays it with the time-dependent growth or decay factor. The term on the right

$$\int_0^t \exp\left[-\int_{t'}^t p(t'')dt''\right] q(t')dt'$$

is a bit more difficult to parse. First the integral over t' can be thought of as a sum of small time intervals where a quantity (e.g., neutrons in a nuclear reactor) is introduced into the system at some time t' by way of the function $q(t')$. The quantity introduced in this time interval about t' must similarly go through exponential growth and decay up until the time t; hence, the limits of integration over the time interval t'' are from t', the time the quantity entered the system, until the time of interest t by way of $p(t'')$.

To illustrate the general form applied to the radioactive decay example, first consider radioactive decay with a constant source:

$$y(t) = y(0)\exp\left[-\int_0^t \lambda dt'\right] + \int_0^t \exp\left[-\int_{t'}^t \lambda dt''\right] Q dt'. \tag{3-8a}$$

The integral in the exponential terms can be carried out simply:

$$y(t) = y(0)e^{-\lambda t} + Q\int_0^t e^{-\lambda(t-t')}dt'. \tag{3-8b}$$

Taking the t term in the exponential out of the integral,

$$y(t) = y(0)e^{-\lambda t} + Qe^{-\lambda t}\int_0^t e^{\lambda t'}dt'. \tag{3-8c}$$

Carrying out the integral gives

$$y(t) = y(0)e^{-\lambda t} + \frac{Q}{\lambda}e^{-\lambda t}\left(e^{\lambda t} - 1\right). \tag{3-8d}$$

Multiplying the exponential through gives

$$y(t) = y(0)e^{-\lambda t} + \frac{Q}{\lambda}\left(1 - e^{-\lambda t}\right), \tag{3-8e}$$

which is equivalent to the result we had before.

3.1.c Operator for First-Order Linear ODE

To connect this to linear algebra, we can write the linear ODE as

$$\left(\frac{d}{dt} + p(t)\right) y(t) = q(t) \tag{3-9}$$

where we factored out $y(t)$ to the right. The term in parentheses is called a *linear operator* and the equation can be written compactly as

$$Ly(t) = q(t) \tag{3-10}$$

where

$$L = \frac{d}{dt} + p(t) \tag{3-11}$$

This is very similar to the form of a linear matrix-vector system $\mathbf{Ax} = \mathbf{b}$. Despite the fact that L involves a derivative, and more generally could involve any number of derivatives or integrals so long as they are linear, the system $Ly = q$ has very similar properties. For this special case, we can determine L^{-1} analytically, which is the general solution we found earlier:

$$L^{-1} = \int_0^t \exp\left[-\int_{t'}^t p(t'')dt''\right](\cdot)dt' \tag{3-12}$$

with the $(\cdot)$ denoting where the function $q(t')$ would go. Note that we still need to be careful and apply the initial condition appropriately.

3.2 Linear ODE Systems

A linear system of ordinary differential equations is a set of coupled linear first-order ODEs. For example, two coupled linear systems are of the form:

$$\frac{dy_1}{dt} + p_{11}(t)y_1(t) + p_{12}y_2(t) = q_1(t), \qquad \frac{dy_2}{dt} + p_{21}(t)y_1(t) + p_{22}y_2(t) = q_1(t).$$

We can generalize this to any number of equations:

$$\begin{aligned}
&\frac{dy_1}{dt} + p_{11}(t)y_1(t) + p_{12}(t)y_2(t) + \cdots + p_{1N}(t)y_N(t) = q_1(t),\\
&\frac{dy_2}{dt} + p_{21}(t)y_1(t) + p_{22}(t)y_2(t) + \cdots + p_{2N}(t)y_N(t) = q_2(t),\\
&\quad\vdots\\
&\frac{dy_N}{dt} + p_{N1}(t)y_1(t) + p_{N2}(t)y_2(t) + \cdots + p_{NN}(t)y_N(t) = q_N(t).
\end{aligned} \tag{3-13}$$

This is often written in matrix-vector form as

$$\frac{d}{dt}\begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_N(t) \end{bmatrix} + \begin{bmatrix} p_{11}(t) & p_{12}(t) & \cdots & p_{1N}(t) \\ p_{21}(t) & p_{22}(t) & \cdots & p_{2N}(t) \\ \vdots & \vdots & \ddots & \vdots \\ p_{N1}(t) & p_{N2}(t) & \cdots & p_{NN}(t) \end{bmatrix} \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_N(t) \end{bmatrix} = \begin{bmatrix} q_1(t) \\ q_2(t) \\ \vdots \\ q_N(t) \end{bmatrix}. \quad (3\text{-}14)$$

Or more compactly as

$$\mathbf{y}'(t) + \mathbf{P}(t)\mathbf{y}(t) = \mathbf{q}(t). \quad (3\text{-}15)$$

In general, this is a very difficult system of equations to solve by hand and usually numerical methods must be used. Fortunately, many systems in science and engineering have a coefficient matrix $\mathbf{P}(t)$ that is independent of time, which we simply call A. Therefore, for the constant coefficient case

$$\mathbf{y}'(t) + \mathbf{A}\mathbf{y}(t) = \mathbf{q}(t). \quad (3\text{-}16)$$

This system is usually solvable analytically so long as we can integrate all of the elements of $\mathbf{q}(t)$ times an integrating factor.

3.2.a Triangular Systems of ODEs with Constant Coefficients

Before discussing the general case of a full matrix of constant coefficients, we will consider the important case where $\mathbf{A}$ is lower triangular. This arises in many applications in health physics involving radioactive decay. A lower triangular system is

$$\begin{aligned} &\frac{dy_1}{dt} + a_{11}y_t(1) = q_1(t), \\ &\frac{dy_2}{dt} + a_{21}y_1(t) + a_{22}y_2(t) = q_2(t), \\ &\quad\vdots \\ &\frac{dy_N}{dt} + a_{N1}y_1(t) + a_{N2}y_2(t) + \cdots + a_{NN}y_N(t) = q_N(t). \end{aligned} \quad (3\text{-}17)$$

As with linear systems of algebraic equations, this can be solved using backward substitution, but now using an integrating factor. The issue is that doing this by hand will yield increasingly complicated terms that need to be integrated, and it is rare that a system with more than a few equations is solved by hand. Nonetheless, we can provide the example of the three-component decay chain.

3.2.b Example: Three Component Decay Chain

Let us consider the scenario where we want to find the population of isotopes $N_1(t)$, $N_2(t)$, and $N_3(t)$ where the first isotope decays to the second, the second decays

to the third and the third is stable, as a function of time given initial populations $N_1(0), N_2(0)$, and $N_3(0)$. The system of equations may be written as follows:

$$\frac{dN_1}{dt} = -\lambda_1 N_1(t), \tag{3-18a}$$

$$\frac{dN_2}{dt} = \lambda_1 N_1(t) - \lambda_2 N_2(t), \tag{3-18b}$$

$$\frac{dN_3}{dt} = \lambda_2 N_2(t). \tag{3-18c}$$

The linear system has the form:

$$\frac{d}{dt}\begin{bmatrix} N_1(t) \\ N_2(t) \\ N_3(t) \end{bmatrix} + \begin{bmatrix} \lambda_1 & 0 & 0 \\ -\lambda_1 & \lambda_2 & 0 \\ 0 & -\lambda_2 & 0 \end{bmatrix} \begin{bmatrix} N_1(t) \\ N_2(t) \\ N_3(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{3-19}$$

The equation for the first isotope is

$$\frac{dN_1}{dt} + \lambda_1 N_1(t) = 0. \tag{3-20a}$$

This could be solved using separation of variables, however, to keep a consistent methodology, let us apply the integrating factor. We can write the equation as:

$$\frac{d}{dt}\left[N_1(t)e^{\lambda_1 t}\right] = 0. \tag{3-20b}$$

Integrating both sides from 0 to t and solving for $N_1(t)$ yields

$$N_1(t) = N_1(0)e^{-\lambda_1 t}. \tag{3-20c}$$

This equation gives the standard result that the initial population of radionuclides decays away with time toward zero.

Continuing with the second equation, we write

$$\frac{dN_2}{dt} + \lambda_2 N_2(t) - \lambda_1 N_1(t) = 0. \tag{3-21a}$$

Substituting in $N_1(t)$ and moving it to the right-hand side

$$\frac{dN_2}{dt} + \lambda_2 N_2(t) = \lambda_1 N_1(0)e^{-\lambda_1 t}. \tag{3-21b}$$

Applying the integrating factor gives

$$\frac{d}{dt}\left[N_2(t)e^{\lambda_2 t}\right] = \lambda_1 N_1(0)e^{(\lambda_2-\lambda_1)t}. \tag{3-21c}$$

Integrating both sides from 0 to t gives

$$N_2(t)e^{\lambda_2 t} = N_2(0) + \frac{\lambda_1 N_1(0)}{\lambda_2 - \lambda_1}\left(e^{(\lambda_2-\lambda_1)t} - 1\right). \tag{3-21d}$$

And solving for $N_2(t)$ gives

$$N_2(t) = N_2(0)e^{-\lambda_2 t} + \frac{\lambda_1 N_1(0)}{\lambda_2 - \lambda_1}\left(e^{-\lambda_1 t} - e^{-\lambda_2 t}\right). \tag{3-21e}$$

As before, the initial population of isotope 2 decays away exponentially; however, there is also a term where isotope 2 builds up from decay of isotope 1 and then subsequently decays.

Finally, we have the equation for the population of the third isotope; inserting $N_2(t)$ yields

$$\frac{dN_3}{dt} = \lambda_2\left[N_2(0)e^{-\lambda_2 t} + \frac{\lambda_2\lambda_1 N_1(0)}{\lambda_2 - \lambda_1}\left(e^{-\lambda_1 t} - e^{-\lambda_2 t}\right)\right]. \tag{3-22a}$$

Observe that there is no dependence on $N_3(t)$, so this equation may be solved by directly integrating from 0 to t:

$$\begin{aligned} N_3(t) = N_3(0) &+ N_2(0)\left(1 - e^{-\lambda_2 t}\right) \\ &+ \frac{\lambda_2\lambda_1 N_1(0)}{\lambda_2 - \lambda_1}\left[\frac{1}{\lambda_1}\left(1 - e^{-\lambda_1 t}\right) - \frac{1}{\lambda_2}\left(1 - e^{-\lambda_2 t}\right)\right]. \end{aligned} \tag{3-22b}$$

The third term can be simplified by

$$\begin{aligned} N_3(t) = N_3(0) &+ N_2(0)\left(1 - e^{-\lambda_2 t}\right) \\ &+ \frac{\lambda_2\lambda_1 N_1(0)}{\lambda_2 - \lambda_1}\left[\frac{\lambda_2}{\lambda_2\lambda_1}\left(1 - e^{-\lambda_1 t}\right) - \frac{\lambda_1}{\lambda_2\lambda_1}\left(1 - e^{-\lambda_2 t}\right)\right], \end{aligned}$$

to yield

$$\begin{aligned} N_3(t) = N_3(0) &+ N_2(0)\left(1 - e^{-\lambda_2 t}\right) \\ &+ \frac{N_1(0)}{\lambda_2 - \lambda_1}\left[\lambda_2\left(1 - e^{-\lambda_1 t}\right) - \lambda_1\left(1 - e^{-\lambda_2 t}\right)\right]. \end{aligned} \tag{3-22c}$$

The first term is the initial population of isotope 3, which is constant because it is stable. The second term denotes the initial population of isotope 2 multiplied by a factor of $1 - e^{-\lambda_2 t}$; this factor goes to one for large times and we can say that all of isotope 2 will eventually decay to isotope 1. The third term can be inspected by letting t become large so the exponential terms go to zero:

$$\frac{N_1(0)}{\lambda_2 - \lambda_1}\left[\lambda_2\left(1 - e^{-\lambda_1 t}\right) - \lambda_1\left(1 - e^{-\lambda_2 t}\right)\right] \to \frac{N_1(0)}{\lambda_2 - \lambda_1}\left(\lambda_2 - \lambda_1\right) = N_1(0),$$

which means that all of isotope 1 will also eventually decay to isotope 3 given a long enough time. These results make logical sense that all radioactive isotopes decay to stable ones eventually.

3.2.c Approximations for Long Time

It is often possible to gain insight about the solution of a triangular system of linear ordinary differential equations for large values of t if the coefficients are of very different magnitudes, as is often the case in radioactive decay where half-lives can vary from milliseconds or less to millions of years or more. In the context of radioactive decay we define the activity of a particular isotope:

$$A_k(t) = \lambda_k N_k(t), \tag{3-23}$$

which represents the instantaneous decay rate of isotope k at time t.

For the case of the three component decay chain, let us consider the cases where the first isotope decays much more rapidly than the second, $\lambda_1 \gg \lambda_2$. In this case, all of isotope 1 will have decayed away for large times,

$$A_1(t) = A_1(0)e^{-\lambda_1 t} \to 0. \tag{3-24}$$

The activity of isotope 2 is

$$\begin{aligned} A_2(t) &= A_2(0)e^{-\lambda_2 t} + \frac{\lambda_2 A_1(0)}{\lambda_2 - \lambda_1}\left(e^{-\lambda_1 t} - e^{-\lambda_2 t}\right) \\ &\to A_2(0)e^{-\lambda_2 t} + \frac{\lambda_2}{\lambda_2 - \lambda_1}A_1(0)\left(0 - e^{-\lambda_2 t}\right) \\ &= \left[A_2(0) - \frac{\lambda_2}{\lambda_2 - \lambda_1}A_1(0)\right]e^{-\lambda_2 t} \\ &= \left[A_2(0) + \frac{\lambda_2}{\lambda_1 - \lambda_2}A_1(0)\right]e^{-\lambda_2 t}. \end{aligned} \tag{3-25}$$

The total activity is the sum of the activities of both radioisotopes, so the total activity at long times is therefore approximately:

$$A(t) \approx \left[A_2(0) + \frac{\lambda_2}{\lambda_1 - \lambda_2}A_1(0)\right]e^{-\lambda_2 t}. \tag{3-26}$$

The other case is where the daughter decays much more quickly than the parent, or $\lambda_2 \gg \lambda_1$. On a long time scale, the decay of radioisotope 1 limits the decrease in activity whereas the decay of radioisotope 2 is effectively instantaneous. The activity of radioisotope 1 is simply

$$A_1(t) = A_1(0)e^{-\lambda_1 t}. \tag{3-27}$$

The activity of radioisotope 2 is

$$\begin{aligned} A_2(t) &= A_2(0)e^{-\lambda_2 t} + \frac{\lambda_2 A_1(0)}{\lambda_2 - \lambda_1}\left(e^{-\lambda_1 t} - e^{-\lambda_2 t}\right) \\ &\to 0 + \frac{\lambda_2}{\lambda_2 - \lambda_1}A_1(0)\left(e^{-\lambda_1 t} - 0\right) \\ &= A_1(0)\frac{\lambda_2}{\lambda_2 - \lambda_1}e^{-\lambda_1 t}. \end{aligned} \tag{3-28}$$

Therefore, the total activity for late times is the approximately

$$\begin{aligned} A(t) &\approx A_1(0)e^{-\lambda_1 t} + A_1(0)\frac{\lambda_2}{\lambda_2 - \lambda_1}e^{-\lambda_1 t} \\ &= A_1(0)\left[1 + \frac{\lambda_2}{\lambda_2 - \lambda_1}\right]e^{-\lambda_1 t}. \end{aligned} \tag{3-29}$$

If λ_2 is very small relative to λ_1, then the activity of the entire sample at late times is effectively twice the activity of $A_1(t)$. This condition is referred to as *secular equilibrium*.

3.3 Matrix Exponential Solution

For the linear system of ODEs with constant coefficients,

$$\mathbf{y}'(t) + \mathbf{A}\mathbf{y}(t) = \mathbf{q}(t), \tag{3-30}$$

it is possible to apply an integrating factor in a similar manner. We may apply a matrix integrating factor that is the exponential of the matrix $\mathbf{A}$:

$$\exp\left(t\mathbf{A}\right). \tag{3-31}$$

We multiply the integrating factor on the left (order now matters because matrix multiplication is non-commutative). After factoring the left-hand side and integrating from 0 to t, we get

$$\int_0^t \frac{d}{dt'}\left[\exp\left(t'\mathbf{A}\right)\mathbf{y}(t')\right]dt' = \int_0^t \exp\left(t'\mathbf{A}\right)\mathbf{q}(t')dt'.$$

Carrying out the integral and solving for $\mathbf{y}(t)$ gives the following result:

$$\mathbf{y}(t) = \exp\left(-t\mathbf{A}\right)\mathbf{y}(0) + \int_0^t \exp\left[-(t-t')\mathbf{A}\right]\mathbf{q}(t')dt'. \tag{3-32}$$

While this equation is the general solution to the a system of linear ordinary differential equations with constant coefficients, we still must define the exponential of a matrix.

3.3.a Properties of the Matrix Exponential

The matrix exponential can be defined in terms of a Taylor series. Recall that for the standard (scalar) exponential:

$$e^{\lambda t} = 1 + \lambda t + \frac{1}{2}(\lambda t)^2 + \cdots + \frac{1}{n!}(\lambda t)^n + \cdots \tag{3-33}$$

It follows that the matrix exponential is

$$e^{t\mathbf{A}} = \mathbf{I} + t\mathbf{A} + \frac{t^2}{2}\mathbf{A}^2 + \cdots + \frac{t^n}{n!}\mathbf{A}^n + \cdots \tag{3-34}$$

where $\mathbf{A}^n$ is the matrix $\mathbf{A}$ multiplied by itself n times using matrix multiplication. Note that if $\mathbf{A}$ is an $N\times N$ matrix, then $e^{\mathbf{A}}$ is also an $N\times N$ matrix. Additionally, the exponential of a zero matrix is the identity matrix $\mathbf{I}$. Finally, matrix exponentials satisfy the property:

$$\exp(\mathbf{A})\exp(\mathbf{B}) = \exp(\mathbf{A}+\mathbf{B}). \tag{3-35}$$

Another important property is that matrix exponential of a diagonal matrix $\mathbf{D}$ is simply the exponential of the diagonal elements:

$$\exp(\mathbf{D}) = \begin{bmatrix} e^{d_1} & 0 & 0 & \cdots \\ 0 & e^{d_2} & 0 & \cdots \\ 0 & 0 & e^{d_3} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{3-36}$$

In all other cases, the matrix exponential is **not** simply the exponential of the matrix elements.

Because the matrix exponential in general requires several matrix multiplications, it is rare that it is actually computed using the Taylor series directly. Rather, we often convert matrix $\mathbf{A}$ into a diagonal matrix $\mathbf{D}$ through a change of basis. Once this is done, then the matrix exponential can be evaluated simply and the change of basis can be reversed in a much simpler manner.

3.4 Diagonalization

To efficiently apply the matrix exponential solution, it is often necessary to devise a change of basis to make the system of equations diagonal. What this change of basis does is remove the coupling between the different functions.

To find this basis recall that if a matrix $\mathbf{A}$ is multiplied on its eigenvector $\mathbf{v}$ then the result is a scalar multiple of the eigenvector $\lambda\mathbf{v}$. It would make sense then to set the new basis vectors as the eigenvectors. Let the transformation matrix $\mathbf{T}$ be a matrix where its columns are the eigenvectors:

$$\mathbf{T} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_N \end{bmatrix}. \tag{3-37}$$

The vector components of $\mathbf{y}$ transform contravariantly, so to get the components in the new new basis

$$\mathbf{z}(t) = \mathbf{T}^{-1}\mathbf{y}(t), \tag{3-38}$$

and therefore can also write write

$$\mathbf{y}(t) = \mathbf{T}\mathbf{z}(t). \tag{3-39}$$

Given the linear system of ODEs,

$$\frac{d\mathbf{y}(t)}{dt} + \mathbf{A}\mathbf{y}(t) = \mathbf{q}(t),$$

we can apply the transformation to obtain

$$\mathbf{T}\frac{d\mathbf{z}(t)}{dt} + \mathbf{A}\mathbf{T}\mathbf{z}(t) = \mathbf{q}(t). \tag{3-40}$$

Let us inspect the product $\mathbf{AT}$. Recall that each column of $\mathbf{T}$ is an eigenvector, so when acted on by $\mathbf{A}$, will multiply each column by a constant being the corresponding eigenvalue:

$$\mathbf{AT} = \begin{bmatrix} \lambda_1\mathbf{v}_1 & \lambda_2\mathbf{v}_2 & \cdots & \lambda_N\mathbf{v}_N \end{bmatrix}. \tag{3-41}$$

Since each column is multiplied by a different scalar constant, we can factor the new matrix as the old matrix times a diagonal matrix with the diagonal elements as the scalar constants, i.e.,

$$\begin{bmatrix} \lambda_1 v_{11} & \lambda_2 v_{12} & \lambda_3 v_{13} & \cdots \\ \lambda_1 v_{21} & \lambda_2 v_{22} & \lambda_3 v_{23} & \cdots \\ \lambda_1 v_{31} & \lambda_2 v_{32} & \lambda_3 v_{33} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} = \begin{bmatrix} v_{11} & v_{12} & v_{13} & \cdots \\ v_{21} & v_{22} & v_{23} & \cdots \\ v_{31} & v_{32} & v_{33} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots \\ 0 & \lambda_2 & 0 & \cdots \\ 0 & 0 & \lambda_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{3-42}$$

Therefore,

$$\mathbf{AT} = \mathbf{TD} \tag{3-43}$$

where

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots \\ 0 & \lambda_2 & 0 & \cdots \\ 0 & 0 & \lambda_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{3-44}$$

Inserting this back into the linear system for $\mathbf{z}(t)$,

$$\mathbf{T}\frac{d\mathbf{z}(t)}{dt} + \mathbf{T}\mathbf{D}\mathbf{z}(t) = \mathbf{q}(t), \tag{3-45}$$

we can multiply the system by $\mathbf{T}^{-1}$ from the left to obtain

$$\frac{d\mathbf{z}(t)}{dt} + \mathbf{D}\mathbf{z}(t) = \mathbf{T}^{-1}\mathbf{q}(t) = \mathbf{r}(t), \tag{3-46}$$

Therefore, applying the integrating factor of the matrix exponential, we obtain the solution as before

$$\mathbf{z}(t) = \exp\left(-t\mathbf{D}\right)\mathbf{z}(0) + \int_0^t \exp\left[-(t-t')\mathbf{D}\right]\mathbf{r}(t')dt'. \tag{3-47}$$

Transforming back to $\mathbf{y}(t)$ gives the result

$$\mathbf{y}(t) = \mathbf{T}\exp\left(-t\mathbf{D}\right)\mathbf{z}(0) + \mathbf{T}\int_0^t \exp\left[-(t-t')\mathbf{D}\right]\mathbf{r}(t')dt'. \tag{3-48}$$

The vectors $\mathbf{z}(0)$ and $\mathbf{r}(t)$ can be found by solving the linear systems

$$\mathbf{T}\mathbf{z}(0) = \mathbf{y}(0), \tag{3-49a}$$
$$\mathbf{T}\mathbf{r}(t) = \mathbf{q}(t). \tag{3-49b}$$

Since the matrix exponential of a diagonal matrix is also a diagonal matrix, we can write

$$\mathbf{T}\exp\left(-t\mathbf{D}\right) = \left[\begin{array}{cccc} e^{-\lambda_1 t}\mathbf{v}_1 & e^{-\lambda_2 t}\mathbf{v}_2 & \cdots \end{array}\right], \tag{3-50}$$

multiplying by the column vector $\mathbf{z}(0)$ gives a sum of terms

$$\mathbf{T}\exp\left(-t\mathbf{D}\right)\mathbf{z}(0) = z_1(0)e^{-\lambda_1 t}\mathbf{v}_1 + z_2(0)e^{-\lambda_2 t}\mathbf{v}_2 + \ldots; \tag{3-51}$$

the inhomogeneous term can be expanded

$$\begin{aligned}
&\mathbf{T}\int_0^t \exp\left[-(t-t')\mathbf{D}\right]\mathbf{r}(t')dt') \\
&= \begin{bmatrix} v_{11} & v_{12} & v_{13} & \cdots \\ v_{21} & v_{22} & v_{23} & \cdots \\ v_{31} & v_{32} & v_{33} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} \int_0^t e^{-\lambda_1(t-t')}r_1(t')dt' \\ \int_0^t e^{-\lambda_2(t-t')}r_2(t')dt' \\ \int_0^t e^{-\lambda_3(t-t')}r_3(t')dt' \\ \vdots \end{bmatrix} \\
&= \left(\int_0^t e^{-\lambda_1(t-t')}r_1(t')dt'\right)\mathbf{v}_1 + \left(\int_0^t e^{-\lambda_2(t-t')}r_2(t')dt'\right)\mathbf{v}_2 + \ldots
\end{aligned} \tag{3-52}$$

This solution can then be written as

$$\mathbf{y}(t) = \sum_{i=1}^{N}\left[z_i(0)e^{-\lambda_i t} + \int_0^t e^{-\lambda_i(t-t')}r_i(t')dt'\right]\mathbf{v}_i. \tag{3-53}$$

3.4.a Example: Mass Flow in Solution Tanks

Suppose there are two interconnected solution tanks containing some fluid as depicted in Fig. 3.1. Tank 1 drains into tank 2. Tank 2 can drain back into tank 1 as well as draining to the outside. At time $t = 0$, there are initially masses $m_1(0)$ and $m_2(0)$ in tanks 1 and 2 respectively. The rate of drainage from each tank is proportional to the mass fluid in each of the tanks. The drainage constant out of tank 1 into tank 2 is 6 hr^{-1}, the drainage constant out of tank 2 and back into tank 1 is 2 hr^{-1}, and the drainage constant out of tank 2 to the outside is 3 hr^{-1}. Also, suppose that fluid is being added to tank 1 at a constant rate Q. Our goal is to find the mass of fluid in each tank at time t.

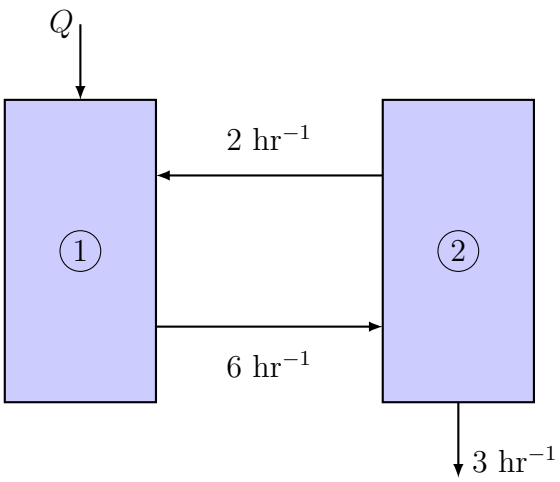

Figure 3.1: Mass flow in solution tank example.

First, we write down the differential equations:

$$\frac{dm_1}{dt} = -6m_1(t) + 2m_2(t) + Q, \tag{3-54a}$$

$$\frac{dm_2}{dt} = (-2-3)m_2(t) + 6m_1(t). \tag{3-54b}$$

Rearranging these into matrix-vector form gives:

$$\frac{d}{dt}\begin{bmatrix} m_1(t) \\ m_2(t) \end{bmatrix} + \begin{bmatrix} 6 & -2 \\ -6 & 5 \end{bmatrix}\begin{bmatrix} m_1(t) \\ m_2(t) \end{bmatrix} + \begin{bmatrix} Q \\ 0 \end{bmatrix}. \tag{3-55}$$

To diagonalize this system, we must find the eigenvalues and eigenvectors of the matrix

$$\mathbf{A} = \begin{bmatrix} 6 & -2 \\ -6 & 5 \end{bmatrix}. \tag{3-56}$$

These are:

$$\lambda = \{2, 9\}\,; \quad \mathbf{v} = \left\{ \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} -2 \\ 3 \end{bmatrix} \right\}. \tag{3-57}$$

The transformation matrix contains the eigenvectors as its columns:

$$\mathbf{T} = \begin{bmatrix} 1 & -2 \\ 2 & 3 \end{bmatrix}. \tag{3-58}$$

The diagonal matrix $\mathbf{D}$ has the corresponding eigenvalues as its elements:

$$\mathbf{D} = \begin{bmatrix} 2 & 0 \\ 0 & 9 \end{bmatrix}. \tag{3-59}$$

Now we find the initial conditions $\mathbf{z}(0)$ and the inhomogeneous term $\mathbf{r}(t)$ in the transformed coordinates. Since the system is only 2×2 we can compute its inverse simply as

$$\mathbf{T}^{-1} = \frac{1}{7}\begin{bmatrix} 3 & 2 \\ -2 & 1 \end{bmatrix}. \tag{3-60}$$

We can then compute these as

$$\mathbf{z}(0) = \mathbf{T}^{-1}\mathbf{m}(0) = \frac{1}{7}\begin{bmatrix} 3 & 2 \\ -2 & 1 \end{bmatrix}\begin{bmatrix} m_1(0) \\ m_2(0) \end{bmatrix} = \frac{1}{7}\begin{bmatrix} 3m_1(0) + 2m_2(0) \\ -2m_1(0) + m_2(0) \end{bmatrix}, \tag{3-61}$$

$$\mathbf{r}(t) = \mathbf{T}^{-1}\mathbf{q}(t) = \frac{1}{7}\begin{bmatrix} 3 & 2 \\ -2 & 1 \end{bmatrix}\begin{bmatrix} Q \\ 0 \end{bmatrix} = \frac{Q}{7}\begin{bmatrix} 3 \\ -2 \end{bmatrix}. \tag{3-62}$$

The solution is

$$\begin{aligned}\mathbf{m}(t) = &\left[\frac{1}{7}(3m_1(0) + 2m_2(0))e^{-2t} + \frac{3Q}{7}\int_0^t e^{-2(t-t')}dt'\right]\begin{bmatrix} 1 \\ 2 \end{bmatrix} \\ &+ \left[\frac{1}{7}(-2m_1(0) + m_2(0))e^{-9t} - \frac{2Q}{7}\int_0^t e^{-9(t-t')}dt'\right]\begin{bmatrix} -2 \\ 3 \end{bmatrix}.\end{aligned} \tag{3-63}$$

Carrying out the integrals gives

$$\begin{aligned}\mathbf{m}(t) = &\left[\frac{1}{7}(3m_1(0) + 2m_2(0))e^{-2t} + \frac{3Q}{14}(1 - e^{-2t})\right]\begin{bmatrix} 1 \\ 2 \end{bmatrix} \\ &+ \left[\frac{1}{7}(-2m_1(0) + m_2(0))e^{-9t} - \frac{2Q}{63}(1 - e^{-9t})\right]\begin{bmatrix} -2 \\ 3 \end{bmatrix}.\end{aligned} \tag{3-64}$$

Writing out the equations explicitly:

$$\begin{aligned}m_1(t) = &\, m_1(0)\left(\frac{3}{7}e^{-2t} + \frac{4}{7}e^{-9t}\right) + m_2(0)\left(\frac{2}{7}e^{-2t} - \frac{2}{7}e^{-9t}\right) \\ &+ \frac{3Q}{14}(1 - e^{-2t}) + \frac{4Q}{63}(1 - e^{-9t}),\end{aligned} \tag{3-65a}$$

$$\begin{aligned}m_2(t) = &\, m_1(0)\left(\frac{6}{7}e^{-2t} - \frac{6}{7}e^{-9t}\right) + m_2(0)\left(\frac{4}{7}e^{-2t} + \frac{3}{7}e^{-9t}\right) \\ &+ \frac{3Q}{7}(1 - e^{-2t}) + \frac{4Q}{21}(1 - e^{-9t}).\end{aligned} \tag{3-65b}$$

3.4.b Complex Eigenvalues

In many situations, the eigenvalues are complex, $\lambda = a + bi$ with $i = \sqrt{-1}$, the imaginary unit. The formulas used before are still valid, however, one additional piece of information is the Euler formula

$$e^{i\theta} = \cos\theta + i\sin\theta, \tag{3-66}$$

which relates the exponential of an imaginary number to trigonometric functions. This implies complex eigenvalues have oscillatory behavior. If the coefficients are all real, then we can assert that if an eigenvalue is complex $a + bi$, then its complex conjugate $a - bi$ is also an eigenvalue. The same can said of the corresponding eigenvectors. In these cases, the mathematics will work out so that the solution is entirely real.

Let us illustrate with an example:

$$\mathbf{y}'(t) + \begin{bmatrix} 4 & 0 & 0 \\ -4 & 1 & 1 \\ 0 & -1 & 1 \end{bmatrix} \mathbf{y}(t) = \mathbf{0}, \quad \mathbf{y}(0) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}. \tag{3-67}$$

The eigenvalues and eigenvectors of this matrix are:

$$\lambda = \{4, 1+i, 1-i\}; \quad \mathbf{v} = \left\{ \begin{bmatrix} 5 \\ -6 \\ 2 \end{bmatrix}, \begin{bmatrix} 0 \\ -i \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ i \\ 1 \end{bmatrix} \right\}. \tag{3-68}$$

The transformation matrix is therefore:

$$\mathbf{T} = \begin{bmatrix} 5 & 0 & 0 \\ -6 & -i & i \\ 2 & 1 & 1 \end{bmatrix}. \tag{3-69}$$

The equation is homogeneous, so we only have the initial condition in the transformed coordinate. This can be obtained by solving

$$\mathbf{T}\mathbf{z}(0) = \mathbf{y}(0)$$

to obtain

$$\mathbf{z}(0) = \frac{1}{5} \begin{bmatrix} 1 \\ -1+3i \\ -1-3i \end{bmatrix}. \tag{3-70}$$

The solution is therefore

$$\begin{aligned} \mathbf{y}(t) = \frac{1}{5} e^{-4t} \begin{bmatrix} 5 \\ -6 \\ 2 \end{bmatrix} + \frac{1}{5}(-1+3i) e^{-(1+i)t} \begin{bmatrix} 0 \\ -i \\ 1 \end{bmatrix} \\ + \frac{1}{5}(-1-3i) e^{-(1-i)t} \begin{bmatrix} 0 \\ i \\ 1 \end{bmatrix}. \end{aligned} \tag{3-71}$$

Applying Euler's formula to the complex exponential gives

$$\begin{aligned} \mathbf{y}(t) = \frac{1}{5} e^{-4t} \begin{bmatrix} 5 \\ -6 \\ 2 \end{bmatrix} + \frac{1}{5}(-1+3i) e^{-t}(\cos(t) - i\sin(t)) \begin{bmatrix} 0 \\ -i \\ 1 \end{bmatrix} \\ + \frac{1}{5}(-1-3i) e^{-t}(\cos(t) + i\sin(t)) \begin{bmatrix} 0 \\ i \\ 1 \end{bmatrix}. \end{aligned} \tag{3-72}$$

Writing each of these out explicitly:

$$y_1(t) = e^{-4t}, \tag{3-73a}$$

$$\begin{aligned} y_2(t) &= -\frac{6}{5}e^{-4t} + \frac{1}{5}(3+i)e^{-t}(\cos(t) - i\sin(t)) \\ &\qquad + \frac{1}{5}(3-i)e^{-t}(\cos(t) + i\sin(t)) \\ &= -\frac{6}{5}e^{-4t} + \left(\frac{6}{5}\cos(t) + \frac{2}{5}\sin(t)\right)e^{-t}, \end{aligned} \tag{3-73b}$$

$$\begin{aligned} y_3(t) &= \frac{2}{5}e^{-4t} + \frac{1}{5}(-1+3i)e^{-t}(\cos(t) - i\sin(t)) \\ &\qquad + \frac{1}{5}(-1-3i)e^{-t}(\cos(t) + i\sin(t)) \\ &= \frac{2}{5}e^{-4t} + \left(-\frac{2}{5}\cos(t) + \frac{6}{5}\sin(t)\right)e^{-t}. \end{aligned} \tag{3-73c}$$

The solution for $y_1(t)$ is pure exponential decay. The solutions for $y_2(t)$ and $y_3(t)$ are damped oscillations. Note the key point that despite imaginary values showing up in the intermediate steps, the solution is indeed real. This will always be the case if the coefficients are entirely real.

3.4.c Defective Matrices

As we saw in the last chapter, sometimes a system can have repeated eigenvalues, which may also have fewer linearly independent eigenvectors than the number of unknowns. This case may also arise in the context of systems of ODEs. What this means is the matrix cannot be diagonalized. Thankfully, this situation seldom arises in engineering applications, so it will only be mentioned briefly here.

When this occurs, we need to find additional information in the form of *generalized eigenvectors*. This is obtained from successive application of the operator $\mathbf{A} - \lambda\mathbf{I}$. Suppose $\mathbf{v}_1$ is the only linearly independent eigenvector for a particular eigenvalue. The generalized eigenvectors may be obtained from

$$\begin{aligned} (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_1 &= \mathbf{0}, \\ (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_2 &= \mathbf{v}_1, \\ (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_3 &= \mathbf{v}_2, \\ &\vdots \end{aligned}$$

Using these generalized eigenvectors to form a transformation matrix $\mathbf{T}$ yields an almost diagonalized matrix with a *Jordan normal form* where the matrix is diagonal with the exception of values of 1 that appear on the superdiagonal where eigenvalues with multiplicity and fewer linearly independent eigenvectors occur. When taking the matrix exponential of such a matrix, we end up with exponentials with polynomial coefficients $t, t^2, \ldots$ analogous to second-order ODEs for which the characteristic equation has a repeated root.

3.5 Numerical Techniques

The matrix exponential and diagonalization methods are only applicable if the ODE is linear and its coefficients are constant. In the cases where they are not, there is no general solution, and usually we revert to performing numerical approximation methods. This section introduces the techniques in the context of linear ODEs with coefficients that are functions of time. Following this, the techniques will be applied to systems of general first-order ODEs.

A system of first-order ODEs may be written generically as

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)), \quad \mathbf{y}(0) = \mathbf{y}_0. \tag{3-74}$$

Here $\mathbf{f}$ is a vector because each equation could involve a different function $f_i(t, \mathbf{y}(t))$. For the first-order linear ODE this becomes

$$\mathbf{y}'(t) = -\mathbf{P}(t)\mathbf{y}(t) + \mathbf{q}(t), \quad \mathbf{y}(0) = \mathbf{y}_0, \tag{3-75}$$

where $\mathbf{P}(t)$ is a general function of time, this system has no analytical solution; however, we could approximate the solution on a time grid: $t_0 = 0, t_1 = \Delta t_1, t_2 = t_1 + \Delta t_2, \ldots$ In most cases, Δt_n will be a fixed interval Δt so that $t_0 = 0, t_1 = \Delta t, t_2 = 2\Delta t, \ldots$ While not necessary, this simplifies the coding as well as the mathematical analysis. The equations herein will use a fixed Δt for simplicity.

The first derivative may be approximated using the finite difference:

$$\mathbf{y}'(t) \approx \frac{1}{\Delta t}\left(\mathbf{y}(t_{n+1}) - \mathbf{y}(t_n)\right) = \frac{1}{\Delta t}\left(\mathbf{y}_{n+1} - \mathbf{y}_n\right). \tag{3-76}$$

Here we introduced the shorthand $\mathbf{y}(t_n) = \mathbf{y}_n$ to keep the notation compact. Now we must map the right-hand side of the equation onto the time grid. There are different options for doing this that have implications on how well the solution is approximated as well as whether the numerical method converges at all. These notes will discuss the three elementary techniques: forward Euler, backward Euler, and improved Euler or the trapezoidal rule.

3.5.a Forward Euler

The forward Euler method takes the right-hand side at the left endpoint of the current time interval. After the finite difference is applied to the first derivative, the system can be written as

$$\frac{1}{\Delta t}\left(\mathbf{y}_{n+1} - \mathbf{y}_n\right) = \mathbf{f}(t_n, \mathbf{y}_n). \tag{3-77}$$

Solving for $\mathbf{y}_{n+1}$ is straightforward:

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \Delta t\mathbf{f}(t_n, \mathbf{y}_n). \tag{3-78}$$

For the linear ODE, this becomes

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \Delta t\left(-\mathbf{P}_n\mathbf{y}_n + \mathbf{q}_n\right). \tag{3-79}$$

Factoring out $\mathbf{y}_n$ to the right on the right-hand side gives the final result:

$$\mathbf{y}_{n+1} = (\mathbf{I} - \Delta t \mathbf{P}_n)\,\mathbf{y}_n + \Delta t \mathbf{q}_n. \tag{3-80}$$

The forward Euler method is advantageous because of its simplicity. Computing the approximate scheme simply involves doing a matrix multiplication at each time step and adding on the inhomogeneous term. Forward Euler is classified as an *explicit method* in that the right-hand side of the original is completely known and therefore the solution at the next time step may be computed directly.

Unfortunately, forward Euler has a couple significant issues limiting its usefulness. The first issue is related to its convergence rate. The error incurred in the solution is proportional to the size of the time step Δt. It is often the case that to get suitable accuracy, one needs to take unacceptably small time steps. The other major issue with forward Euler is how the errors tend to compound. If Δt is too large, we can get into a situation where the errors tend to geometrically grow with time. This means the solution becomes progressively worse with time until the solution "blows up". We say that the forward Euler method has the potential to become unstable if the time step is too large.

To illustrate some of these issues, let us consider the example from Sec. 3.4.b. There we had the linear system:

$$\mathbf{y}'(t) + \begin{bmatrix} 4 & 0 & 0 \\ -4 & 1 & 1 \\ 0 & -1 & 1 \end{bmatrix} \mathbf{y}(t) = \mathbf{0}, \quad \mathbf{y}(0) = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

and worked out the solution. For illustration purposes, let us compare against the reference solution $y_2(t)$, which is

$$y_2(t) = -\frac{6}{5}e^{-4t} + \left(\frac{6}{5}\cos(t) + \frac{2}{5}\sin(t)\right)e^{-t}.$$

We may use forward Euler to approximate the solution as follows:

$$\begin{bmatrix} y_{1,n+1} \\ y_{2,n+1} \\ y_{3,n+1} \end{bmatrix} = \begin{bmatrix} 1-4\Delta t & 0 & 0 \\ 4\Delta t & 1-\Delta t & -\Delta t \\ 0 & \Delta t & 1-\Delta t \end{bmatrix} \begin{bmatrix} y_{1,n} \\ y_{2,n} \\ y_{3,n} \end{bmatrix}. \tag{3-81}$$

Let us consider the time interval $0 \leq t \leq 5$, with 10, 20, and 40 time intervals having $\Delta t = 0.5, 0.25, 0.125$ respectively. The analytic solution is compared with that obtained from the forward Euler method for various time steps in Fig. 3.2. The analytical solution is plotted as the black smooth curve. The red line is for $\Delta t = 0.5$. This line is very jagged and oscillates wildly. It does not approximate the solution well at all. This is a consequence of the time step being too large and the errors growing. The blue curve gives the result for $\Delta t = 0.25$. This curve does not exhibit the large oscillations, and is somewhat representative of the shape of the analytical solution. The orange curve gives the result for $\Delta t = 0.125$. This curve is closer to the analytical solution. Indeed, if Δt is made increasingly small, the result will approach the actual analytical solution.

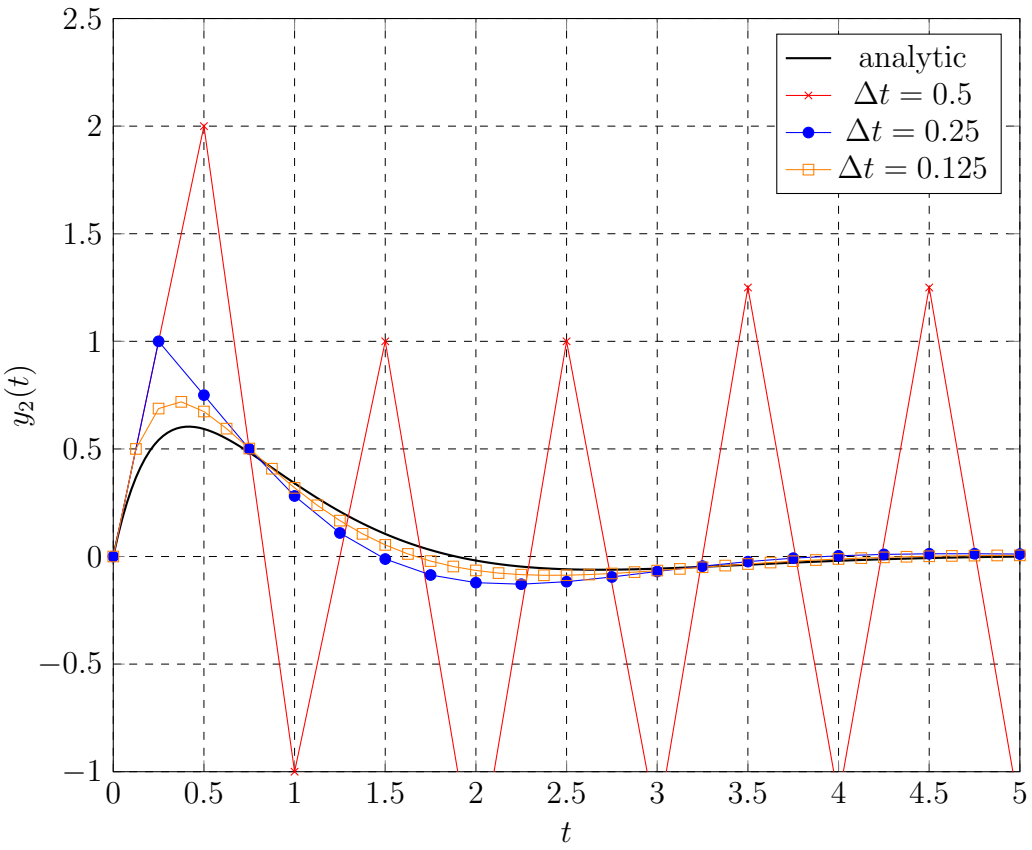

Figure 3.2: Example of the forward Euler method.

3.5.b Backward Euler

An alternative is to take the right-hand side at the right endpoint of the time interval:

$$\frac{1}{\Delta t}\left(\mathbf{y}_{n+1} - \mathbf{y}_n\right) = \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}). \tag{3-82}$$

For the linear ODE this becomes

$$\frac{1}{\Delta t}\left(\mathbf{y}_{n+1} - \mathbf{y}_n\right) = -\mathbf{P}_{n+1}\mathbf{y}_{n+1} + \mathbf{q}_{n+1}. \tag{3-83}$$

Moving the $\mathbf{y}$ terms at time step n to the right-hand side and $n+1$ terms to the left-hand side gives:

$$\mathbf{y}_{n+1} + \Delta t\mathbf{P}_{n+1}\mathbf{y}_{n+1} = \mathbf{y}_n + \Delta t\mathbf{q}_{n+1}. \tag{3-84}$$

Factoring the left-hand side gives the linear system for the unknown value of $\mathbf{y}$ at the following time step:

$$\left(\mathbf{I} + \Delta t\mathbf{P}_{n+1}\right)\mathbf{y}_{n+1} = \mathbf{y}_n + \Delta t\mathbf{q}_{n+1}. \tag{3-85}$$

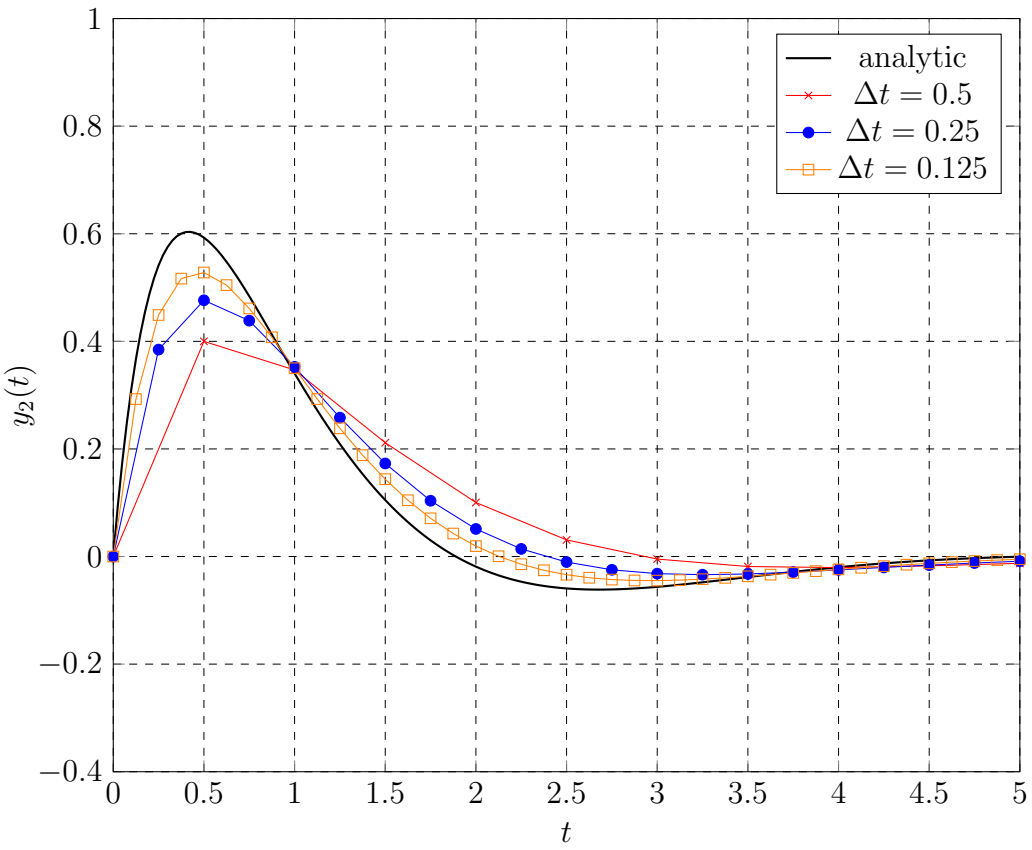

Figure 3.3: Example of the backward Euler method.

The linear system is solved using various techniques of linear algebra, e.g., Gaussian elimination.

The backward Euler scheme has a significant addition in complexity over forward Euler in that there is now a linear system of equations to solve. This is a consequence of the right-hand side being taken at the following time step. We call this an *implicit method* in that we need to simultaneously solve for the right-hand side and the solution vector.

As with forward Euler, the backward Euler scheme has an error that is proportional to the size of the time step Δt. Backward Euler does, however, have one major advantage over forward Euler. The errors that occur during each time step tend to offset each time step and cancel one another out, which implies that the error terms will dampen with time. If time steps are too large, then we will observe false oscillatory behavior each time step indicative of this trend.

The same example that was used for forward Euler is solved again, but this time with backward Euler. The linear system solved is

$$\begin{bmatrix} 1+4\Delta t & 0 & 0 \\ -4\Delta t & 1+\Delta t & \Delta t \\ 0 & -\Delta t & 1+\Delta t \end{bmatrix} \begin{bmatrix} y_{1,n+1} \\ y_{2,n+1} \\ y_{3,n+1} \end{bmatrix} = \begin{bmatrix} y_{1,n} \\ y_{2,n} \\ y_{3,n} \end{bmatrix}. \tag{3-86}$$

The results are given in Fig. 3.3. This time the $\Delta t = 0.5$ solution follows the general shape of the solution and, unlike forward Euler, does not wildly oscillate. This is because of the fact that backward Euler has increased stability. The level of agreement otherwise is about the same for forward and backward Euler, which is to be expected as they have the same rate of error decay.

3.5.c Improved Euler (Trapezoidal Rule)

Previously, we discussed the explicit forward Euler scheme as well as the backward Euler scheme. Both of these methods have an error proportional to the width of the time step Δt. It reasons to ask whether one could devise a scheme that can converge more rapidly with the time step. The answer to this is yes, and the simplest approach is based on the trapezoidal rule called improved Euler. This takes the right-hand side at the midpoint of the time interval using the arithmetic mean:

$$\frac{1}{\Delta t}\left(\mathbf{y}_{n+1} - \mathbf{y}_n\right) = \frac{1}{2}\left[\mathbf{f}(t_n, \mathbf{y}_n) + \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1})\right]. \tag{3-87}$$

As before, we bring all of the terms for $\mathbf{y}_{n+1}$ to the left-hand side and everything else to the right-hand side. After factoring, the expression becomes

$$\left(\mathbf{I} + \frac{\Delta t}{2}\mathbf{P}_{n+1}\right)\mathbf{y}_{n+1} = \left(\mathbf{I} - \frac{\Delta t}{2}\mathbf{P}_n\right)\mathbf{y}_n + \frac{\Delta t}{2}\left(\mathbf{q}_n + \mathbf{q}_{n+1}\right). \tag{3-88}$$

As with backward Euler, we must solve a linear system of equations to obtain the solution, making this another implicit scheme. While more complicated than both forward and backward Euler, the trapezoidal rule has two very attractive properties. First, the error converges as $(\Delta t)^2$, meaning that as the time step gets smaller, the error gets smaller at a much faster rate. Second, like backward Euler, the error terms tend to offset and dampen with time, leading to an algorithm that is stable.

The same example as with forward and backward Euler is repeated. This time the system of equations is

$$\begin{aligned}
&\begin{bmatrix} 1 + 4(\Delta t/2) & 0 & 0 \\ -4(\Delta t/2) & 1 + (\Delta t/2) & (\Delta t/2) \\ 0 & -(\Delta t/2) & 1 + (\Delta t/2) \end{bmatrix} \begin{bmatrix} y_{1,n+1} \\ y_{2,n+1} \\ y_{3,n+1} \end{bmatrix} \\
&= \begin{bmatrix} 1 - 4(\Delta t/2) & 0 & 0 \\ 4(\Delta t/2) & 1 - (\Delta t/2) & -(\Delta t/2) \\ 0 & (\Delta t/2) & 1 - (\Delta t/2) \end{bmatrix} \begin{bmatrix} y_{1,n} \\ y_{2,n} \\ y_{3,n} \end{bmatrix}.
\end{aligned} \tag{3-89}$$

The results are shown in Fig. 3.4. As with backward Euler, none of the results exhibit any serious problems seen for the forward Euler $\Delta t = 0.5$ case. The improved Euler method shows better convergence properties compared to both the forward and backward Euler, with even the $\Delta t = 0.5$ case showing good agreement except for early in the solution. This is consistent with the rate of decay of the error being proportional to $(\Delta t)^2$.

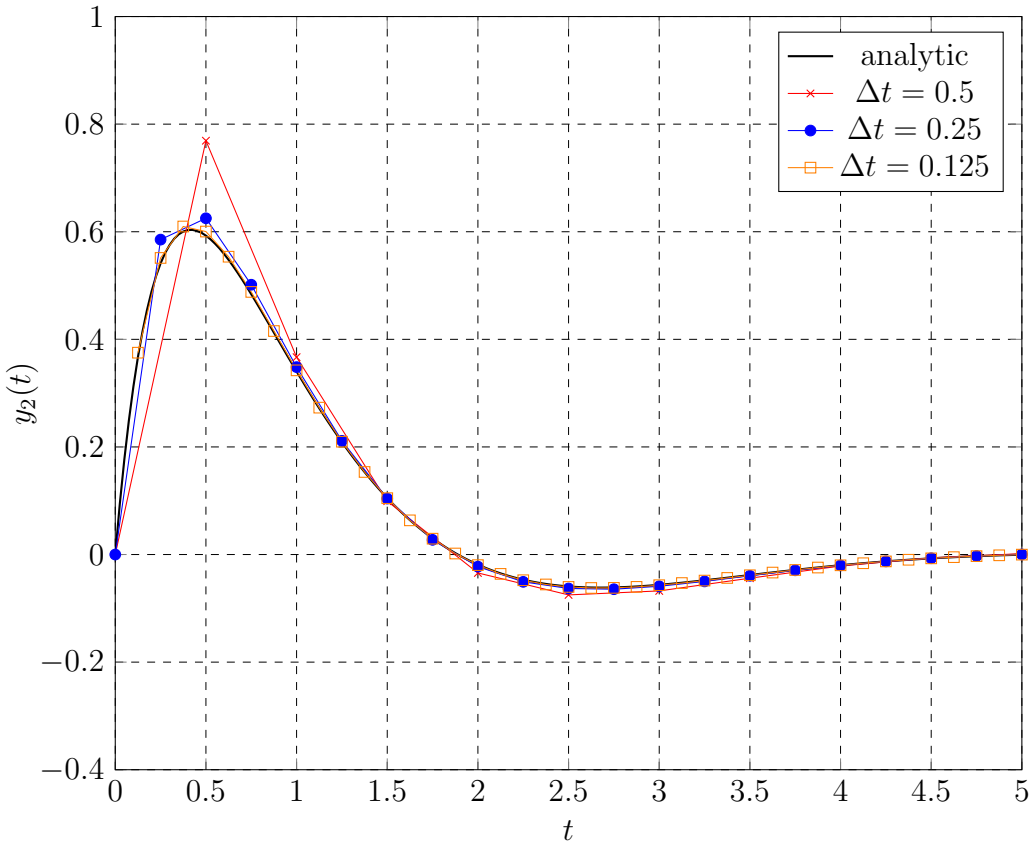

Figure 3.4: Example of the improved Euler method.

It is natural to ask whether one can get a faster convergence rate than $(\Delta t)^2$, and the answer is yes for both explicit and implicit methods. There are increasingly complicated methods with faster convergence rates. These methods are very powerful and used in practice, however, there is an important consideration regarding the stability of the solution method, i.e., whether or not the method will indeed converge. For linear ODEs the backward and improved Euler methods are guaranteed to be stable (nonlinear systems cannot necessarily guarantee this). Unfortunately, any method with a convergence rate of faster than $(\Delta t)^2$ is not unconditionally stable and care must be taken to ensure convergence. These more advanced methods could easily fill part of a numerical methods course and will not be explored here.

3.5.d Error Comparisons

The last few sections made assertions about the rate of convergence of the forward, backward, and improved Euler methods. The forward and backward Euler methods have an error that scales as Δt and the improved Euler method has an error that

scales as $(\Delta t)^2$. If the error goes as

$$\epsilon = c(\Delta t)^k = c\left(\frac{T}{N}\right)^k \to \frac{c}{N^k},$$

where c is some constant describing the magnitude of the error, T is the time interval that the simulation is run that gets pulled into the constant, N is the number of time steps, and k is the convergence rate. It is illustrative to take the logarithm of both sides of the equation

$$\log \epsilon = \log c - k \log(N).$$

The base of the logarithm is not particularly relevant. This expression shows that the logarithm of the error is linear with respect to the logarithm of the number of time steps. The slope of the line is negative, $-k$, meaning the error should be inversely proportional to the number of time steps. The magnitude of this slope gives the rate of convergence.

From our previous example, we have a known analytical reference solution. Therefore, we can do a quantitative comparison of the error for each method. (It is often the case an analytical solution is not available, so the reference solution is usually one generated with a very small time interval.) The error is computed for a single equation using some measure, which is often the root-mean-squared error (which is similar to the Euclidian or L2 norm). Suppose we have a known function $y(t)$ and an approximate version of that function $\widetilde{y}(t)$ defined over some time interval $0 \leq t \leq T$. The root-mean-squared error is

$$\epsilon = \left[\frac{1}{T}\int_0^T (\widetilde{y}(t) - y(t))^2 dt\right]^{1/2}. \tag{3-90}$$

This integral gives some measure that is proportional to the magnitude of the area between the curves $y(t)$ and $\widetilde{y}(t)$. (We could instead integrate over $|\widetilde{y}(t) - y(t)|$ and dispense with the square root; however, in many applications we wish to minimize the error and, since the integrand would not be smooth, we could not take a derivative. Because of this the root-mean-squared error is preferred.) Since we only have the function mapped on some time grid with N steps and fixed interval Δt, this integral can be approximated by a sum

$$\epsilon = \left[\frac{\Delta t}{T}\sum_{n=1}^{N}(\widetilde{y}_n - y_n)^2\right]^{1/2} = \left[\frac{1}{N}\sum_{n=1}^{N}(\widetilde{y}_n - y_n)^2\right]^{1/2}. \tag{3-91}$$

Here $\widetilde{y}_n$ is our approximate solution at time t_n and y_n is the reference solution at the respective time. If we have K functions each with error ϵ_k, we may combine them in a similar manner with the L2 norm:

$$\epsilon = \left[\sum_{k=1}^{K}\epsilon_k^2\right]^{1/2}. \tag{3-92}$$

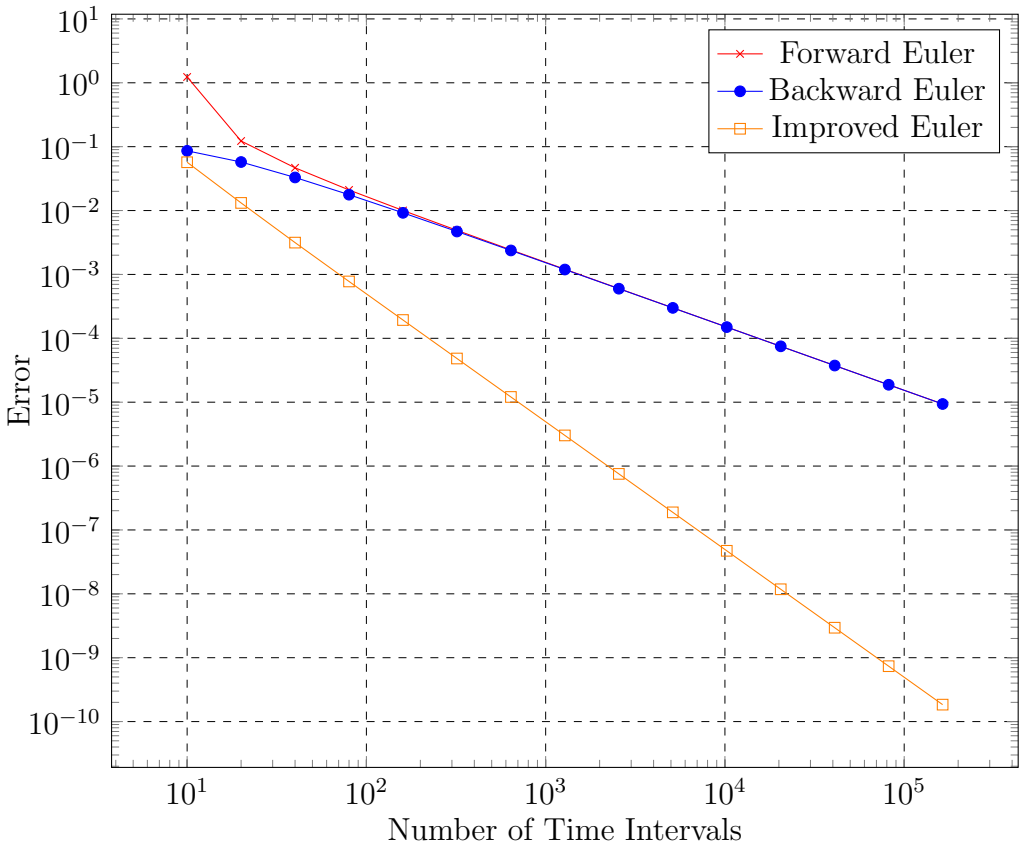

Figure 3.5: Comparison of error for iterative methods.

The error for the three different methods is plotted in Fig. 3.5. The forward and backward Euler methods eventually converge to the same line, meaning that so long as Δt is small enough, they both have the same error. By inspecting the plot the slope is -1, which is the expected result since the error scales as Δt. For large Δt the error of the forward Euler method is greater, which is a consequence of the instability of the explicit method where the error tends to compound. Once the threshold for stability is met, the two methods have the small error.

From the plot it is clear that the improved Euler method always outperforms both the forward and backward Euler methods. The slope of the line is -2, confirming that the error indeed scales of $(\Delta t)^2$. As an added bonus, like backward Euler, it does not exhibit any instability problems that forward Euler has.

These comparisons would not be complete without mentioning the computational time to generate these solutions. Much of this depends upon the specific problem, how well the solvers are implemented (many advanced solvers can take advantage of sparsity or other structures of the matrices), and if the compiler can make any optimizations. For a relatively simplistic implementation of these methods with this problem, for the same number of time steps, the backward Euler takes about twice as long to run compared to forward Euler, and improved Euler takes about three times

longer to run than forward Euler. Results may vary depending on various factors, but the trend is typical.

Using a fixed time step is not a fair comparison, because the end user would run as long as necessary to get the required accuracy. Suppose an accuracy of $\epsilon \approx 10^{-6}$ is desired. For backward and forward Euler this is about 1.5×10^6 time steps on this problem. For improved Euler, this only requires about 2000 time steps. The run time estimate to get the same level of accuracy has the improved Euler method running about *200 times faster* than forward Euler, which is effectively instantaneous on a modern computer.

3.5.e Application to Nonlinear ODEs

The numerical techniques we discussed may be applied to non-linear systems of ordinary differential equations as well where the vector of functions $\mathbf{f}$ in

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}(t)), \quad \mathbf{y}(0) = \mathbf{y}_0,$$

is any function of $\mathbf{y}(t)$.

The forward Euler method is simple to adapt. We can apply the forward difference to the derivative to get

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \Delta t \mathbf{f}(t_n, \mathbf{y}_n). \tag{3-93}$$

Since $\mathbf{y}_n$ is known, $\mathbf{f}(t_n, \mathbf{y}_n)$ may be evaluated explicitly and we can advance the solution in time.

For the implicit schemes, the solution technique is more difficult. For backward Euler, we have

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \Delta t \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}); \tag{3-94}$$

and for improved Euler

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{\Delta t}{2} \left(\mathbf{f}(t_n, \mathbf{y}_n) + \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}) \right). \tag{3-95}$$

The issue is that $\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1})$ depends upon y_{n+1}, which is unknown. If $\mathbf{f}(t, \mathbf{y})$ is linear in $\mathbf{y}$, then we can solve a linear system as before. If it is not, then we need to use an interative numerical root finding algorithm such as the Newton-Raphson method.

Before discussing the Newton-Raphson method, it is worth commenting that the conclusions regarding convergence and stability assume a *linear* system of equations. If this condition is not met, then there is little that can be said in general about the rate of convergence of particular algorithms, or even whether a particular algorithm will converge. Despite this, the general trends of convergence and stability tend to hold reasonably well and methods such as improved Euler tend to perform well in practice.

3.5.f Newton-Raphson Iteration

Newton-Raphson is a classic iterative technique to find a root of a function

$$g(x) = 0.$$

The idea is given a guess for the root, call it $x^{(k)}$, we use the first derivative to write an equation for a line about that point

$$y = g(x^{(k)}) + g'(x^{(k)})(x - x^{(k)}). \tag{3-96}$$

If we set $y = 0$, we can use the equation for a line to solve for the unknown value x that we set to the value for the next iteration or $x^{(k+1)}$:

$$x^{(k+1)} = x^{(k)} - \frac{g(x^{(k)})}{g'(x^{(k)})}. \tag{3-97}$$

This process may be repeated again and again, and, if the method converges, we will eventually find a good approximation for a root of $g(x)$. A key point here is *if the method converges*. Being a nonlinear iteration, the Newton-Raphson is not guaranteed to converge. However, when it does, it tends to converge in a way that the error decreases geometrically fast. (If Newton's method fails, in practice, many applications switch over to a slower, but more robust, bisection method.)

The Newton-Raphson method may be generalized to find the roots of a vector of functions

$$\mathbf{g}(\mathbf{x}) = \mathbf{0}$$

or in vector form

$$\begin{bmatrix} g_1(x_1, x_2, \cdots, x_N) \\ g_2(x_1, x_2, \cdots, x_N) \\ \vdots \\ g_N(x_1, x_2, \cdots, x_N) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

In a similar fashion, we write a series of equations for lines and find where those lines converge to zero. The resulting iteration scheme has the following form:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \mathbf{J}^{-1}(\mathbf{x}^{(k)})\mathbf{g}(\mathbf{x}^{(k)}). \tag{3-98}$$

The matrix $\mathbf{J}(\mathbf{x})$ is called the *Jacobian matrix* and is the multidimensional version of the first derivative:

$$\mathbf{J}(\mathbf{x}) = \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_N} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} & \cdots & \frac{\partial g_2}{\partial x_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_N}{\partial x_1} & \frac{\partial g_N}{\partial x_2} & \cdots & \frac{\partial g_N}{\partial x_N} \end{bmatrix}. \tag{3-99}$$

To avoid computing the inverse of the Jacobian matrix we often rewrite the Newton-Raphson iteration as a linear system

$$\mathbf{J}(\mathbf{x}^{(k)})\mathbf{x}^{(k+1)} = \mathbf{J}(\mathbf{x}^{(k)})\mathbf{x}^{(k)} - \mathbf{g}(\mathbf{x}^{(k)}). \tag{3-100}$$

Since $\mathbf{x}^{(k)}$ is known, the right-hand side is completely known, and the solution vector $\mathbf{x}^{(k+1)}$ can be obtained by solving the linear system using techniques such as Gaussian elimination.

The iteration begins with some initial guess $\mathbf{x}^{(0)}$ and continues until some user-defined convergence criterion is met, e.g.,

$$|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}| < \epsilon. \tag{3-101}$$

Applied to backward Euler, the Newton-Raphson iteration is

$$\mathbf{g}(\mathbf{y}_{n+1}) = \mathbf{y}_{n+1} - \mathbf{y}_n - \Delta t\mathbf{f}(t_{n+1}, \mathbf{y}_{n+1}) = \mathbf{0}; \tag{3-102}$$

and for improved Euler,

$$\mathbf{g}(\mathbf{y}_{n+1}) = \mathbf{y}_{n+1} - \mathbf{y}_n - \frac{\Delta t}{2}\left(\mathbf{f}(t_n, \mathbf{y}_n) + \mathbf{f}(t_{n+1}, \mathbf{y}_{n+1})\right) = \mathbf{0}. \tag{3-103}$$

A Newton-Raphson iteration must be performed each time step, similar to how a linear solve was done in the linear case. Note that each time step may require several Newton-Raphson iterations, for which each requires the solution of a linear system of equations. Usually the initial guess for the Newton-Raphson iteration is $\mathbf{y}_n$, the vector at the previous time step. A potential concern is the nonconvergence of the Newton-Raphson iteration. Fortunately, for sufficiently small time steps, Newton-Raphson does tend to converge to a solution and do as such in few iterations.

3.5.g Example: Fission Reactor Kinetics

An application of a nonlinear system of equations is analysis of nuclear reactor kinetics. The criticality of a nuclear reactor drives the time behavior of the neutron population in the reactor. If the reactor is critical ($k = 1$), the power stays constant; if the reactor is subcritical ($k < 1$) the reactor power tends to fall exponentially; and if the reactor is supercritical ($k > 1$), the reactor power tends to rise exponentially. If the power reaches a certain level, there is sufficient energy release from nuclear fission to significantly increase the temperature of the reactor. As the temperature changes, so do the nuclear interaction properties, e.g., changes in density moving the nuclei further apart or enhanced thermal motion leading to Doppler broadening of the nuclear resonances in the cross sections, and the criticality of the system changes, which, in turn, leads to a change in the rate of power increase, affecting the temperature, and so on.

The primary method of energy production is from nuclear fission, which is directly proportional to the number of neutrons in the reactor. However, there is one added complication in that a small fraction (a few fractions of a percent) of the fission

products (called delayed neutron precursors) undergo β^- decay and subsequently emit a neutron (called a delayed neutron) through its de-excitation process, which must be considered in the neutron balance. The time scale for a fission neutron producing another fission neutron (microseconds in a light-water reactor) is a few orders of magnitude shorter than the time scale for the radioactive decay emission of delayed neutrons (milliseconds to minutes). This means they need to be treated as separate differential equations.

The equation for the fission rate density is

$$\frac{dp}{dt} = \left(\frac{\rho(t,T) - \beta}{\Lambda}\right) p(t) + \frac{\lambda}{\Lambda}\zeta(t) + q(t). \tag{3-104}$$

Here $p(t)$ is the fission rate density; $\rho(t,T)$ is the reactivity, which is related to the criticality of the system through

$$\rho(t,T) = \frac{k(t,T) - 1}{k(t,T)}, \tag{3-105}$$

and is also a function of temperature; β is the effective fraction of neutrons that emerge from decay of fission products, Λ is the effective lifetime of the neutron, λ is some averaged decay constant for the delayed neutron precursors; $\zeta(t)$ is related to the density of delayed neutron precursors; and $q(t)$ is a source density of neutrons. The delayed neutron precursors have the balance equation

$$\frac{d\zeta}{dt} = \beta p(t) - \lambda\zeta(t). \tag{3-106}$$

For the reactivity we assume that it is constant with the addition of term that is linearly dependent upon temperature T:

$$\rho(t,T) = \rho_0 + \gamma(T(t) - T_0), \tag{3-107}$$

here γ is a temperature feedback coefficient. The temperature is proportional to the fission rate and can be described with the equation

$$Dc_p\frac{dT}{dt} = \kappa p(t) - h(T(t) - T_0), \tag{3-108}$$

here the temperature T is defined to be T_0 at the ambient temperature, D is the mass density (often ρ but this was used for reactivity), c_p is the specific heat of the material, κ is some fission to energy conversion factor, and h is a heat transfer or cooling coefficient. We have explicitly assumed D, c_p, and h are not functions of temperature, which is a simplification we can justify if the temperature change is small. We are also neglecting the heat generation from fission products, which assumes the fission product inventory is small (not a good assumption in a power reactor). Also, for short time transients on the order of milliseconds, we often neglect losses from cooling, which has a timescale of minutes.

Once the reactivity is inserted into the rate equation for fission, we end up with a nonlinear ODE:

$$\frac{dp}{dt} = \alpha p(t) + \frac{\gamma}{\Lambda} T(t)p(t) + \frac{\lambda}{\Lambda}\zeta(t) + q(t). \tag{3-109}$$

where

$$\alpha = \frac{\rho_0 - \beta}{\Lambda}.$$

The system is nonlinear because it contains a product term, namely $T(t)p(t)$.

To summarize, the equations to be solved are:

$$\frac{dp}{dt} = \left(\alpha + \frac{\gamma}{\Lambda}T(t)\right) p(t) + \frac{\lambda}{\Lambda}\zeta(t) + q(t), \tag{3-110a}$$

$$\frac{d\zeta}{dt} = \beta p(t) - \lambda\zeta(t), \tag{3-110b}$$

$$\frac{dT}{dt} = \frac{\kappa}{Dc_p} p(t) - \frac{h}{Dc_p} T(t). \tag{3-110c}$$

Applying forward Euler to this system yields the system of algebraic equations

$$p_{n+1} = \left(1 + \alpha\Delta t + \frac{\gamma\Delta t}{\Lambda} T_n\right) p_n + \frac{\lambda\Delta t}{\Lambda}\zeta_n + \Delta t q_n, \tag{3-111a}$$

$$\zeta_{n+1} = (1 - \lambda\Delta t)\zeta_n + \beta p_n, \tag{3-111b}$$

$$T_{n+1} = \left(1 - \frac{h\Delta t}{Dc_p}\right) T_n + \frac{\kappa\Delta t}{Dc_p} p_n. \tag{3-111c}$$

Since everything on the right-hand side is known, we can solve for the fission rate, precursor population, and temperature at the next time step directly.

Applying backward Euler yields the system

$$g_1 = \left(1 - \alpha\Delta t - \frac{\gamma\Delta t}{\Lambda} T_{n+1}\right) p_{n+1} - \frac{\lambda\Delta t}{\Lambda}\zeta_{n+1} - \Delta t q_{n+1} - p_n = 0, \tag{3-112a}$$

$$g_2 = (1 + \lambda\Delta t)\zeta_{n+1} - \beta\Delta_t p_{n+1} - \zeta_n = 0, \tag{3-112b}$$

$$g_3 = \left(1 + \frac{h\Delta t}{Dc_p}\right) T_{n+1} - \frac{\kappa\Delta t}{Dc_p} p_{n+1} - T_n = 0. \tag{3-112c}$$

This system of equations is in the form necessary for the Newton-Raphson iteration. Now, we must compute the terms of the Jacobian matrix:

$$\frac{\partial g_1}{\partial p_{n+1}} = 1 - \alpha\Delta t - \frac{\gamma\Delta t}{\Lambda} T_{n+1}, \tag{3-113a}$$

$$\frac{\partial g_1}{\partial \zeta_{n+1}} = -\frac{\lambda\Delta t}{\Lambda}, \tag{3-113b}$$

$$\frac{\partial g_1}{\partial T_{n+1}} = -\frac{\gamma\Delta t}{\Lambda} p_{n+1}, \tag{3-113c}$$

$$\frac{\partial g_2}{\partial p_{n+1}} = -\beta\Delta t, \tag{3-113d}$$

$$\frac{\partial g_2}{\partial \zeta_{n+1}} = 1 + \lambda\Delta t, \tag{3-113e}$$

$$\frac{\partial g_2}{\partial T_{n+1}} = 0, \tag{3-113f}$$

$$\frac{\partial g_3}{\partial p_{n+1}} = -\frac{\kappa\Delta t}{Dc_p}, \tag{3-113g}$$

$$\frac{\partial g_3}{\partial \zeta_{n+1}} = 0, \tag{3-113h}$$

$$\frac{\partial g_3}{\partial T_{n+1}} = 1 + \frac{h\Delta t}{Dc_p}. \tag{3-113i}$$

This gives the Jacobian matrix

$$\mathbf{J}(\mathbf{x}^{(k)}) = \begin{bmatrix} 1 - \alpha\Delta t - \dfrac{\gamma\Delta t}{\Lambda}T_{n+1}^{(k)} & -\dfrac{\lambda\Delta t}{\Lambda} & -\dfrac{\gamma\Delta t}{\Lambda}p_{n+1}^{(k)} \\ -\beta\Delta t & 1 + \lambda\Delta t & 0 \\ -\dfrac{\kappa\Delta t}{Dc_p} & 0 & 1 + \dfrac{h\Delta t}{Dc_p} \end{bmatrix} \tag{3-114}$$

where

$$\mathbf{x}^{(k)} = \begin{bmatrix} p_{n+1}^{(k)} \\ \zeta_{n+1}^{(k)} \\ T_{n+1}^{(k)} \end{bmatrix}. \tag{3-115}$$

The Newton-Raphson iteration is now done by solving the linear system

$$\mathbf{J}(\mathbf{x}^{(k)})\mathbf{x}^{(k+1)} = \mathbf{J}(\mathbf{x}^{(k)})\mathbf{x}^{(k)} - \mathbf{g}(\mathbf{x}^{(k)}).$$

A logical initial guess is

$$\mathbf{x}^{(0)} = \begin{bmatrix} p_n \\ \zeta_n \\ T_n \end{bmatrix}, \tag{3-116}$$

which are the quantities at the previous time step. The Newton-Raphson iteration proceeds until some convergence criterion is met. In practice, sometimes a maximum number of iterations needs to be set, because the method can end up bouncing back and forth changing only a small amount each time to avoid the code locking. Generally speaking, if convergence fails, this means that a smaller time step should be taken. Once the iteration completes, the quantities p, ζ, T for the next step are known and the time stepping algorithm continues by doing another Newton-Raphson iteration until the end of the simulation.

The equations for improved Euler are more complicated. These are

$$g_1 = \left(1 - \frac{\alpha\Delta t}{2} - \frac{\gamma\Delta t}{2\Lambda}T_{n+1}\right)p_{n+1} - \frac{\lambda\Delta t}{2\Lambda}\zeta_{n+1} - \left(1 + \frac{\alpha\Delta t}{2} + \frac{\gamma\Delta t}{2\Lambda}T_n\right)p_n - \frac{\lambda\Delta t}{2\Lambda}\zeta_n - \frac{\Delta t}{2}(q_n + q_{n+1}) = 0, \tag{3-117a}$$

$$g_2 = \left(1 + \frac{\lambda\Delta t}{2}\right)\zeta_{n+1} - \frac{\beta\Delta t}{2}p_{n+1} - \left(1 - \frac{\lambda\Delta t}{2}\right)\zeta_n - \frac{\beta\Delta t}{2}p_n = 0, \tag{3-117b}$$

$$g_3 = \left(1 + \frac{h\Delta t}{2Dc_p}\right)T_{n+1} - \frac{\kappa\Delta t}{2Dc_p}p_{n+1} - \left(1 - \frac{h\Delta t}{2Dc_p}\right)T_n - \frac{\kappa\Delta t}{2Dc_p}p_n = 0. \tag{3-117c}$$

The Jacobian matrix is mostly unchanged except for an extra factor of $^1/_2$ on the time steps:

$$\mathbf{J}(\mathbf{x}^{(k)}) = \begin{bmatrix} 1 - \frac{\alpha\Delta t}{2} - \frac{\gamma\Delta t}{2\Lambda}T_{n+1}^{(k)} & -\frac{\lambda\Delta t}{2\Lambda} & -\frac{\gamma\Delta t}{2\Lambda}p_{n+1}^{(k)} \\ -\frac{\beta\Delta t}{2} & 1 + \frac{\lambda\Delta t}{2} & 0 \\ -\frac{\kappa\Delta t}{2Dc_p} & 0 & 1 + \frac{h\Delta t}{2Dc_p} \end{bmatrix}. \tag{3-118}$$

Otherwise the iteration sequence is unchanged from backward Euler.

The improved Euler method is applied to solve this problem numerically. The following constant values are used:

$$\begin{aligned} \Lambda &= 2.5 \times 10^{-6}\ \text{s}, \\ \beta &= 0.0076, \\ \rho_0 &= 5\beta = 0.038, \\ \gamma &= -1 \times 10^{-4}\ \text{K}^{-1}, \\ \lambda &= 0.25\ \text{s}^{-1}, \\ c_p &= 4.18\ \text{J g}^{-1}\ \text{K}^{-1}, \\ D &= 1.0\ \text{g cm}^{-3}, \\ \kappa &= 3.2 \times 10^{-11}\ \text{J/fission}, \\ h &= 4 \times 10^{-3}\ \text{J cm}^{-3}\ \text{K}^{-1}\ \text{s}^{-1}\ , \\ p(0) &= 1000\ \text{fissions cm}^{-3}\ \text{s}^{-1}, \\ \zeta(0) &= 0\ \text{precursors cm}^{-3}, \\ T(0) &= T_0\ \text{K (note } T = T_0 \text{ corresponds to ambient temperature)}, \\ q(t) &= 0.01\ \text{neutrons cm}^{-3}\ \text{s}^{-1}\ \text{(constant)}. \end{aligned}$$

At time $t = 0$ a instantaneous insertion of reactivity $\rho_0 = 5\$$ is made. The fission rate density is plotted on log-log scale in Fig. 3.6 and the delayed neutron precursor density is plotted in log-log scale in Fig. 3.7. The temperature and effective multiplication factor are plotted in Figs. 3.8 and 3.9 respectively.

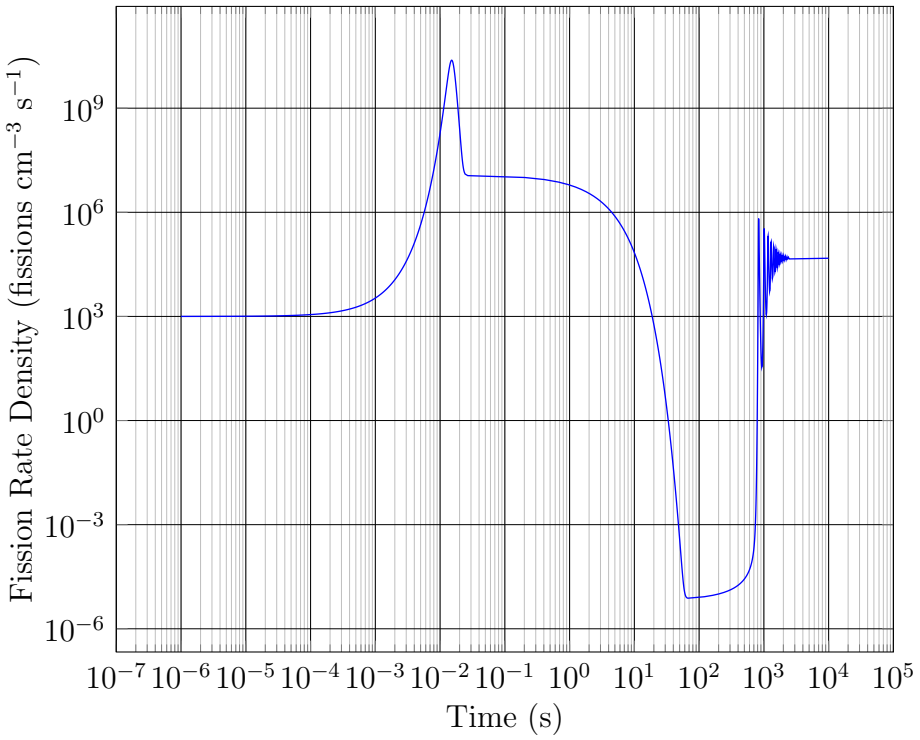

Figure 3.6: Fission rate density for fission reactor kinetics example.

Since the system is immediately prompt supercritical ($\rho > \beta; \alpha > 0$) we can expect the fission density to rise exponentially at a rapid pace on the time scale of milliseconds, since the neutron lifetime is 25 microseconds. Eventually, we observe that the fission rate density rapidly experiences a decrease, which is a result of the temperature feedback with negative κ. As the temperature increases for the fission energy release, the reactivity deceases until the system becomes subcritical and the fission rate rapidly drops.

However, during the prompt transient there was a significant buildup of fission product precursors. These do not simply disappear when the reactor goes subcritical. Rather, there is an exponentially decaying source of neutrons that is driving the subcritical reactor and producing more fission. Since the half-life of these fission products is on the order of a second, the time scale for this source term is on the order of a few tens of seconds. During this time, the reactor continues to heat up as the resulting fissions release more energy.

Finally after most of the precursors have decayed away, the reactor begins to cool steadily, which causes the reactivity to increase again. The cooling process is on the order of minutes. After several minutes, the reactor becomes critical again and another, but less severe, transient will occur. This transient is in the delayed supercritical regime, which means the power rise is on the order of tens of seconds. Eventually the temperature increases, driving the reactor subcritical, leading to the power to fall rapidly with the decay of fission products. The process repeats again

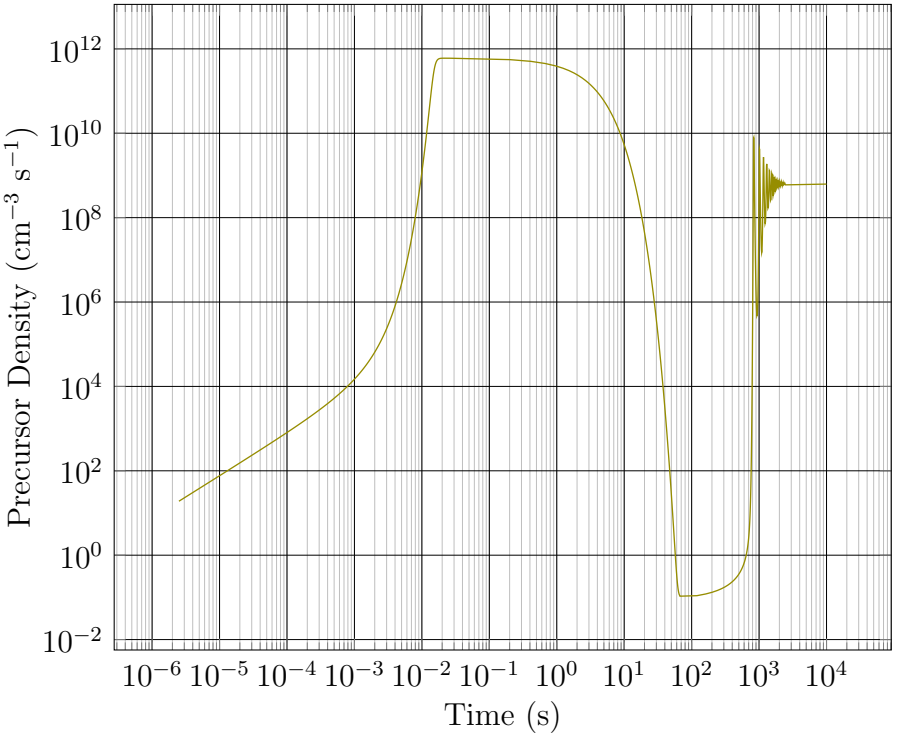

Figure 3.7: Precursor density for fission reactor kinetics example.

and again in a damped oscillation until an equilibrium power is reached.

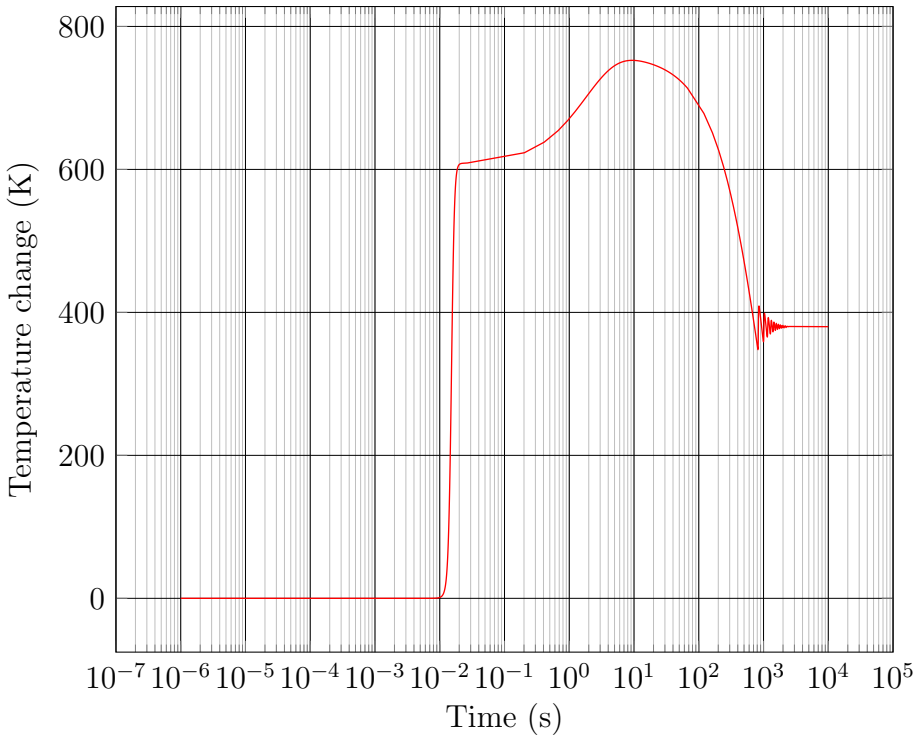

Figure 3.8: Temperature change for fission reactor kinetics example.

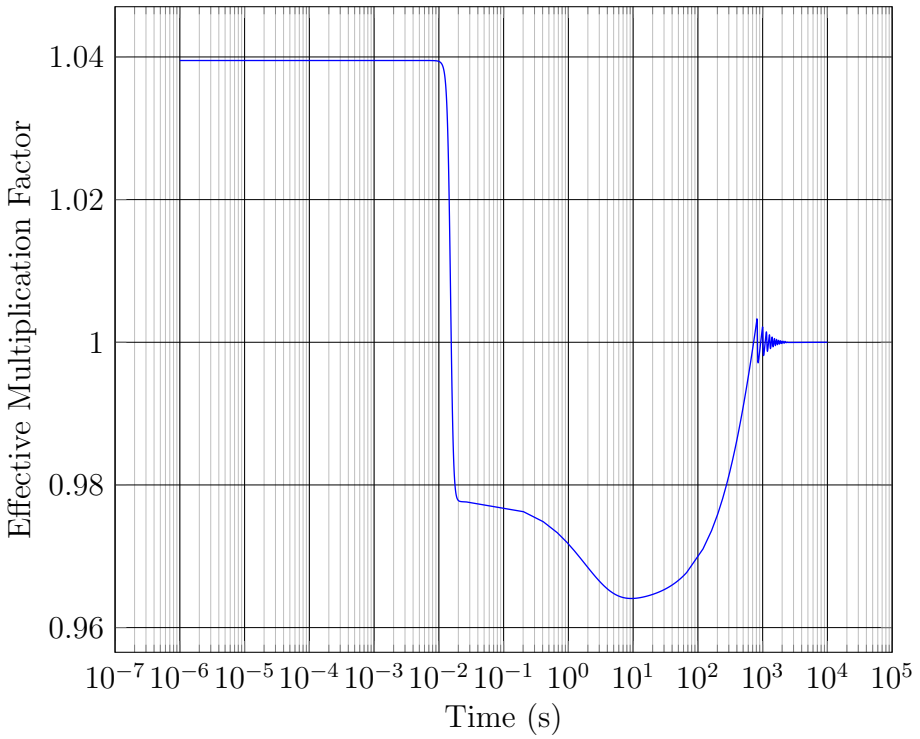

Figure 3.9: Effective multiplication factor for fission reactor kinetics example.

3.6 Second-Order Linear ODEs

Until now, we discussed systems of first-order linear ordinary differential equations. The presence of only the first derivative led to a set of solution techniques involving the integrating factor. Once second-derivatives are added, such techniques no longer apply.

Additionally, first-order ODEs often describe *initial-value problems* that takes a known value at some point and propagates it forward in whatever variable is involved, usually time. Second-order ODEs can also describe initial-value problems and are often used in the equations of motion for dynamical systems relating position and velocity where both are known at an initial time. It turns out many of the same techniques we used in first-order ODEs are directly applicable for this class of problems.

Another class of problems that are also associated with second-order ODEs are *boundary value problems* that describe the solution of a field (such as temperature) in the interior of an object where some constraints, called boundary conditions, are made on the exterior of the region. This class of problems has fundamentally different properties, and will be the primary focus of the remainder of this chapter.

The second-order linear ODE takes the form:

$$\frac{d^2y}{dx^2} + p(x)\frac{dy}{dx} + q(x)y(x) = r(x). \tag{3-119}$$

When $p(x)$ and $q(x)$ are not constant, it is difficult, if not impossible, to obtain an analytical solution and the equations then need to be solved numerically. Fortunately, many problems of practical interest have constant coefficients, which are amenable to analytical techniques. The simplified version with constant coefficients is written as

$$\frac{d^2y}{dx^2} + a\frac{dy}{dx} + by(x) = r(x). \tag{3-120}$$

To solve linear second-order ODEs with constant coefficients, we often split the problem into two major steps. First, we solve the homogeneous version by setting the right-hand side equation to zero:

$$\frac{d^2y_h}{dx^2} + a\frac{dy_h}{dx} + by_h(x) = 0. \tag{3-121}$$

Once we have obtained a solution, we then find another solution called the particular solution

$$\frac{d^2y_p}{dx^2} + a\frac{dy_p}{dx} + by_p(x) = r(x). \tag{3-122}$$

These two solutions can be combined to obtain the generation solution to the problem:

$$y(x) = y_h(x) + y_p(x). \tag{3-123}$$

These are discussed in the remainder of this section.

3.6.a Solution of the Homogeneous Problem

To solve for $y_h(x)$ we make a guess that the solution follows an exponential form:

$$y_h(x) = Ce^{\lambda x} \tag{3-124}$$

with C as some scaling constant (which we find using boundary conditions) and λ some unknown parameter that we hope to find. Inserting this guess into the homogeneous second-order ODE gives:

$$C\lambda^2 e^{\lambda x} + Ca\lambda e^{\lambda x} + Cbe^{\lambda x} = 0. \tag{3-125}$$

The factor

$$Ce^{\lambda x}$$

is common on all terms and, because the right-hand side is zero, they are simply scaling constants and can be divided out. This leaves the quadratic polynomial

$$\lambda^2 + a\lambda + b = 0, \tag{3-126}$$

which has the solution via the quadratic formula of

$$\lambda = \frac{1}{2}\left(-a \pm \sqrt{a^2 - 4b}\right). \tag{3-127}$$

Assuming that the two roots are distinct, call them λ_1 and λ_2, we can write the solution to the homogeneous problem as a linear combination of those solutions:

$$y_h(x) = C_1 e^{\lambda_1 x} + C_2 e^{\lambda_2 x}. \tag{3-128}$$

When $\lambda_1 = \lambda_2$, we have the case where we need to find another solution. It turns out that when this is the case,

$$Cxe^{\lambda x}$$

also satisfies the differential equation. Therefore, when $\lambda_1 = \lambda_2 = \lambda$ we can write the solution as

$$y_h(x) = (C_1 + C_2 x)e^{\lambda x}. \tag{3-129}$$

Note that there are two coefficients C_1 and C_2, which are the consequence of "integrating" twice. These will be solved using the initial or boundary conditions, depending on the class of problem.

When λ is real and of the form $\lambda = \pm\mu$, we can express the homogeneous solution in terms of exponentials or hyperbolic trigonometric functions:

$$\begin{aligned} y_h(x) &= C_1 e^{\mu x} + C_2 e^{-\mu x}, \\ y_h(x) &= A_1 \sinh(\mu x) + A_2 \cosh(\mu x). \end{aligned} \tag{3-130}$$

Here the constant coefficients are different. The hyperbolic trigonometric functions are related to the exponential by

$$\sinh(x) = \frac{e^x - e^{-x}}{2}, \tag{3-131a}$$

$$\cosh(x) = \frac{e^x + e^{-x}}{2}. \tag{3-131b}$$

The choice of exponentials or hyperbolic trigonometric functions is arbitrary and motivated by whatever makes the math easier. For example, in some boundary value problems, the hyperbolic trigonometric functions can be crafted in a way to eliminate one of the constant coefficients and simplify the form of the solution.

When λ is imaginary or complex, we can either work in complex exponentials (often more convenient) or apply Euler's formula to express the solution as trigonometric functions. For purely imaginary $\lambda = \pm i\omega$ functions the homogeneous solution takes either form:

$$\begin{aligned} y_h(x) &= C_1 e^{i\omega x} + C_2 e^{-i\omega x}, \\ y_h(x) &= A_1 \sin(\omega x) + A_2 \cos(\omega x). \end{aligned} \tag{3-132}$$

For the more general case of a complex $\lambda = \mu \pm i\omega$, the homogeneous solution may be written as

$$y_h(x) = A_1 e^{\mu x} \sin(\omega x) + A_2 e^{\mu x} \cos(\omega x).$$

When working with trigonometric or hyperbolic trigonometric functions, it can also be beneficial to apply a *phase shift*. For example, the homogeneous solutions for the real problem with $\lambda = \pm\mu$ could also be written as

$$y_h(x) = B_1 \sinh(\mu x + \varphi) + B_2 \cosh(\mu x + \varphi),$$

and for $\lambda = \pm i\omega$, the homogeneous solution can be written as

$$y_h(x) = B_1 \sin(\omega x + \varphi) + B_2 \cos(\omega x + \varphi).$$

3.6.b Linear Independence and the Wronskian

In the previous section, we provided various examples of homogeneous solutions that are possible; however, it would be good to offer a more general set of rules. When we are solving a homogeneous second-order ordinary differential equations, there are two derivatives and we must find two linearly independent functions that satisfy the homogeneous differential equation. Any two will work, so long as they satisfy this criterion.

Assessing linear independence can be done using an object called the *Wronskian*. Suppose we have two solutions $y_1(x)$ and $y_2(x)$ that satisfy the homogeneous differential equation, the Wronskian is given by the following determinant:

$$W(x) = \begin{vmatrix} y_1(x) & y_2(x) \\ y_1'(x) & y_2'(x) \end{vmatrix} = y_1(x)y_2'(x) - y_2(x)y_1'(x). \tag{3-133}$$

The test for linear independence is

$$W(x) \neq 0 \text{ for some values of } x \text{ in the domain.} \tag{3-134}$$

Conversely, if $y_1(x)$ and $y_2(x)$ satisfy the homogeneous differential equation and $W(x) = 0$ for all x in the domain, then the solutions are linearly dependent and we require a different linearly independent solution.

As an example, consider the homogeneous differential equation:

$$y''(x) + 4y(x) = 0. \tag{3-135a}$$

Suppose we wish to determine if

$$y_1(x) = \sin(2x) \tag{3-135b}$$
$$y_2(x) = \sin(1 - 2x) \tag{3-135c}$$

are linearly independent solutions. First, we plug in the solutions into the differential equation:

$$-4\sin(2x) + 4\sin(2x) = 0, \tag{3-135d}$$
$$-4\sin(1 - 2x) + 4\sin(1 - 2x) = 0. \tag{3-135e}$$

These solutions satisfy the differential equation. The Wronskian is

$$\begin{aligned} W(x) &= \begin{vmatrix} \sin(2x) & \sin(1-2x) \\ 2\cos(2x) & -2\cos(1-2x) \end{vmatrix} \\ &= -2\sin(2x)\cos(1-2x) - 2\cos(2x)\sin(1-2x) = -2\sin(1). \end{aligned} \tag{3-135f}$$

Since $-2\sin(1) \neq 0$, then we know both $y_1(x)$ and $y_2(x)$ are linearly independent solutions of the homogeneous differential equation. Therefore, we can write the solution as

$$y(x) = C_1\sin(2x) + C_2\sin(1 - 2x). \tag{3-135g}$$

Now that the homogeneous solution is known, we must find the particular solution. This is done with one of two approaches: method of undetermined coefficients and variation of parameters.

3.6.c Method of Undetermined Coefficients

An approach that is relatively simple, but only works in a narrow set of situations for specific forms of the inhomogeneous term $r(x)$ is called the method of undetermined coefficients. The basic approach is to "guess" a functional form for $y_p(x)$ and, should that form be reasonable, we can solve for the coefficients. We now outline a few common cases:

For the case where the inhomogeneous term is a polynomial,

$$r(x) = r_0 + r_1x + r_2x^2 + \ldots \tag{3-136}$$

we can guess that the particular solution is a polynomial

$$f_h(x) = a_0 + a_1 x + a_2 x^2 + \dots \tag{3-137}$$

plug it into the equation and match the coefficients.

To illustrate, consider the ODE:

$$y''(x) + 4y'(x) + 3y(x) = 1 + 3x^2. \tag{3-138a}$$

We can guess the particular solution as a quadratic polynomial

$$y_p(x) = a_0 + a_1 x + a_2 x^2. \tag{3-138b}$$

Inserting this into the differential equation gives

$$2a_2 + 4(2a_2 x + a_1) + 3(a_0 + a_1 x + a_2 x^2) = 1 + 0x^2 + 3x^2. \tag{3-138c}$$

Rearranging to combine terms gives

$$(3a_0 + 4a_1 + 2a_2) + (3a_1 + 8a_2)x + 3a_2 x^2 = 1 + 0x^2 + 3x^2. \tag{3-138d}$$

By matching coefficients we can determine

$$\begin{aligned} 3a_0 + 4a_1 + 2a_2 &= 1, \\ 3a_1 + 8a_2 &= 0, \\ 3a_2 &= 3; \end{aligned} \tag{3-138e}$$

or as a linear system

$$\begin{bmatrix} 3 & 4 & 2 \\ 0 & 3 & 8 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 3 \end{bmatrix}. \tag{3-138f}$$

Solving this linear system gives

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} {}^{29}/_{9} \\ -{}^{8}/_{3} \\ 1 \end{bmatrix} \tag{3-138g}$$

Therefore, the particular solution is

$$y_p(x) = \frac{29}{9}x^2 - \frac{8}{3}x + 1. \tag{3-138h}$$

Another form is where the inhomogeneous term has the function of an exponential:

$$r(x) = k_0 e^{\beta x}. \tag{3-139}$$

For this, we guess a particular solution of the form

$$y_p(x) = a_0 e^{-\beta x}. \tag{3-140}$$

More generally, if we have an inhomogeneous solution of the form

$$r(x) = (k_0 + k_1 x + k_2 x^2 + \ldots)e^{-\beta x}. \tag{3-141}$$

we guess a solution of the form

$$y_p(x) = (a_0 + a_1 x + a_2 x^2)e^{-\beta x}. \tag{3-142}$$

Another form for the inhomogeneous term where method of coefficients will work is where the inhomogeneous term is a trigonometric function:

$$\begin{aligned} r(x) &= k_0 \sin(\alpha x), \\ r(x) &= k_0 \cos(\alpha x). \end{aligned} \tag{3-143}$$

In either of these cases, we guess a particular solution of the form

$$y_p(x) = a_0 \sin(\alpha x) + b_0 \cos(\alpha x). \tag{3-144}$$

As with the exponential case, these can be generalized to account for polynomial coefficients

$$\begin{aligned} r(x) &= (k_0 + k_1 x + k_2 x^2 + \ldots)\sin(\alpha x), \\ r(x) &= (k_0 + k_1 x + k_2 x^2 + \ldots)\cos(\alpha x); \end{aligned} \tag{3-145}$$

we guess a solution of the form

$$y_p(x) = (a_0 + a_1 x + a_2 x^2 + \ldots)\sin(\alpha x) + (b_0 + b_1 x + b_2 x^2 + \ldots)\cos(\alpha x). \tag{3-146}$$

If we encounter a linear combination of polynomials, exponentials, and/or trigonometric functions, we can guess that a solution is the linear combination of the guesses with the combined coefficients. As with the polynomial case, the goal is to take the guess, plug it into the differential equation, collect like terms, and solve a linear system for the coefficients.

To illustrate, consider the example

$$y''(x) + 4y'(x) + 3y(x) = 1 + 2x + 3x\cos x. \tag{3-147a}$$

For this we guess a function of five coefficients

$$y_p(x) = a_0 + a_1 x + b_0 \sin x + b_1 x \sin x + c_0 \cos x + c_1 x \cos x. \tag{3-147b}$$

The derivatives are

$$\begin{aligned} y_p'(x) &= a_1 + b_0 \cos x + b_1(\sin x + x\cos x) - c_0 \sin x + c_1(\cos x - x \sin x) \\ &= a_1 + (b_1 - c_0)\sin x - c_1 x \sin x + (b_0 + c_1)\cos x + b_1 x \cos x; \end{aligned} \tag{3-147c}$$

$$\begin{aligned} y_p''(x) &= (b_1 - c_0)\cos x - c_1(\sin x + x\cos x) - (b_0 + c_1)\sin x + b_1(\cos x - x\sin x) \\ &= -(2c_1 + b_0)\sin x - b_1 x \sin x + (2b_1 - c_0)\cos x - c_1 x \cos x. \end{aligned} \tag{3-147d}$$

Inserting this into the differential equation gives

$$\begin{aligned}
&-(2c_1+b_0)\sin x - b_1 x\sin x + (2b_1-c_0)\cos x - c_1 x\cos x \\
&4(a_1+(b_1-c_0)\sin x - c_1 x\sin x + (b_0+c_1)\cos x + b_1 x\cos x) \\
&3(a_0+a_1x+b_0\sin x + b_1 x\sin x + c_0\cos x + c_1 x\cos x) \qquad = 1+2x+3x\cos x.
\end{aligned}$$

Matching coefficients, this yields a system of equations

$$\begin{aligned}
3a_0+4a_1 &= 1,\\
3a_1x &= 2x,\\
(2b_0+4b_1-4c_0-2c_1)\sin x &= 0\sin x,\\
(2b_1-4c_1)x\sin x &= 0x\sin x,\\
(4b_0+2b_1+2c_0+4c_1)\cos x &= 0\cos x,\\
(4b_1+2c_1)x\cos x &= 3x\cos x,
\end{aligned} \tag{3-147e}$$

which written as a linear system

$$\begin{bmatrix} 3 & 4 & 0 & 0 & 0 & 0\\ 0 & 3 & 0 & 0 & 0 & 0\\ 0 & 0 & 2 & 4 & -4 & -2\\ 0 & 0 & 0 & 2 & 0 & -4\\ 0 & 0 & 4 & 2 & 2 & 4\\ 0 & 0 & 0 & 4 & 0 & 2 \end{bmatrix} \begin{bmatrix} a_0\\ a_1\\ b_0\\ b_1\\ c_0\\ c_1 \end{bmatrix} = \begin{bmatrix} 1\\ 2\\ 0\\ 0\\ 0\\ 3 \end{bmatrix}. \tag{3-147f}$$

Solving this system yields

$$\begin{bmatrix} a_0\\ a_1\\ b_0\\ b_1\\ c_0\\ c_1 \end{bmatrix} = \begin{bmatrix} -5/9\\ 2/3\\ -33/50\\ 3/5\\ 3/25\\ 3/10 \end{bmatrix}. \tag{3-147g}$$

The particular solution is therefore:

$$y_p(x) = -\frac{5}{9} + \frac{2}{3}x - \frac{33}{50}\sin x + \frac{3}{5}x\sin x + \frac{3}{25}\cos x + \frac{3}{10}x\cos x. \tag{3-147h}$$

3.6.d Variation of Parameters

A more general method for solving ordinary differential equations is the method of variation of parameters. Applied to second-order ODEs we assume the general solution to our problem,

$$y''(x) + ay'(x) + by(x) = r(x)$$

can be expressed as

$$y(x) = u(x)y_1(x) + v(x)y_2(x), \tag{3-148}$$

where $y_1(x)$ and $y_2(x)$ are the solutions we found to the homogeneous solution. Since we have two unknown functions $u(x)$ and $v(x)$ satisfying the relationship above, we need a second equation. For this, we assume the second constraint

$$u'(x)y_1(x) + v'(x)y_2(x) = 0. \tag{3-149}$$

While not obvious here, this assumption turns out to be consistent with the mathematics to follow and will make the subsequent steps easier.

Taking the first derivative of the proposed form, we obtain

$$y'(x) = u(x)y_1'(x) + v(x)y_2'(x) + u'(x)y_1(x) + v'(x)y_2(x);$$

however, the second two terms are zero on the count of the second constraint. Therefore,

$$y'(x) = u(x)y_1'(x) + v(x)y_2'(x). \tag{3-150}$$

Taking the second derivative yields

$$y''(x) = u'(x)y_1'(x) + v'(x)y_2'(x) + u(x)y_1''(x) + v(x)y_2''. \tag{3-151}$$

Inserting the derivatives into the original ODE gives

$$\begin{aligned}&\Big[u'(x)y_1'(x) + v'(x)y_2'(x) + u(x)y_1''(x) + v(x)y_2''\Big]\\ &+ a\Big[u(x)y_1'(x) + v(x)y_2'(x)\Big] + b\Big[u(x)y_1(x) + v(x)y_2(x)\Big] = r(x)\end{aligned}$$

Now regrouping the terms to factor in terms of $u(x)$ and $v(x)$:

$$\begin{aligned}&\Big[u'(x)y_1'(x) + v'(x)y_2'(x)\Big] + u(x)\Big[y_1''(x) + ay_1'(x) + by_1(x)\Big]\\ &+ v(x)\Big[y_2''(x) + ay_2'(x) + by_2(x)\Big] = r(x)\end{aligned} \tag{3-152}$$

The second and third terms in square brackets satisfy the homogeneous equation:

$$\begin{aligned}y_1''(x) + ay_1'(x) + by_1(x) = 0,\\ y_2''(x) + ay_2'(x) + by_2(x) = 0,\end{aligned}$$

and are therefore zero. We can now simplify to write

$$u'(x)y_1'(x) + v'(x)y_2'(x) = r(x). \tag{3-153}$$

This equation with our constraint gives a system of equations for $u'(x)$ and $v'(x)$, which is

$$\begin{bmatrix} y_1(x) & y_2(x) \\ y_1'(x) & y_2'(x) \end{bmatrix} \begin{bmatrix} u'(x) \\ v'(x) \end{bmatrix} = \begin{bmatrix} 0 \\ r(x) \end{bmatrix}. \tag{3-154}$$

We can solve for $u'(x)$ and $v'(x)$ by inverting the 2×2 matrix:

$$\begin{bmatrix} u'(x) \\ v'(x) \end{bmatrix} = \frac{1}{W(x)} \begin{bmatrix} y_2'(x) & -y_2(x) \\ -y_1'(x) & y_1(x) \end{bmatrix} \begin{bmatrix} 0 \\ r(x) \end{bmatrix}. \tag{3-155}$$

where $W(x)$ is the Wronskian.

We can now write out two equations for $u'(x)$ and $v'(x)$ and integrate to obtain an expression for the unknown functions:

$$u(x) = -\int W^{-1}(x) y_2(x) r(x) dx, \tag{3-156a}$$

$$v(x) = \int W^{-1}(x) y_1(x) r(x) dx. \tag{3-156b}$$

To illustrate variation of parameters, consider the differential equation

$$y''(t) + 4y(t) = \cot t. \tag{3-157a}$$

The homogeneous solution yields two linearly independent solutions of

$$y_1(t) = \sin(2t), \tag{3-157b}$$

$$y_2(t) = \cos(2t). \tag{3-157c}$$

Using variation of parameters, we have the solution

$$y(t) = u(t)\sin(2t) + v(t)\cos(2t). \tag{3-157d}$$

To find $u(t)$ and $v(t)$ first compute the Wronskian as

$$W(t) = \begin{vmatrix} \sin(2t) & \cos(2t) \\ 2\cos(2t) & -2\sin(2t) \end{vmatrix} = -2. \tag{3-157e}$$

Evaluating for the functions gives

$$\begin{aligned} u(t) &= -\int \frac{1}{-2}\cos(2t)\cot(t)dt \\ &= \frac{1}{4}\Big[\cos(2t) + 2\ln(\sin(t))\Big] + C_1. \end{aligned} \tag{3-157f}$$

$$\begin{aligned} v(t) &= \int \frac{1}{-2}\sin(2t)\cot(t)dt \\ &= -\frac{1}{2}\Big[t + \sin(t)\cos(t)\Big] + C_2. \end{aligned} \tag{3-157g}$$

Inserting these into the proposed solution gives

$$y(t) = C_1 \sin(2t) + C_2 \cos(2t) + \frac{1}{4}\Big[\cos(2t) + 2\ln(\sin(t))\Big]\sin(2t) - \frac{1}{2}\Big[t + \sin(t)\cos(t)\Big]\cos(2t). \tag{3-157h}$$

Expanding these out gives

$$y(t) = C_1 \sin(2t) + C_2 \cos(2t) + \frac{1}{4}\cos(2t)\sin(2t) + \frac{1}{2}\sin(2t)\ln(\sin(t)) - \frac{1}{2}\cos(t)\sin(t)\cos(2t) - \frac{t}{2}\cos(2t). \tag{3-157i}$$

Now we can apply the double angle formula

$$\frac{1}{2}\sin(2t) = \cos(t)\sin(t)$$

to the fifth term to get

$$-\frac{1}{2}\cos(t)\sin(t)\cos(2t) = -\frac{1}{4}\cos(2t)\sin(2t). \tag{3-157j}$$

This then cancels out with the third term and we are left with the solution:

$$y(t) = C_1 \sin(2t) + C_2 \cos(2t) + \frac{1}{2}\sin(2t)\ln(\sin(t)) - \frac{t}{2}\cos(2t). \tag{3-157k}$$

3.6.e Example: Vibrational Resonance

Consider a mass-spring system with an applied forcing function as follows:

$$y''(t) + \omega^2 y(t) = \cos(\omega t), \quad y(0) = 0, y'(0) = 0. \tag{3-158}$$

Let's try variation of parameters by assuming a general solution of the form

$$y(t) = u(t)\sin(\omega t) + v(t)\cos(\omega t). \tag{3-159}$$

Where $y_1(t) = \sin(\omega t)$ and $y_2(t) = \cos(\omega t)$. The Wronskian is

$$W(t) = \begin{vmatrix} \sin(\omega t) & \cos(\omega t) \\ \omega\cos(\omega t) & -\omega\sin(\omega t) \end{vmatrix} = -\omega. \tag{3-160}$$

The coefficients $u(t)$ and $v(t)$ are

$$\begin{aligned} u(t) &= -\int \left(-\frac{1}{\omega}\right)\cos(\omega t)\cos(\omega t)dt \\ &= \frac{t}{2\omega} + \frac{1}{4\omega^2}\sin(2\omega t) + C_1. \end{aligned} \tag{3-161a}$$

$$v(t) = \int \left(-\frac{1}{\omega}\right) \sin(\omega t)\cos(\omega t) dt$$
$$= \frac{1}{4\omega^2}\cos(2\omega t) + C_2. \tag{3-161b}$$

Inserting this into our general solution gives

$$y(t) = C_1 \sin(\omega t) + C_2 \cos(\omega t) + \frac{1}{4\omega^2}\sin(2\omega t)\sin(\omega t)$$
$$+ \frac{1}{4\omega^2}\cos(2\omega t)\cos(\omega t) + \frac{t}{2\omega}\sin(\omega t). \tag{3-162}$$

Using the double angle formulas

$$\sin(2\omega t) = 2\sin(\omega t)\cos(\omega t), \tag{3-163a}$$
$$\cos(2\omega t) = 2\cos^2(\omega t) - 1, \tag{3-163b}$$

Inspecting the third and fourth terms and inserting the double angle formulas

$$\frac{1}{4\omega^2}\sin(2\omega t)\sin(\omega t) + \frac{1}{4\omega^2}\cos(2\omega t)\cos(\omega t)$$
$$= \frac{1}{4\omega^2}[2\sin(\omega t)\cos(\omega t)]\sin(\omega t) + \frac{1}{4\omega^2}[2\cos^2(\omega t) - 1]\cos(\omega t)$$
$$= \frac{1}{2\omega^2}\cos(\omega t)\left[\sin^2(\omega t) + \cos^2(\omega t)\right] - \frac{1}{4\omega^2}\cos(\omega t)$$
$$= \frac{1}{2\omega^2}\cos(\omega t) - \frac{1}{4\omega^2}\cos(\omega t) = \frac{1}{4\omega^2}\cos(\omega t).$$

Inserting this back into the general solution

$$y(t) = C_1 \sin(\omega t) + C_2 \cos(\omega t) + \frac{1}{4\omega^2}\cos(\omega t) + \frac{t}{2\omega}\sin(\omega t). \tag{3-164}$$

Since the second and third terms are simply $\cos(\omega t)$ with constant multipliers, we can combine them into the a constant by redefining C_2 to get the final answer:

$$y(t) = C_1 \sin(\omega t) + C_2 \cos(\omega t) + \frac{t}{2\omega}\sin(\omega t). \tag{3-165}$$

Observing the form of the solution, we end up with a term that goes as $t\sin(\omega t)$. This implies the solution is a growing oscillation. Generally speaking, in mechanical designs if a system is driven with a forcing function governed by one of its natural frequencies that would arise in the homogeneous equation, the system will become unstable and experience ever increasing stresses. These types of mechanical stresses have led to catastrophic failures of bridges and early aircraft and must be avoided.

3.7 Initial Value Problems

An important class of problems are called initial value problems. In an initial value problem described by a second-order ODE, the domain of the problem is defined as

$0 \leq t < \infty$ and we are provided with initial conditions on the functions and its first derivative at $t = 0$. A common application is writing an equation of motion via Newton's second law where we know an initial position and speed and the object trajectory changes based upon applied external forces.

For example a mass-spring damper problem under the influence of gravity takes the form:

$$F = mx''(t) = -kx(t) - mg, \quad x(0) = 0, x'(0) = 0. \tag{3-166}$$

Rewriting this differential equation as

$$x''(t) + \omega^2 x(t) = -g \tag{3-167}$$

where

$$\omega = \sqrt{\frac{k}{m}}, \tag{3-168}$$

we obtain the solution

$$x(t) = C_1 \sin(\omega t) + C_2 \cos(\omega t) - \frac{g}{\omega^2}. \tag{3-169}$$

Evaluating the derivative for the speed gives

$$v(t) = x'(t) = C_1 \omega \cos(\omega t) - C_2 \omega \sin(\omega t). \tag{3-170}$$

Applying the initial conditions gives

$$0 = C_2 - \frac{g}{\omega^2}, \tag{3-171a}$$

$$0 = C_1 \omega. \tag{3-171b}$$

Therefore we can see that $C_1 = 0$ and $C_2 = g/\omega^2$, so the general solution is

$$x(t) = \frac{g}{\omega^2} \cos(\omega t) - \frac{g}{\omega^2}. \tag{3-172}$$

The mass therefore starts at its peak and oscillates in time. A notional plot of the solution is given in Fig. 3.10.

What makes an initial value problem as such is its domain $0 \leq t < \infty$ and that the initial conditions are given at $t = 0$. This gives the property that for a linear second-order ODE, the solution always exists. Furthermore, if we find a general solution (two linearly independent solutions for the homogeneous equation and a particular solution) that satisfies the differential equation and the boundary conditions, we are guaranteed that the solution is unique. In the other class of problems we will study, called boundary value problems, it is possible to construct boundary conditions that are inconsistent with the differential equation, which would yield no solution.

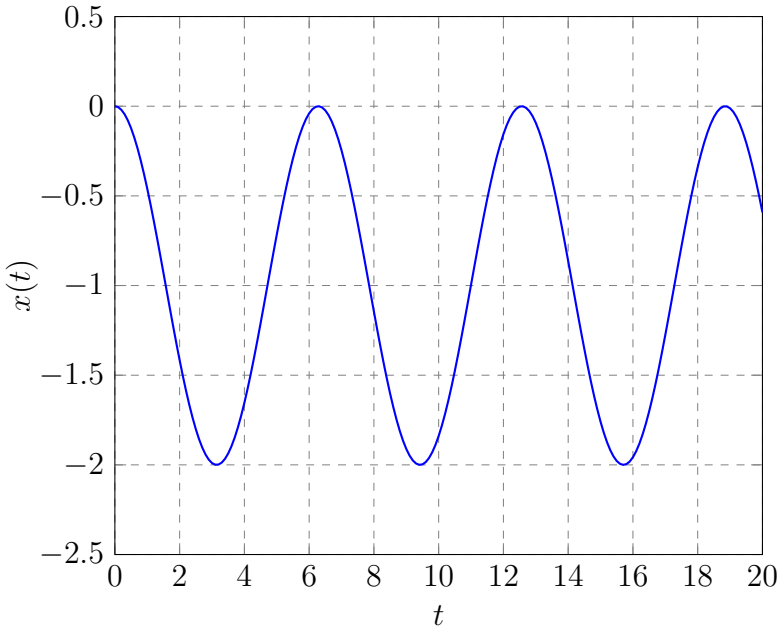

Figure 3.10: Solution of a simple spring problem.

3.7.a Coupled Systems of Initial Value Problems

As with first-order ODEs we can have coupled systems of second-order initial value problems. For example, two coupled initial value problems can be written as

$$\begin{aligned} y_1''(t) + a_{11}y_1'(t) + b_{11}y_1(t) + a_{12}y_2'(t) + b_{12}y_2(t) &= r_1(t), \\ y_2''(t) + a_{21}y_1'(t) + b_{21}y_1(t) + a_{22}y_2'(t) + b_{22}y_2(t) &= r_2(t). \end{aligned} \tag{3-173}$$

with initial conditions on $y_1(0)$, $y_1'(0)$, $y_2(0)$, and $y_2'(0)$ given.

There are two approaches that can be used to address this system of equations. The first method that works only for initial value problems (not boundary value problems), is to reduce the system of N second order equations to a set of first order equations. If we let $v(t) = y'(t)$. We can write the above system as follows:

$$\begin{aligned} y_1'(t) - v_1(t) &= 0, \\ y_2'(t) - v_2(t) &= 0, \\ v_1'(t) + a_{11}v_1(t) + b_{11}y_1(t) + a_{12}v_2(t) + b_{12}y_2(t) &= r_1(t), \\ v_2'(t) + a_{21}v_1(t) + b_{21}y_1(t) + a_{22}v_2(t) + b_{22}y_2(t) &= r_2(t), \end{aligned} \tag{3-174}$$

or as an equivalent linear system as

$$\begin{bmatrix} y_1'(t) \\ y_2'(t) \\ v_1'(t) \\ v_2'(t) \end{bmatrix} + \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ b_{11} & b_{12} & a_{11} & a_{12} \\ b_{21} & b_{22} & a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} y_1(t) \\ y_2(t) \\ v_1(t) \\ v_2(t) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ r_1(t) \\ r_2(t) \end{bmatrix}. \tag{3-175}$$

This linear system of first-order differential equations can be solved using the techniques discussed previously. The advantage to this approach is that there are relatively system techniques available to solve this problem. The disadvantage is that we now need to solve twice as many equations. Additionally, this technique will not work for boundary value problems, which are very common in science and engineering.

The second solution method for the system of equations is similar to what is done for one equation. First, we solve the homogeneous problem and then attempt to find a particular solutions. For systems of second-order ordinary differential equations, we will only use the method of undetermined coefficients to find the particular solutions.

3.7.b Example: Coupled Mass-Spring System

To illustrate, consider the scenario with two blocks of equal mass m that are connected to the walls and each other by a series of springs with equal spring constants. First we will consider the case where there is no external force. Then we will consider another case with some forcing function.

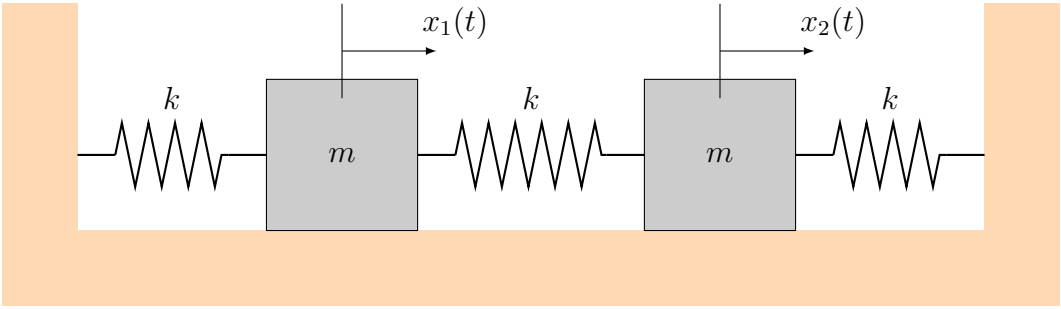

Figure 3.11: Illustration of coupled mass-spring system.

The equations of motion describing the position of the blocks for the homogeneous case are as follows:

$$mx_1''(t) = -kx_1(t) + k(x_2(t) - x_1(t)), \tag{3-176a}$$
$$mx_2''(t) = -kx_2(t) - k(x_2(t) - x_1(t)). \tag{3-176b}$$

These equations can be rewritten as

$$x_1''(t) + \frac{2k}{m}x_1(t) - \frac{k}{m}x_2(t) = 0, \tag{3-177a}$$
$$x_2''(t) - \frac{k}{m}x_1(t) + \frac{2k}{m}x_2(t) = 0. \tag{3-177b}$$

Since this is an oscillatory system, let us guess the following solutions:

$$x_1(t) = A_1 e^{i\omega t}, \tag{3-178a}$$
$$x_2(t) = A_2 e^{i\omega t}. \tag{3-178b}$$

(We could guess real values in the exponential and would get to imaginary values anyway, so this simplifies matters.) Inserting this into the differential equation gives the following:

$$-A_1\omega^2 + \frac{2k}{m}A_1 - \frac{k}{m}A_2 = 0, \tag{3-179}$$

$$-A_2\omega^2 - \frac{k}{m}A_1 + \frac{2k}{m}A_2 = 0. \tag{3-180}$$

This linear system can be written in matrix-vector form as

$$\begin{bmatrix} \frac{2k}{m} - \omega^2 & -\frac{k}{m} \\ -\frac{k}{m} & \frac{2k}{m} - \omega^2 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{3-181}$$

To satisfy this relationship, we know the coefficients A_1, A_2 are zero (the trivial solution) or the determinant of the matrix is zero. This system is equivalent to an eigenvalue problem where the eigenvalue is ω^2. Solving for the eigenvalues we get

$$\omega^2 = \left\{\omega_1^2, \omega_2^2\right\} = \left\{\frac{k}{m}, \frac{3k}{m}\right\}. \tag{3-182}$$

Inserting each eigenvalue and finding the eigenvectors gives the corresponding eigenvectors

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{3-183}$$

This implies that for the first eigenvalue ω_1^2, $A_1 = A_2$, and for the second eigenvalue ω_2^2, $A_1 = -A_2$.

Recall that our proposed solution to differential equation was in terms of ω and not ω^2, so we have

$$\omega_1 = \pm\sqrt{\frac{k}{m}}, \quad \omega_2 = \pm\sqrt{\frac{3k}{m}}. \tag{3-184}$$

The $\pm$ terms give us two linearly independent solutions $e^{i\omega t}$ and $e^{-i\omega t}$ with different coefficients that we will call A_i and B_i with i corresponding to each eigenvalue. We can now write our solution as a linear combination of all terms

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \left(A_1 e^{i\omega_1 t} + B_1 e^{-i\omega_1 t}\right) \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \left(A_2 e^{i\omega_2 t} + B_2 e^{-i\omega_2 t}\right) \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{3-185}$$

Expanding these out gives the homogeneous solution:

$$x_1(t) = A_1 e^{i\omega_1 t} + A_2 e^{i\omega_2 t} + B_1 e^{-i\omega_1 t} + B_2 e^{-i\omega_2 t} \tag{3-186a}$$

$$x_2(t) = A_1 e^{i\omega_1 t} - A_2 e^{i\omega_2 t} + B_1 e^{-i\omega_1 t} - B_2 e^{-i\omega_2 t}. \tag{3-186b}$$

Now let us consider a specific case for the homogeneous solution. Suppose we are given the initial conditions

$$\begin{aligned} x_1(0) &= 0, \\ x_2(0) &= 0, \\ x_1'(0) &= 1, \\ x_2'(0) &= 0. \end{aligned} \tag{3-187}$$

We give the first mass a "kick" at time $t = 0$, but everything else is initially stationary. To apply the initial conditions we also need to velocities, which are

$$x_1'(t) = i\omega_1 A_1 e^{i\omega_1 t} + i\omega_2 A_2 e^{i\omega_2 t} - i\omega_1 B_1 e^{-i\omega_1 t} - i\omega_2 B_2 e^{-i\omega_2 t}, \tag{3-188a}$$

$$x_2'(t) = i\omega_1 A_1 e^{i\omega_1 t} - i\omega_2 A_2 e^{i\omega_2 t} - i\omega_1 B_1 e^{-i\omega_1 t} + i\omega_2 B_2 e^{-i\omega_2 t}. \tag{3-188b}$$

Inserting each initial condition into the appropriate equations gives the system:

$$A_1 + A_2 + B_1 + B_2 = 0 \tag{3-189a}$$

$$A_1 - A_2 + B_1 - B_2 = 0 \tag{3-189b}$$

$$\omega_1 A_1 + \omega_2 A_2 - \omega_1 B_1 - \omega_2 B_2 = -i, \tag{3-189c}$$

$$\omega_1 A_1 - \omega_2 A_2 - \omega_1 B_1 + \omega_2 B_2 = 0. \tag{3-189d}$$

In matrix-vector form this is

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ \omega_1 & \omega_2 & -\omega_1 & -\omega_2 \\ \omega_1 & -\omega_2 & -\omega_1 & \omega_1 \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -i \\ 0 \end{bmatrix}. \tag{3-190}$$

Let $\omega_1 = \omega$ and it follows from the eigenvalues that $\omega_2 = \sqrt{3}\omega_1 = \sqrt{3}\omega$. Solving the above system yields the solution vector

$$\begin{bmatrix} A_1 \\ A_2 \\ B_1 \\ B_2 \end{bmatrix} = \frac{i}{4\omega} \begin{bmatrix} -1 \\ -1/\sqrt{3} \\ 1 \\ 1/\sqrt{3} \end{bmatrix} \tag{3-191}$$

To provide numbers, suppose that $k/m = 1$; which means $\omega = 1$. Inserting the parameters into the equation gives

$$x_1(t) = \frac{i}{4}\left(-e^{it} - \frac{1}{\sqrt{3}}e^{i\sqrt{3}t} + e^{-it} + \frac{1}{\sqrt{3}}e^{-i\sqrt{3}t}\right) \tag{3-192a}$$

$$x_2(t) = \frac{i}{4}\left(-e^{it} + \frac{1}{\sqrt{3}}e^{i\sqrt{3}t} + e^{-it} - \frac{1}{\sqrt{3}}e^{-i\sqrt{3}t}\right). \tag{3-192b}$$

While these equations involve the imaginary unit i, they are indeed real. If we apply Euler's formula, we can express the complex exponentials in terms of trigonometric

functions:

$$x_1(t) = \frac{1}{2}\sin(t) + \frac{1}{2\sqrt{3}}\sin(\sqrt{3}t), \tag{3-193a}$$

$$x_2(t) = \frac{1}{2}\sin(t) - \frac{1}{2\sqrt{3}}\sin(\sqrt{3}t). \tag{3-193b}$$

These functions are plotted in Fig. 3.12.

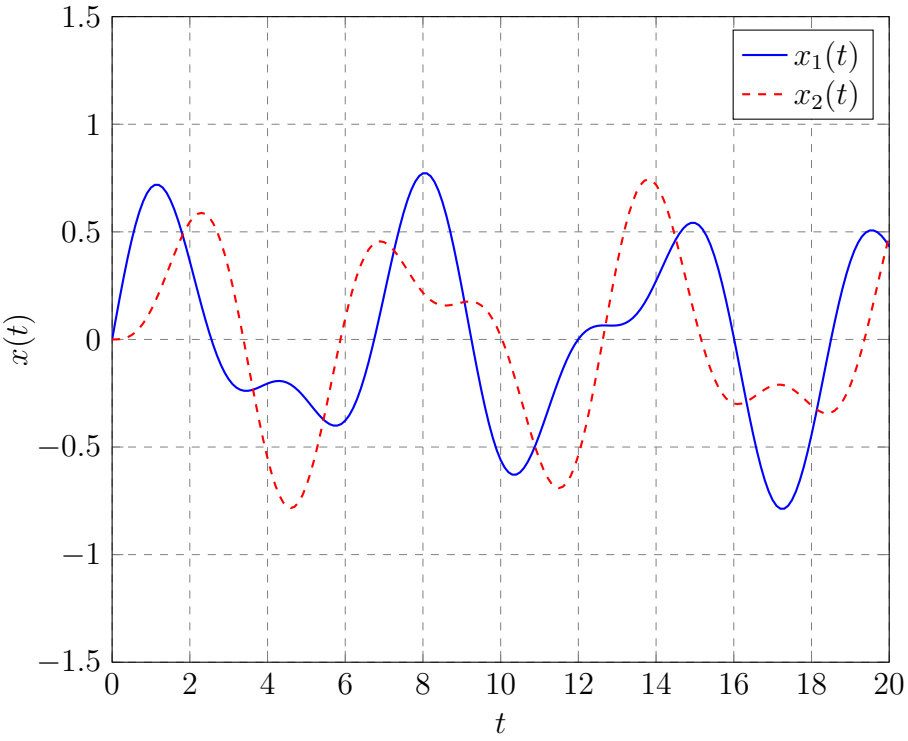

Figure 3.12: Solution of a homogeneous coupled mass-spring problem.

Now let us suppose we apply an exponential forcing function to the block 1. This leads to a new set of differential equations as follows:

$$x_1''(t) + \frac{2k}{m}x_1(t) - \frac{k}{m}x_2(t) = e^{-\gamma t}, \tag{3-194a}$$

$$x_2''(t) - \frac{k}{m}x_1(t) + \frac{2k}{m}x_2(t) = 0. \tag{3-194b}$$

The homogeneous solution is identical to what we found before. To find the particular solution, we apply the method of underdetermined coefficients. Since the inhomogeneous solution is exponential, we guess an exponential:

$$x_1(t) = c_1 e^{-\gamma t}, \tag{3-195a}$$

$$x_2(t) = c_2 e^{-\gamma t}. \tag{3-195b}$$

Plugging these into the differential equations and canceling out the exponential yields the system

$$\gamma^2 c_1 + \frac{2k}{m}c_1 - \frac{k}{m}c_2 = 1, \tag{3-196a}$$

$$\gamma^2 c_2 - \frac{k}{m}c_1 + \frac{2k}{m}_2 = 0. \tag{3-196b}$$

We can write this in matrix-vector form as

$$\begin{bmatrix} \gamma^2 + \frac{2k}{m} & -\frac{k}{m} \\ -\frac{k}{m} & \gamma^2 + \frac{2k}{m} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{3-197}$$

Solving this system yields the solution vector

$$\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} \dfrac{\gamma^2 + {}^{2k}/_{m}}{(\gamma^2 + {}^{2k}/_{m})^2 - {}^{k^2}/_{m^2}} \\ \dfrac{{}^{k}/_{m}}{(\gamma^2 + {}^{2k}/_{m})^2 - {}^{k^2}/_{m^2}} \end{bmatrix}. \tag{3-198}$$

The general solution is therefore (leaving in terms of constants c_1 and c_2 since it is rather unwieldy):

$$x_1(t) = A_1 e^{i\omega_1 t} + A_2 e^{i\omega_2 t} + B_1 e^{-i\omega_1 t} + B_2 e^{-i\omega_2 t} + c_1 e^{-\gamma t}, \tag{3-199a}$$

$$x_2(t) = A_1 e^{i\omega_1 t} - A_2 e^{i\omega_2 t} + B_1 e^{-i\omega_1 t} - B_2 e^{-i\omega_2 t} + c_2 e^{-\gamma t}. \tag{3-199b}$$

Taking derivatives to get the velocity terms gives

$$x_1'(t) = i\omega_1 A_1 e^{i\omega_1 t} + i\omega_2 A_2 e^{i\omega_2 t} - i\omega_1 B_1 e^{-i\omega_1 t} - i\omega_2 B_2 e^{-i\omega_2 t} - \gamma c_1 e^{-\gamma t}, \tag{3-200a}$$

$$x_2'(t) = i\omega_1 A_1 e^{i\omega_1 t} - i\omega_2 A_2 e^{i\omega_2 t} - i\omega_1 B_1 e^{-i\omega_1 t} + i\omega_2 B_2 e^{-i\omega_2 t} - \gamma c_2 e^{-\gamma t}. \tag{3-200b}$$

Before moving onto discussing an example with an inhomogeneous term and obtaining the particular solution, let us connect this to eigenvalues and eigenvectors from linear algebra. To begin solving the problem, we calculated the eigenvalues of the system. In the context of mass-spring systems or harmonic oscillators, these eigenvalues correspond to the natural frequencies of the system. For each of these eigenvalues (natural frequencies) we obtained a vector of *eigenfunctions* that are proportional to $e^{\pm i\omega_k t}$ that could then be expressed in terms of trigonometric functions. Eigenfunctions are the continuous analog in calculus to the eigenvectors in linear algebra. As we applied the initial conditions, note that the homogeneous solution is a linear combination of the eigenfunctions, which here can be expressed as either complex exponentials or trigonometric functions. The ability to express a homogeneous solution as a linear combinations of its eigenfunctions is a general result to differential equations and an important idea.

Next let us consider a case with an inhomogeneous term. Suppose that again $k/m = 1$ and now $\gamma = 1$. The coefficients for the particular solution are $c_1 = {}^{3}/_{8}$

and $c_2 = {}^1\!/_8$. Furthermore, we will assume that the positions and velocities are now initially zero:

$$\begin{aligned} x_1(0) &= 0, \\ x_2(0) &= 0, \\ x_1'(0) &= 0, \\ x_2'(0) &= 0. \end{aligned} \tag{3-201}$$

Applying these initial conditions yields the system

$$A_1 + A_2 + B_1 + B_2 = -\frac{3}{8} \tag{3-202a}$$

$$A_1 - A_2 + B_1 - B_2 = -\frac{1}{8} \tag{3-202b}$$

$$A_1 + \sqrt{3}A_2 - B_1 - \sqrt{3}B_2 = -\frac{3i}{8}, \tag{3-202c}$$

$$A_1 - \sqrt{3}A_2 - B_1 + \sqrt{3}B_2 = -\frac{i}{8}. \tag{3-202d}$$

The corresponding matrix-vector form is

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & \sqrt{3} & -1 & -\sqrt{3} \\ 1 & -\sqrt{3} & -1 & \sqrt{3} \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} -{}^3\!/_8 \\ -{}^1\!/_8 \\ -{}^{3i}\!/_8 \\ -{}^i\!/_8 \end{bmatrix}. \tag{3-203}$$

The solution vector is

$$\begin{bmatrix} A_1 \\ A_2 \\ B_1 \\ B_2 \end{bmatrix} = -\frac{1}{8} \begin{bmatrix} 1+i \\ \frac{1}{2} + \frac{i}{2\sqrt{3}} \\ 1-i \\ \frac{1}{2} - \frac{i}{2\sqrt{3}} \end{bmatrix}. \tag{3-204}$$

Plugging in the values and applying Euler's formula gives the result:

$$x_1(t) = \frac{1}{4}\sin(t) - \frac{1}{4}\cos(t) + \frac{1}{8\sqrt{3}}\sin(\sqrt{3}t) - \frac{1}{8}\cos(\sqrt{3}t) + \frac{3}{8}e^{-t}, \tag{3-205a}$$

$$x_2(t) = \frac{1}{4}\sin(t) - \frac{1}{4}\cos(t) - \frac{1}{8\sqrt{3}}\sin(\sqrt{3}t) + \frac{1}{8}\cos(\sqrt{3}t) + \frac{1}{8}e^{-t}. \tag{3-205b}$$

This solution is plotted in Fig. 3.13. As before, we have the homogeneous solution expressed as a linear combination of eigenfunctions (this time with different coefficients) with an added term because of the forcing function.

3.8 Boundary Value Problems

The other major and important class of problems for differential equations are boundary value problems. Boundary value problems involving second-order ordinary differential equation are ubiquitous and appear in fields ranging from: fluid dynamics,

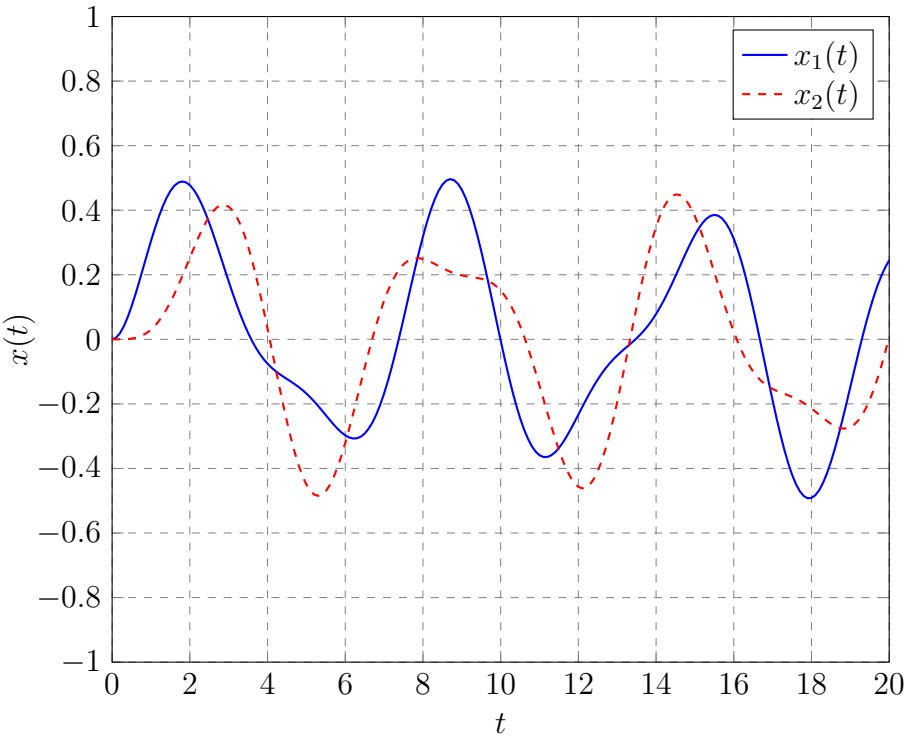

Figure 3.13: Solution of a coupled mass-spring problem with an exponential forcing function.

heat transfer, neutron transport and reactor physics, plasma physics, materials science, and quantum mechanics to name a few.

The difference of a boundary value problem versus an initial value problem is now the problem is defined over some domain with boundary conditions applied at the exterior as opposed to from $0 \leq t \leq \infty$ with initial conditions given at $t = 0$ The system could be infinite $-\infty < x < \infty$, semi-infinite similar to initial value problems $a \leq x < \infty$, or finite $a \leq x \leq b$. In boundary value problems, boundary conditions are provided, as the name implies, at the exterior boundaries of the system and these can be in terms of the function, its derivative, or a linear combination thereof. It is also common that for many problems in science and engineering the interior of the problem has discrete regions, often representing different objects or materials, and these are treated mathematically with interface conditions. Once we have the boundary and interface conditions, we can set up a system of equations for the boundary value problem and attempt to solve.

3.8.a Boundary Conditions

Boundary conditions are specified at the edges of the problem domain and are necessary to have a well-posed problem. There are three major types of boundary conditions that appear in most science and engineering applications: the Dirichlet boundary

condition, the Neumann boundary condition, and the Robbin boundary condition.

The Dirichlet boundary condition is the simplest type and specifies the solution of the field at some boundary point x_b:

$$f(x_b) = c. \tag{3-206}$$

For example, in heat conduction problems we are often interested in solving for the distribution of temperatures, or the temperature field. In many cases the temperature at one of the boundaries is known. Also in fluids problems, we often specify the "no-slip condition" that says the velocity of the fluid is zero at the walls.

The Neumann boundary condition specifies the first derivative of the solution of the field at some boundary point x_b:

$$f'(x_b) = c. \tag{3-207}$$

A common example occurs in heat condition where rather than the temperature T being specified on the boundary, we have a known heat flux q, which is proportional to the derivative of the temperature,

$$q = -k\frac{dT}{dx}. \tag{3-208}$$

Another important place where the Neumann boundary condition arises is when we are applying symmetry to a problem. It is often the case that it is easier to solve a portion of a much larger problem where that portion is replicated in a way that we have symmetry. In these cases, it is often the case that derivatives are zero at boundaries for symmetry. We often refer to this as a "reflecting boundary condition" or "symmetry boundary condition".

The third important type of boundary condition is the Robin boundary condition, which specifies a linear combination of the field and its first derivative are given on the boundary:

$$af(x_b) + bf'(x_b) = c. \tag{3-209}$$

A common example of this in heat conduction involves a convective boundary condition, where energy is transferred to the ambient medium. This is often written as

$$q = h(T - T_\infty), \tag{3-210}$$

where h is a heat transfer coefficient and T_∞ is the ambient temperature "infinitely" far away from the heated object being analyzed. To see this is equivalent to the form of the Robin boundary condition, let us expand the heat flux and rearrange:

$$\frac{k}{h}T'(x_b) + T(x_b) = T_\infty. \tag{3-211}$$

The other place in nuclear engineering where this comes up is neutron diffusion in reactors, which is called the Marshak boundary condition. Here we wish to know about

the neutron path-length rate density (or neutron scalar flux, which is proportional to the nuclear reaction rate) $\phi(x)$ at the boundary. This is given as

$$\frac{1}{4}\phi(x_b) \mp \frac{D}{2}\phi'(x_b) = J^{\pm}(x_b). \tag{3-212}$$

Here $J^{+}(x_b)$ and $J^{-}(x_b)$ are the given neutron currents (flow rate of neutrons) given on the left and right sides of the problem respectively; D is the neutron diffusion coefficient describing the magnitude of the ability for neutrons to spread out within a given medium. Note that the $\pm$ and $\mp$ symbols mean that when one side uses the $+$ the other uses the $-$ and vice versa.

There is another class of boundary condition that often occurs. It is often the case that we must assert that the field is finite or tends toward zero as the field goes out toward infinity (in other words, the field does not "blow up" for large magnitude in x). Also, when we are in cylindrical or spherical geometry, we often must assert that the field is finite at the boundary.

3.8.b Interface Conditions

Within the interior of the problem, we must specify how the field and its derivative are connected to adjacent regions. It is often the case that the field is continuous. If we have an interface point a between regions 1 and 2, the interface condition for the field is often

$$f_1(a) = f_2(a). \tag{3-213}$$

Note that it is possible in certain situations, e.g., shock waves that are important in inertial confinement fusion, to have a discontinuity in the field; this is called a jump condition.

We also must often specify some condition on the derivative at the interface. Physically, this means that the flow rate of some physical quantity (e.g., energy, neutrons, etc.) related to the first derivative of the field (e.g., heat flux, neutron current, respectively) is continuous across the boundary. Because the flow rates of these properties often depend directly upon some property in the system, the derivative of the field is often discontinuous and exhibits a "kink". This interface condition is often written as

$$k_1 f_1'(a) = k_2 f_2'(a). \tag{3-214}$$

Sometimes there can be an added inhomogeneous constant value as well (analogous to a jump condition). This can, for example, denote a source of energy at the boundary or a surface charge in electrostatics.

The derivation of these conditions involves integrating the differential equation over some small range $\pm\epsilon$ around the interface and taking the limit as $\epsilon \to 0$. Consider the following diffusion equation at an interface

$$-\frac{d}{dx}\left(D(x)\frac{d\phi}{dx}\right) + \Sigma_a(x)\phi(x) = Q(x), \tag{3-215}$$

where the diffusion coefficient is

$$D(x) = \begin{cases} D_1 & x < a, \\ D_2 & x > a, \end{cases} . \tag{3-216}$$

If we wish to derive an interface condition, we integrate from $a - \epsilon$ to $a + \epsilon$:

$$-\int_{a-\epsilon}^{a+\epsilon} \frac{d}{dx}\left(D(x)\frac{d\phi}{dx}\right) + \int_{a-\epsilon}^{a+\epsilon} \Sigma_a(x)\phi(x) = \int_{a-\epsilon}^{a+\epsilon} Q(x). \tag{3-217}$$

Using the second fundamental theorem of calculus on the first term yields

$$-\left[D_2\phi_2'(a+\epsilon) - D_1\phi_1'(a-\epsilon)\right] + \int_{a-\epsilon}^{a+\epsilon} \Sigma_a(x)\phi(x) = \int_{a-\epsilon}^{a+\epsilon} Q(x). \tag{3-218}$$

Now, we take the limit as $\epsilon \to 0$. The integrals that remain go to zero and we are left with:

$$-\left[D_2\phi_2'(a) - D_1\phi_1'(a)\right] = 0. \tag{3-219}$$

3.8.c Example: Flow Between Two Parallel Plates

Consider a fluid flowing between two plates in a rectangular duct with a width $2L$ such that $-L \leq x \leq L$ and very wide in the y and z directions so that we can ignore the effects of the boundaries in those directions. The flow is driven by a constant pressure gradient in the y direction driven by gravity:

$$\frac{dp}{dy} = \rho g. \tag{3-220}$$

The equations describing the velocity profile in the y direction, $v_y(x)$ is given as

$$\mu\frac{d^2v_y}{dx^2} = \frac{dp}{dy} = \rho g; \tag{3-221}$$

Here μ is the fluid viscosity that represents the friction generated within the fluid. The fluid velocity profile is assumed to satisfy the no-slip condition, so

$$v_y(-L) = 0, \tag{3-222a}$$

$$v_y(L) = 0. \tag{3-222b}$$

The differential equation is simple to solve, as it can be integrated twice. This yields a solution in terms of two constants of iteration:

$$v_y(x) = \frac{\rho g x^2}{2\mu} + C_1 x + C_2. \tag{3-223}$$

Inserting in the boundary conditions gives

$$v_y(-L) = \frac{\rho g L^2}{2\mu} - C_1 L + C_2 = 0, \tag{3-224a}$$

$$v_y(L) = \frac{\rho g L^2}{2\mu} + C_1 L + C_2 = 0. \tag{3-224b}$$

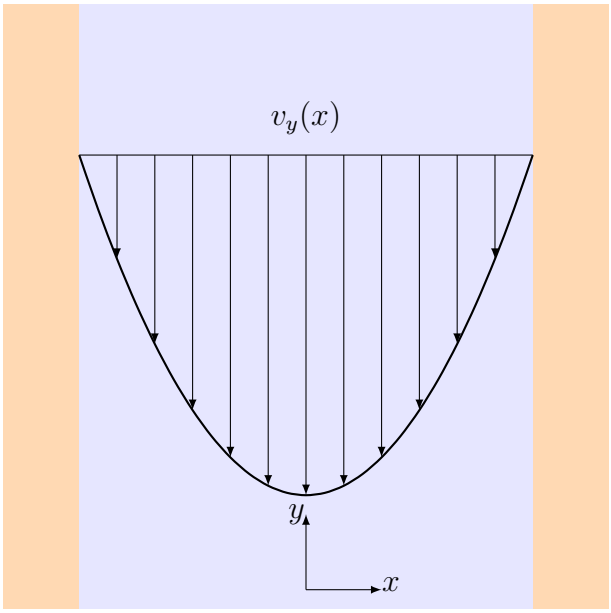

Figure 3.14: Illustration of velocity profile between two parallel plates.

Writing a linear system in matrix-vector form gives

$$\begin{bmatrix} -L & 1 \\ L & 1 \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = -\frac{\rho g L^2}{2\mu} \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{3-225}$$

Solving this system yields

$$C_1 = 0. \tag{3-226a}$$

$$C_2 = -\frac{\rho g L^2}{2\mu}. \tag{3-226b}$$

Inserting this into the equation gives the final result

$$v_y(x) = -\frac{\rho g L^2}{2\mu} \left[1 - \left(\frac{y}{L} \right)^2 \right]. \tag{3-227}$$

The flow is negative because gravity points in the downward direction and has a parabolic shape with the maximum being at the centerline $x = 0$. The velocity profile is depicted in Fig. 3.14.

3.8.d Example: Heat Conduction in a Nuclear Fuel Rod

An important parameter for nuclear fission reactor analysis is modeling the energy transport in the reactor. Part of this is finding the temperature field within a fuel pin to ensure that the temperature throughout the fuel and cladding stays below materials limit that could damage the cladding and lead to a subsequent release of radioactive fission products into the coolant.

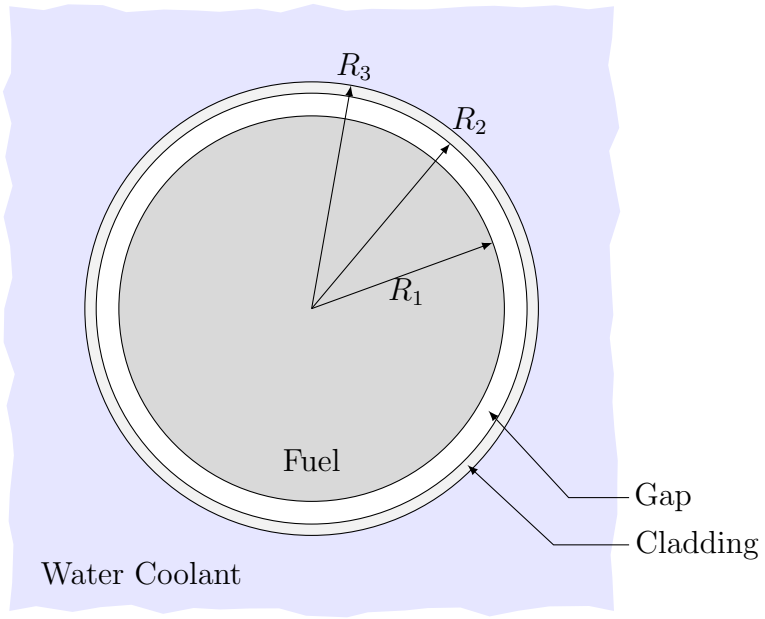

Figure 3.15: Illustration of a nuclear fuel rod in coolant.

A fuel pin is a cylindrical object (depicted in Fig. 3.15) with a diameter of about 1 cm with the uranium dioxide fuel being about 0.8 cm in diameter, cladding on the outside being about 0.05 cm in thickness, and the area between being a fill gas that has a low thermal conductivity. The fuel is surrounded by water where the heat is removed through forced convection.

We will solve a simplified version of this problem where we only consider the radial dependence and neglect any axial power dependence. In cylindrical geometry, the equation for heat conduction within a region is given by

$$-\frac{k}{r}\frac{d}{dr}\left(r\frac{dT}{dr}\right) = Q(r), \tag{3-228}$$

where r is the radial coordinate, k is the thermal conductivity that we assume to be independent of temperature and constant within each region, and $Q(r)$ is the heat generation distribution.

Within the fuel, which we denote as region 1, we assume the heat generation is spatially constant $Q_1(r) = Q_1$. The differential equation becomes:

$$-\frac{k_1}{r}\frac{d}{dr}\left(r\frac{dT}{dr}\right) = Q_1. \tag{3-229a}$$

In the gap and cladding (regions 2 and 3 respectively), there is no heat generation $Q_2(r) = 0, Q_3(r) = 0$. This leads to the equations:

$$-\frac{k_2}{r}\frac{d}{dr}\left(r\frac{dT}{dr}\right) = 0, \tag{3-229b}$$

$$-\frac{k_3}{r}\frac{d}{dr}\left(r\frac{dT}{dr}\right) = 0. \tag{3-229c}$$

We have three second-order ordinary differential equations that all require a total of six constraints. For the fuel region, we require that at the origin, $r = 0$, is finite:

$$T_1(0) < \infty. \tag{3-230a}$$

Next, we require two interface conditions between the fuel and gap being that the temperature and heat flux are continuous across the interface:

$$T_1(R_1) = T_2(R_1), \tag{3-230b}$$
$$q_1(R_1) = q_2(R_1). \tag{3-230c}$$

Likewise, the same interface conditions apply between the gap and cladding:

$$T_2(R_2) = T_3(R_2), \tag{3-230d}$$
$$q_2(R_2) = q_3(R_2). \tag{3-230e}$$

Finally, on the outer surface of the cladding, the energy is removed by the coolant via the convection process. Here we apply the convective boundary condition:

$$q_3(R_3) = h(T_3(R_3) - T_\infty). \tag{3-230f}$$

As with the previous example, these equations can be integrated twice directly. The results are:

$$T_1(r) = -\frac{Q_1}{4k_1}r^2 + A_1\ln(r) + B_1, \tag{3-231a}$$
$$T_2(r) = A_2\ln(r) + B_2, \tag{3-231b}$$
$$T_3(r) = A_3\ln(r) + B_3. \tag{3-231c}$$

Applying the condition that $T_1(0)$ is finite requires that $A_1 = 0$ since $\ln(r) \to -\infty$ as $r \to 0$. Given that simplification, let us compute the heat fluxes:

$$q_1(r) = -k_1\frac{dT_1}{dr} = \frac{Q_1}{2}r, \tag{3-232a}$$
$$q_2(r) = -k_2\frac{dT_2}{dr} = -k_2A_2\frac{1}{r}, \tag{3-232b}$$
$$q_3(r) = -k_3\frac{dT_2}{dr} = -k_3A_3\frac{1}{r}. \tag{3-232c}$$

Applying the interface condition for the temperature between the fuel and gap gives

$$-\frac{Q_1}{4k_1}R_1^2 + B_1 = A_2\ln(R_1) + B_2. \tag{3-233}$$

Moving all the coefficients to the left-hand side and everything else to the right-hand side gives

$$B_1 - \ln(R_1)A_2 - B_2 = \frac{Q_1R_1^2}{4k_1}. \tag{3-234}$$

Applying the corresponding heat flux equation gives

$$-k_2 A_2 \frac{1}{R_1} = \frac{Q_1 R_1}{2} \tag{3-235}$$

Solving for A_2 explicitly gives

$$A_2 = -\frac{Q_1 R_1^2}{2k_2}. \tag{3-236}$$

Next, applying the temperature interface condition to gap/clad interface gives

$$A_2 \ln(R_2) + B_2 = A_3 \ln(R_2) + B_3. \tag{3-237}$$

Grouping terms gives

$$\ln(R_2)A_2 + B_2 - \ln(R_2)A_3 - B_3 = 0. \tag{3-238}$$

The corresponding heat flux condition is

$$k_2 A_2 - k_3 A_3 = 0 \tag{3-239}$$

Solving for A_3 is straightforward:

$$A_3 = -\frac{Q_1 R_1^2}{2k_3}. \tag{3-240}$$

Finally, we apply the convective boundary condition to the outer surface of the cladding:

$$-k_3 A_3 \frac{1}{R_3} = h\left(A_3 \ln(R_3) + B_3 - T_\infty\right). \tag{3-241}$$

Rearranging terms gives

$$\left(\ln(R_3) + \frac{k_3}{hR_3}\right) A_3 + B_3 = T_\infty. \tag{3-242}$$

Collecting these into a linear system gives

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -\ln(R_1) & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & \ln(R_2) & 1 & -\ln(R_2) & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \ln(R_3) + \frac{k_3}{hR_3} & 1 \end{bmatrix} \begin{bmatrix} A_1 \\ B_1 \\ A_2 \\ B_2 \\ A_3 \\ B_3 \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{Q_1 R_1^2}{4k_1} \\ -\frac{Q_1 R_1^2}{2k_2} \\ 0 \\ -\frac{Q_1 R_1^2}{2k_3} \\ T_\infty \end{bmatrix}. \tag{3-243}$$

We can then apply Gaussian elimination to solve this system. The following coefficients are obtained:

$$A_1 = 0, \tag{3-244a}$$

$$B_1 = T_\infty + \frac{Q_1 R_1^2}{2}\left[\frac{1}{2k_1} + \frac{1}{k_2}\ln\left(\frac{R_2}{R_1}\right) + \frac{1}{k_3}\ln\left(\frac{R_3}{R_2}\right) + \frac{1}{hR_3}\right], \tag{3-244b}$$

$$A_2 = -\frac{Q_1 R_1^2}{2k_2}, \tag{3-244c}$$

$$B_2 = T_\infty + \frac{Q_1 R_1^2}{2}\left[\frac{1}{k_2}\ln(R_2) + \frac{1}{k_3}\ln\left(\frac{R_3}{R_2}\right) + \frac{1}{hR_3}\right]. \tag{3-244d}$$

$$A_3 = -\frac{Q_1 R_1^2}{2k_3}, \tag{3-244e}$$

$$B_3 = T_\infty + \frac{Q_1 R_1^2}{2}\left[\frac{1}{k_3}\ln(R_3) + \frac{1}{hR_3}\right]. \tag{3-244f}$$

The temperature field solution is therefore:

$$\begin{aligned} T_1(r) = T_\infty &+ \frac{Q_1 R_1^2}{4k_1}\left[1 - \left(\frac{r}{R_1}\right)^2\right] \\ &+ \frac{Q_1 R_1^2}{2}\left[\frac{1}{k_2}\ln\left(\frac{R_2}{R_1}\right) + \frac{1}{k_3}\ln\left(\frac{R_3}{R_2}\right) + \frac{1}{hR_3}\right], \end{aligned} \tag{3-245a}$$

$$T_2(r) = T_\infty + \frac{Q_1 R_1^2}{2}\left[\frac{1}{k_2}\ln\left(\frac{R_2}{r}\right) + \frac{1}{k_3}\ln\left(\frac{R_3}{R_2}\right) + \frac{1}{hR_3}\right], \tag{3-245b}$$

$$T_3(r) = T_\infty + \frac{Q_1 R_1^2}{2}\left[\frac{1}{k_3}\ln\left(\frac{R_3}{r}\right) + \frac{1}{hR_3}\right]. \tag{3-245c}$$

Using the following values:

$$\begin{aligned} R_1 &= 0.4 \text{ cm}, \\ R_2 &= 0.475 \text{ cm}, \\ R_3 &= 0.5 \text{ cm}, \\ Q_1 &= 5 \times 10^6 \text{ W}\cdot\text{m}^{-3}, \\ k_1 &= 0.25 \text{ W}\cdot\text{m}^{-1}\cdot\text{K}^{-1}, \\ k_2 &= 0.02 \text{ W}\cdot\text{m}^{-1}\cdot\text{K}^{-1}, \\ k_3 &= 20 \text{ W}\cdot\text{m}^{-1}\cdot\text{K}^{-1}, \\ h &= 30 \text{ W}\cdot\text{m}^{-2}\cdot\text{K}^{-1}, \\ T_\infty &= 300 \text{ K}, \end{aligned}$$

we solve the linear system and insert the numerical values into the coefficients. The temperature distribution is plotted in Fig. 3.16. The inner region has the fuel with a modest thermal conductivity and heat generation. This leads to a parabolic temperature shape with a small change in temperature change across the region. Moving

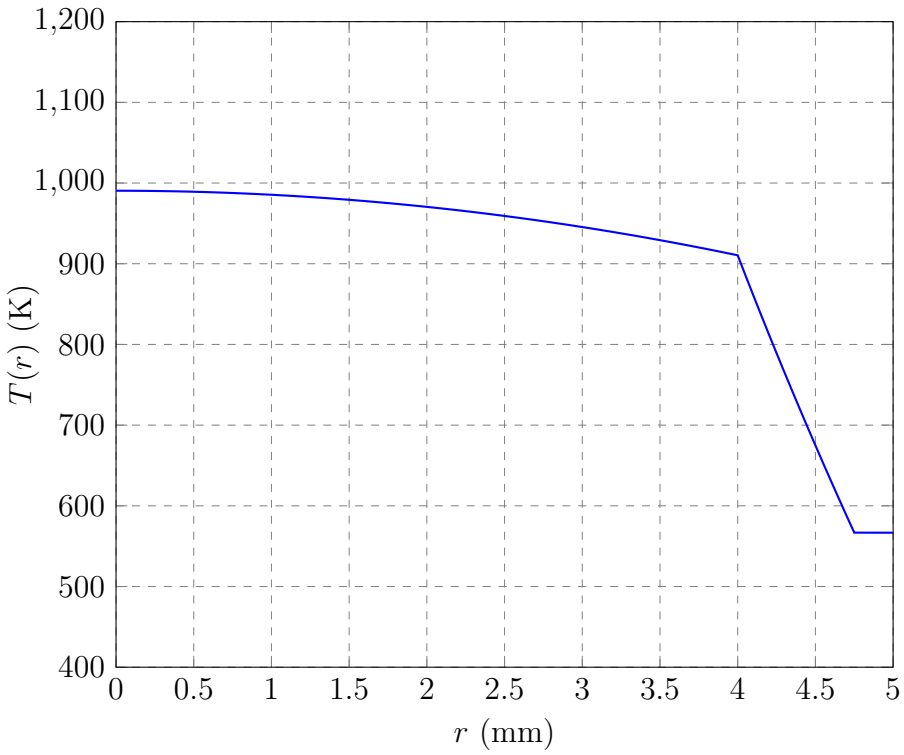

Figure 3.16: Temperature distribution in nuclear fuel rod.

outward, the next region is the gas-filled gap, with no heat generation and low thermal conductivity. Because the heat transfer is minimal in this region, the temperature field drops significantly. Finally, we have the zircaloy cladding, with a high thermal conductivity. Because of this, the temperature change in the clad is small.

Notice that in the solution that while the temperature field is continuous, its derivatives at the interfaces are not. Recall that the heat flux is continuous across interfaces, which is proportional to the derivative times the thermal conductivity, which varies quite significantly between the fuel, gap, and cladding.

3.8.e Example: Neutron Diffusion in a Planar Lattice

The theory of neutron diffusion is often used to model the neutron field distribution in nuclear reactor analysis. Neutron diffusion is an approximation to the more accurate model of neutron transport, but tends to work well in situations where the neutrons in the field are not too biased in a particular direction (cf., a particle beam). In this problem we consider the case where we model the spatial distribution of neutrons that have completely slowed down following moderation, i.e. thermal neutrons. The differential equation for neutron diffusion in 1-D Cartesian geometry with a single energy group is

$$-D\frac{d^2\phi}{dx^2} + \Sigma_a\phi(x) = Q(x). \tag{3-246}$$

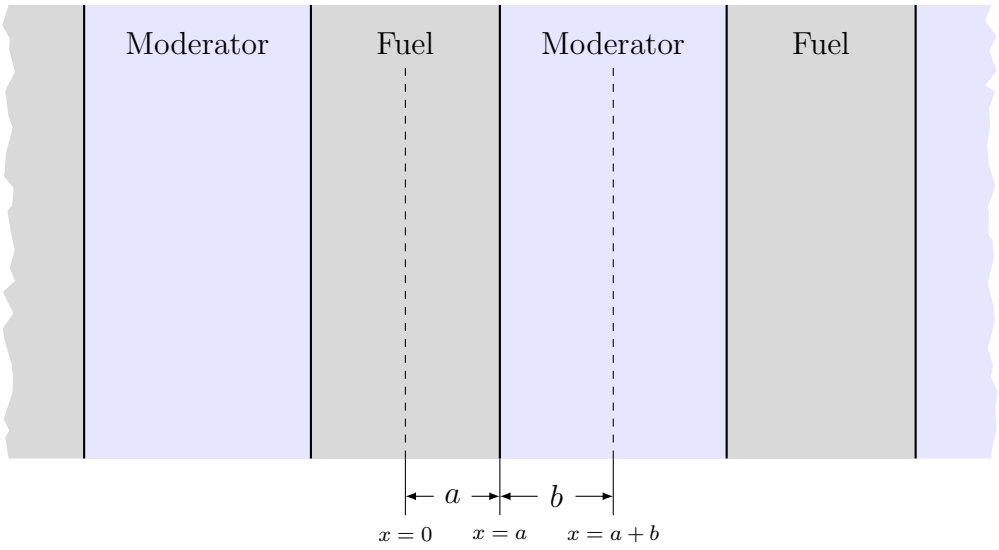

Figure 3.17: Illustration of a planar fuel-moderator lattice.

Here $\phi(x)$ is the scalar flux or path-length rate density of the neutrons (units of neutrons per area per time), D is the neutron diffusion coefficient (units of length), Σ_a is the macroscopic absorption cross section (units of per length), and $Q(x)$ is the inhomogeneous neutron source term (units of neutrons per volume per time). We often divide by $-D$ and write this as

$$\frac{d^2\phi}{dx^2} - \frac{1}{L^2}\phi(x) = -\frac{Q(x)}{D}, \tag{3-247}$$

where L is called the diffusion length (units of length),

$$L = \sqrt{\frac{D}{\Sigma_a}}. \tag{3-248}$$

We also define the net neutron current, or net flow rate of neutrons as

$$J(x) = -D\frac{d\phi}{dx}. \tag{3-249}$$

The problem geometry will be an infinite 1-D planar lattice of fuel and moderator regions with fuel as region 1 and moderator as region 2 (see Fig. 3.17) . Since the problem exhibits symmetry we can analyze a single unit cell with symmetry boundary conditions at the reflection planes, namely that the net flow rates of neutrons (neutron current) $J(x)$ are zero across the reflection boundaries. The interface conditions require that both the neutron scalar flux $\phi(x)$ and current $J(x)$ are continuous across the interface. The neutron source will be spatially constant in region 2, which is the moderator and represents where neutrons slow down into the thermal group. The

equations for the fuel and moderator are respectively,

$$\phi_1''(x) - \frac{1}{L_1^2}\phi_1(x) = 0, \quad 0 \le x \le a; \tag{3-250a}$$

$$\phi_2''(x) - \frac{1}{L_2^2}\phi_2(x) = -\frac{Q}{D_2}, \quad a \le x \le a+b. \tag{3-250b}$$

The boundary and interface conditions are

$$J_1(0) = 0, \tag{3-251a}$$

$$\phi_1(a) = \phi_2(a), \tag{3-251b}$$

$$J_1(a) = J_2(a), \tag{3-251c}$$

$$J_2(a+b) = 0. \tag{3-251d}$$

The solutions to the differential equations involve real exponentials that we elect to write in terms of the hyperbolic trigonometric functions. This is motivated by the fact that the $\sinh(0) = 0$, which we can use to simplify some of the coefficients. For the moderator region, we also introduce a translation by $a+b$ so we get the sinh term disappear when we apply the boundary condition. These are

$$\phi_1(x) = A_1 \sinh\left(\frac{x}{L_1}\right) + B_1 \cosh\left(\frac{x}{L_2}\right), \tag{3-252a}$$

$$\phi_2(x) = A_2 \sinh\left(\frac{a+b-x}{L_2}\right) + B_2 \cosh\left(\frac{a+b-x}{L_2}\right) + \frac{QL_2^2}{D_2}. \tag{3-252b}$$

The net currents are

$$J_1(x) = -\frac{D_1 A_1}{L_1}\cosh\left(\frac{x}{L_1}\right) - \frac{D_1 B_1}{L_1}\sinh\left(\frac{x}{L_2}\right), \tag{3-253a}$$

$$J_2(x) = \frac{D_2 A_2}{L_2}\cosh\left(\frac{a+b-x}{L_2}\right) + \frac{D_2 B_2}{L_2}\sinh\left(\frac{a+b-x}{L_2}\right). \tag{3-253b}$$

Applying the symmetry conditions $J_1(0) = 0$ and $J_2(a+b) = 0$, gives the equations

$$-\frac{D_1 A_1}{L_1} = 0, \tag{3-254a}$$

$$\frac{D_2 A_2}{L_2} = 0, \tag{3-254b}$$

which implies that $A_1 = A_2 = 0$. Applying the interface conditions, we obtain

$$B_1 \cosh\left(\frac{a}{L_1}\right) = B_2 \cosh\left(\frac{b}{L_2}\right) + \frac{QL_2^2}{D_2}, \tag{3-254c}$$

$$-\frac{D_1 B_1}{L_1}\sinh\left(\frac{a}{L_1}\right) = \frac{D_2 B_2}{L_2}\sinh\left(\frac{b}{L_2}\right). \tag{3-254d}$$

We can then write these two equations as a linear system in terms of the coefficients

$$\begin{bmatrix} \cosh\left(\frac{a}{L_1}\right) & -\cosh\left(\frac{b}{L_2}\right) \\ \frac{D_1}{L_1}\sinh\left(\frac{a}{L_1}\right) & \frac{D_2}{L_2}\sinh\left(\frac{b}{L_2}\right) \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} \frac{QL_2^2}{D_2} \\ 0 \end{bmatrix}. \tag{3-255}$$

Solving this system for the coefficients yields

$$B_1 = \frac{QL_2^2}{D_2}\left[\frac{\frac{D_2}{L_2}\sinh\left(\frac{b}{L_2}\right)}{\frac{D_2}{L_2}\cosh\left(\frac{a}{L_1}\right)\sinh\left(\frac{b}{L_2}\right) + \frac{D_1}{L_1}\sinh\left(\frac{a}{L_1}\right)\cosh\left(\frac{b}{L_2}\right)}\right], \tag{3-256a}$$

$$B_2 = -\frac{QL_2^2}{D_2}\left[\frac{\frac{D_1}{L_1}\sinh\left(\frac{a}{L_1}\right)}{\frac{D_2}{L_2}\cosh\left(\frac{a}{L_1}\right)\sinh\left(\frac{b}{L_2}\right) + \frac{D_1}{L_1}\sinh\left(\frac{a}{L_1}\right)\cosh\left(\frac{b}{L_2}\right)}\right]. \tag{3-256b}$$

Therefore the solution is

$$\phi_1(x) = \frac{QL_2^2}{D_2}\left[\frac{\frac{D_2}{L_2}\sinh\left(\frac{b}{L_2}\right)\cosh\left(\frac{x}{L_2}\right)}{\frac{D_2}{L_2}\cosh\left(\frac{a}{L_1}\right)\sinh\left(\frac{b}{L_2}\right) + \frac{D_1}{L_1}\sinh\left(\frac{a}{L_1}\right)\cosh\left(\frac{b}{L_2}\right)}\right], \tag{3-257a}$$

$$\phi_2(x) = \frac{QL_2^2}{D_2}\left[1 - \frac{\frac{D_1}{L_1}\sinh\left(\frac{a}{L_1}\right)\cosh\left(\frac{a+b-x}{L_2}\right)}{\frac{D_2}{L_2}\cosh\left(\frac{a}{L_1}\right)\sinh\left(\frac{b}{L_2}\right) + \frac{D_1}{L_1}\sinh\left(\frac{a}{L_1}\right)\cosh\left(\frac{b}{L_2}\right)}\right]. \tag{3-257b}$$

The following numbers are representative for a uranium dioxide fuel and light water moderator:

$$\begin{aligned} a &= 0.5 \text{ cm}, \\ b &= 0.9 \text{ cm}, \\ Q &= 10^{10} \text{ neutrons}\cdot\text{cm}^{-3}\cdot\text{s}^{-1}, \\ D_1 &= 0.615 \text{ cm}, \\ D_2 &= 0.144 \text{ cm}, \\ L_1 &= 1.908 \text{ cm}, \\ L_2 &= 2.685 \text{ cm}. \end{aligned}$$

Using those numbers the scalar flux is plotted in Fig. 3.18. The thermal neutron scalar flux is highest in the moderator region on the right, as this is where the neutrons thermalize. The scalar flux falls as the field gets closer to the fuel as neutrons that enter the fuel tend to be absorbed (some of these absorptions cause fission) and therefore fewer neutrons exit the fuel than exit. As one may expect, the minimum neutron scalar flux is at the center of the fuel.

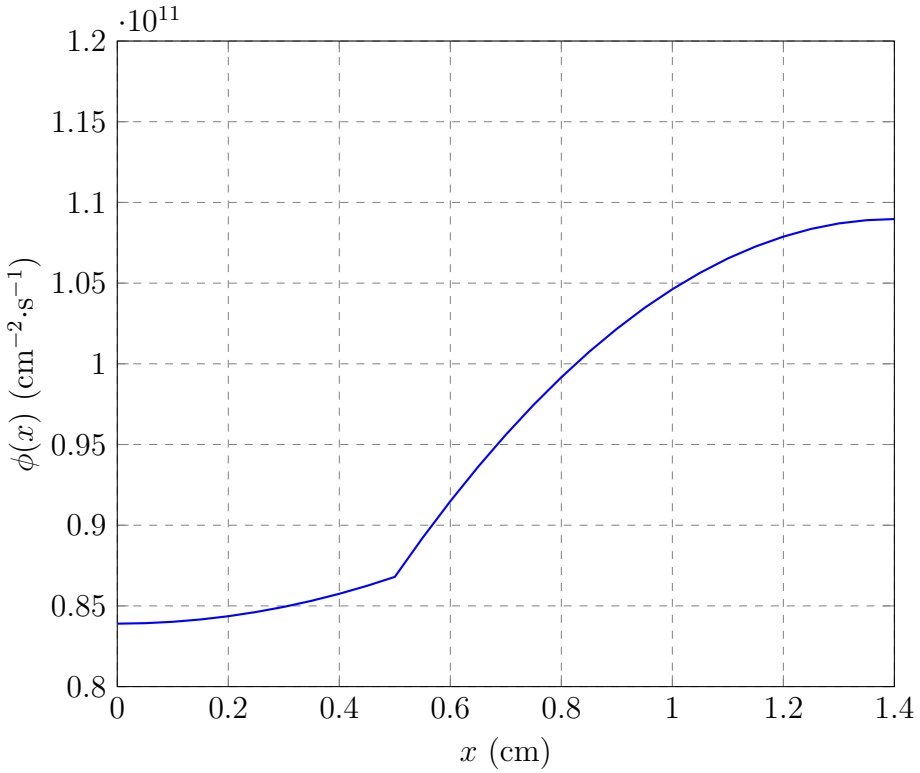

Figure 3.18: Neutron scalar flux (path-length rate density) in a planar fuel-moderator lattice.

3.8.f Example: Quantum Particle in a Finite Potential Well

In quantum mechanics, particles are described by wavefunctions using the Schrödinger equation. The steady-state, 1-D Cartesian form is

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} + V(x)\psi(x) = E\psi(x). \tag{3-258}$$

Here $\psi(x)$ is the wavefunction of the particle, $\hbar = h/(2\pi)$ or the reduced Planck's constant, m is the mass of the particle, $V(x)$ is a prescribed potential energy field, and E is the energy of the particle. The wavefunction $\psi(x)$ is generally a complex function that is related to the probability of a particle being at a certain point in space. The probability per unit length (probability density) is given by

$$f(x) = \psi(x)\psi^*(x), \tag{3-259}$$

where $\psi^*(x)$ is the complex conjugate of $\psi(x)$. Because of the nature of probabilities, the wavefunction has the normalization

$$\int_{-\infty}^{\infty} \psi(x)\psi^*(x)dx = 1. \tag{3-260}$$

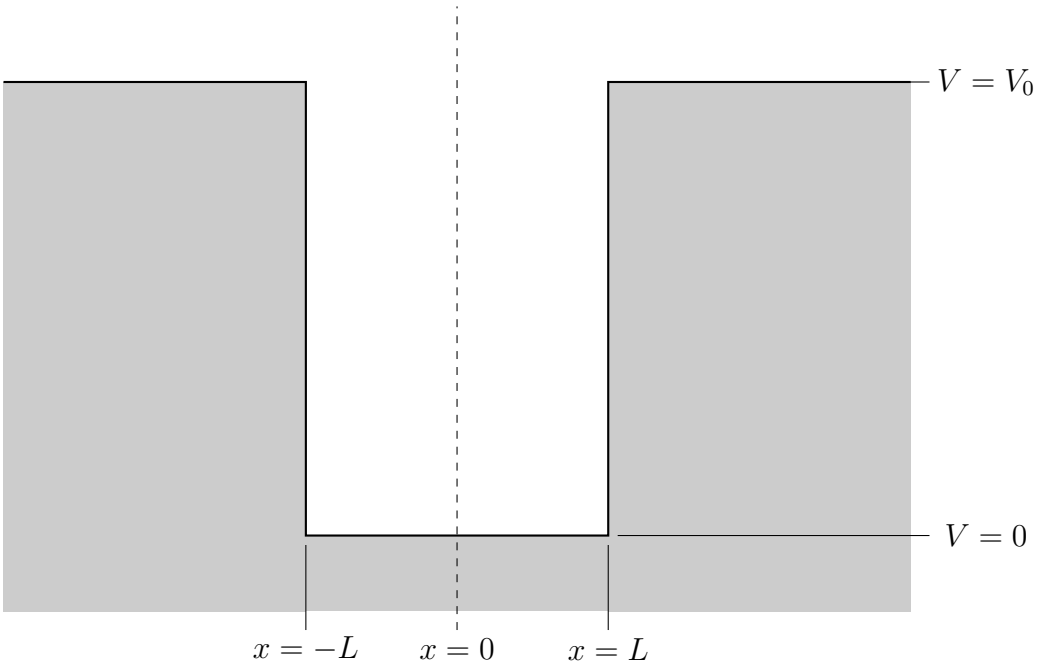

Figure 3.19: Illustration of a finite potential (square) well.

The wavefunction and its first derivative must be continuous at interfaces and must be finite everywhere.

Here we consider the case of a symmetric finite potential well with width $2L$ (see Fig. 3.19) with the potential function

$$V(x) = \begin{cases} V_0 & x < -L \\ 0 & -L \leq x \leq L \\ V_0 & x > L \end{cases} . \tag{3-261}$$

Because of symmetry we can simplify this problem and only solve the right half of the problem $x \geq 0$ with region 1 being $0 \leq x \leq L$ and region 2 being $x > L$. The differential equations become

$$\frac{d^2\psi_1}{dx^2} + \frac{2mE}{\hbar^2}\psi_1(x) = 0, \tag{3-262a}$$

$$\frac{d^2\psi_2}{dx^2} - \frac{2m(V_0 - E)}{\hbar^2}\psi_2(x) = 0. \tag{3-262b}$$

As we can see from the second equation, that we will get different behavior if $E < V_0$ and $E > V_0$. Since the case with $E < V_0$ is more interesting, let us consider that case. For this, we define the following coefficients:

$$k_1^2 = \frac{2mE}{\hbar^2}, \tag{3-263a}$$

$$k_2^2 = \frac{2m(V_0 - E)}{\hbar^2}. \tag{3-263b}$$

The equations are therefore

$$\frac{d^2\psi_1}{dx^2} + k_1^2\psi_1(x) = 0, \tag{3-264a}$$
$$\frac{d^2\psi_2}{dx^2} - k_2^2\psi_2(x) = 0, \tag{3-264b}$$

which have the solutions

$$\psi_1(x) = A_1 \sin(k_1 x) + B_1 \cos(k_1 x), \tag{3-265a}$$
$$\psi_2(x) = A_2 e^{k_2 x} + B_2 e^{-k_2 x}. \tag{3-265b}$$

Since we require that the wavefunction be finite everywhere, we require that $A_2 = 0$. Therefore,

$$\psi_1(x) = A_1 \sin(k_1 x) + B_1 \cos(k_1 x), \tag{3-266a}$$
$$\psi_2(x) = B_2 e^{-k_2 x}. \tag{3-266b}$$

The derivatives are

$$\frac{d\psi_1}{dx} = A_1 k_1 \cos(k_1 x) - B_1 k_1 \sin(k_1 x), \tag{3-267a}$$
$$\frac{d\psi_2}{dx} = -B_2 k_2 e^{-k_2 x}. \tag{3-267b}$$

To help with the boundary and interface conditions, we make an observation that there are two types of symmetry that the wavefunction may possess. First, the wavefunction may be reflected about the y-axis, which are the *even parity* solutions denoted by $\psi^+(x)$. Second, since the wavefunction may be both positive and negative, we must also permit *odd parity* solutions reflected about the origin, which we denote by $\psi^-(x)$. Because the Schrödinger equation is linear, we can apply the superposition principle and solve for the even and odd solutions separately and then add them together,

$$\psi(x) = \psi^+(x) + \psi^-(x). \tag{3-268}$$

The even parity solutions must be reflected about the y-axis and since its derivative must be smooth, we demand that

$$\left.\frac{d\psi_1^+}{dx}\right|_{x=0} = 0. \tag{3-269a}$$

For the odd-parity solutions, we require symmetry with respect to the origin so we enforce that

$$\psi_1^-(0) = 0. \tag{3-269b}$$

Both the even and odd parity solutions and their respective first derivatives must be continuous across the interface at $x = L$:

$$\psi_1^{\pm}(L) = \psi_2^{\pm}(L), \tag{3-269c}$$

$$\left.\frac{d\psi_1^{\pm}}{dx}\right|_{x=L} = \left.\frac{d\psi_2^{\pm}}{dx}\right|_{x=L} \tag{3-269d}$$

Let us begin by solving for the even parity wavefunction. Applying the boundary condition that the first-derivative at $x = 0$ is zero, gives

$$A_1 k_1 = 0, \tag{3-270}$$

meaning that $A_1 = 0$. The interface conditions are therefore

$$B_1 \cos(k_1 L) = B_2 e^{-k_2 L}, \tag{3-271a}$$

$$-B_1 k_1 \sin(k_1 L) = -B_2 k_2 e^{-k_2 L}. \tag{3-271b}$$

This yields the following linear system:

$$\begin{bmatrix} \cos(k_1 L) & -e^{-k_2 L} \\ -k_1 \sin(k_1 L) & k_2 e^{-k_2 L} \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{3-272}$$

For this system to have a solution, either the solution vector must be zero or the determinant of the matrix must be zero. The former case is of no interest, because it describes the case where there is no particle. The latter case yields the following determinant:

$$\begin{vmatrix} \cos(k_1 L) & -e^{-k_2 L} \\ -k_1 \sin(k_1 L) & k_2 e^{-k_2 L} \end{vmatrix} = k_2 e^{-k_2 L} \cos(k_1 L) - k_1 e^{-k_2 L} \sin(k_1 L) = 0. \tag{3-273}$$

We obtain the relationship:

$$\frac{k_2}{k_1} - \tan(k_1 L) = 0. \tag{3-274}$$

Unfortunately, this equation cannot be solved directly and solutions must be obtained from an iterative root finding algorithm.

Repeating this process for the odd parity wavefunction, we apply the boundary condition requiring the wavefunction to go to zero at $x = 0$ yielding

$$B_1 = 0. \tag{3-275}$$

The interface conditions are then

$$A_1 \sin(k_1 L) = B_2 e^{-k_2 L}, \tag{3-276a}$$

$$A_1 k_1 \cos(k_1 L) = -B_2 k_2 e^{-k_2 L}, \tag{3-276b}$$

which has the linear system

$$\begin{bmatrix} \sin(k_1 L) & -e^{-k_2 L} \\ k_1 \cos(k_1 L) & k_2 e^{-k_2 L} \end{bmatrix} \begin{bmatrix} A_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{3-277}$$

As before, we find where the determinant is zero,

$$\begin{vmatrix} \sin(k_1 L) & -e^{-k_2 L} \\ k_1 \cos(k_1 L) & k_2 e^{-k_2 L} \end{vmatrix} = k_2 e^{-k_2 L} \sin(k_1 L) + k_1 e^{-k_2 L} \cos(k_1 L) = 0. \tag{3-278}$$

This gives the solution

$$\frac{k_2}{k_1} + \cot(k_1 L) = 0, \tag{3-279}$$

which also must be solved numerically.

Putting the resulting transcendental expressions in terms of energy gives

$$g_1(E) = \sqrt{\frac{V_0 - E}{E}} - \tan\left(\frac{\sqrt{2mE}L}{\hbar}\right) = 0, \tag{3-280a}$$

$$g_2(E) = \sqrt{\frac{V_0 - E}{E}} + \cot\left(\frac{\sqrt{2mE}L}{\hbar}\right) = 0. \tag{3-280b}$$

At this point, let us introduce some numerical values corresponding to an electron in a 20 eV potential well with a width of 0.4 nm:

$$\begin{aligned} \hbar &= 1.055 \times 10^{-34} \text{ J·s}, \\ m &= 9.109 \times 10^{-31} \text{ kg}, \\ L &= 2 \times 10^{-10} \text{ m}, \\ V_0 &= 20 \text{ eV} = 3.204 \times 10^{-18} \text{ J}. \end{aligned}$$

The functions $g_1(E)$ and $g_2(E)$ are plotted in the domain $0 \leq E \leq V_0$ in Fig. 3.20. As we can see from the plot the even (symmetric) case with the solid blue line cross the x-axis twice and the odd (antisymmetric) case with the dashed red line crosses the x-axis once. The roots of these functions correspond to the discrete (or quantized) energy states that satisfy the Schrödinger equation and its associated boundary conditions. Using a numerical root finder such as Newton-Raphson, the resulting energies are

$$E_1 = 1.58 \text{ eV}, \quad E_2 = 6.20 \text{ eV}, \quad E_3 = 13.41 \text{ eV}.$$

These are the eigenvalues of the Schrödinger equation corresponding to this boundary value problem.

To obtain the normalized wavefunctions, we apply the normalization condition,

$$\int_{-\infty}^{\infty} \psi(x)\psi^*(x)dx,$$

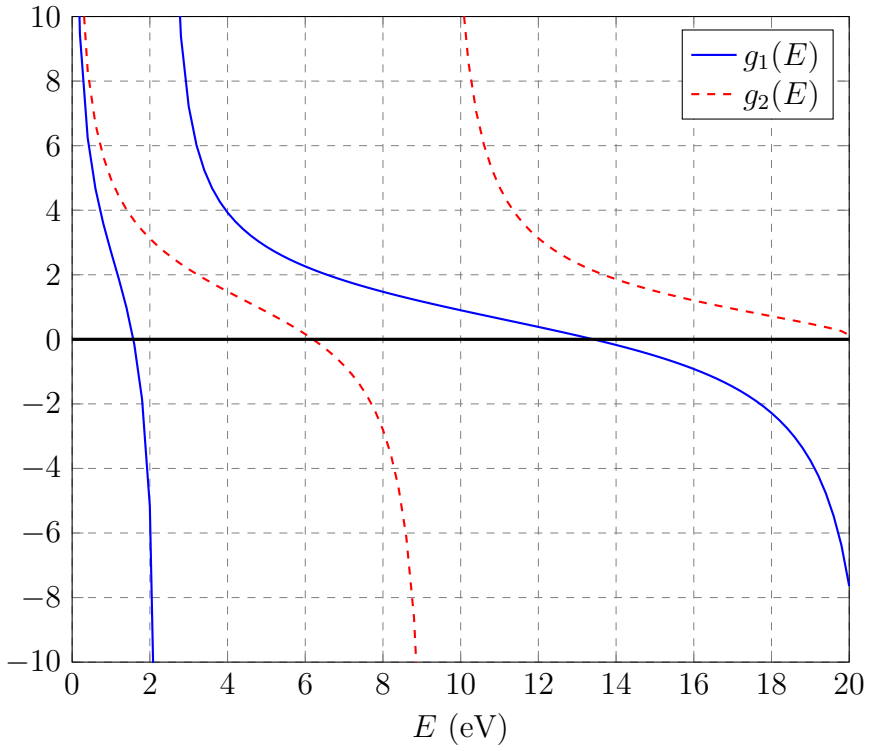

Figure 3.20: Plot of transcendental equations for finite potential well.

using each of the eigenvalues. Since we are applying symmetry, we know that the particle has a probability of being in the $x > 0$ range equal to $^1/_2$. For the even parity (symmetric) cases, we can apply the continuity of the wavefunction to relate the coefficients

$$B_2 = B_1 e^{k_2 L} \cos(k_1 L); \tag{3-281a}$$

and in the odd parity (antisymmetric) cases, these are

$$B_2 = A_1 e^{k_2 L} \sin(k_1 L); \tag{3-281b}$$

The corresponding coefficient that remains for the even and odd cases respectively can be found by evaluating

$$B_1^2 = \frac{1}{2}\left[\int_0^L \cos^2(k_1 x)dx + \cos^2(k_1 L)\int_L^\infty e^{-2k_2(x-L)}dx\right]^{-1}, \tag{3-282a}$$

$$A_1^2 = \frac{1}{2}\left[\int_0^L \sin^2(k_1 x)dx + \sin^2(k_1 L)\int_L^\infty e^{-2k_2(x-L)}dx\right]^{-1}. \tag{3-282b}$$

Evaluating the integrals and solving for the coefficients gives

$$B_1 = \frac{1}{\sqrt{2}}\left[\frac{L}{2} + \frac{1}{4k_1}\sin(2k_1L) + \frac{1}{2k_2}\cos^2(k_1L)\left(1 - e^{-2k_2L}\right)\right]^{-1/2}, \tag{3-283a}$$

$$A_1 = \frac{1}{\sqrt{2}}\left[\frac{L}{2} - \frac{1}{4k_1}\sin(2k_1L) + \frac{1}{2k_2}\sin^2(k_1L)\left(1 - e^{-2k_2L}\right)\right]^{-1/2}. \tag{3-283b}$$

The wavefunctions with the numerical values of the coefficients are plotted in Fig. 3.21. The solid blue curve corresponds to the lowest energy state, which is referred to as the ground state. The wavefunction for the ground state is strictly positive and symmetric about the y axis. The dashed red curve corresponds to the next energy and is symmetric about the origin or antisymmetric. The third wavefunction is the dash-dot orange curve and is symmetric. Note that as the energy is higher, the wavefunction tends to be more spread out, with a higher probability that the particle will find itself on top of the well, which is forbidden from classical physics. Note that the finite potential well only has three solutions (or eigenvalues) that satisfy the restriction of $E < V_0$. This means that the finite potential well only has three allowed quantized energy states for which particles can be bound in the well.

While we did not analyze the case where $E > V_0$, this would lead to an oscillatory solution to the Schrödinger equation within region 1. This solution corresponds to a free particle, as it has sufficient energy to not be trapped within the potential well.

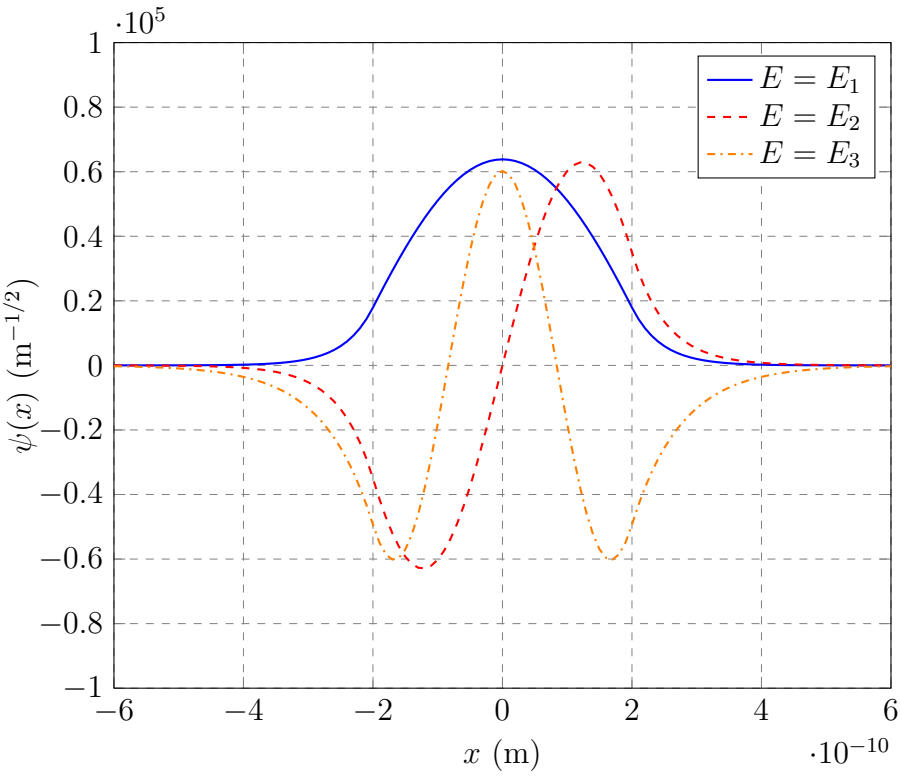

Figure 3.21: Plot of wavefunctions for finite potential well.

Note that the free particle does not have a normalizable wavefunction.

3.9 Finite Difference Method

In many practical applications, the solution of boundary value problems cannot be obtained analytically, or at least not in a practical manner. Instead, computational methods are used to approximate the solutions of these boundary value problems. There are a handful of techniques that have been developed over the past several decades. The simplest approach among these is called the *finite difference method*. This method approximates the field, which we shall denote generically as $u(x)$, on a set of spatial grid points x_i where $i = 0, \ldots, N$. This approach maps the differential equation onto this spatial grid and then solves a system of (usually tridiagonal) linear equations to obtain the solution.

3.9.a Approximations of Derivatives

In our previous discussion on numerical solutions of first-order differential equations, we stated based on the definition of the derivative that it can be approximated as

$$\frac{dy}{dx} \approx \frac{y_{i+1} - y_i}{\Delta x}. \tag{3-284}$$

This is called the *forward-difference approximation*. This definition made sense in the context of initial value problems where the derivatives are with respect to time since time always moves forward. In boundary value problems, the field can vary spatially in either direction. Here we show how we can quantify the accuracy of this approximation in boundary value problems.

The error can be written as the difference of the actual derivative and the approximation:

$$\Delta x \left(\frac{dy}{dx}\right)_i - (y_{i+1} - y_i) = R\Delta x. \tag{3-285}$$

Here we moved Δx to the numerator and R is the error or residual. Next, we can perform a Taylor series expansion on the $y_{i+i} = y(x_i + \Delta x)$ about x_i. This is

$$\begin{aligned} y_{i+1} &= y_i + (x + \Delta x - x)\left(\frac{dy}{dx}\right)_i + \frac{1}{2}(x + \Delta x - x)^2 \left(\frac{d^2y}{dx^2}\right)_i \\ &\quad + \frac{1}{6}(x + \Delta x - x)^3 \left(\frac{d^3y}{dx^3}\right)_i + \ldots \\ &= y_i + (\Delta x)\left(\frac{dy}{dx}\right)_i + \frac{(\Delta x)^2}{2}\left(\frac{d^2y}{dx^2}\right)_i + \frac{(\Delta x)^3}{6}\left(\frac{d^3y}{dx^3}\right)_i + \ldots \end{aligned} \tag{3-286}$$

Inserting this into the above expression and grouping terms gives

$$\begin{aligned} &\Delta x \left(\frac{dy}{dx}\right)_i + y_i - \left(y_i + (\Delta x)\left(\frac{dy}{dx}\right)_i + \frac{(\Delta x)^2}{2}\left(\frac{d^2y}{dx^2}\right)_i + \ldots\right) = R\Delta x, \\ &- \frac{(\Delta x)^2}{2}\left(\frac{d^2y}{dx^2}\right)_i - \frac{(\Delta x)^3}{6}\left(\frac{d^3y}{dx^3}\right)_i + \ldots = R\Delta x. \end{aligned} \tag{3-287}$$

Therefore, we see that the residual or error is

$$R = -\frac{(\Delta x)}{2}\left(\frac{d^2y}{dx^2}\right)_i - \frac{(\Delta x)^2}{6}\left(\frac{d^3y}{dx^3}\right)_i + \ldots = \mathcal{O}(\Delta x). \tag{3-288}$$

Here we introduced the notation $\mathcal{O}(\Delta x)$ to indicate the terms are of order Δx or higher powers. Therefore, we expect the error to diminish as Δx as the spatial grid becomes small.

We could also define the *backward-difference approximation* as:

$$\frac{dy}{dx} \approx \frac{y_i - y_{i-1}}{\Delta x}. \tag{3-289}$$

This instead takes the slope using the point to the left as opposed to the right. We could repeat the analysis above and see that the error is also $\mathcal{O}(\Delta x)$.

As we saw with forward and backward Euler versus improved Euler, a natural question is whether one can do better than error being $\mathcal{O}(\Delta x)$. It turns out we can. Suppose we wish to approximate the first derivative by using three points, y_{i-1}, y_i, and y_{i+1}, the central point where the derivative is being taken and the two neighboring grid points. We can write this using the same error formula but in terms of unknown coefficients:

$$\Delta x\left(\frac{dy}{dx}\right)_i - (ay_{i-1} + by_i + cy_{i+1}) = R\Delta x. \tag{3-290}$$

The Taylor series expansions for the y_{i-1} and y_{i+1} terms are

$$\begin{aligned} y_{i-1} &= y_i - (\Delta x)\left(\frac{du}{dx}\right)_i + \frac{(\Delta x)^2}{2}\left(\frac{d^2y}{dx^2}\right)_i \\ &\quad - \frac{(\Delta x)^3}{6}\left(\frac{d^3y}{dx^3}\right)_i + \frac{(\Delta x)^4}{24}\left(\frac{d^4y}{dx^4}\right)_i + \ldots \end{aligned} \tag{3-291a}$$

$$\begin{aligned} y_{i+1} &= y_i + (\Delta x)\left(\frac{dy}{dx}\right)_i + \frac{(\Delta x)^2}{2}\left(\frac{d^2y}{dx^2}\right)_i \\ &\quad + \frac{(\Delta x)^3}{6}\left(\frac{d^3y}{dx^3}\right)_i + \frac{(\Delta x)^4}{24}\left(\frac{d^4y}{dx^4}\right)_i + \ldots \end{aligned} \tag{3-291b}$$

Inserting these into the expression and grouping terms gives

$$\begin{aligned} &-(a+b+c)y_i + (1+a-c)\Delta x\left(\frac{dy}{dx}\right)_i - (a+c)\frac{(\Delta x)^2}{2}\left(\frac{d^2y}{dx^2}\right)_i \\ &+ (a-c)\frac{(\Delta x)^3}{6}\left(\frac{d^3y}{dx^3}\right)_i - (a+c)\frac{(\Delta x)^4}{24}\left(\frac{d^4y}{dx^4}\right)_i \ldots = R\Delta x. \end{aligned} \tag{3-292}$$

Since we have three coefficients, we can now attempt to make the $(\Delta x)^2$ terms go to zero in the hopes that we can get an approximation that is $\mathcal{O}(\Delta x^2)$ in error. We equate the first three terms then with zero and get the system of equations:

$$a + b + c = 0, \tag{3-293a}$$

$$a - c = -1, \tag{3-293b}$$

$$a + c = 0. \tag{3-293c}$$

These can be solved using methods of linear algebra to obtain

$$a = -\frac{1}{2}, \quad b = 0, \quad c = \frac{1}{2}.$$

Therefore, our approximation of the first derivative becomes

$$\frac{dy}{dx} \approx \frac{y_{i+1} - y_{i-1}}{2\Delta x}. \tag{3-294}$$

This is referred to as the *central-difference approximation.* Note that this is an average of the forward and backward difference. It is analogous to improved Euler in the sense that we average the two together to get an improved approximation.

The central-difference approximation is guaranteed to be *at least* $\mathcal{O}(\Delta x^2)$. Here it is at least because it is possible that fortuitously the Δx^3 terms or higher could go to zero as well. We check this term and see that unfortunately this is not the case,

$$a - c = -\frac{1}{2} - \frac{1}{2} = -1 \neq 0.$$

Therefore, the error term $R\Delta x$ when dividing by Δx gives terms that are

$$R = -\frac{(\Delta x)^2}{6}\left(\frac{d^3y}{dx^3}\right)_i + \ldots = \mathcal{O}(\Delta x^2). \tag{3-295}$$

It is possible to go further and use more points and have increasingly accurate approximations of the first derivative. However, this comes at a cost that more grid points are needed to evaluate the equation and makes the solution more expensive. There are also penalties regarding stability of the solution for Δx that are too large. Therefore, the most common approximation for the first derivative is the central-difference approximation.

It is also possible to approximate higher derivatives. The nth derivative requires at least $n+1$ points to approximate consistently. For example, the first derivative finds the slope of the line, which requires $1+1=2$ points. The second derivative measures curvature and requires at least three points. Let us attempt to approximate the second derivative using the same set of three points we used in the central difference approximation:

$$(\Delta x)^2\left(\frac{d^2y}{dx^2}\right)_i - (ay_{i-1} + by_i + cy_{i+1}) = R(\Delta x)^2. \tag{3-296}$$

Here we moved a factor of Δx^2 to the numerator. As before, we insert the Taylor series expansions and set the first three order terms in Δx to zero. This gives the equation:

$$\begin{aligned} &-(a+b+c)y_i + (a-c)\Delta x\left(\frac{dy}{dx}\right)_i - (1-a+c)\frac{(\Delta x)^2}{2}\left(\frac{d^2y}{dx^2}\right)_i \\ &+(a-c)\frac{(\Delta x)^3}{6}\left(\frac{d^3y}{dx^3}\right)_i - (a+c)\frac{(\Delta x)^4}{24}\left(\frac{d^4y}{dx^4}\right)_i \ldots = R\Delta x. \end{aligned} \tag{3-297}$$

We can now form the linear system:

$$a + b + c = 0, \tag{3-298a}$$
$$a - c = 0, \tag{3-298b}$$
$$a + c = 2. \tag{3-298c}$$

These equations are slightly different because instead of the inhomogeneous term appearing on the first derivative, it is now on the second derivative or third equation. Note the right-hand side is 2 because of the factor of one-half from the Taylor series expansion. Solving this system of equations gives the values

$$a = 1, \quad b = -2, \quad c = 1.$$

This means the second derivative can be approximated by

$$\frac{d^2y}{dx^2} \approx \frac{y_{i-1} - 2y_i + y_{i+1}}{(\Delta x)^2}. \tag{3-299}$$

To analyze the error, we inspect the third derivative term:

$$(a - c)\frac{(\Delta x)^3}{6}\left(\frac{d^3y}{dx^3}\right)_i = (1 - 1)\frac{(\Delta x)^3}{6}\left(\frac{d^3y}{dx^3}\right)_i = 0.$$

This means the Δx^3 terms vanish automatically. Looking at the fourth derivative term, we have

$$-(a + c)\frac{(\Delta x)^4}{24}\left(\frac{d^4y}{dx^4}\right)_i = -2\frac{(\Delta x)^4}{24}\left(\frac{d^4y}{dx^4}\right)_i \neq 0.$$

This means that our error term can be found by taking the residual and dividing by Δx^2:

$$R = -\frac{(\Delta x)^2}{12}\left(\frac{d^4u}{dx^4}\right)_i + \ldots = \mathcal{O}(\Delta x^2). \tag{3-300}$$

3.9.b Reaction-Diffusion Equation

A common boundary value problem that arises in engineering applications is the reaction-diffusion equation:

$$-\frac{d}{dx}\left(D(x)\frac{d\phi}{dx}\right) + \lambda(x)\phi(x) = Q(x), \tag{3-301}$$

where $\phi(x)$ is some spatially dependent physical quantity and $Q(x)$ is an internal source. We define the flow rate across some interface as

$$J(x) = -D(x)\frac{d\phi}{dx}. \tag{3-302}$$

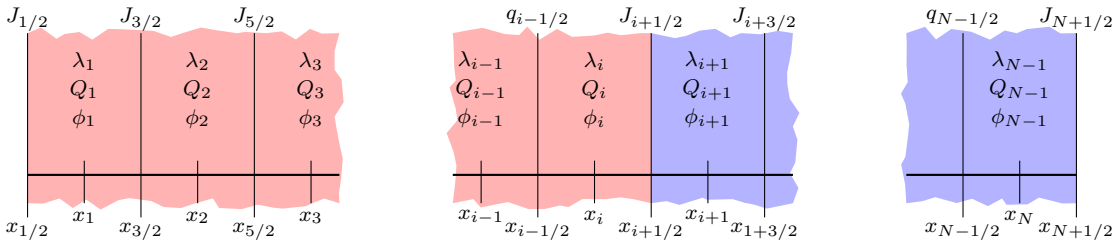

Figure 3.22: Grid or stencil for the cell-centered finite differencing scheme for the reaction-diffusion equation.

This equation arises in applications including the analysis of viscous fluids in laminar flow, structural mechanics, heat conduction, and neutron diffusion. Here $D(x)$ is a diffusion coefficient and $\lambda(x)$ is a reaction coefficient. The former describes the local rate that the physical quantity of the field spreads out, and the latter describes the local rate that the field is created or destroyed in proportion to itself. The right-hand side has an inhomogeneous term for the field being created (e.g., heat generation from nuclear reactions or the creation of neutrons from fission).

To apply the finite difference approximation we need to define a spatial grid or stencil. There are many strategies employed and here we present a scheme that is intended to conserve physical quantities within each spatial cell, i.e. the rate of production of some quantity within a cell plus the rate that the quantity flows in equals the rate that it is destroyed within the cell plus the rate it flows out. This sounds rather obvious and we had not had to worry about it because the differential equations (either neutron transport or diffusion) naturally support particle balance by virtue of how they were derived. The problem is that the process of mapping these differential equations onto a spatial discretization, which converts a calculus problem into algebra one, may or may not preserve the balance of particles depending on how this is done.

To this end, we introduce the *cell-centered* spatial discretization, which is depicted in Fig. 3.22. The cell-centered spatial discretization breaks the problem into N spatial regions indexed from $i = 1, 2, \ldots, N$ (note that many programming language index from zero, so implementations need to subtract one from each index) each having a width $\Delta_i = x_{i+1} - x_i$. The point x_i is the center of the ith spatial zone. This implies that $x_{1/2} = 0$ is the left edge of the slab and $x_{N+1/2}$ is the right edge. Material properties and sources are taken to be constant within each spatial zone. Ideally, we choose the spatial discretization to line up with the geometric regions of the problem. If we cannot do this, or do not wish to for some reason, then we need to find some average value of the material properties.

Discretion of the Continuity Equation

To derive the discretization scheme, we first insert the definition of the flow rate into the reaction-diffusion equation to obtain a *continuity equation*:

$$\frac{dJ}{dx} + \lambda(x)\phi(x) = Q(x). \tag{3-303}$$

A continuity equation states that the local net outflow rate dJ/dx (proportional to the spatial gradient of a physical quantity) plus the internal loss rate equals the gain rate.

We then integrate over a spatial region:

$$\begin{aligned}\int_{x_{i-1/2}}^{x_{i+1/2}} \frac{dJ}{dx} + \lambda(x)\phi(x)dx &= \int_{x_{i-1/2}}^{x_{i+1/2}} Q(x)dx, \\ J_{i+1/2} - J_{i-1/2} + \int_{x_{i-1/2}}^{x_{i+1/2}} \lambda(x)\phi(x)dx &= \int_{x_{i-1/2}}^{x_{i+1/2}} Q(x)dx.\end{aligned} \tag{3-304}$$

Since the reaction coefficient is assumed to be spatially constant over the region and equal to λ_i, we can pull it out of the integral. Next, we define the cell-average quantity and source to be at the center of the cell,

$$\phi_i = \frac{1}{\Delta_i}\int_{x_{i-1/2}}^{x_{i+1/2}} \phi(x)dx, \tag{3-305a}$$

$$Q_i = \frac{1}{\Delta_i}\int_{x_{i-1/2}}^{x_{i+1/2}} Q(x)dx. \tag{3-305b}$$

Inserting these cell-averaged quantities and sources we get

$$J_{i+1/2} - J_{i-1/2} + \lambda_i\Delta_i\phi_i = Q_i\Delta_i. \tag{3-306}$$

The next task is to figure out the flow rates at the cell-edges.

Interface Conditions for Internal Regions

We handle the flow rates differently if the cell is an internal region versus a boundary region. First we explore the internal regions, which require us to use the interface condition that the net current must be continuous. Beginning on the left-hand side of a cell centered at x_i, we apply both the forward and backward differencing schemes to Eq. (3-302) to relate the flow rates J on each edge with the physical quantities u,

$$J_{i-1/2} = -D_i\frac{\phi_i - \phi_{i-i/2}}{\Delta_i/2}, \tag{3-307a}$$

$$J_{i-1/2} = -D_{i-1}\frac{\phi_{i-1/2} - \phi_{i-1}}{\Delta_{i-1}/2}, \tag{3-307b}$$

for the element centered about x_i and x_{i-1} respectively. Note the factor of one-half on the width of the cell is because we take the finite difference to approximate the derivative from the center to the edge, which is half the width.

Both of these equations involve the quantity of interest at the cell edge $\phi(x_{i-1/2}) = \phi_{i-1/2}$, which is not part of the discretized continuity equation. However, since we have two equations in terms of the cell-edge quantity and the flow rate at that edge, we can eliminate this extra variable by equating the two equations and solving for $\phi_{i-1/2}$. After a bit of algebra, we get

$$\phi_{i-1/2} = \left[\frac{D_{i-1}/\Delta_{i-1}}{D_{i-1}/\Delta_{i-1} + D_i/\Delta_i}\right]\phi_{i-1} + \left[\frac{D_i/\Delta_i}{D_{i-1}/\Delta_{i-1} + D_i/\Delta_i}\right]\phi_i. \tag{3-308}$$

If we inspect the terms of brackets, we can see see that if we were to add them together, we would get one. Therefore, we define the term on the left as ω_{i-1} and have $\omega_i = 1 - \omega_{i-1}$. This allows us to write this more compactly as

$$\phi_{i-1/2} = \omega_{i-1}\phi_{i-1} + (1 - \omega_{i-1})\phi_i. \tag{3-309}$$

Inserting this into Eq. (3-307a), we can then solve for the flow rate at the left interface:

$$\begin{aligned} J_{i-1/2} &= -\frac{2D_i}{\Delta_i}\left[\phi_i - \omega_{i-1}\phi_{i-1} - (1 - \omega_{i-1})\phi_i\right] \\ &= -\frac{2D_i\omega_{i-1}}{\Delta_i}(\phi_i - \phi_{i-1}) \\ &= -2\frac{(D_{i-1}/\Delta_{i-1})(D_i/\Delta_i)}{D_{i-1}/\Delta_{i-1} + D_i/\Delta_i}(\phi_i - \phi_{i-1}) \\ &= -\widetilde{D}_{i-1/2}(\phi_i - \phi_{i-1}). \end{aligned} \tag{3-310}$$

Here we defined as shorthand the edge-averaged diffusion coefficient

$$\widetilde{D}_{i-1/2} = 2\frac{(D_{i-1}/\Delta_{i-1})(D_i/\Delta_i)}{D_{i-1}/\Delta_{i-1} + D_i/\Delta_i}. \tag{3-311}$$

At first glance, this may not look like much of an average, but if we set the cell widths to be equal, we can show that this is the harmonic mean of the diffusion coefficients. (Note that this is different than what is sometimes used for the diffusion coefficients in standard finite difference methods where the they are taken in an ad hoc manner to be the arithmetic mean.)

We can repeat the process for the flow rate on the right edge $J_{1+1/2}$ by consider the cell at x_i and the one to the right centered about x_{i+1}. Applying the finite difference on both sides of the edge gives

$$J_{i+1/2} = -D_i\frac{\phi_{i+1/2} - \phi_i}{\Delta_i/2}, \tag{3-312a}$$

$$J_{i+1/2} = -D_{i+1}\frac{\phi_i - \phi_{i-1/2}}{\Delta_{i+1}/2}. \tag{3-312b}$$

Setting these equal and solving for the cell-edge scalar flux gives

$$J_{i+1/2} = \left[\frac{D_i/\Delta_i}{D_{i+1}/\Delta_{i+1} + D_i/\Delta_i}\right]\phi_i + \left[\frac{D_{i+1}/\Delta_{i+1}}{D_{i+1}/\Delta_{i+1} + D_i/\Delta_i}\right]\phi_{i+1}$$

$$= \omega_i\phi_i + (1-\omega_i)\phi_{i+1}, \tag{3-313}$$

again noting the terms in brackets sum to one. Inserting this into Eq. (3-312a) to eliminate the cell-edge quantity, we get

$$\begin{aligned} J_{i+1/2} &= -\frac{2D_i}{\Delta_i}\left[\omega_i\phi_i + (1-\omega_i)\phi_{i+1} - \phi_i\right] \\ &= -\frac{2D_i(1-\omega_{i-1})}{\Delta_i}(\phi_{i+1}-\phi_i) \\ &= -2\frac{(D_i/\Delta_i)(D_{i+1}/\Delta_{i+1})}{D_i/\Delta_i + D_{i+1}/\Delta_{i+1}}(\phi_{i+1}-\phi_i) \\ &= -\widetilde{D}_{i+1/2}(\phi_{i+1}-\phi_i). \end{aligned} \tag{3-314}$$

Here again, the edge-averaged diffusion coefficient

$$\widetilde{D}_{i+1/2} = 2\frac{(D_i/\Delta_i)(D_{i+1}/\Delta_{i+1})}{D_i/\Delta_i + D_{i+1}/\Delta_{i+1}}. \tag{3-315}$$

Note that this coefficient is identical to the one in Eq. (3-311) by letting $i \to i+1$, which is comforting.

Plugging these edge net currents back into the discretized continuity equation, we have

$$-\widetilde{D}_{i+1/2}(\phi_{i+1}-\phi_i) + \widetilde{D}_{i-1/2}(\phi_i - \phi_{i-1}) + \lambda_i\Delta_i\phi_i = Q_i\Delta_i. \tag{3-316}$$

Grouping terms we get

$$-\widetilde{D}_{i-1/2}u_{i-1} + \left(\widetilde{D}_{i-1/2} + \widetilde{D}_{i+1/2} + \lambda_i\Delta_i\right)\phi_i - \widetilde{D}_{i+1/2}\phi_{i+1} = Q_i\Delta_i. \tag{3-317}$$

If we inspect this expression for the cell centered at x_i, we notice that its average quantity u_i is coupled only to the average quantities to its immediate neighbors. If we write out the equations for each of the interior mesh cells, $i = 1, \ldots, N-2$, (we address the boundary cells next) we notice that this forms a tridiagonal system of equations. We can therefore write the coefficients as the subdiagonal (lower) element ℓ_i, diagonal element d_i, the superdiagonal (upper) element u_i, and the right-hand side vector element r_i as

$$\ell_i = -\widetilde{D}_{i-1/2}, \tag{3-318a}$$

$$d_i = \widetilde{D}_{i-1/2} + \widetilde{D}_{i+1/2} + \lambda_i\Delta_i, \tag{3-318b}$$

$$u_i = -\widetilde{D}_{i+1/2}, \tag{3-318c}$$

$$r_i = Q_i\Delta_i. \tag{3-318d}$$

These interior cell equations may then be written generically as

$$\ell_i\phi_{i-1} + d_i\phi_i + u_i\phi_{i+1} = r_i. \tag{3-319}$$

It turns out that we get pretty much the exact same results on the boundaries as well. One difference is that on the left and right sides at $i = 0$ and $i = N - 1$ we have

$$\ell_0 = 0, \tag{3-320a}$$

$$u_{N-1} = 0, \tag{3-320b}$$

which would be outside the bounds of the matrix. We also can get an additional term on the right-hand side vector should there be an inward partial current. Of course, we still need to actually compute the edge diffusion coefficients, which is the next topic.

Boundary Conditions

We can write the boundary condition generically as

$$\alpha\phi + \beta J = \gamma, \tag{3-321}$$

where α, β, and γ are given coefficients. On the left boundary we give them an ℓ subscript and an r subscript on the right boundary.

We first derive the boundary condition on the left side. The continuity equation is

$$J_{3/2} - J_{1/2} + \lambda_1 \Delta_1 \phi_1 = Q_1 \Delta_1. \tag{3-322}$$

From the analysis for interior elements, we have

$$J_{3/2} = -\widetilde{D}_{3/2}(\phi_2 - \phi_1). \tag{3-323}$$

The task then is to figure out the flow rate at the boundary $J_{1/2}$ in terms of the quantities of interest and the boundary condition.

Applying the finite difference rule to connect the left edge to the interior, we have

$$J_{1/2} = -D_1 \frac{\phi_1 - \phi_{1/2}}{\Delta_1/2}. \tag{3-324}$$

Next, we perform some algebra to obtain

$$(2D_1/\Delta_1)\phi_{1/2} = J_{1/2} + (2D_1/\Delta_1)\phi_1. \tag{3-325}$$

Next, we multiply the left boundary condition by $2D_1/\Delta_1$,

$$(2D_1/\Delta_1)\alpha_\ell \phi_{1/2} + (2D_1/\Delta_1)\beta_\ell J_{1/2} = (2D_1/\Delta_1)\gamma_\ell, \tag{3-326}$$

substitute in for the quantity at the boundary and solve for the flow rate as

$$J_{1/2} = -\frac{2D_1/\Delta_1}{\alpha_\ell + \beta_\ell(2D_1/\Delta_1)}\alpha_\ell \phi_1 + \frac{2D_1/\Delta_1}{\alpha_\ell + \beta_\ell(2D_1/\Delta_1)}\gamma_\ell. \tag{3-327}$$

We can define the edge-averaged diffusion coefficient on the left boundary as

$$\widetilde{D}_{1/2} = \frac{2D_1/\Delta_1}{\alpha_\ell + \beta_\ell(2D_1/\Delta_1)}. \tag{3-328}$$

We then have the flow rate on the left boundary as

$$J_{1/2} = -\widetilde{D}_{1/2}\alpha_\ell\phi_1 + \widetilde{D}_{1/2}\gamma_\ell. \tag{3-329}$$

This result appears similar to what we found for the interior element except there is an additional factor of α_ℓ on the edge-average diffusion coefficient and there is another term for the inflow of the quantity given by γ.

Inserting the flow rates into the continuity equation gives

$$-\widetilde{D}_{3/2}(\phi_2 - \phi_1) + \widetilde{D}_{1/2}\alpha_\ell\phi_1 - \widetilde{D}_{1/2}\gamma_\ell + \lambda_1\Delta_1\phi_1 = Q_1\Delta_1. \tag{3-330}$$

After some rearrangement, we obtain

$$\left(\widetilde{D}_{1/2}\alpha_\ell + \widetilde{D}_{3/2} + \lambda_1\Delta_1\right)\phi_1 - \widetilde{D}_{3/2}\phi_2 = Q_1\Delta_1 + \widetilde{D}_{1/2}\gamma_\ell. \tag{3-331}$$

We can repeat the analysis on the right boundary. The continuity equation is

$$J_{N+1/2} - J_{N-1/2} + \lambda_N\Delta_N\phi_N = Q_N\Delta_N. \tag{3-332}$$

Since $J_{N-1/2}$ is for an interior element,

$$J_{N-1/2} = -\widetilde{D}_{N-1/2}(\phi_N - \phi_{N-1}). \tag{3-333}$$

Apply the finite difference rule on the right edge,

$$J_{N+1/2} = -D_N\frac{\phi_{N+1/2} - \phi_N}{\Delta_N/2}, \tag{3-334}$$

and obtain

$$(2D_N/\Delta_N)\phi_{N+1/2} = -J_{N+1/2} + (2D_N/\Delta_N)\phi_N. \tag{3-335}$$

Multiplying the boundary condition as before by $2D_{N-1}/\Delta_{N-1}$,

$$(2D_N/\Delta_N)\alpha_r\phi_{N+1/2} + (2D_N/\Delta_N)\beta_r J_{N+1/2} = (2D_N/\Delta_N)\gamma_r, \tag{3-336}$$

and making the substitution gives

$$J_{N+1/2} = \widetilde{D}_{N+1/2}\alpha_r\phi_N - \widetilde{D}_{N+1/2}\gamma_r, \tag{3-337}$$

where the edge-averaged diffusion coefficient is

$$\widetilde{D}_{N+1/2} = \frac{2D_N/\Delta_N}{\alpha_r - \beta_r(2D_N/\Delta_N)}. \tag{3-338}$$

Substituting into the continuity equation and rearranging gives a very similar result to that of the left boundary

$$-\widetilde{D}_{N-1/2}\phi_{N-1} + \left(\widetilde{D}_{N-1/2} + \widetilde{D}_{N+1/2}\alpha_r + \lambda_N\Delta_N\right)\phi_N = Q_N\Delta_N + \widetilde{D}_{N+1/2}\gamma_r. \quad (3\text{-}339)$$

We can then use these boundary results to define the coefficients of the tridiagonal matrix and right-hand side vector elements. On the left boundary (top row) we have

$$d_1 = \widetilde{D}_{1/2}\alpha_\ell + \widetilde{D}_{3/2} + \lambda_1\Delta_1, \quad (3\text{-}340a)$$
$$u_1 = -\widetilde{D}_{3/2}, \quad (3\text{-}340b)$$
$$r_1 = Q_1\Delta_1 + \widetilde{D}_{1/2}\gamma_\ell. \quad (3\text{-}340c)$$

And on the right boundary (bottom row) we have

$$\ell_N = -\widetilde{D}_{N-1/2}, \quad (3\text{-}340d)$$
$$d_N = \widetilde{D}_{N-1/2} + \widetilde{D}_{N+1/2}\alpha_r + \lambda_N\Delta_N, \quad (3\text{-}340e)$$
$$r_N = Q_N\Delta_N + \widetilde{D}_{N+1/2}\gamma_r \quad (3\text{-}340f)$$

Comparing these with the interior elements, we note two key differences. First, for the diagonal element the edge-average diffusion coefficient has an additional factor of α. Second, there is an extra term on the right-hand side vector γ that typically denotes flow of the quantity of interest into the problem domain from the exterior.

We also consider a couple special cases. The first is the Dirichlet boundary condition where $\alpha = 1, \beta = 0$, and $\gamma = \phi_b$ where ϕ_b is some known boundary value. The edge-average diffusion coefficient at the left boundary becomes

$$\widetilde{D}_{1/2} = \frac{2D_1}{\Delta_1}, \quad (3\text{-}341)$$

and similarly on the right.

Another special case is the Neumann boundary condition where there is no net inflow of the quantity. This is $\alpha = 0$, $\beta = 1$, $\gamma = 0$. Since $\alpha = \gamma = 0$, the edge-average diffusion coefficient vanishes from the expressions.

3.9.c Example: Heat Conduction in a Nuclear Fuel Plate

To illustrate the finite difference method, let us revisit a the heat conduction example in Sec. 3.8.d, except this time we will solve it using both finite difference and compare the result with an analytic solution.

The plate fuel element will have its midplane at $x = 0$ so that we can apply symmetry and solve half of the problem by asserting there is no net flow of thermal energy at this point. The fuel plate has a half thickness of R_f, a gap ranging from $R_f \leq x < R_g$, and cladding ranging from $R_g \leq x < R_c$. The edge of the clad will be treated with a convective boundary condition to model the removal of heat via a

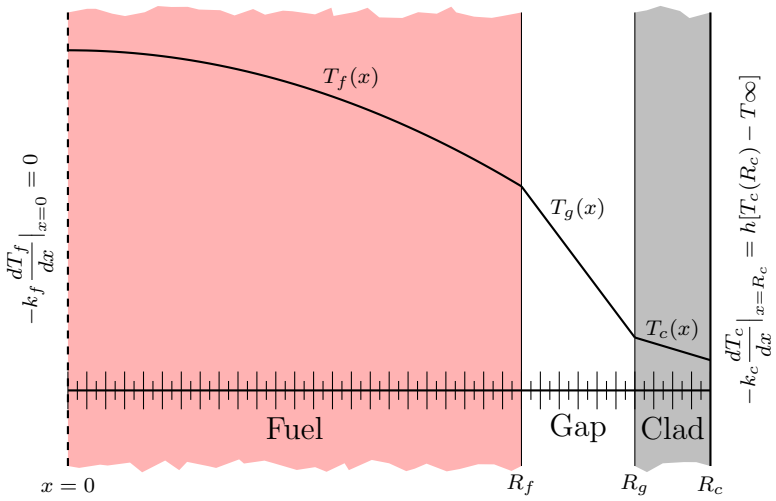

Figure 3.23: Depiction of the example finite difference problem for the temperature distribution in a nuclear fuel plate.

coolant. The geometry depicted in Fig. 3.23. A finite difference grid is also shown with large ticks being the grid points where the temperature is obtained and the small ticks are the midpoints where the thermal conductivities and heat source are specified.

The differential equations describing the heat transfer in this nuclear fuel plate are

$$-k_f \frac{d^2T_f}{dx^2} = Q, \quad 0 \leq x \leq R_f, \tag{3-342a}$$

$$-k_g \frac{d^2T_g}{dx^2} = 0, \quad R_f < x \leq R_g, \tag{3-342b}$$

$$-k_c \frac{d^2T_c}{dx^2} = 0, \quad R_g < x \leq R_c, \tag{3-342c}$$

subject to the boundary conditions

$$q_f(0) = 0, \tag{3-342d}$$

$$T_f(R_f) = T_g(R_f), \tag{3-342e}$$

$$q_f(R_f) = q_g(R_f), \tag{3-342f}$$

$$T_g(R_g) = T_c(R_g), \tag{3-342g}$$

$$q_g(R_g) = q_c(R_g), \tag{3-342h}$$

$$q(R_c) = h(T_c(R_c) - T_\infty). \tag{3-342i}$$

Recall that the heat flux q is related to the temperature T by

$$q_i = -k_i \frac{dT_i}{dx}.$$

The process to obtain an analytic reference solution is much the same as before and will not be repeated. The solution is

$$T_f(x) = T_\infty + QR_f\left(\frac{R_f^2 - x^2}{2k_f R_f} + \frac{R_g - R_f}{k_g} + \frac{R_c - R_g}{k_c} + \frac{1}{h}\right), \tag{3-343a}$$

$$T_g(x) = T_\infty + QR_f\left(\frac{R_g - x}{k_g} + \frac{R_c - R_g}{k_c} + \frac{1}{h}\right), \tag{3-343b}$$

$$T_c(x) = T_\infty + QR_f\left(\frac{R_c - x}{k_c} + \frac{1}{h}\right). \tag{3-343c}$$

This is sketched on Fig. 3.23.

Application of Finite Difference

Now we will apply the finite difference method to approximate this solution for the temperature field. This involves mapping terms to the formulas for the tridiagonal system in Sec. 3.9.b to the context of this problem. The diffusion coefficients are the thermal conductivities k. The reaction coefficient $\lambda(x)$ is zero in this problem since there is no sink of thermal energy such as an endothermic chemical reaction.

For the boundary conditions, on the left-hand side we have a Neumann boundary condition such that

$$\alpha_\ell = 0, \quad \beta_\ell = 1, \quad \gamma_\ell = 0.$$

The right-hand side is a Robin boundary condition that can be written as

$$hT - q = hT_\infty.$$

This implies that

$$\alpha_r = h, \quad \beta_r = -1, \quad \gamma_r = hT_\infty.$$

The edge-averaged diffusion coefficient or thermal conductivity for an interior region is

$$\widetilde{D}_{i+1/2} = 2\frac{(k_i/\Delta_i)(k_{i+1}/\Delta_{i+1})}{(k_i/\Delta_i) + (k_{i+1}/\Delta_{i+1})}, \tag{3-344}$$

Note that for an interior cell where both sides are the same region can be simplified to

$$\widetilde{D}_{i+1/2} = 2\frac{(k/\Delta)(k/\Delta)}{(k/\Delta) + (k/\Delta)} = \frac{k}{\Delta}. \tag{3-345}$$

The left-boundary edge-average diffusion coefficient is unity, $\widetilde{D}_{1/2} = 1$, but it does not appear in the matrix. On the right-boundary, we can obtain an edge-average diffusion coefficient:

$$\widetilde{D}_{N+1/2} = \frac{2(k_c/\Delta_c)}{h + 2(k_c/\Delta_c)}. \tag{3-346}$$

We now have on the left side, the tridiagonal elements are

$$d_1 = \frac{k_f}{\Delta_f}, \quad u_1 = -\frac{k_f}{\Delta_f}, \quad r_1 = Q\Delta_f, \tag{3-347a}$$

assuming the fuel spatial grid has more than one element. At an interface between cells, we apply the general expressions:

$$\begin{aligned}
\ell_i &= -2\frac{(k_{i-1}/\Delta_{i-1})(k_i/\Delta_i)}{(k_{i-1}/\Delta_{i-1}) + (k_i/\Delta_i)}, \\
d_i &= 2\frac{(k_{i-1}/\Delta_{i-1})(k_i/\Delta_i)}{(k_{i-1}/\Delta_{i-1}) + (k_i/\Delta_i)} + 2\frac{(k_i/\Delta_i)(k_{i+1}/\Delta_{i+1})}{(k_i/\Delta_i) + (k_{i+1}/\Delta_{i+1})}, \\
u_i &= -2\frac{(k_i/\Delta_i)(k_{i+1}/\Delta_{i+1})}{(k_i/\Delta_i) + (k_{i+1}/\Delta_{i+1})}, \\
r_i &= Q_i\Delta_i.
\end{aligned} \tag{3-347b}$$

And at the right boundary cell we have

$$\ell_N = -\frac{k_g}{\Delta_g}, \quad d_N = \frac{k_g}{\Delta_g} + \frac{2(k_g/\Delta_g)h}{h + 2(k_g/\Delta_g)}, \quad r_N = \frac{2(k_g/\Delta_g)h}{h + 2(k_g/\Delta_g)}T_\infty, \tag{3-347c}$$

assuming the spatial grid for the gap has more than one element. These can then be inserted into a tridiagonal solver to obtain the solution for cell-average temperatures.

To demonstrate the quality of the solution, the following numerical values are used:

$$\begin{aligned}
R_f &= 0.2 \text{ cm}, \\
R_g &= 0.275 \text{ cm}, \\
R_c &= 0.3 \text{ cm}, \\
k_f &= 0.25 \text{ W}\cdot\text{m}^{-1}\cdot\text{K}^{-1}, \\
k_g &= 0.02 \text{ W}\cdot\text{m}^{-1}\cdot\text{K}^{-1}, \\
k_c &= 20 \text{ W}\cdot\text{m}^{-1}\cdot\text{K}^{-1}, \\
Q &= 1 \times 10^7 \text{ W}\cdot\text{m}^{-3}, \\
h &= 100 \text{ W}\cdot\text{m}^{-2}\cdot\text{K}^{-1}, \\
T_\infty &= 300 \text{ K}.
\end{aligned}$$

The problem is solved using a cell-centered finite difference method with a $\Delta x = 0.005$ cm. The results of the finite difference solution are compared with the analytic solution in Fig. 3.24. For this problem, the finite difference solution agrees with the reference analytic solution with the cell-average quantities lining up well within a degree K compared to the analytical solution.

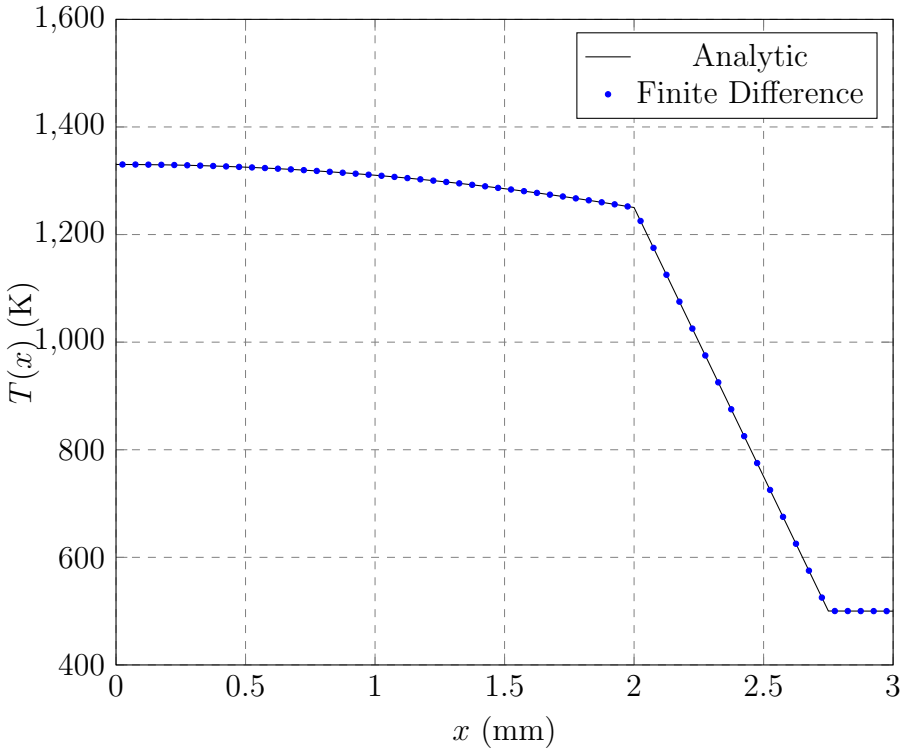

Figure 3.24: Temperature distribution obtained from the finite difference method on the nuclear fuel plate heat conduction example.

Chapter 4

Vector Calculus

In Chapter 2 we introduced the concept of a vector, without reference to Calculus. Vectors do not only have to consist of scalar values, but may be functions as well. Because of this, it is meaningful to bring in the concepts of Calculus and discuss instantaneous rates of change and integrals of quantities. The study of vector calculus is vital to understanding many area of physics relevant to nuclear engineering including mechanics, fluid dynamics, and electrodynamics. This chapter is separated into two parts: differential and integral vector calculus.

For notation, we define the concept of a *multivariablescalar field* $f(x, y, z)$, which we sometimes write in shorthand as $f(\boldsymbol{x})$ where $\boldsymbol{x} = (x, y, z)$. A scalar field takes values of (x, y, z) as input and provides a number as output. Generally speaking, we will use unbolded characters to denote scalar fields or numbers. We can also introduce the concept of a *vector field*

$$\mathbf{F}(x, y, z) = F_x(x, y, z)\hat{\mathbf{i}} + F_y(x, y, z)\hat{\mathbf{j}} + F_z(x, y, z)\hat{\mathbf{k}} \tag{4-1}$$

that takes values of (x, y, z) as input and returns a vector as output. We will generally use capital bolded letters in this chapter to denote vector fields with components with non-bolded capital letters with subscripts. Constant vectors will generally be given by lowercase bold letters, unless denoted otherwise. In defining these terms, we explicitly used Cartesian coordinates. One of the powerful features of vectors, however, is that they can be written in a manner that is independent of the coordinate system. This chapter will discuss different coordinate systems such as cylindrical and spherical coordinates as well. To this end, this chapter derives the concepts of basis vectors in curvilinear and explains how they are used with coordinate transformations.

In the second part of this chapter, integral vector calculus is discussed. The topics include line, surface, and volume integrals; fundamental theorems of vector calculus including the divergence and Stokes' theorem; and finally we provide an introduction to scalar and vector potential functions.

4.1 Vector Derivatives

First, we will study the methods of assessing different rates of change of vector fields. In addition to the standard total and partial derivatives, we also have an operator called a gradient that converts a scalar field into a vector field. We can also apply this gradient on vector fields using the multiplication operations of dot products, cross products, and outer products to produce different quantities with different physical meanings.

4.1.a Derivative of a Vector Field

The derivative of a vector field is the simplest of the operations, being the simple distribution of the derivative of each component. The derivative of a vector field with respect to some parameter t is

$$\frac{\partial \mathbf{F}}{\partial t} = \frac{\partial F_x}{\partial t}\hat{\mathbf{i}} + \frac{\partial F_y}{\partial t}\hat{\mathbf{j}} + \frac{\partial F_z}{\partial t}\hat{\mathbf{k}}. \tag{4-2}$$

The parameter t could be one of the position variables (x, y, z) as well.

4.1.b Gradient

The gradient operator in Cartesian coordinates is defined as

$$\nabla = \frac{\partial}{\partial x}\hat{\mathbf{i}} + \frac{\partial}{\partial y}\hat{\mathbf{j}} + \frac{\partial}{\partial z}\hat{\mathbf{k}}. \tag{4-3}$$

Normally the gradient operator acts upon scalar field $f(x, y, z)$ and outputs a vector field $\mathbf{F}(x, y, z)$:

$$\mathbf{F} = \nabla f = \frac{\partial f}{\partial x}\hat{\mathbf{i}} + \frac{\partial f}{\partial y}\hat{\mathbf{j}} + \frac{\partial f}{\partial z}\hat{\mathbf{k}}. \tag{4-4}$$

The vector described by the vector ∇f at each point (x, y, z) points along the direction of steepest ascent.

An example of the gradient for a 2-D scalar field $f(x, y) = 2x + y^2$ is shown in Fig. 4.1. The color plot provides the scalar field where the blue values denote negative scalar values of the field and red denote positive values. When the gradient is applied, we get $\nabla f = 2\hat{\mathbf{i}} + 2y\hat{\mathbf{j}}$, a vector field, denoted by the arrow plot on the right. At each coordinate, a vector is given with a relative (not absolute) magnitude and an orientation. As we can see, along the x axis, the gradient points only along the x-axis, denoting that it is the direction of steepest ascent. Off the x axis, the gradient points away from the origin, denoting that the direction of steepest ascent is at some angle.

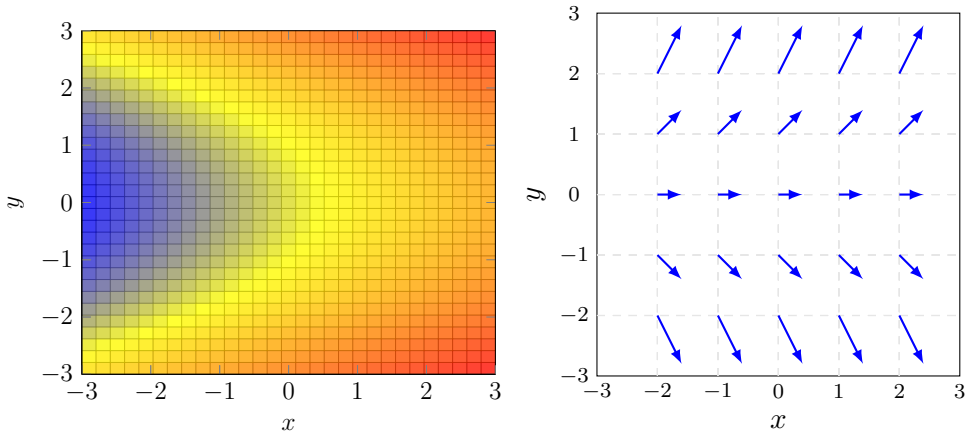

Figure 4.1: Depiction of a scalar field $f(x, y) = 2x + y^2 - 1$ and its associated gradient vector field.

4.1.c Directional Derivative

Sometimes we are interested in the instantaneous rate of change of a scalar field along a particular direction. For this purpose, we define the directional derivative:

$$\nabla_{\mathbf{v}} f = \mathbf{v} \cdot \nabla f. \tag{4-5}$$

This gives the instantaneous rate of the function $f(x, y, z)$ moving through some point (x, y, z) with a "velocity" vector $\mathbf{v}$. The directional derivative takes a scalar field and returns another scalar field.

It is often the case that the vector $\mathbf{v}$ is a unit vector $\hat{\mathbf{\Omega}}$. We also sometimes write the directional derivative as

$$\frac{df}{ds} = \hat{\mathbf{\Omega}} \cdot \nabla f, \tag{4-6}$$

where s is a coordinate axis defined along the direction unit vector $\hat{\mathbf{\Omega}}$.

We can show that the gradient gives the vector of steepest ascent by taking the directional derivative with the unit vector $\hat{\mathbf{\Omega}}$:

$$\hat{\mathbf{\Omega}} \cdot \nabla f = |\hat{\mathbf{\Omega}}||\nabla f| \cos\theta = |\nabla f| \cos\theta. \tag{4-7}$$

Note that $|\hat{\mathbf{\Omega}}| = 1$ since $\hat{\mathbf{\Omega}}$ is a unit vector. The directional derivative is maximized where $\cos\theta = 1$.

$$\hat{\mathbf{\Omega}} \cdot \nabla f = |\nabla f|.$$

This occurs when the unit vector is the normalized gradient

$$\hat{\mathbf{\Omega}} = \frac{\nabla f}{|\nabla f|}. \tag{4-8}$$

Therefore, the directional derivative is maximized along the gradient.

4.1.d Divergence

We can take the dot product of the gradient operator with a vector field. In Cartesian coordinates, the divergence of a vector field is defined as

$$\nabla \cdot \mathbf{F} = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z}. \tag{4-9}$$

The divergence operator takes a vector field and returns a scalar field as output.

The motivation for applying the divergence operator is that it has the physical significance of being the density of the flux of a vector field out of some differential volume at each point (x, y, z). In other words, the divergence is a local measure of the net outward flow of a vector field.

In the context of fluid dynamics, flows with fluid velocity $\mathbf{u}$ for which $\nabla \cdot \mathbf{u} = 0$ are called incompressible. In this case, the velocity field vector points are not allowed to point toward or away from one another, as it would involve the fluid becoming more or less dense. In the context of electromagnetism, the magnetic field follows $\nabla \cdot \mathbf{B} = 0$, which is referred to as a solenoidal field. The divergence gives information about how much each point in space behaves like a localized source of a field. Since there are no point sources of magnetic field (called magnetic monopoles), no point behaves like a localized point source and therefore the magnetic field exhibits no divergence.

4.1.e Curl

The curl is the cross product of the gradient operator and a vector field. In Cartesian coordinates this is given as

$$\begin{aligned} \nabla \times \mathbf{F} &= \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_x & F_y & F_z \end{vmatrix} \\ &= \left(\frac{\partial F_z}{\partial y} - \frac{\partial F_y}{\partial z} \right) \hat{\mathbf{i}} - \left(\frac{\partial F_z}{\partial x} - \frac{\partial F_x}{\partial z} \right) \hat{\mathbf{j}} + \left(\frac{\partial F_y}{\partial x} - \frac{\partial F_x}{\partial y} \right) \hat{\mathbf{k}}. \end{aligned} \tag{4-10}$$

The curl takes a vector field and returns another vector field.

The curl of a vector field describes the local circulation of a vector field about each point (x, y, z). One can envision an infinitesimal surface about each point (x, y, z) with an outward normal vector oriented such that the direction of rotation is clockwise with respect to that vector, which is $\nabla \times \mathbf{F}$. The magnitude of the curl gives a measure of the degree of circulation.

In fluid dynamics, the curl is used to measure the vorticity of the flow, which is the rate that the fluid rotates at a given point. For this reason, we often refer to fluid flows where $\nabla \times \mathbf{F} = \mathbf{0}$ as irrotational. Later, we will see that when the curl of a force field is zero, we can call that force conservative and the work done by such a force along a trajectory only depends upon the endpoints. This is important, because the fundamental forces of nature (e.g., gravity) are conservative, even though macroscopic descriptions of nature used in engineering include nonconservative forces such as friction or viscosity.

4.1.f Convective (Material) Derivative

In fluid dynamics, we often consider a fixed differential element of fluid with scalar property $f(x, y, z, t)$ (e.g., the concentration of some radioisotope tracer) moving in a velocity field $\mathbf{u}(x, y, z, t)$. This is described by the convective or material derivative:

$$\frac{Df}{Dt} = \frac{\partial f}{\partial t} + \mathbf{u} \cdot \nabla f. \tag{4-11}$$

We can also define the convective derivative for a vector field $\mathbf{F}$ (e.g., the fluid momentum) as well:

$$\frac{D\mathbf{F}}{Dt} = \frac{\partial \mathbf{F}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{F}. \tag{4-12}$$

The first term is simply the derivative of a vector field. The second term, however, involves taking the gradient of a vector field. This may be defined using the outer product. For Cartesian coordinates, this is

$$\nabla \mathbf{F} = \nabla \otimes \mathbf{F} = \left(\frac{\partial}{\partial x}\hat{\mathbf{i}} + \frac{\partial}{\partial y}\hat{\mathbf{j}} + \frac{\partial}{\partial x}\hat{\mathbf{k}} \right) \otimes \left(F_x\hat{\mathbf{i}} + F_y\hat{\mathbf{j}} + F_z\hat{\mathbf{k}} \right). \tag{4-13}$$

Expanding the outer product gives

$$\begin{aligned} \nabla \mathbf{F} = &\frac{\partial F_x}{\partial x}\hat{\mathbf{i}}\hat{\mathbf{i}} + \frac{\partial F_y}{\partial x}\hat{\mathbf{i}}\hat{\mathbf{j}} + \frac{\partial F_z}{\partial x}\hat{\mathbf{i}}\hat{\mathbf{k}} \\ &+ \frac{\partial F_x}{\partial y}\hat{\mathbf{j}}\hat{\mathbf{i}} + \frac{\partial F_y}{\partial y}\hat{\mathbf{j}}\hat{\mathbf{j}} + \frac{\partial F_z}{\partial y}\hat{\mathbf{j}}\hat{\mathbf{k}} \\ &+ \frac{\partial F_x}{\partial z}\hat{\mathbf{k}}\hat{\mathbf{i}} + \frac{\partial F_y}{\partial z}\hat{\mathbf{k}}\hat{\mathbf{j}} + \frac{\partial F_z}{\partial z}\hat{\mathbf{k}}\hat{\mathbf{k}}. \end{aligned} \tag{4-14}$$

Taking the dot product with the fluid velocity gives

$$\begin{aligned} \mathbf{u} \cdot \nabla \mathbf{F} = &u_x \left(\frac{\partial F_x}{\partial x}\hat{\mathbf{i}} + \frac{\partial F_y}{\partial x}\hat{\mathbf{j}} + \frac{\partial F_z}{\partial x}\hat{\mathbf{k}} \right) \\ &+ u_y \left(\frac{\partial F_x}{\partial y}\hat{\mathbf{i}} + \frac{\partial F_y}{\partial y}\hat{\mathbf{j}} + \frac{\partial F_z}{\partial y}\hat{\mathbf{k}} \right) \\ &+ u_z \left(\frac{\partial F_x}{\partial z}\hat{\mathbf{i}} + \frac{\partial F_y}{\partial z}\hat{\mathbf{j}} + \frac{\partial F_z}{\partial z}\hat{\mathbf{k}} \right). \end{aligned} \tag{4-15}$$

Combining the vector components yields

$$\begin{aligned} \mathbf{u} \cdot \nabla \mathbf{F} = &\left(u_x\frac{\partial F_x}{\partial x} + u_y\frac{\partial F_x}{\partial y} + u_z\frac{\partial F_x}{\partial z} \right)\hat{\mathbf{i}} \\ &+ \left(u_x\frac{\partial F_y}{\partial x} + u_y\frac{\partial F_y}{\partial y} + u_z\frac{\partial F_y}{\partial z} \right)\hat{\mathbf{j}} \\ &+ \left(u_x\frac{\partial F_z}{\partial x} + u_y\frac{\partial F_z}{\partial y} + u_z\frac{\partial F_z}{\partial z} \right)\hat{\mathbf{k}}. \end{aligned} \tag{4-16}$$

Putting this altogether the convective derivative expressed as components becomes

$$\begin{aligned}\frac{D\mathbf{F}}{Dt} &= \left(\frac{\partial F_x}{\partial t} + u_x\frac{\partial F_x}{\partial x} + u_y\frac{\partial F_x}{\partial y} + u_z\frac{\partial F_x}{\partial z}\right)\hat{\mathbf{i}} \\ &+ \left(\frac{\partial F_y}{\partial t} + u_x\frac{\partial F_y}{\partial x} + u_y\frac{\partial F_y}{\partial y} + u_z\frac{\partial F_y}{\partial z}\right)\hat{\mathbf{j}} \\ &+ \left(\frac{\partial F_z}{\partial t} + u_x\frac{\partial F_z}{\partial x} + u_y\frac{\partial F_z}{\partial y} + u_z\frac{\partial F_z}{\partial z}\right)\hat{\mathbf{k}}. \end{aligned} \tag{4-17}$$

4.1.g Laplacian

We can also take second derivatives using the vector operations. The most common and useful of these is the divergence of the gradient, or the Laplacian. In Cartesian coordinates, the Laplacian of a scalar field (the scalar Laplacian) is given by

$$\nabla\cdot\nabla f = \nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}. \tag{4-18}$$

The Laplacian of a scalar field is another scalar field.

The Laplacian measures how a quantity given by a scalar field f spreads out from a point (x, y, z). The Laplacian is commonly encountered in problems related to diffusion, which describe heat conduction and neutrons diffusion. Related to the scalar Laplacian is the diffusion operator, which is

$$\nabla\cdot D(\boldsymbol{x})\nabla, \tag{4-19}$$

where $D(\boldsymbol{x})$ is a *diffusion coefficient* that may be a function of position. Where $D(\boldsymbol{x}) = D =$ constant, the diffusion operator takes the form of the generic scalar Laplacian multiplied by the constant diffusion coefficient.

In addition to being able to take the Laplacian of a scalar field, we may also take the Laplacian of a vector field $\nabla\cdot\nabla\mathbf{F}$. As we saw when discussing the convective derivative, we must take the gradient of a vector field, which, for Cartesian coordinates, is

$$\begin{aligned}\nabla\mathbf{F} &= \frac{\partial F_x}{\partial x}\hat{\mathbf{i}}\hat{\mathbf{i}} + \frac{\partial F_y}{\partial x}\hat{\mathbf{i}}\hat{\mathbf{j}} + \frac{\partial F_z}{\partial x}\hat{\mathbf{i}}\hat{\mathbf{k}} \\ &+ \frac{\partial F_x}{\partial y}\hat{\mathbf{j}}\hat{\mathbf{i}} + \frac{\partial F_y}{\partial y}\hat{\mathbf{j}}\hat{\mathbf{j}} + \frac{\partial F_z}{\partial y}\hat{\mathbf{j}}\hat{\mathbf{k}} \\ &+ \frac{\partial F_x}{\partial z}\hat{\mathbf{k}}\hat{\mathbf{i}} + \frac{\partial F_y}{\partial z}\hat{\mathbf{k}}\hat{\mathbf{j}} + \frac{\partial F_z}{\partial z}\hat{\mathbf{k}}\hat{\mathbf{k}}. \end{aligned}$$

Next, taking the divergence gives

$$\begin{aligned}\nabla\cdot\nabla\mathbf{F} &= \frac{\partial^2 F_x}{\partial x^2}\hat{\mathbf{i}} + \frac{\partial^2 F_y}{\partial x^2}\hat{\mathbf{j}} + \frac{\partial^2 F_z}{\partial x^2}\hat{\mathbf{k}} \\ &+ \frac{\partial^2 F_x}{\partial y^2}\hat{\mathbf{i}} + \frac{\partial^2 F_y}{\partial y^2}\hat{\mathbf{j}} + \frac{\partial^2 F_z}{\partial y^2}\hat{\mathbf{k}} \end{aligned}$$

$$+ \frac{\partial^2 F_x}{\partial z^2}\hat{\mathbf{i}} + \frac{\partial^2 F_y}{\partial z^2}\hat{\mathbf{j}} + \frac{\partial^2 F_z}{\partial z^2}\hat{\mathbf{k}}. \tag{4-20}$$

Collecting the different terms

$$\begin{aligned}\nabla \cdot \nabla \mathbf{F} = &\left(\frac{\partial^2 F_x}{\partial x^2} + \frac{\partial^2 F_x}{\partial y^2} + \frac{\partial^2 F_x}{\partial z^2}\right)\hat{\mathbf{i}} \\ &+ \left(\frac{\partial^2 F_y}{\partial x^2} + \frac{\partial^2 F_y}{\partial y^2} + \frac{\partial^2 F_y}{\partial z^2}\right)\hat{\mathbf{j}} \\ &+ \left(\frac{\partial^2 F_z}{\partial x^2} + \frac{\partial^2 F_z}{\partial y^2} + \frac{\partial^2 F_z}{\partial z^2}\right)\hat{\mathbf{k}}.\end{aligned} \tag{4-21}$$

This can now be written in terms of the scalar Laplacian:

$$\nabla^2 \mathbf{F} = \nabla^2 F_x \hat{\mathbf{i}} + \nabla^2 F_y \hat{\mathbf{j}} + \nabla^2 F_z \hat{\mathbf{k}}. \tag{4-22}$$

This shows that for Cartesian coordinates the vector Laplacian is a sum of the scalar Laplacian of each component and describes how much each component of the vector field $\mathbf{F}$ spreads out from (x, y, z).

An important identity that generalizes the vector Laplacian to any coordinate system is

$$\nabla^2 \mathbf{F} = \nabla(\nabla \cdot \mathbf{F}) - \nabla \times (\nabla \times \mathbf{F}). \tag{4-23}$$

In other words, the vector Laplacian is the gradient of the divergence minus the curl of the curl. There are other combinations of second derivatives that appear throughout the application of vector calculus, but these will not be discussed in detail here.

4.1.h Vector Derivative Identities

There are numerous vector identities that can be employed to manipulate equations and cast them in simpler or more physically meaningful forms.

Two common vector identities for vector fields $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are as follows: The first involve interchanging the vectors when we take dot product of the cross product of vectors:

$$\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) = \mathbf{B} \cdot (\mathbf{C} \times \mathbf{A}) = \mathbf{C} \cdot (\mathbf{A} \times \mathbf{B}). \tag{4-24}$$

The three vectors can be reordered in a cyclic manner.

The second vector identity that simplifies the cross product of the cross product is

$$\mathbf{A} \times (\mathbf{B} \times \mathbf{C}) = \mathbf{B}(\mathbf{A} \cdot \mathbf{C}) - \mathbf{C}(\mathbf{A} \cdot \mathbf{B}). \tag{4-25}$$

This allows for the expression of the cross products in terms of dot products. Note that this vector identity is often referred to be its mnemonic, the BAC-CAB identity, which follows the ordering of the vectors on the right-hand side.

We can also write six different product rules involving the gradient, divergence, and curl. The gradient of the product of two scalar fields f and g has a similar form as the standard product rule:

$$\nabla(fg) = f\nabla g + g\nabla f. \tag{4-26}$$

The gradient of the dot product of two vector fields $\mathbf{A}$ and $\mathbf{B}$ is

$$\nabla(\mathbf{A}\cdot\mathbf{B}) = \mathbf{A}\times(\nabla\times\mathbf{B}) + \mathbf{B}\times(\nabla\times\mathbf{A}) + (\mathbf{A}\cdot\nabla)\mathbf{B} + (\mathbf{B}\cdot\nabla)\mathbf{A}. \tag{4-27}$$

The divergence of the product of a scalar field $\mathbf{f}$ and vector field $\mathbf{A}$ is

$$\nabla\cdot(f\mathbf{A}) = f(\nabla\cdot\mathbf{A}) + \mathbf{A}\cdot\nabla f. \tag{4-28}$$

The divergence of the cross product of two vector fields $\mathbf{A}$ and $\mathbf{B}$ is

$$\nabla\cdot(\mathbf{A}\times\mathbf{B}) = \mathbf{B}\cdot(\nabla\times\mathbf{A}) - \mathbf{A}\cdot(\nabla\times\mathbf{B}). \tag{4-29}$$

The curl of the product of a scalar field $\mathbf{f}$ and vector field $\mathbf{A}$ is

$$\nabla\times(f\mathbf{A}) = f(\nabla\times\mathbf{A}) - \mathbf{A}\times(\nabla f). \tag{4-30}$$

Finally, the curl of the cross product of two vector fields $\mathbf{A}$ and $\mathbf{B}$ is

$$\nabla\times(\mathbf{A}\times\mathbf{B}) = (\mathbf{B}\cdot\nabla)\mathbf{A} - (\mathbf{A}\cdot\nabla)\mathbf{B} + \mathbf{A}(\nabla\cdot\mathbf{B}) - \mathbf{B}(\nabla\cdot\mathbf{A}). \tag{4-31}$$

Additionally, there are identities involving the second derivatives. The first is that the divergence of the curl is zero

$$\nabla\cdot(\nabla\times\mathbf{A}) = 0. \tag{4-32}$$

The curl of the gradient is also zero

$$\nabla\times(\nabla\mathbf{A}) = \mathbf{0}. \tag{4-33}$$

Finally, we can rearrange the definition of the vector Laplacian to get the curl of the curl as

$$\nabla\times(\nabla\times\mathbf{A}) = \nabla(\nabla\cdot\mathbf{A}) - \nabla^2\mathbf{A}. \tag{4-34}$$

Finally, we can list identities for third derivatives, i.e., those involving the Laplacian. The gradient of the scalar Laplacian can be written as the vector Laplacian of the gradient

$$\nabla(\nabla^2 f) = \nabla^2(\nabla f). \tag{4-35}$$

The divergence of vector Laplacian is the scalar Laplacian of the divergence

$$\nabla\cdot(\nabla^2\mathbf{A}) = \nabla^2(\nabla\cdot\mathbf{A}). \tag{4-36}$$

Finally, the curl of the vector Laplacian is the vector Laplacian of the curl

$$\nabla\times(\nabla^2\mathbf{A}) = \nabla^2(\nabla\times\mathbf{A}). \tag{4-37}$$

4.1.i Example: Vorticity Equation for an Incompressible Fluid

An incompressible fluid satisfies a simplified version of the Navier-Stokes equation:

$$\rho\frac{D\mathbf{u}}{Dt} = -\nabla p + \mu\nabla^2\mathbf{u}. \tag{4-38}$$

Here $\mathbf{u}(x, y, z, t)$ is the fluid velocity vector field, p is the scalar pressure field, ρ is the constant fluid density, and μ is the constant fluid viscosity. Since the fluid is incompressible, we know that the divergence of the fluid velocity field is zero,

$$\nabla \cdot \mathbf{u} = 0. \tag{4-39}$$

We define the vorticity as the curl of the velocity field,

$$\boldsymbol{\omega} = \nabla \times \mathbf{u}. \tag{4-40}$$

The vorticity, as the name implies, describes the local circulation of the fluid.

The goal is to derive a conservation equation for the vorticity. To begin, we first expand the convective derivative:

$$\rho\frac{D\mathbf{u}}{Dt} = \rho\frac{\partial\mathbf{u}}{\partial t} + \rho\mathbf{u}\cdot\nabla\mathbf{u} = -\nabla p + \mu\nabla^2\mathbf{u}. \tag{4-41}$$

Recall the vector identity for the gradient of a dot product of two vector fields:

$$\nabla(\mathbf{A}\cdot\mathbf{B}) = \mathbf{A}\times(\nabla\times\mathbf{B}) + \mathbf{B}\times(\nabla\times\mathbf{A}) + (\mathbf{A}\cdot\nabla)\mathbf{B} + (\mathbf{B}\cdot\nabla)\mathbf{A}.$$

Here we let $\mathbf{A} = \mathbf{B} = \mathbf{u}$ to get

$$\nabla(\mathbf{u}\cdot\mathbf{u}) = 2\mathbf{u}\times(\nabla\times\mathbf{u}) + 2(\mathbf{u}\cdot\nabla)\mathbf{u}. \tag{4-42}$$

Solving for $\mathbf{u}\cdot\nabla\mathbf{u}$:

$$(\mathbf{u}\cdot\nabla)\mathbf{u} = \frac{1}{2}\nabla(\mathbf{u}\cdot\mathbf{u}) - \mathbf{u}\times(\nabla\times\mathbf{u}). \tag{4-43}$$

Inserting this into the fluid differential equation gives

$$\rho\frac{\partial\mathbf{u}}{\partial t} + \frac{\rho}{2}\nabla(\mathbf{u}\cdot\mathbf{u}) - \rho\mathbf{u}\times(\nabla\times\mathbf{u}) = -\nabla p + \mu\nabla^2\mathbf{u}. \tag{4-44}$$

Note that the term

$$\frac{\rho}{2}\nabla(\mathbf{u}\cdot\mathbf{u})$$

is related to the kinetic energy of the fluid.

Next, we take the curl of the resulting equation to get:

$$\rho\frac{\partial\boldsymbol{\omega}}{\partial t} + \frac{\rho}{2}\nabla\times\nabla(\mathbf{u}\cdot\mathbf{u}) - \rho\nabla\times(\mathbf{u}\times\boldsymbol{\omega}) = -\nabla\times\nabla p + \mu\nabla\times(\nabla^2\mathbf{u}). \tag{4-45}$$

Because the curl of the gradient is zero, we can eliminate the kinetic energy and pressure terms:

$$\rho\frac{\partial \boldsymbol{\omega}}{\partial t} - \rho\nabla \times (\mathbf{u} \times \boldsymbol{\omega}) = \mu\nabla \times (\nabla^2\mathbf{u}). \tag{4-46}$$

We can apply the identity for the curl of the cross product:

$$\nabla \times (\mathbf{u} \times \boldsymbol{\omega}) = (\boldsymbol{\omega} \cdot \nabla)\mathbf{u} - (\mathbf{u} \cdot \nabla)\boldsymbol{\omega} + \mathbf{u}(\nabla \cdot \boldsymbol{\omega}) - \boldsymbol{\omega}(\nabla \cdot \mathbf{u}). \tag{4-47}$$

The third term is zero because the divergence of the curl is zero

$$\nabla \cdot \boldsymbol{\omega} = \nabla \cdot (\nabla \times \mathbf{u}) = 0. \tag{4-48}$$

The fourth term is zero because the fluid is incompressible. Therefore, we now have

$$\rho\frac{\partial \boldsymbol{\omega}}{\partial t} + \rho(\mathbf{u} \cdot \nabla)\boldsymbol{\omega} - \rho(\boldsymbol{\omega} \cdot \nabla)\mathbf{u} = \mu\nabla \times (\nabla^2\mathbf{u}). \tag{4-49}$$

The first two terms can be regrouped into the convective derivative of the vorticity. For the right-hand side, we can equate the curl of the vector Laplacian with the Laplacian of the curl:

$$\nabla \times (\nabla^2\mathbf{u}) = \nabla^2(\nabla \times \mathbf{u}) = \nabla^2\boldsymbol{\omega}. \tag{4-50}$$

The vorticity equation becomes

$$\frac{D\boldsymbol{\omega}}{Dt} = (\boldsymbol{\omega} \cdot \nabla)\mathbf{u} + \frac{\mu}{\rho}\nabla^2\boldsymbol{\omega}. \tag{4-51}$$

The equation states that the change of the vorticity of a fluid element is given by the stretching of the vortex because of gradients in the flow velocity and is dissipated by the viscosity of the fluid.

4.2 Curvilinear Coordinates

In calculus we are permitted to have basis vectors that are continuous functions of space. This allows us to define useful coordinate systems that make solving practical engineering problems much simpler. The most common of these are curvilinear coordinates. In 2-D the most common curvilinear coordinate system is polar coordinates, where the basis vectors in (x, y) are replaced with radial and polar angle basis vectors (r, θ). In 3-D, we have two common curvilinear coordinate systems. The first is cylindrical coordinates (r, θ, z) with radial, polar, and axial variables and spherical basis vectors. The second is spherical coordinates (r, θ, ϕ) with radial, polar, and azimuthal basis vectors respectively. This section reviews these coordinate systems. In this section we will also show the that basis vectors can be expressed as partial derivatives.

4.2.a Cartesian Basis Vectors

To begin, we start with Cartesian coordinates. Here we will show the results in 2D, as it is easier to visualize graphically, but the results easily extend to 3D. Let us consider some position vector

$$\mathbf{R}(x,y) = x\hat{\mathbf{i}} + y\hat{\mathbf{j}}. \tag{4-52}$$

Additionally, we have the position vector evaluated at $\mathbf{R}(x+\Delta x, y)$ for some offset Δx while y is left constant. $\mathbf{R}(x+\Delta x, y)$ is shifted in the x direction by Δx. Next, we define a vector as the difference of these $\mathbf{R}(x+\Delta x, y) - \mathbf{R}(x,y)$. These vectors are illustrated in Fig. 4.2.

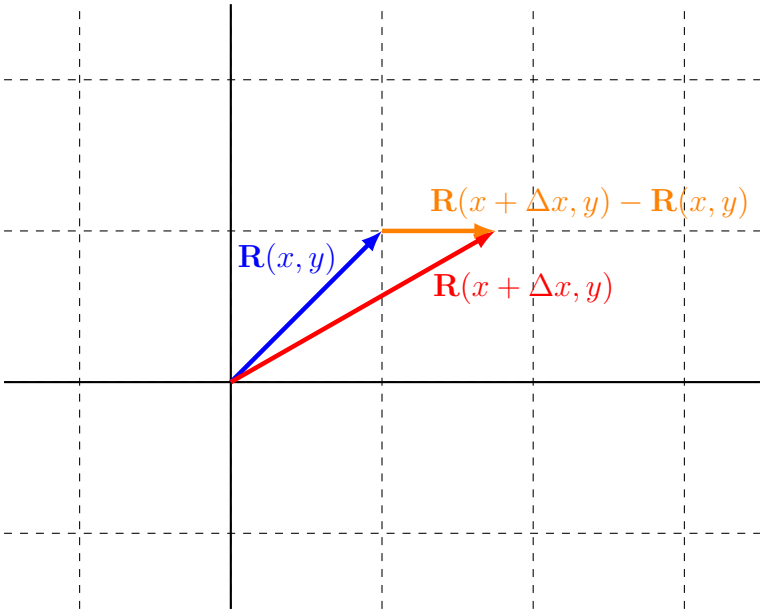

Figure 4.2: Illustration of derivative along x direction.

The derivative is then

$$\frac{\partial \mathbf{R}}{\partial x} \approx \frac{\mathbf{R}(x+\Delta x, y) - \mathbf{R}(x,y)}{\Delta x}. \tag{4-53}$$

The magnitude of the vector $\mathbf{R}(x+\Delta x, y) - \mathbf{R}(x,y)$ is Δx. Therefore, the magnitude of the derivative is one. Then the limit gives a unit vector in the x direction:

$$\frac{\partial \mathbf{R}}{\partial x} = \lim_{\Delta x \to 0} \frac{\mathbf{R}(x+\Delta x, y) - \mathbf{R}(x,y)}{\Delta x} = \lim_{\Delta x \to 0} \frac{\Delta x \hat{\mathbf{i}}}{\Delta x} = \hat{\mathbf{i}} = \mathbf{e}_x. \tag{4-54}$$

By similar arguments, a small change in y gives

$$\frac{\partial \mathbf{R}}{\partial y} = \lim_{\Delta y \to 0} \frac{\mathbf{R}(x, y+\Delta y) - \mathbf{R}(x,y)}{\Delta y} = \lim_{\Delta y \to 0} \frac{\Delta x \hat{\mathbf{j}}}{\Delta x} = \hat{\mathbf{j}} = \mathbf{e}_y. \tag{4-55}$$

Likewise, in 3-D, a small change in z gives

$$\frac{\partial \mathbf{R}}{\partial z} = \lim_{\Delta z \to 0} \frac{\mathbf{R}(x, y, z+\Delta z) - \mathbf{R}(x,y,z)}{\Delta z} = \lim_{\Delta z \to 0} \frac{\Delta z \hat{\mathbf{k}}}{\Delta z} = \hat{\mathbf{k}} = \mathbf{e}_z. \tag{4-56}$$

4.2.b Polar Basis Vectors

In 2-D we often analyze problems with circular symmetry, and in these cases polar coordinates are ideal. Rather than describe the coordinates using x and y positions, we instead use the distance from the origin, the radius r, and some angle with respect to the x axis, the polar angle θ. The radial coordinate has a range of $[0, \infty)$ and the polar angle coordinate has a range of $[0, 2\pi)$. The relationships to go from Cartesian to polar coordinates are

$$r = \sqrt{x^2 + y^2}, \tag{4-57a}$$

$$\theta = \tan^{-1}\left(\frac{y}{x}\right); \tag{4-57b}$$

and the conversions from polar to Cartesian coordinates are

$$x = r\cos\theta, \tag{4-58a}$$

$$y = r\sin\theta. \tag{4-58b}$$

To get the basis vectors we do similar to what we did for Cartesian coordinates, but using a position vector $\mathbf{R}(r, \theta)$. For the radial coordinate, we have vectors $\mathbf{R}(r, \theta)$ and $\mathbf{R}(r + \Delta r, \theta)$. As before, we take the difference between the two $\mathbf{R}(r + \Delta r, \theta) - \mathbf{R}(r, \theta)$ and observe that its magnitude is Δr. Taking the limit, we get the same result as with Cartesian coordinates:

$$\frac{\partial \mathbf{R}}{\partial r} = \lim_{\Delta r \to 0} \frac{\mathbf{R}(r + \Delta r, \theta) - \mathbf{R}(r, \theta)}{\Delta r} == \lim_{\Delta r \to 0} \frac{\Delta r \hat{\mathbf{r}}}{\Delta r} = \hat{\mathbf{r}} = \mathbf{e}_r. \tag{4-59}$$

An aspect of $\mathbf{e}_r = \hat{\mathbf{r}}$ is that, unlike the Cartesian basis vector, it always points away from the origin, and therefore depends explicitly on the θ coordinate.

For the polar basis vector, we take the position vectors $\mathbf{R}(r, \theta)$ and $\mathbf{R}(r, \theta + \Delta\theta)$. These two vectors are illustrated in blue and red respectively in Fig. 4.3 for two different sets with the same $\Delta\theta$ but different radii. The difference between the two is given in the same plot in orange. To find the magnitude of the difference vector $\ell = |\mathbf{R}(r, \theta + \Delta\theta) - \mathbf{R}(r, \theta)|$, we can use the law of cosines

$$\ell^2 = r^2 + r^2 - 2rr\cos(\Delta\theta) = 2r^2[1 - \cos(\Delta\theta)]. \tag{4-60}$$

Since $\Delta\theta$ is small, we can approximate it with a Taylor series expansion about $\Delta\theta = 0$:

$$\cos(\Delta\theta) \approx \cos(0) - \Delta\theta\sin(0) - \frac{(\Delta\theta)^2}{2}\cos(0) = 1 - \frac{(\Delta\theta)^2}{2}. \tag{4-61}$$

Substituting this expansion into the law of cosines

$$\ell^2 \approx 2r^2[1 - (1 - \frac{(\Delta\theta)^2}{2})] = r^2(\Delta\theta)^2, \quad \ell \approx r\Delta\theta. \tag{4-62}$$

Note that the magnitude of the difference is proportional to the radius r. Also observe that the difference vector for a finite $\Delta\theta$ is not tangent to the radial coordinate;

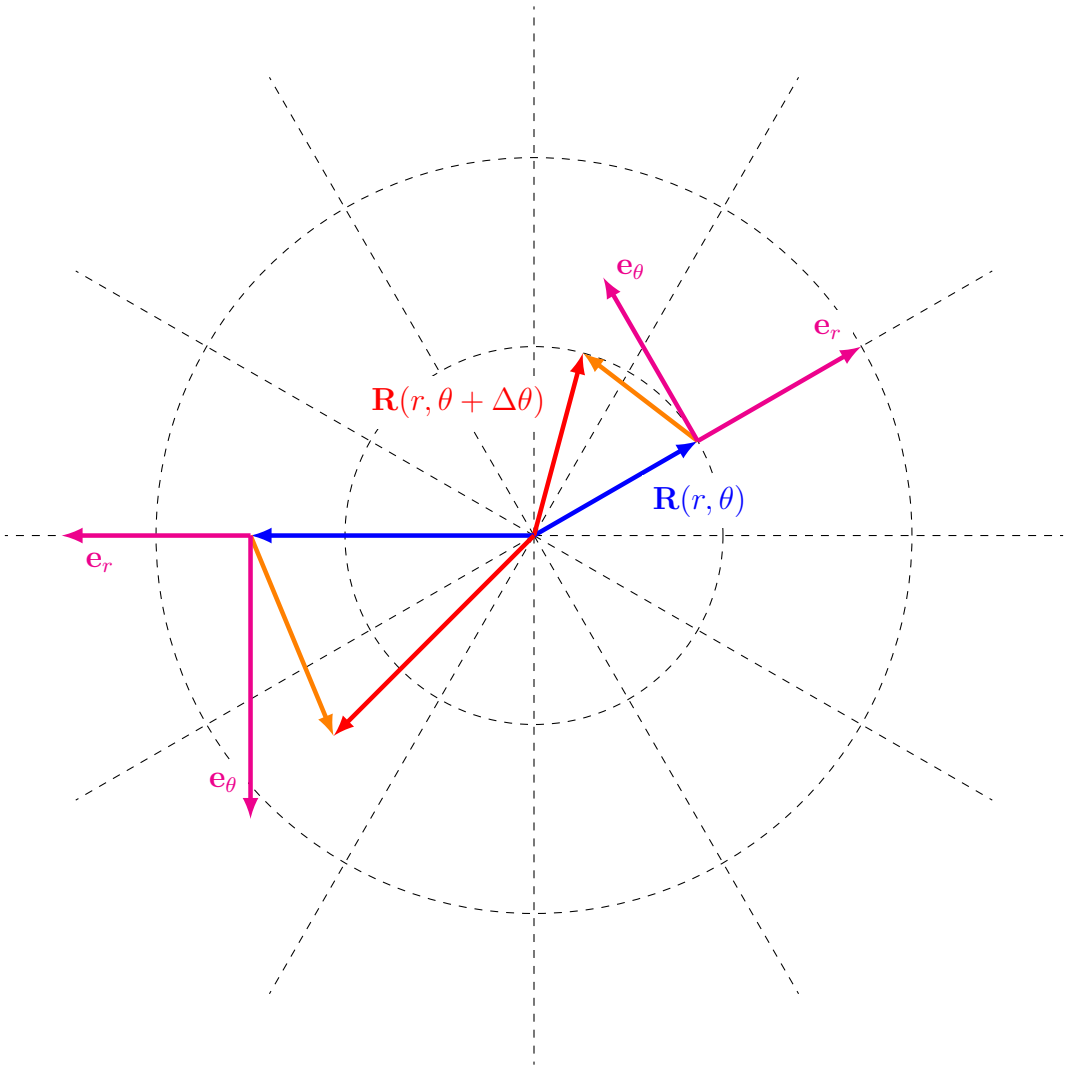

Figure 4.3: Illustration of radial and polar unit vectors $\mathbf{e}_r$ and $\mathbf{e}_\theta$ for two different radii and polar angles, but same $\Delta\theta$.

however, as $\Delta\theta$ becomes smaller the difference vector becomes increasingly tangent and limits to a vector as such. The limit as $\Delta\theta$ goes to zero is

$$\frac{\partial \mathbf{R}}{\partial \theta} = \lim_{\Delta\theta\to 0} \frac{\mathbf{R}(r,\theta+\Delta\theta)-\mathbf{R}(r,\theta)}{\Delta\theta} = \lim_{\Delta\theta\to 0} \frac{r\Delta\theta\hat{\boldsymbol{\theta}}}{\Delta\theta} = r\hat{\boldsymbol{\theta}} = \mathbf{e}_\theta. \tag{4-63}$$

Here we note that, unlike for the Cartesian or radial basis vectors, the unit polar basis vector $\hat{\boldsymbol{\theta}}$ is not the same as the unnormalized polar basis vector $\mathbf{e}_\theta$, with the difference being a factor of the radius. Here $\mathbf{e}_\theta$ scales in proportion to the radius. We use both basis vectors here because each has a particular value depending upon the application.

4.2.c Cylindrical Basis Vectors

We now proceed to the 3-D coordinate systems. The first is cylindrical coordinates, which, as the name implies is useful for problems that have cylindrical characteristics. Similar to the polar coordinate system, we have a radial coordinate r, polar angle coordinate θ, and, now in addition, an axial coordinate z. The radial coordinate has a range of $[0, \infty)$, the polar angle coordinate has a range of $[0, 2\pi)$, and the axial coordinate ranges from $(-\infty, \infty)$. Note that some mathematics textbooks use different symbols for these coordinate variables; here we have chosen the most common convention used in physics.

The coordinate transformations are almost identical to polar coordinates except for the addition of the axial coordinate, which is just the z coordinate in Cartesian coordinates. The transformation for Cartesian to cylindrical is

$$r = \sqrt{x^2 + y^2}, \tag{4-64a}$$
$$\theta = \tan^{-1}\left(\frac{y}{x}\right), \tag{4-64b}$$
$$z = z. \tag{4-64c}$$

and the conversions from polar to Cartesian coordinates are

$$x = r\cos\theta, \tag{4-65a}$$
$$y = r\sin\theta, \tag{4-65b}$$
$$z = z. \tag{4-65c}$$

The basis vectors are largely identical to the polar coordinate system with the axial coordinate behaving identically to the Cartesian coordinate basis vector. For completeness, these are

$$\frac{\partial \mathbf{R}}{\partial r} = \hat{\mathbf{r}} = \mathbf{e}_r, \tag{4-66a}$$
$$\frac{\partial \mathbf{R}}{\partial \theta} = r\hat{\boldsymbol{\theta}} = \mathbf{e}_\theta, \tag{4-66b}$$
$$\frac{\partial \mathbf{R}}{\partial z} = \hat{\mathbf{z}} = \mathbf{e}_z. \tag{4-66c}$$

Here we use $\hat{\mathbf{z}}$ instead of $\hat{\mathbf{k}}$ to distinguish between the use of cylindrical coordinates versus Cartesian coordinates; however, they are, for all purposes, identical.

4.2.d Spherical Basis Vectors

The other 3-D curvilinear coordinate that is common is the spherical coordinate system. The spherical coordinate system has a radial coordinate r, a polar angle coordinate θ that moves along a line of constant longitude from the north to south pole, and azimuthal angle ϕ that moves eastward along a line of constant latitude around the sphere. Note that radial coordinate goes from $[0, \infty)$, the polar angle

coordinate goes from $[0, \pi]$, and the azimuthal angle coordinate goes from $[0, 2\pi)$. As with the cylindrical coordinate system, we are using the most common convention used in physics, which differs from many mathematical texts.

The coordinate transformation rules from Cartesian to spherical coordinates are

$$r = \sqrt{x^2 + y^2 + z^2}, \tag{4-67a}$$

$$\theta = \cos^{-1}\left(\frac{z}{\sqrt{x^2 + y^2 + z^2}}\right), \tag{4-67b}$$

$$\phi = \tan^{-1}\left(\frac{y}{x}\right). \tag{4-67c}$$

The transformation rules from spherical to Cartesian coordinates are

$$x = r\sin\theta\cos\phi, \tag{4-68a}$$

$$y = r\sin\theta\sin\phi, \tag{4-68b}$$

$$z = r\cos\theta. \tag{4-68c}$$

The radial basis vector is identical to the polar and cylindrical coordinates:

$$\frac{\partial \mathbf{R}}{\partial r} = \hat{\mathbf{r}} = \mathbf{e}_r. \tag{4-69a}$$

The polar angle in spherical coordinates has a different range, but the polar basis vector is identical to its analogs in polar and cylindrical coordinates:

$$\frac{\partial \mathbf{R}}{\partial \theta} = r\hat{\boldsymbol{\theta}} = \mathbf{e}_\theta. \tag{4-69b}$$

To derive the azimuthal basis vector, we again take a position vector $\mathbf{R}(r, \theta, \phi)$ and compare it to another position vector with azimuthal offset $\Delta\phi$ for fixed radial coordinate r and polar angle coordinate θ, $\mathbf{R}(r, \theta, \phi + \Delta\phi)$. We then have to consider how the length of the difference $\mathbf{R}(r, \theta, \phi + \Delta\phi) - \mathbf{R}(r, \theta, \phi)$ scales with r and θ. As r changes, the circumference of the circular path of constant latitude grows in proportion to the radius, so the magnitude of the difference scales in direct proportion to r, which is the same as the polar coordinate. For variable polar angle θ, however, the circumference also changes. Near the north pole, the distance around the sphere along a line of constant latitude is shorter than near the equator. Therefore, for a fixed azimuthal offset $\Delta\phi$, we travel proportionately further near the equator than near the poles. Doing a bit of trigonometry can show that the circumference is proportional to $\sin\theta$. Therefore, the azimuthal basis vector is

$$\frac{\partial \mathbf{R}}{\partial \phi} = r\sin\theta\hat{\boldsymbol{\phi}} = \mathbf{e}_\phi. \tag{4-69c}$$

A note about all of the major coordinate systems—Cartesian, polar, cylindrical, and spherical—are orthogonal, and, if the unit basis vectors are used, orthonormal. The coordinate systems are defined such that at any position all of the coordinate basis vectors are always oriented such that they are 90° from each other.

This completes our discussion of the major coordinate systems used in a large majority of scientific and engineering applications. There are other coordinate systems that are occasionally used for specialized applications including ellipsoidal and toroidal coordinate systems; however, given that they are seldom encountered, they will not be discussed here.

4.3 Coordinate Transformations

In the context of linear algebra we discussed the concept of changing a basis. A change of basis takes a set of basis vectors, which define a coordinate system, and transforms them into a new set of basis vectors, which define a new coordinate system. Recall that we developed a transformation matrix $\mathbf{T}$ and the basis vectors transformed covariantly with respect to themselves (the definition of a coordinate transformation), whereas components of a vector in one coordinate system transform contravariantly using $\mathbf{T}^{-1}$ with respect to the basis vectors.

4.3.a Jacobian and Transformation Matrix

The Jacobian matrix provides a method of changing basis vectors from one coordinate system to another and is directly analogous to the transformation matrices from linear algebra. Because we saw in the previously that the basis vectors are described by partial derivatives, we can use the chain rule from multivariable calculus to relate them. In 2-D Cartesian to polar we have:

$$\frac{\partial \mathbf{R}}{\partial r} = \frac{\partial x}{\partial r}\frac{\partial \mathbf{R}}{\partial x} + \frac{\partial y}{\partial r}\frac{\partial \mathbf{R}}{\partial y}, \tag{4-70a}$$

$$\frac{\partial \mathbf{R}}{\partial \theta} = \frac{\partial x}{\partial \theta}\frac{\partial \mathbf{R}}{\partial x} + \frac{\partial y}{\partial \theta}\frac{\partial \mathbf{R}}{\partial y}. \tag{4-70b}$$

Equivalently,

$$\mathbf{e}_r = \frac{\partial x}{\partial r}\mathbf{e}_x + \frac{\partial y}{\partial r}\mathbf{e}_y, \tag{4-71a}$$

$$\mathbf{e}_\theta = \frac{\partial x}{\partial \theta}\mathbf{e}_x + \frac{\partial y}{\partial \theta}\mathbf{e}_y. \tag{4-71b}$$

Since the transformations of $x = r\cos\theta$ and $y = r\sin\theta$ are given, we can explicitly evaluate the derivatives to obtain

$$\mathbf{e}_r = \cos\theta\mathbf{e}_x + \sin\theta\mathbf{e}_y, \tag{4-72a}$$

$$\mathbf{e}_\theta = -r\sin\theta\mathbf{e}_x + r\cos\theta\mathbf{e}_y. \tag{4-72b}$$

Since a vector is a column vector in matrix notation, we can express the transformation matrix with columns corresponding to those vectors,

$$\mathbf{J} = \frac{\partial(x,y)}{\partial(r,\theta)} = \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix} = \begin{bmatrix} {}^{x}/_{r} & -y \\ {}^{y}/_{r} & x \end{bmatrix}. \tag{4-73}$$

Here we used the symbol $\mathbf{J}$ as opposed to $\mathbf{T}$. Sometimes the notation of a partial derivative of multiple arguments with respect to multiple arguments is used as well to explicitly specify the coordinate transform performed by the matrix. This matrix is called the *Jacobian matrix* that transforms the basis vectors from Cartesian to polar coordinates in terms of $\mathbf{e}_r$ and $\mathbf{e}_\theta$. The Jacobian matrix is also used to derive the expressions for length, area, and volume, which will be discussed later in this chapter.

Recall that the unit vector $\mathbf{e}_\theta$ is unnormalized, and therefore increases in magnitude in proportion to the distance from the origin. It is often more convenient to transform to a coordinate system where the basis vectors are unnormalized. Recall $\mathbf{e}_r = \hat{\mathbf{r}}$ and $\mathbf{e}_\theta = r\hat{\boldsymbol{\theta}}$; the basis unit vectors of the polar coordinate system become is

$$\hat{\mathbf{r}} = \cos\theta\hat{\mathbf{i}} + \sin\theta\hat{\mathbf{j}}, \tag{4-74a}$$

$$\hat{\boldsymbol{\theta}} = -\sin\theta\hat{\mathbf{i}} + \cos\theta\hat{\mathbf{j}}. \tag{4-74b}$$

This gives the transformation matrix

$$\mathbf{T} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} {}^{x}\!/_{r} & -{}^{y}\!/_{r} \\ {}^{y}\!/_{r} & {}^{x}\!/_{r} \end{bmatrix}. \tag{4-75}$$

Obtaining the Jacobian matrix to convert from Cartesian to cylindrical coordinates is largely identical, so this is given as

$$\mathbf{J} = \frac{\partial(x,y,z)}{\partial(r,\theta,z)} = \begin{bmatrix} \cos\theta & -r\sin\theta & 0 \\ \sin\theta & r\cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} {}^{x}\!/_{r} & -y & 0 \\ {}^{y}\!/_{r} & x & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{4-76}$$

As with polar coordinates, the Jacobian matrix transforms into a coordinate system with an unnormalized basis vector $\mathbf{e}_\theta$. To get the more conventional unit basis vectors $(\hat{\mathbf{r}}, \hat{\boldsymbol{\theta}}, \hat{\mathbf{z}})$ we use the following transformation matrix:

$$\mathbf{T} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} {}^{x}\!/_{r} & -{}^{y}\!/_{r} & 0 \\ {}^{y}\!/_{r} & {}^{x}\!/_{r} & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{4-77}$$

To go from Cartesian to spherical coordinates we apply the partial derivatives and obtain

$$\mathbf{e}_r = \hat{\mathbf{r}} = \sin\theta\cos\phi\hat{\mathbf{i}} + \sin\theta\sin\phi\hat{\mathbf{j}} + \cos\theta\hat{\mathbf{k}}, \tag{4-78a}$$

$$\mathbf{e}_\theta = r\hat{\boldsymbol{\theta}} = r\cos\theta\cos\phi\hat{\mathbf{i}} + r\cos\theta\sin\phi\hat{\mathbf{j}} - r\sin\theta\hat{\mathbf{k}}, \tag{4-78b}$$

$$\mathbf{e}_\phi = r\sin\theta\hat{\boldsymbol{\phi}} = -r\sin\theta\sin\phi\hat{\mathbf{i}} + r\sin\theta\cos\phi\hat{\mathbf{j}}. \tag{4-78c}$$

The Jacobian matrix for the transformation of Cartesian to spherical coordinates is

$$\mathbf{J} = \frac{\partial(x,y,z)}{\partial(r,\theta,\phi)} = \begin{bmatrix} \sin\theta\cos\phi & r\cos\theta\cos\phi & -r\sin\theta\sin\phi \\ \sin\theta\sin\phi & r\cos\theta\sin\phi & r\sin\theta\cos\phi \\ \cos\theta & -r\sin\theta & 0 \end{bmatrix}. \tag{4-79}$$

The basis vectors $\mathbf{e}_\theta$ and $\mathbf{e}_\phi$ are unnormalized such that $\mathbf{e}_\theta = r\hat{\boldsymbol{\theta}}$ and $\mathbf{e}_\phi = r\sin\theta\hat{\boldsymbol{\phi}}$. Therefore, the unit basis vectors for the spherical coordinate system are given by

$$\hat{\mathbf{r}} = \sin\theta\cos\phi\hat{\mathbf{i}} + \sin\theta\sin\phi\hat{\mathbf{j}} + \cos\theta\hat{\mathbf{k}}, \tag{4-80a}$$

$$\hat{\boldsymbol{\theta}} = \cos\theta\cos\phi\hat{\mathbf{i}} + \cos\theta\sin\phi\hat{\mathbf{j}} - \sin\theta\hat{\mathbf{k}}, \tag{4-80b}$$

$$\hat{\boldsymbol{\phi}} = -\sin\phi\hat{\mathbf{i}} + \cos\phi\hat{\mathbf{j}}. \tag{4-80c}$$

The transformation matrix is

$$\mathbf{T} = \begin{bmatrix} \sin\theta\cos\phi & \cos\theta\cos\phi & -\sin\phi \\ \sin\theta\sin\phi & \cos\theta\sin\phi & \cos\phi \\ \cos\theta & -\sin\theta & 0 \end{bmatrix}. \tag{4-81}$$

For a general coordinate transformation from (x, y, z) to (u, v, w), we can use the multivariable chain rule to write

$$\mathbf{e}_u = \frac{\partial x}{\partial u}\mathbf{e}_x + \frac{\partial y}{\partial u}\mathbf{e}_y + \frac{\partial z}{\partial u}\mathbf{e}_z, \tag{4-82a}$$

$$\mathbf{e}_v = \frac{\partial x}{\partial v}\mathbf{e}_x + \frac{\partial y}{\partial v}\mathbf{e}_y + \frac{\partial z}{\partial v}\mathbf{e}_z, \tag{4-82b}$$

$$\mathbf{e}_w = \frac{\partial x}{\partial w}\mathbf{e}_x + \frac{\partial y}{\partial w}\mathbf{e}_y + \frac{\partial z}{\partial w}\mathbf{e}_z. \tag{4-82c}$$

Jacobian matrix has these vector components as the columns

$$\mathbf{J} = \frac{\partial(x, y, z)}{\partial(u, v, w)} = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} & \frac{\partial y}{\partial w} \\ \frac{\partial z}{\partial u} & \frac{\partial z}{\partial v} & \frac{\partial z}{\partial w} \end{bmatrix}. \tag{4-83}$$

Note that in this formulation, the basis vectors are not necessarily orthogonal nor are they necessarily normalized so that they are unit vectors.

Transforming the basis vectors in the reverse direction (e.g., spherical to Cartesian) could be done by recomputing the derivatives and forming a new Jacobian matrix. Alternatively, one can also compute the inverse of the Jacobian matrix $\mathbf{J}^{-1}$. As we have in linear algebra, the basis vectors transform covariantly with respect to the transformation of the basis vectors (by definition) using $\mathbf{J}$. Conversely, vector components transform contravariantly using $\mathbf{J}^{-1}$.

4.3.b Example: Straight Trajectory in Polar Coordinates

While it is often convenient to express the differential equations in a coordinate system compatible with the geometry of the problem, certain aspects of the physics may become more complicated. For example, neutral particles (e.g., photons and neutrons) travel in essentially straight-line trajectories, and it becomes important to rewrite those trajectories in the new coordinate systems. Here we will consider the example of a straight line trajectory in the x direction, $\hat{\boldsymbol{\Omega}} = \hat{\mathbf{i}}$, expressed in polar coordinates along a point $(x, y = 1)$.

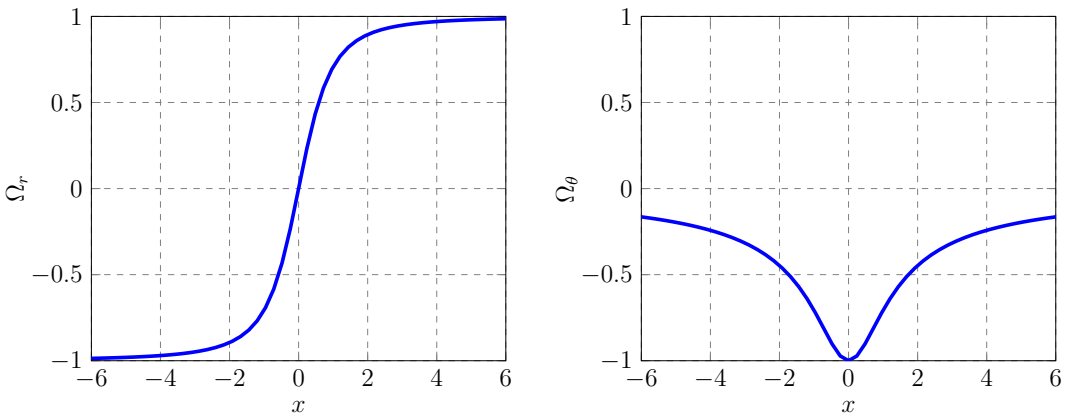

Figure 4.4: Vector components of straight line trajectory in x direction along $(x, 1)$ in polar coordinates.

Recall that the transformation matrix from Cartesian to polar coordinates is

$$\mathbf{T} = \begin{bmatrix} {}^{x}/_{r} & -{}^{y}/_{r} \\ {}^{y}/_{r} & {}^{x}/_{r} \end{bmatrix}.$$

As with algebraic vectors, the components transform contravariantly with respect to the basis vectors, and we need to use the inverse or backward transform. This is

$$\mathbf{T}^{-1} = \begin{bmatrix} {}^{x}/_{r} & {}^{y}/_{r} \\ -{}^{y}/_{r} & {}^{x}/_{r} \end{bmatrix} = \begin{bmatrix} \dfrac{x}{\sqrt{1+x^2}} & \dfrac{1}{\sqrt{1+x^2}} \\ -\dfrac{1}{\sqrt{1+x^2}} & \dfrac{x}{\sqrt{1+x^2}} \end{bmatrix}. \tag{4-84}$$

(To derive this, recall that $x^2 + y^2 = r^2$.) Applying the inverse transformation to the unit vector gives

$$\hat{\mathbf{\Omega}}' = \begin{bmatrix} \dfrac{x}{\sqrt{1+x^2}} & \dfrac{1}{\sqrt{1+x^2}} \\ -\dfrac{1}{\sqrt{1+x^2}} & \dfrac{x}{\sqrt{1+x^2}} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \dfrac{x}{\sqrt{1+x^2}} \\ -\dfrac{1}{\sqrt{1+x^2}} \end{bmatrix}. \tag{4-85}$$

Or expressed in component form,

$$\hat{\mathbf{\Omega}}' = \frac{x}{\sqrt{1+x^2}}\hat{\mathbf{r}} - \frac{1}{\sqrt{1+x^2}}\hat{\boldsymbol{\theta}}. \tag{4-86}$$

The components are plotted in Fig. 4.4. Moving from left to right, the radial component Ω_r starts near -1 crosses the origin at $x = 0$ and proceeds to 1, whereas the polar angle component Ω_θ is always negative, starting toward zero, reaching a maximum of -1 at $x = 0$, and going again toward zero.

To help understand this behavior for $y = 1$, two vectors $\hat{\mathbf{\Omega}}$ at different values of x are plotted in Fig. 4.5 in blue along with the corresponding basis vectors in

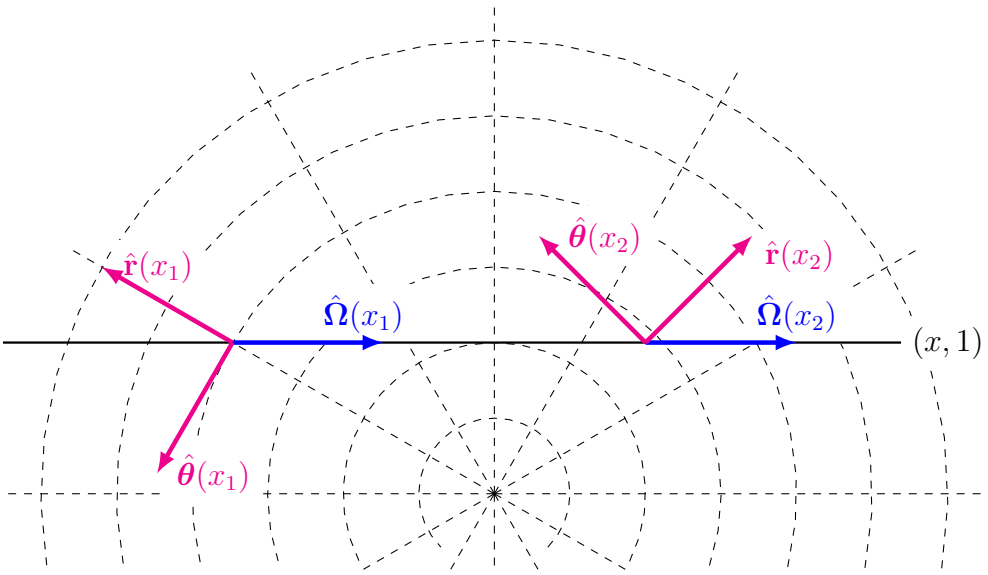

Figure 4.5: Illustration of a couple vector locations for the straight-line trajectory alongside the corresponding basis vectors.

magenta. For the negative values of x, the cosine of the angle between $\hat{\mathbf{r}}$ and the vector $\hat{\boldsymbol{\Omega}}$ is negative, meaning the radial component of the vector is also negative. When the vector $\hat{\boldsymbol{\Omega}}$ crosses the y axis, it is perpendicular to the $\hat{\mathbf{r}}$ and therefore the radial coordinate is zero. Correspondingly, when x is positive, the cosine of the angle between $\hat{\boldsymbol{\Omega}}$ and $\hat{\mathbf{r}}$ is also positive. As x gets large in either the positive or negative direction, the vector $\hat{\boldsymbol{\Omega}}$ becomes increasingly aligned with $\hat{\mathbf{r}}$.

In terms of $\hat{\boldsymbol{\theta}}$, for negative values of x, the dot product is negative. The same is true for positive x, as $\hat{\boldsymbol{\theta}}$ is always aligned in a direction that is counterclockwise and tangent to a circle about the origin. Note that as x approaches the origin from the left, the angle between $\hat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\theta}}$ approaches 180°. As x moves away from the points y axis, the angle between $\hat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\theta}}$ continues to decrease and tends toward 90°. Since the range of the angle between $\hat{\boldsymbol{\Omega}}$ and $\hat{\boldsymbol{\theta}}$ ranges from 270° to 90°, the cosine is always negative and Ω_θ is always negative. Note that if we send the trajectory for a fixed value of $-y$, then the Ω_θ component would always be positive.

4.3.c Gradient in Non-Cartesian Coordinates

One of the advantages of the vector notation is that it is independent of coordinate system. Thus far, we have discussed how to perform the gradient and its associated operators in Cartesian coordinates.

Here we show how to obtain the gradient in general coordinates. Suppose we have a coordinate system with coordinate indices (u, v, w) and respective basis vectors $\mathbf{e}_u$, $\mathbf{e}_v$, and $\mathbf{e}_w$. Suppose we express the gradient in terms of the basis vectors with unknown coefficients

$$\nabla f = A_u \mathbf{e}_u + A_v \mathbf{e}_v + A_w \mathbf{e}_w. \tag{4-87}$$

Next we take the directional derivative along each of the directions

$$\frac{\partial f}{\partial u} = \mathbf{e}_u \cdot \nabla f = A_u(\mathbf{e}_u \cdot \mathbf{e}_u) + A_v(\mathbf{e}_u \cdot \mathbf{e}_v) + A_w(\mathbf{e}_u \cdot \mathbf{e}_w), \tag{4-88a}$$

$$\frac{\partial f}{\partial v} = \mathbf{e}_v \cdot \nabla f = A_u(\mathbf{e}_v \cdot \mathbf{e}_u) + A_v(\mathbf{e}_v \cdot \mathbf{e}_v) + A_w(\mathbf{e}_v \cdot \mathbf{e}_w), \tag{4-88b}$$

$$\frac{\partial f}{\partial w} = \mathbf{e}_w \cdot \nabla f = A_u(\mathbf{e}_w \cdot \mathbf{e}_u) + A_v(\mathbf{e}_w \cdot \mathbf{e}_v) + A_w(\mathbf{e}_w \cdot \mathbf{e}_w). \tag{4-88c}$$

This can be written as the following linear system:

$$\begin{bmatrix} \mathbf{e}_u \cdot \mathbf{e}_u & \mathbf{e}_u \cdot \mathbf{e}_v & \mathbf{e}_u \cdot \mathbf{e}_w \\ \mathbf{e}_v \cdot \mathbf{e}_u & \mathbf{e}_v \cdot \mathbf{e}_v & \mathbf{e}_v \cdot \mathbf{e}_w \\ \mathbf{e}_w \cdot \mathbf{e}_u & \mathbf{e}_w \cdot \mathbf{e}_v & \mathbf{e}_w \cdot \mathbf{e}_w \end{bmatrix} \begin{bmatrix} A_u \\ A_v \\ A_w \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial u} \\ \frac{\partial f}{\partial v} \\ \frac{\partial f}{\partial w} \end{bmatrix} \tag{4-89}$$

Recall that this matrix is the metric tensor $\mathbf{g}$ from linear algebra. To compute the elements of the metric tensor, we express the basis vectors in terms of Cartesian coordinate partial derivatives using the multivariable chain rule. For example:

$$\begin{aligned} \mathbf{e}_u \cdot \mathbf{e}_v &= \left(\frac{\partial x}{\partial u}\mathbf{e}_x + \frac{\partial y}{\partial u}\mathbf{e}_y + \frac{\partial z}{\partial u}\mathbf{e}_z\right) \cdot \left(\frac{\partial x}{\partial v}\mathbf{e}_x + \frac{\partial y}{\partial v}\mathbf{e}_y + \frac{\partial z}{\partial v}\mathbf{e}_z\right) \\ &= \frac{\partial x}{\partial u}\frac{\partial x}{\partial v} + \frac{\partial y}{\partial u}\frac{\partial y}{\partial v} + \frac{\partial z}{\partial u}\frac{\partial z}{\partial v}. \end{aligned} \tag{4-90}$$

The cross terms vanish because $\mathbf{e}_x \cdot \mathbf{e}_y = 0$, $\mathbf{e}_x \cdot \mathbf{e}_z = 0$, and $\mathbf{e}_y \cdot \mathbf{e}_z = 0$. Once these products of partial derivatives have been computed, we can find each element of the metric tensor and solve the linear system of equations by finding the inverse metric tensor $\mathbf{g}^{-1}$ to obtain the coefficients.

4.3.d Example: Gradient in Spherical Coordinates

To illustrate this, let us compute the gradient in spherical coordinates. Recall that

$$\begin{aligned} x &= r\sin\theta\cos\phi, \\ y &= r\sin\theta\sin\phi, \\ z &= r\cos\theta. \end{aligned}$$

Taking the partial derivatives and evaluating the products of basis vectors

$$\mathbf{e}_r \cdot \mathbf{e}_r = \sin^2\theta\cos^2\phi + \sin^2\theta\sin^2\phi + \cos^2\theta = 1, \tag{4-91a}$$

$$\mathbf{e}_r \cdot \mathbf{e}_\theta = r\sin\theta\cos\theta\cos^2\phi + r\cos\theta\cos\theta\sin^2\phi - r\cos\theta\sin\theta = 0, \tag{4-91b}$$

$$\mathbf{e}_r \cdot \mathbf{e}_\phi = -r\sin^2\theta\sin\phi\cos\phi + r\sin^2\theta\sin\phi\cos\phi = 0, \tag{4-91c}$$

$$\mathbf{e}_\theta \cdot \mathbf{e}_\theta = r^2\cos^2\theta\cos^2\phi + r^2\cos^2\theta\sin^2\phi + r^2\sin^2\theta = r^2, \tag{4-91d}$$

$$\mathbf{e}_\theta \cdot \mathbf{e}_\phi = -r^2\sin\theta\cos\theta\sin\phi\cos\phi + r^2\sin\theta\cos\theta\sin\phi\cos\phi = 0, \tag{4-91e}$$

$$\mathbf{e}_\phi \cdot \mathbf{e}_\phi = r^2\sin^2\theta\sin^2\phi + r^2\sin^2\theta\cos^2\phi = r^2\sin^2\theta. \tag{4-91f}$$

Note that the cross terms are all zero; this is a consequence of the spherical coordinate system being orthogonal. The linear system with the metric tensor becomes:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & r^2 \sin^2\theta \end{bmatrix} \begin{bmatrix} A_r \\ A_\theta \\ A_\phi \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial r} \\ \frac{\partial f}{\partial \theta} \\ \frac{\partial f}{\partial \phi} \end{bmatrix}. \tag{4-92}$$

Since the metric tensor for this coordinate system is diagonal, we can easily invert to solve the system:

$$\begin{bmatrix} A_r \\ A_\theta \\ A_\phi \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{r^2} & 0 \\ 0 & 0 & \frac{1}{r^2 \sin^2\theta} \end{bmatrix} \begin{bmatrix} \frac{\partial f}{\partial r} \\ \frac{\partial f}{\partial \theta} \\ \frac{\partial f}{\partial \phi} \end{bmatrix}. \tag{4-93}$$

Therefore, the gradient is

$$\nabla f = \frac{\partial f}{\partial r}\mathbf{e}_r + \frac{1}{r^2}\frac{\partial f}{\partial \theta}\mathbf{e}_\theta + \frac{1}{r^2 \sin^2\theta}\frac{\partial f}{\partial \phi}\mathbf{e}_\phi. \tag{4-94}$$

We generally express the gradient in terms of the unit basis vectors to make the coordinate system orthonormal. The gradient with the orthonormal spherical basis vectors is

$$\nabla f = \frac{\partial f}{\partial r}\hat{\mathbf{r}} + \frac{1}{r}\frac{\partial f}{\partial \theta}\hat{\boldsymbol{\theta}} + \frac{1}{r\sin\theta}\frac{\partial f}{\partial \phi}\hat{\boldsymbol{\phi}}. \tag{4-95}$$

4.3.e Metric Tensor and Scale Factors

In the derivation for the gradient, we found the metric tensor,

$$\mathbf{g} = \begin{bmatrix} \mathbf{e}_u \cdot \mathbf{e}_u & \mathbf{e}_u \cdot \mathbf{e}_v & \mathbf{e}_u \cdot \mathbf{e}_w \\ \mathbf{e}_v \cdot \mathbf{e}_u & \mathbf{e}_v \cdot \mathbf{e}_v & \mathbf{e}_v \cdot \mathbf{e}_w \\ \mathbf{e}_w \cdot \mathbf{e}_u & \mathbf{e}_w \cdot \mathbf{e}_v & \mathbf{e}_w \cdot \mathbf{e}_w \end{bmatrix}, \tag{4-96}$$

arise. Recall from linear algebra (where the basis vectors are constant everywhere) the metric tensor is used to measure distance and angles between vectors. In the context of calculus, where we allow the basis vectors to change continuously with position, the metric tensor takes on the meaning that allows us to consistently measure differential length in each coordinate direction at each point in space.

The curvilinear coordinate systems all have the property of orthogonality, which means that the basis vectors are orthogonal at all points in space. When this occurs, the off-diagonal terms of the metric tensor goes to zero, as we saw for spherical coordinates. Therefore, for orthogonal coordinate systems, we can write the metric tensor as

$$\mathbf{g} = \begin{bmatrix} \mathbf{e}_u \cdot \mathbf{e}_u & 0 & 0 \\ 0 & \mathbf{e}_v \cdot \mathbf{e}_v & 0 \\ 0 & 0 & \mathbf{e}_w \cdot \mathbf{e}_w \end{bmatrix} = \begin{bmatrix} h_u^2 & 0 & 0 \\ 0 & h_v^2 & 0 \\ 0 & 0 & h_w^2 \end{bmatrix}. \tag{4-97}$$

Here the h factors are called the *scale factors*, which are the square root of the diagonal elements of the metric tensor in an orthogonal coordinate system:

$$h_i = \sqrt{g_{ii}}. \tag{4-98}$$

The scale factors allow us to develop general formulas for the gradient, divergence, and curl. They also permit us to put the coordinate systems into an orthonormal form having unit basis vectors. Note that it is important to emphasize that this only applies in orthogonal coordinate systems. Fortunately, non-orthogonal curvilinear coordinates (called skew coordinates) are seldom used in scientific and engineering applications, so knowing the forms for orthogonal coordinate systems is almost always sufficient.

The scale factors for Cartesian coordinates are:

$$h_x = 1, \tag{4-99a}$$
$$h_y = 1, \tag{4-99b}$$
$$h_z = 1. \tag{4-99c}$$

The scale factors for cylindrical coordinates are

$$h_r = 1, \tag{4-100a}$$
$$h_\theta = r, \tag{4-100b}$$
$$h_z = 1. \tag{4-100c}$$

Finally, the scale factors for spherical coordinates are

$$h_r = 1, \tag{4-101a}$$
$$h_\theta = r, \tag{4-101b}$$
$$h_\phi = r \sin\theta. \tag{4-101c}$$

4.3.f Gradient, Divergence, and Curl in Curvilinear Coordinates

Using the scale factors, we can develop expressions for the gradient, divergence, curl, and Laplacian that apply to any orthonormal coordinate system. The gradient is

$$\nabla f = \frac{1}{h_u}\frac{\partial f}{\partial u}\hat{\mathbf{u}} + \frac{1}{h_v}\frac{\partial f}{\partial v}\hat{\mathbf{v}} + \frac{1}{h_w}\frac{\partial f}{\partial w}\hat{\mathbf{w}}, \tag{4-102}$$

where $\hat{\mathbf{u}}$, $\hat{\mathbf{v}}$, and $\hat{\mathbf{w}}$ are normalized unit basis vectors. The divergence is

$$\nabla \cdot \mathbf{F} = \frac{1}{h_u h_v h_w}\left[\frac{\partial}{\partial u}(h_v h_w F_u) + \frac{\partial}{\partial v}(h_u h_w F_v) + \frac{\partial}{\partial w}(h_u h_v F_w)\right]. \tag{4-103}$$

The curl is given by

$$\nabla \times \mathbf{F} = \frac{1}{h_u h_v h_w}\begin{vmatrix} h_u\hat{\mathbf{u}} & h_v\hat{\mathbf{v}} & h_w\hat{\mathbf{w}} \\ \dfrac{\partial}{\partial u} & \dfrac{\partial}{\partial v} & \dfrac{\partial}{\partial w} \\ h_u F_u & h_v F_v & h_w F_w \end{vmatrix}. \tag{4-104}$$

The scalar Laplacian is

$$\nabla^2 f = \frac{1}{h_u h_v h_w}\left[\frac{\partial}{\partial u}\left(\frac{h_v h_w}{h_u}\frac{\partial f}{\partial u}\right) + \frac{\partial}{\partial v}\left(\frac{h_u h_w}{h_v}\frac{\partial f}{\partial v}\right) + \frac{\partial}{\partial w}\left(\frac{h_u h_v}{h_w}\frac{\partial f}{\partial w}\right)\right]. \quad (4\text{-}105)$$

Applying these to cylindrical coordinates, we obtain the following expressions for the gradient, divergence, curl, and Laplician:

$$\nabla f = \frac{\partial f}{\partial r}\hat{\mathbf{r}} + \frac{1}{r}\frac{\partial f}{\partial \theta}\hat{\boldsymbol{\theta}} + \frac{\partial f}{\partial z}\hat{\mathbf{z}}, \quad (4\text{-}106a)$$

$$\nabla \cdot \mathbf{F} = \frac{1}{r}\frac{\partial}{\partial r}(rF_r) + \frac{1}{r}\frac{\partial F_\theta}{\partial \theta} + \frac{\partial F_z}{\partial z}, \quad (4\text{-}106b)$$

$$\nabla \times \mathbf{F} = \frac{1}{r}\left(\frac{\partial F_z}{\partial \theta} - \frac{\partial}{\partial z}(rF_\theta)\right)\hat{\mathbf{r}} + \left(\frac{\partial F_r}{\partial z} - \frac{\partial F_z}{\partial r}\right)\hat{\boldsymbol{\theta}} + \frac{1}{r}\left(\frac{\partial}{\partial r}(rF_\theta) - \frac{\partial F_r}{\partial \theta}\right)\hat{\mathbf{z}}, \quad (4\text{-}106c)$$

$$\nabla^2 f = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial f}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 f}{\partial \theta^2} + \frac{\partial^2 f}{\partial z^2}. \quad (4\text{-}106d)$$

The gradient, divergence, curl, and Laplician in spherical coordinates are:

$$\nabla f = \frac{\partial f}{\partial r}\hat{\mathbf{r}} + \frac{1}{r}\frac{\partial f}{\partial \theta}\hat{\boldsymbol{\theta}} + \frac{1}{r\sin\theta}\frac{\partial f}{\partial \phi}\hat{\boldsymbol{\phi}}, \quad (4\text{-}107a)$$

$$\nabla \cdot \mathbf{F} = \frac{1}{r^2}\frac{\partial}{\partial r}(r^2 F_r) + \frac{1}{r\sin\theta}\frac{\partial}{\partial \theta}(\sin\theta F_\theta) + \frac{1}{r\sin\theta}\frac{\partial F_\phi}{\partial \phi}, \quad (4\text{-}107b)$$

$$\nabla \times \mathbf{F} = \frac{1}{r\sin\theta}\left(\frac{\partial}{\partial \theta}(\sin\theta F_\phi) - \frac{\partial F_\theta}{\partial \phi}\right)\hat{\mathbf{r}} + \frac{1}{r}\left(\frac{1}{\sin\theta}\frac{\partial F_r}{\partial \phi} - \frac{\partial}{\partial r}(rF_\phi)\right)\hat{\boldsymbol{\theta}} + \frac{1}{r}\left(\frac{\partial}{\partial r}(rF_\theta) - \frac{\partial F_r}{\partial \theta}\right)\hat{\boldsymbol{\phi}}, \quad (4\text{-}107c)$$

$$\nabla^2 f = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial f}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta\frac{\partial f}{\partial \theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 f}{\partial \phi^2}. \quad (4\text{-}107d)$$

4.4 Covector Fields

The gradient operator takes a scalar field and converts it into a vector field. Recall for linear algebra that column vectors are vectors, whereas row vectors are covectors. It is therefore reasonable to ask whether there is an analogous covector field. The answer is yes, and it is related to the differential of a scalar field df. To understand that the differential operator takes a vector field and produces a covector field, we must define the differential or "d" operator in a specific way. Given this new interpretation of the differential df, we can develop a connection for line integration to linear algebra.

4.4.a Directional Derivative and the Differential Operator

Traditionally we think of dx as a small change in the variable x. Consider the consider simple the scalar field $f = x$. This scalar field simply takes on the x value at each

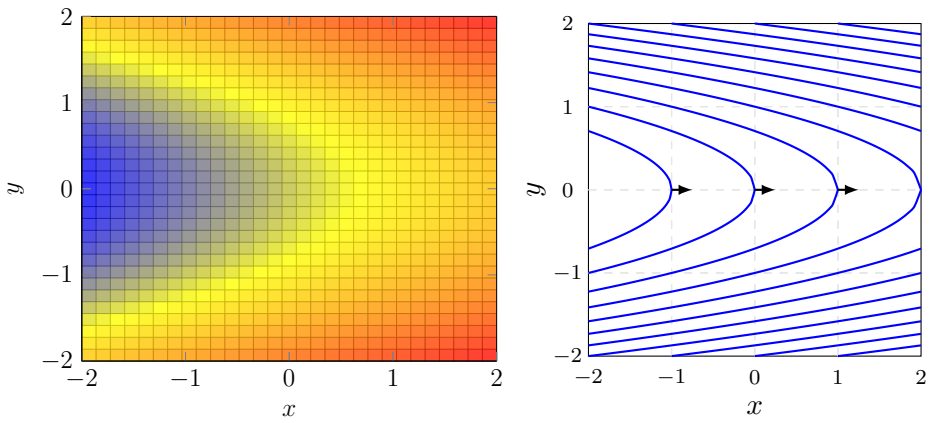

Figure 4.6: Depiction of a scalar field $f(x, y) = x + y^2 - 1$ and its associated covector field $df = dx + 2ydy$.

point. Now suppose we take the directional derivative along the $\mathbf{v}$ direction of this scalar field that we equate to some dx operating upon $\mathbf{v}$:

$$dx(\mathbf{v}) = \nabla_{\mathbf{v}} x = \mathbf{v} \cdot \nabla x = \mathbf{v} \cdot \mathbf{e}_x = v_x, \tag{4-108}$$

which is simply the x component of the vector $\mathbf{v}$. We may think of the operator dx as taking the x component of some vector. An equivalent formulation is to write dx as a row vector

$$dx = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} = \epsilon_x, \tag{4-109}$$

which is equivalent to the Cartesian basis covector ϵ_x. Recall that the basis covector is a set of vertical parallel lines (in 2-D space) or planes (in 3-D space) oriented in the x direction. Likewise, $dy = \epsilon_y$ and $dz = \epsilon_z$ which are lines or planes with the respective y and z orientations.

We can now take the differential of a scalar field f by applying the multivariable chain rule:

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy + \frac{\partial f}{\partial z} dz. \tag{4-110}$$

This says that the differential change in f is equal to the differential change in x times the partial derivative of f with respect to x, plus the differential change in y times the partial derivative of f with respect to y, and similarly for z. In the context of the directional derivative df acting on vector $\mathbf{v}$ gives the instantaneous rate of change of f along $\mathbf{v}$. From the multivariable chain rule, this can be written as the sum of Cartesian basis covector fields.

To illustrate this geometrically, consider the 2-D scalar field

$$f(x, y) = x + y^2 - 1. \tag{4-111}$$

The differential operator acting upon f gives

$$df = dx + 2ydy. \tag{4-112}$$

Writing these as covectors:

$$df = \begin{bmatrix} 1 & 0 \end{bmatrix} + 2y \begin{bmatrix} 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2y \end{bmatrix}. \tag{4-113}$$

As we saw in linear algebra, the covectors can be thought of as lines of constant "elevation". Let us suppose df now acts on some position vector $\mathbf{R} = x\mathbf{e}_x + y\mathbf{e}_y$:

$$df(\mathbf{R}) = \begin{bmatrix} 1 & 2y \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = x + 2y^2 = r. \tag{4-114}$$

Here r represents some scalar number. Represented graphically, a fixed value of r gives a curve. (Recall in linear algebra, since things were not functions of space, the curves were simple lines). Several such curves are plotted in Fig. 4.6 along with the associated scalar field. These curves can be thought of as contour lines of the scalar field. The density of these lines along a direction gives a measure of how much the scalar field is changing along that direction, which is similar to lines of constant elevation of a terrain map.

To be more precise, the differential operator d acting upon f maps a scalar field to a covector field, which is analogous to how the gradient operator ∇ acting upon f maps a scalar field maps to a vector field. As with vector fields, a covector field assigns a covector to every point in space. To compute the directional derivative with respect to a constant vector $\mathbf{v}$ at some point (x, y) we evaluate the covector field and get a stack of parallel lines (or parallel planes in 3-D space) and then make a count (in a continuous sense) of the number of lines that vector $\mathbf{v}$ pierces.

Returning to the example, if we wish to evaluate the directional derivative at point $(1, -1/2)$, we evaluate the covector field at that point

$$df = \begin{bmatrix} 1 & -1 \end{bmatrix}.$$

Suppose now this covector acts upon a vector $\mathbf{v}_1 = 2\mathbf{e}_x + \mathbf{e}_y$. This gives

$$df(\mathbf{v}_1) = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = 1.$$

This is illustrated graphically in Fig. 4.7. The vector $\mathbf{v}_1$ pierces one covector line in the positive direction and therefore the directional derivative of $\mathbf{v}_1$ at the point $(1, -1/2)$ point is 1. The sign is positive since the vector is moving along with the direction indicated by the covector planes (this could be thought of as increasing elevation).

We can also take the directional derivative along $\mathbf{v}_2 = -1/2\mathbf{e}_x + 3\mathbf{e}_y$ at $(1, -1/2)$. The result of this multiplication gives

$$df(\mathbf{v}_2) = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} -1/2 \\ 3 \end{bmatrix} = -\frac{7}{2}.$$

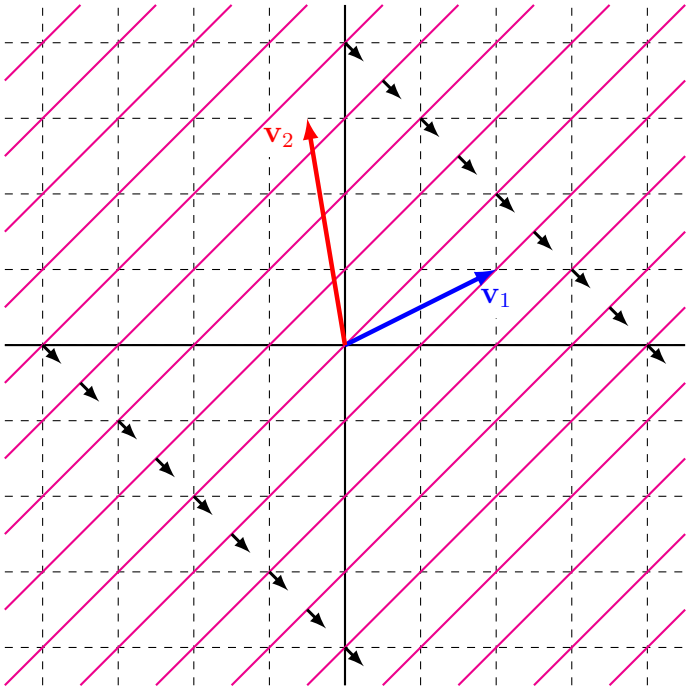

Figure 4.7: Depiction of the directional derivative using the covector field $df = dx + 2ydy$ evaluated at $(1, -1/2)$ with a vectors $\mathbf{v}_1 = 2\mathbf{e}_x + \mathbf{e}_y$ and $\mathbf{v}_2 = -1/2\mathbf{e}_x + 3\mathbf{e}_y$.

This is because the vector goes through 3.5 covector lines, but in the direction of decreasing contour lines.

The covector field is often referred to as a differential form, or more precisely a differential 1-form. The concept of a 1-form is used to define integration along a curved path, where each differential unit of length traversed along the path is multiplied by the covector field or 1-form at that path. While we will not go into this in these notes, we can also write 2-forms and 3-forms, which are used to define surface and volume integrals.

4.5 Line Integrals

Some problems in physics and engineering involve evaluating integrals of functions along trajectories. Herein we refer to the idea as taking a line integral and the trajectory is given by path $P(a, b)$ with endpoints given by locations a and b along the path. We may then take the integral of some scalar field $f(x, y, z)$. It is common to parameterize the path with some common variable t such that the variables (x, y, z) can be written in terms of the parameter t, which we often take to be time. The integrals in terms of that parameter t can be written as

$$\int_{P(a,b)} f(t)dt. \tag{4-115}$$

In many applications (e.g., finding the work done by a force field) we take the line integral with respect to a vector field, which is given as

$$\int_{P(a,b)} \mathbf{F} \cdot d\boldsymbol{\ell}. \tag{4-116}$$

Here $d\boldsymbol{\ell}$ is the differential line vector, which is discussed in the subsequent section.

It is also conventional to take integrals over closed paths, i.e., those that start and end at the same point. For these cases, we denote this with a circle in the integral as such

$$\oint_{P} \mathbf{F} \cdot d\boldsymbol{\ell}. \tag{4-117}$$

4.5.a Differential Line Vector and Length

The differential line vector for a general orthogonormal coordinate system (u, v, w) in terms of the scale factors is given by

$$d\boldsymbol{\ell} = h_u du\hat{\mathbf{u}} + h_v dv\hat{\mathbf{v}} + h_w dw\hat{\mathbf{w}}. \tag{4-118}$$

Inserting the scale factors for Cartesian, cylindrical, and spherical coordinates gives

$$d\boldsymbol{\ell} = dx\hat{\mathbf{i}} + dy\hat{\mathbf{j}} + dz\hat{\mathbf{k}}, \tag{4-119a}$$
$$d\boldsymbol{\ell} = dr\hat{\mathbf{r}} + rd\theta\hat{\boldsymbol{\theta}} + dz\hat{\mathbf{z}}, \tag{4-119b}$$
$$d\boldsymbol{\ell} = dr\hat{\mathbf{r}} + rd\theta\hat{\boldsymbol{\theta}} + r\sin\theta d\phi\hat{\boldsymbol{\phi}}. \tag{4-119c}$$

The differential-length squared for a general orthogonal coordinate system (u, v, w) in terms of the scale factors is given by the dot product of the differential line vector with itself

$$(d\ell)^2 = d\boldsymbol{\ell} \cdot d\boldsymbol{\ell} = (h_u du)^2 + (h_v dv)^2 + (h_w dw)^2. \tag{4-120}$$

Inserting the scale factors for Cartesian, cylindrical, and spherical coordinates gives

$$(d\ell)^2 = (dx)^2 + (dy)^2 + (dz)^2, \tag{4-121a}$$
$$(d\ell)^2 = (dr)^2 + r^2(d\theta)^2 + (dz)^2, \tag{4-121b}$$
$$(d\ell)^2 = (dr)^2 + r^2(d\theta)^2 + r^2\sin^2\theta(d\phi)^2. \tag{4-121c}$$

The differential unit of length $d\ell$ is therefore the square root of $(d\ell)^2$.

4.5.b Example: Moment of Inertia of a Circular Arc

Suppose we have a wire of fixed radius R constant density ρ shaped into a circular arc. The orientation of the wire is given in Fig. 4.8. The wire is oriented such that the wire starts at an angle θ with respect to the x-axis and ends at an angle ϕ as shown.

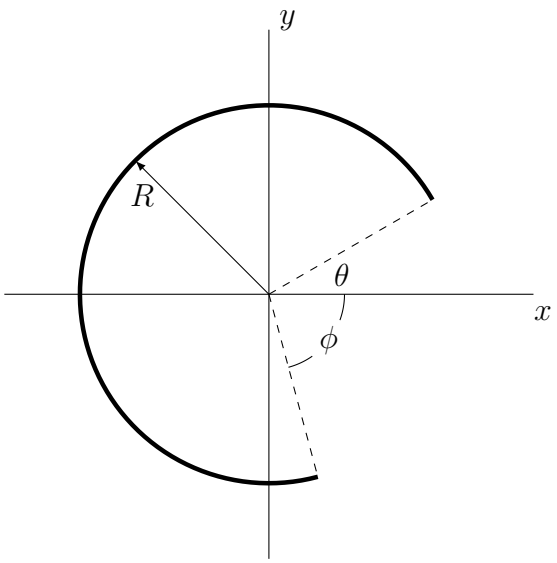

Figure 4.8: Illustration of a circular arc wire.

The moment of inertia describes the resistance of an object to rotation about a specific axis. Using the Cartesian axes, the moments of inertia are given by

$$I_x = \int_{P(a,b)} (y^2 + z^2)\rho(x, y, z)dt, \tag{4-122a}$$

$$I_y = \int_{P(a,b)} (x^2 + z^2)\rho(x, y, z)dt, \tag{4-122b}$$

$$I_z = \int_{P(a,b)} (x^2 + y^2)\rho(x, y, z)dt. \tag{4-122c}$$

We may parameterize the circular arc as follows:

$$x = R\cos(t), \tag{4-123a}$$

$$y = R\sin(t), \tag{4-123b}$$

$$z = 0, \tag{4-123c}$$

where the parameter t ranges from θ to $2\pi - \phi$. Inserting the parameterization into the moment of inertia equations and carrying out the integration gives

$$I_x = \rho R^2 \int_\theta^{2\pi-\phi} \sin^2(t)dt = \frac{\rho R^2}{2}\left[2\pi - \theta - \phi + \frac{1}{2}\left(\sin(2\theta) + \sin(2\phi)\right)\right], \tag{4-124a}$$

$$I_y = \rho R^2 \int_\theta^{2\pi-\phi} \cos^2(t)dt = \frac{\rho R^2}{2}\left[2\pi - \theta - \phi - \frac{1}{2}\left(\sin(2\theta) + \sin(2\phi)\right)\right], \tag{4-124b}$$

$$I_z = \rho R^2 \int_\theta^{2\pi-\phi} dt = \rho R^2(2\pi - \phi - \theta). \tag{4-124c}$$

4.5.c Example: Work on a Charged Particle by an Electric Field

Suppose we have a point charged particle (e.g., an electron) with charge q moving in a straight trajectory. The charged particle is in the presence of an electric field generated by a stationary point charge Q at the origin. Coulomb's law for electrostatic point charges in spherical coordinates relates the

$$\mathbf{F} = \frac{1}{4\pi\epsilon_0}\frac{qQ}{r^2}\hat{\mathbf{r}}, \tag{4-125}$$

where r is the distance between the two point charges and ϵ_0 is a constant called the electric permittivity of free space. The work done by a force is given by

$$W = \int_{P[a,b]} \mathbf{F} \cdot d\boldsymbol{\ell}. \tag{4-126}$$

The most natural coordinate system to use in this case is the spherical coordinate system, for which

$$d\boldsymbol{\ell} = dr\hat{\mathbf{r}} + rd\theta\hat{\boldsymbol{\theta}} + r\sin\theta d\phi\hat{\boldsymbol{\phi}}.$$

Taking the dot product gives the work as

$$W = \frac{qQ}{4\pi\epsilon_0}\int_{P[a,b]}\frac{dr}{r^2}. \tag{4-127}$$

Since the problem is spherically symmetric, we are free to orient the coordinate system as we wish. The particle moves in a straight-line trajectory, and we can take this to be the z direction with x and y fixed. The trajectory starts at some location $z = a$ and moves to $z = b$. In spherical coordinates

$$r = \sqrt{x^2 + y^2 + z^2}.$$

Differentiating with respect to z

$$dr = \frac{z}{\sqrt{x^2 + y^2 + z^2}}dz. \tag{4-128}$$

Inserting this into the equation for work gives

$$W = \frac{qQ}{4\pi\epsilon_0}\int_a^b \frac{z}{(x^2 + y^2 + z^2)^{3/2}}dz. \tag{4-129}$$

Evaluating the integral gives

$$W = \frac{qQ}{4\pi\epsilon_0}\left[\frac{1}{\sqrt{x^2 + y^2 + a^2}} - \frac{1}{\sqrt{x^2 + y^2 + b^2}}\right]. \tag{4-130}$$

4.5.d Example: Circumference of an Ellipse

To find the circumference of an ellipse, we derive an equation for the arc length. The differential length squared in Cartesian coordinates is

$$(d\ell)^2 = (dx)^2 + (dy)^2. \tag{4-131}$$

We may apply the chain rule to dx and dy to put in terms of some parameter t:

$$(d\ell)^2 = \left[\left(\frac{\partial x}{\partial t}\right)^2 + \left(\frac{\partial y}{\partial t}\right)^2\right](dt)^2. \tag{4-132}$$

Taking the square root and integrating gives the equation for the arc length

$$\ell = \int_{P(a,b)} \sqrt{\left(\frac{\partial x}{\partial t}\right)^2 + \left(\frac{\partial y}{\partial t}\right)^2}\, dt. \tag{4-133}$$

The ellipse has semi-major axis along x with length $2a$ and the semi-minor axis along y with length $2b$. A valid parameterization of the ellipse is

$$x = a\sin t, \tag{4-134a}$$
$$y = b\cos t. \tag{4-134b}$$

where t ranges from 0 to 2π. Taking the derivatives with respect to t gives

$$\frac{\partial x}{\partial t} = a\cos t, \tag{4-135a}$$
$$\frac{\partial y}{\partial t} = -b\sin t. \tag{4-135b}$$

Inserting these into the equation for arc length gives

$$\ell = \int_0^{2\pi} \sqrt{a^2\cos^2 t + b^2\sin^2 t}\, dt. \tag{4-136}$$

Using the trigonometric identity $\sin^2 t + \cos^2 t = 1$, we can write

$$\ell = \int_0^{2\pi} \sqrt{a^2(1-\sin^2 t) + b^2\sin^2 t}\, dt.$$

Rearranging and factoring out a^2 gives

$$\ell = a\int_0^{2\pi} \sqrt{1 - \left(1 - \frac{b^2}{a^2}\right)\sin^2 t}\, dt.$$

We can then write the term in the square root as the eccentricity

$$\epsilon = \sqrt{1 - \frac{b^2}{a^2}} \tag{4-137}$$

to get

$$\ell = a \int_0^{2\pi} \sqrt{1 - \epsilon^2 \sin^2 t} dt. \tag{4-138}$$

Unfortunately, this integral does not have an analytic form in terms of the typical functions. Nonetheless, we can proceed if we permit ourselves to use *special functions*. Because of the symmetry of an ellipse, we can write the integral as

$$\ell = 4a \int_0^{\pi/2} \sqrt{1 - \epsilon^2 \sin^2 t} dt = 4aE(\epsilon). \tag{4-139}$$

Here $E(\epsilon)$ is called the *complete elliptic integral of the second kind* that is defined as

$$E(x) = \int_0^{\pi/2} \sqrt{1 - x^2 \sin^2 t} dt. \tag{4-140}$$

Special functions such as the elliptic integrals are available in most mathematical software and there are standard numerical libraries to compute them for most programming languages.

4.6 Surface Integrals

In addition to line integrals, we often encounter surface integrals as well. Surface integrals often arise when we study flow rates of physical quantities (e.g., fluid mass, thermal energy, electric fields, particles) across interfaces. Surface integrals are directly connected to both line and volume integrals. Later in this chapter, we will show Stokes theorem, which relates the line integral around a closed path to a surface integral of a physical quantity crossing that surface, and the divergence theorem, which relates the flow rate across a closed surface to the total production (plus loss) rate inside the bounded volume.

The most common form of surface integral takes the form

$$\int_S \mathbf{F} \cdot d\mathbf{S}. \tag{4-141}$$

Here $d\mathbf{S}$ is called the differential surface vector, which will be discussed in the next section. The surface integral is an integration over two dimensions and

We often also encounter integrals over closed, convex surfaces. For this case we denote, as with the line integral, the integral with a circle:

$$\oint_S \mathbf{F} \cdot d\mathbf{S}. \tag{4-142}$$

4.6.a Differential Surface Vector and Area

We define the differential surface vector to be

$$d\mathbf{S} = \hat{\mathbf{n}} dS. \tag{4-143}$$

Here dS is the scalar differential area of a surface element, and $\hat{\mathbf{n}}$ is the unit vector normal to the surface at some point on the surface. Note that these quantities generally often upon the location of the surface.

In general orthogonal (u, v, w) coordinates using the scale factors, we can define the differential surface vector in the u-v plane as

$$d\mathbf{S} = (h_u du\hat{\mathbf{u}}) \times (h_v dv\hat{\mathbf{v}}). \tag{4-144}$$

Because the unit vectors are orthogonal, $\hat{\mathbf{u}} \times \hat{\mathbf{v}} = \hat{\mathbf{w}}$. Therefore, the differential surface vectors is also

$$d\mathbf{S} = h_u h_v du dv \hat{\mathbf{w}}. \tag{4-145}$$

The other two orientations of the differential surface vectors can be found by interchanging the u, v, and w coordinates. It follows that the scalar differential area is simply

$$dS = h_u h_v du dv. \tag{4-146}$$

The differential surface vectors in Cartesian coordinates in the x-y, x-z, and y-z planes are, respectively:

$$d\mathbf{S} = dxdy\hat{\mathbf{k}}, \tag{4-147a}$$
$$d\mathbf{S} = dxdz\hat{\mathbf{j}}, \tag{4-147b}$$
$$d\mathbf{S} = dydz\hat{\mathbf{i}}. \tag{4-147c}$$

In cylindrical coordinates, the differential surface vectors in Cartesian coordinates in the r-θ, r-z, and θ-z planes are:

$$d\mathbf{S} = rdrd\theta\hat{\mathbf{z}}, \tag{4-148a}$$
$$d\mathbf{S} = drdz\hat{\boldsymbol{\theta}}, \tag{4-148b}$$
$$d\mathbf{S} = rd\theta dz\hat{\mathbf{r}}. \tag{4-148c}$$

Finally, in spherical coordinates, the differential surface vectors in Cartesian coordinates in the r-θ, r-ϕ, and θ-ϕ planes are:

$$d\mathbf{S} = rdrd\theta\hat{\boldsymbol{\phi}}, \tag{4-149a}$$
$$d\mathbf{S} = r\sin\theta drd\phi\hat{\boldsymbol{\theta}}, \tag{4-149b}$$
$$d\mathbf{S} = r^2\sin\theta d\theta d\phi\hat{\mathbf{r}}. \tag{4-149c}$$

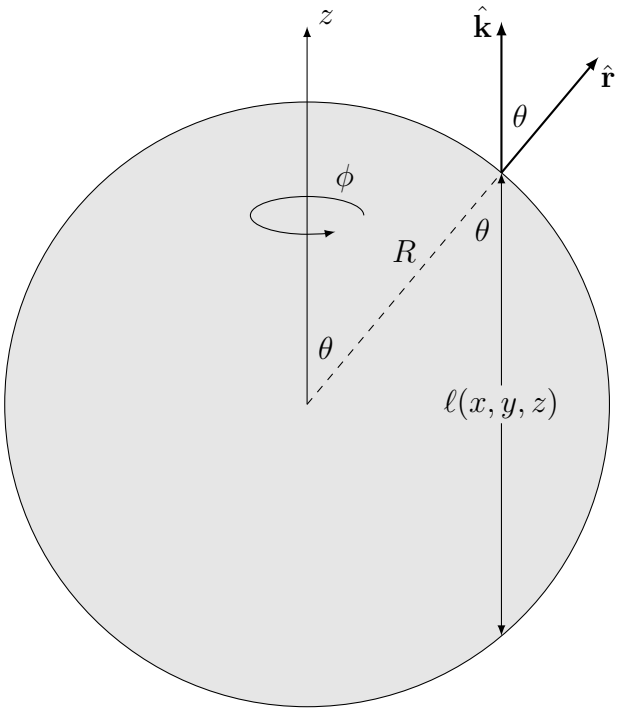

Figure 4.9: Illustration of the problem setup for the escape probability from a radioactive sphere.

4.6.b Example: Photon Escape Probability

Suppose we have a uniform sphere of radius R that emits photon (gamma) radiation uniformly and isotropically (equal probability in all directions) within a sphere illustrated in Fig. 4.9. The sphere consists of a high-Z material such that photoelectric absorption is dominant and any photon that interacts with an atom is immediately absorbed. We wish to calculate the probability that an emitted photon escapes the sphere to assess its radiological hazard.

The escape probability can be written as the ratio of the rate photons escape the sphere divided by the total rate they are emitted in the sphere. If the emission rate density is Q_0 photons per unit volume per unit time and the sphere has a volume V, then the total emission rate is $Q_0 V$. The rate photons escape is given by

$$\int_S \mathbf{J} \cdot d\mathbf{S}. \tag{4-150}$$

Here $\mathbf{J}$ is a vector field describing the photon current. Because of spherical symmetry, we may always work the problem such that the z axis is always aligned in the direction of photon flight. When we do this, we may express the photon current vector field as

$$\mathbf{J} = \frac{Q_0}{\Sigma_t} \left[1 - \exp\left(-\Sigma_t \ell(x, y, z)\right)\right] \hat{\mathbf{k}}. \tag{4-151}$$

Here $\ell(x, y, z)$ is the length of a chord along the z direction from one end of the sphere to the other. The escape probability becomes

$$\begin{aligned} p_{esc} &= \frac{1}{Q_0 V} \int_S \frac{Q_0}{\Sigma_t} \left[1 - \exp\left(-\Sigma_t \ell(x, y, z)\right)\right] \hat{\mathbf{k}} \cdot d\mathbf{S} \\ &= \frac{1}{\Sigma_t V} \int_S \left[1 - \exp\left(-\Sigma_t \ell(x, y, z)\right)\right] \hat{\mathbf{k}} \cdot d\mathbf{S} \end{aligned} \tag{4-152}$$

Since all of the photons travel in the $+z$ direction, the photons only leak out of the northern hemisphere of the sphere. For this, we will integrate over the θ-ϕ plane in spherical coordinates. The integral over the northern hemisphere becomes:

$$p_{esc} = \frac{1}{\Sigma_t V} \int_0^{2\pi} \int_0^{\pi/2} \left[1 - \exp\left(-\Sigma_t \ell(x, y, z)\right)\right] \hat{\mathbf{k}} \cdot (R^2 \sin\theta \hat{\mathbf{r}}) d\theta d\phi. \tag{4-153}$$

From the diagram, we can see that

$$\hat{\mathbf{k}} \cdot \hat{\mathbf{r}} = \cos\theta. \tag{4-154}$$

The chord length is

$$\ell(x, y, z) = 2z = 2R\cos\theta. \tag{4-155}$$

Inserting this into the integral gives

$$p_{esc} = \frac{R^2}{\Sigma_t V} \int_0^{2\pi} \int_0^{\pi/2} \left[1 - \exp\left(-2\Sigma_t R\cos\theta\right)\right] \cos\theta \sin\theta d\theta d\phi. \tag{4-156}$$

The azimuthal angle can be integrated trivially:

$$p_{esc} = \frac{2\pi R^2}{\Sigma_t V} \int_0^{\pi/2} \left[1 - \exp\left(-2\Sigma_t R\cos\theta\right)\right] \cos\theta \sin\theta d\theta. \tag{4-157}$$

To carry out the polar integral, introduce the transformation

$$\mu = \cos\theta, \tag{4-158a}$$

$$d\mu = -\sin\theta d\theta. \tag{4-158b}$$

After making this transformation:

$$p_{esc} = \frac{2\pi R^2}{\Sigma_t V} \int_0^1 \left[1 - \exp\left(-2\Sigma_t R\mu\right)\right] \mu d\mu. \tag{4-159}$$

After carrying out the integrals (integrate by parts) we get

$$p_{esc} = \frac{2\pi R^2}{\Sigma_t V} \left[\frac{1}{2} - \frac{1 - (1 + 2\Sigma_t R)e^{-2\Sigma_t R}}{4\Sigma_t^2 R^2}\right]. \tag{4-160}$$

After some rearrangement and expanding out the volume of a sphere,

$$p_{esc} = \frac{3}{8\Sigma_t^3 R^3}\left[2\Sigma_t^2 R^2 - 1 + (1 + 2\Sigma_t R)e^{-2\Sigma_t R}\right]. \tag{4-161}$$

It is often convenient to express the result in terms of the optical radius

$$\tau = \Sigma_t R, \tag{4-162}$$

which gives the final result

$$p_{esc} = \frac{3}{8\tau^3}\left[(1 + 2\tau)e^{-2\tau} + 2\tau^2 - 1\right]. \tag{4-163}$$

4.7 Volume Integrals

The final related quantity to vector analysis is the concept of a volume integral. Most commonly, volume integrals are performed upon some density field of some quantity per unit volume to find the number of that quantity within some region. This is given as

$$\int_V \rho(\boldsymbol{x})dV. \tag{4-164}$$

Here dV is the differential volume, which gets expanded out as a triple differential.

As mentioned in the previous section on surface integrals, the volume integral of the divergence of a vector field can also be related to the surface integral of the net outgoing flux of a physical quantity for that same field. This will be discussed later in the chapter.

4.7.a Differential Volume

The differential volume dV is a scalar quantity that can be obtained generally by taking the determinant of the Jacobian matrix

$$dV = \det(\mathbf{J})\, dudvdw. \tag{4-165}$$

An equivalent formulation for orthogonal coordinate systems is the product of the scale factors times the differential elements of each coordinate:

$$dV = h_u h_v h_w dudvdw. \tag{4-166}$$

For Cartesian, cylindrical, and spherical coordinates, the differential volume element becomes

$$dV = dxdydz, \tag{4-167a}$$
$$dV = rdrd\theta dz, \tag{4-167b}$$
$$dV = r^2 \sin\theta drd\theta d\phi. \tag{4-167c}$$

4.7.b Example: Enclosed Charge in a Cylinder

In electrostatics calculations of the electric field, we often compute the charge enclosed by a given volume Q_{enc} with a prescribed charge density ρ. The approach is to simply integrate the charge density over the region.

Suppose we have a charge density given in a cylindrical region:

$$\rho(r,\theta,z) = \begin{cases} \rho_0 e^{-\kappa r^2}, & 0 \le r \le R,\ 0 \le \theta < 2\pi,\ 0 \le z \le H \\ 0, & \text{otherwise} \end{cases}. \tag{4-168}$$

The charge enclosed by this region may be obtained by integrating over the region (or any region larger than it). This is:

$$\begin{aligned} Q_{enc} &= \int_V \rho_0 e^{-\kappa r^2} dV \\ &= \int_0^H \int_0^{2\pi} \int_0^R \rho_0 e^{-\kappa r^2} r dr d\theta dz. \end{aligned} \tag{4-169}$$

Carrying out the θ and z integrals is trivial, yielding

$$Q_{enc} = 2\pi H \rho_0 \int_0^R r e^{-\kappa r^2} dr. \tag{4-170}$$

Performing the integral over r using integration by parts gives the final result:

$$Q_{enc} = \frac{\pi H \rho_0}{\kappa}(1 - e^{-\kappa R^2}). \tag{4-171}$$

4.8 Integral Theorems

In vector calculus we have three major integral theorems that, in many circumstances, allow us to greatly simplify problems or show equivalences between quantities of interest. The first is the gradient theorem, which is useful in defining conservative forces or scalar potential functions, which allows us to take a line integral of a vector field as simply the scalar potential evaluated at the endpoints. The second is the divergence theorem, which relates volume and surface integrals. And the third is Stokes' theorem, which relates surface and line integrals.

4.8.a Gradient Theorem

The gradient theorem, which is analogous to the second fundamental theorem of calculus, but allied to line integrals states that

$$\int_{P(a,b)} \nabla f \cdot d\boldsymbol{\ell} = f(P(b)) - f(P(a)). \tag{4-172}$$

This states that if a vector field is described by a gradient of a function, then we can assert that the line integral only depends upon the endpoints. Note that since the

curl of the gradient is always zero, this implies that if we have some vector field $\mathbf{F}$ and if $\nabla \times \mathbf{F} = \mathbf{0}$, then the line integral of that vector field is path independent. We state that the vector field $\mathbf{F}$ is conservative. This result will be used in our discussion on potential functions in the following section.

4.8.b Divergence Theorem

The divergence theorem relates volume and surface integrals. It states that the divergence of some vector field integrated over a convex volume is equal to the surface integral of the vector field over the surface bounding that volume. In other words,

$$\oint_S \mathbf{F} \cdot d\mathbf{S} = \int_V \nabla \cdot \mathbf{F} dV. \tag{4-173}$$

The idea with the divergence theorem starts from the interpretation of $(\nabla \cdot \mathbf{F})dV$ is the net outflow of a physical quantity given by vector field $\mathbf{F}$ out of some convex differential volume dV about (x, y, z). For a six sided cubic differential volume, we may express this as:

$$(\nabla \cdot \mathbf{F})dV_1 = \text{ outflow across faces of } dV_1 \ - \ \text{inflow across faces of } dV_1 \ . \tag{4-174}$$

If we then stack a differential volume dV_2 next to dV_1 such that the face $\textcircled{2}$ of dV_1 is adjacent to $\textcircled{1}$ of dV_2 and add the two together we get

$$\begin{aligned}
&(\nabla \cdot \mathbf{F})dV_1 + (\nabla \cdot \mathbf{F})dV_2 \\
&= \text{ outflow across face } \textcircled{2} \text{ of } dV_1 \\
&- \text{ inflow across face } \textcircled{2} \text{ of } dV_1 \\
&+ \text{ outflow across exterior faces } dV_1 \ - \ \text{inflow across exterior faces of } dV_1 \\
&+ \text{ outflow across face } \textcircled{1} \text{ of } dV_2 \\
&- \text{ inflow across face } \textcircled{1} \text{ of } dV_2 \\
&+ \text{ outflow across exterior faces } dV_2 \ - \ \text{inflow across exterior faces of } dV_2 \ .
\end{aligned} \tag{4-175}$$

Since there are no particles created on the face, we can assert the continuity condition that

$$\text{outflow across face } \textcircled{2} \text{ of } dV_1 \ = \ \text{inflow across face } \textcircled{1} \text{ of } dV_2 \ , \tag{4-176a}$$

$$\text{inflow across face } \textcircled{2} \text{ of } dV_1 \ = \ \text{outflow across face } \textcircled{1} \text{ of } dV_2 \ . \tag{4-176b}$$

Therefore, the terms cancel and we are left with

$$\begin{aligned}
&(\nabla \cdot \mathbf{F})dV_1 + (\nabla \cdot \mathbf{F})dV_2 \\
&= \text{ outflow across exterior faces } dV_1 \ - \ \text{inflow across exterior faces of } dV_1 \\
&+ \text{ outflow across exterior faces } dV_2 \ - \ \text{inflow across exterior faces of } dV_2 \ .
\end{aligned} \tag{4-177}$$

If we add another set of differential volumes such that the net volume remains convex so as not to permit reentry, we have the same result that only the exterior faces remain. Since adding over numerous differentials limits to an integral, we get the result that

$$\begin{aligned}\lim_{k\to\infty}\sum_k(\nabla\cdot\mathbf{F})dV_k &= \int_V \nabla\cdot\mathbf{F}dV \\ &= \text{ outflow across the exterior boundary of } V \\ &- \text{ inflow across the exterior boundary of } V\ .\end{aligned} \tag{4-178}$$

Now, if we inspect the surface integral, we note that the differential surface vector $d\mathbf{S} = \hat{\mathbf{n}}dS$, where the normal vector $\hat{\mathbf{n}}$ is defined by convention to be pointed outward from the surface. When $\mathbf{F}\cdot\hat{\mathbf{n}} > 0$ we have outward directed flow (positive) across the surface and when $\mathbf{F}\cdot\hat{\mathbf{n}} < 0$ we have inward directed flow (negative) across, which is equivalent to the volume integral of the divergence.

4.8.c Stokes' (Curl) Theorem

Stoke's theorem states the line integral of a vector field along a closed path is equal to the curl of that vector field integrated over some some smooth surface that is bounded by the path. This is stated mathematically as

$$\oint_P \mathbf{F}\cdot d\boldsymbol{\ell} = \int_S(\nabla\times\mathbf{F})\cdot d\mathbf{S}. \tag{4-179}$$

While the idea is more abstract than the divergence theorem, the argument proceeds along similar lines. The curl of a vector field gives the local circulation of some differential patch of area about some point (x, y, z) with an outward normal vector $\hat{\mathbf{n}}$. This circulation is often represented with some outward unit normal that describes counter-clockwise rotation with respect to that outward normal. If we take another path of area adjacent, we have another counter-clockwise rotation. As with the divergence theorem, the counter-clockwise rotation of one side is canceled by the counter-clockwise rotation of the other side since the rotation vectors are opposite of one another. This leaves only the rotation about the exterior of the combined surface. As with the divergence theorem, if we continue to stack patches of adjacent area and add them up, the limiting result is a line integral over the bounding curve.

4.8.d Example: Maxwell's Equations of Electromagnetism

Maxwell's equations of electromagnetism can be written in equivalent differential and integral forms. Depending upon the context, working with one of these formulations over the other may produce a problem that is easier to solve.

The first of Maxwell's equations states that the divergence of the electric field is equal to the local charge density divided by the electric permittivity:

$$\nabla\cdot\mathbf{E} = \frac{\rho}{\epsilon_0}. \tag{4-180}$$

If we integrate the left hand side over a closed, convex volume, we can write

$$\int_V (\nabla \cdot \mathbf{E}) dV = \int_V \frac{\rho}{\epsilon_0} dV = \frac{Q_{enc}}{\epsilon_0}. \tag{4-181}$$

The right-hand side becomes the total net charge enclosed by the volume. By applying the divergence theorem we get

$$\int_V (\nabla \cdot \mathbf{E}) dV = \oint_S \mathbf{E} \cdot d\mathbf{S} = \frac{Q_{enc}}{\epsilon_0}. \tag{4-182}$$

Faraday's law states that the curl of the electric field is equal to the negative time rate of change of the magnetic field:

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}. \tag{4-183}$$

Taking the integral over some non-closed surface of this gives

$$\int_S (\nabla \times \mathbf{E}) \cdot d\mathbf{S} = -\frac{\partial}{\partial t} \int_S \mathbf{B} \cdot d\mathbf{S} = -\frac{\partial \Phi_B}{\partial t}. \tag{4-184}$$

The right-hand side becomes the time-rate of change of the magnetic flux. Applying Stokes' theorem gives

$$\int_S (\nabla \times \mathbf{E}) \cdot d\mathbf{S} = \oint_P \mathbf{E} \cdot d\boldsymbol{\ell} = -\frac{\partial \Phi_B}{\partial t}. \tag{4-185}$$

This states that the integral over a closed path of the electric field vector is equivalent to minus to the time rate of change of the magnetic flux across any smooth surface bounded by that curve.

The nonexistence of magnetic monopoles (point sources of magnetic charge) implies that the divergence of the magnetic field is zero:

$$\nabla \cdot \mathbf{B} = 0. \tag{4-186}$$

As with Gauss' law for electric fields, we integrate this equation over the volume and apply the divergence theorem:

$$\int_V (\nabla \cdot \mathbf{B}) dV = \oint_S \mathbf{B} \cdot d\mathbf{S} = 0. \tag{4-187}$$

This states that if we integrate the magnetic field over any closed surface, we will get zero. In other words, there is never a net outward directed magnetic flux out of a volume.

Finally, we have Ampere's law (with Maxwell's correction) that relates the curl of the magnetic field with the current and time rate of change of the electric field:

$$\nabla \times \mathbf{B} = \mu_0 \left(\mathbf{J} + \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right). \tag{4-188}$$

Here μ_0 is the magnetic permeability of free space. If, as we did with Faraday's law, we proceed to take the surface integral over some non-closed surface, we get

$$\int_S (\nabla \times \mathbf{B}) \cdot d\mathbf{S} = \mu_0 \left(\int_S \mathbf{J} \cdot d\mathbf{S} + \epsilon_0 \frac{\partial}{\partial t} \left(\int_S \mathbf{E} \cdot d\mathbf{S} \right) \right). \tag{4-189}$$

Applying Stoke's theorem gives

$$\int_S (\nabla \times \mathbf{B}) \cdot d\mathbf{S} = \oint_P \mathbf{B} \cdot d\boldsymbol{\ell} = \mu_0 \left(\int_S \mathbf{J} \cdot d\mathbf{S} + \epsilon_0 \frac{\partial \Phi_E}{\partial t} \right). \tag{4-190}$$

This states that the line integral over a closed path of the magnetic field is equal to the sum of the current flowing across the surface plus the time rate of change of the electric flux.

4.9 Potential Functions

A useful mathematical object for describing vector fields is the potential function. These come in two flavors, a scalar potential Φ and a vector potential $\mathbf{A}$. Note that while these are sometimes similar to the concept of potential energy from classical mechanics, the two concepts are fundamentally different and should not be confused. This is an unfortunate naming convention, but is unavoidable. The advantage of casting vector fields in terms of potentials is that they can make the mathematics of performing line and surface integrals much simpler. In the context of line integration, it also allows us to express the idea in terms of curves within covector fields.

4.9.a Helmholtz Decomposition

The Helmholtz decomposition theorem (also called the fundamental theorem of vector analysis) states that any vector field that is at least differentiable twice everywhere (corresponding to most vector fields in physics) may be written as the the sum of the gradient of a scalar function and the curl of another function:

$$\mathbf{F} = -\nabla\Phi + \nabla \times \mathbf{A}. \tag{4-191}$$

Here we call Φ the scalar potential function and $\mathbf{A}$ the vector potential function. Note that the minus sign on the $\nabla\Phi$ term is purely because of an arbitrary convention and not inherent in the mathematics.

Why this decomposition is useful is that we can often write differential equations or integrals involving unknown vector fields in terms of the potential functions instead. The resulting differential equations or integrals are often much easier to solve. Once we know the potential functions, we can use the relationship to obtain the unknown vector field.

4.9.b Scalar Potential Function

Vector fields satisfying $\nabla \times \mathbf{F} = \mathbf{0}$ are called irrotational or conservative and can be described using the scalar potential Φ alone, i.e.,

$$\mathbf{F} = -\nabla\Phi, \quad \text{if } \nabla \times \mathbf{F} = \mathbf{0}. \tag{4-192}$$

Examples of irrotational vector fields include the gravitational field and the electric field in the absence of time-varying magnetic fields.

Then the curl of a vector field zero, we can also state that the line integral depends only upon the potential function at the endpoints:

$$\int_{P(a,b)} \mathbf{F} \cdot d\boldsymbol{\ell} = \Phi(a) - \Phi(b). \tag{4-193}$$

Correspondingly, it follows that if the path is over a closed loop then,

$$\oint_{P} \mathbf{F} \cdot d\boldsymbol{\ell} = 0. \tag{4-194}$$

To demonstrate this, the line integral may be written as

$$-\int_{P(a,b)} \nabla\Phi \cdot d\boldsymbol{\ell}. \tag{4-195}$$

Suppose the path is parameterized by t. We can then use the chain rule to write

$$-\int_{P(a,b)} \nabla\Phi \cdot \frac{d\boldsymbol{\ell}}{dt} dt. \tag{4-196}$$

The gradient of the potential field can be expanded in Cartesian coordinates as

$$\nabla\Phi = \frac{\partial\Phi}{\partial x}\hat{\mathbf{i}} + \frac{\partial\Phi}{\partial y}\hat{\mathbf{j}} + \frac{\partial\Phi}{\partial x}\hat{\mathbf{k}}. \tag{4-197}$$

(The result applies to any coordinate system, but the choice of Cartesian coordinates simplifies the derivation.) We can then expand the derivative of the line vector with respect to the parameter t ranging along the path from $a \leq t \leq b$ using the multivariable chain rule:

$$\frac{d\boldsymbol{\ell}}{dt} = \frac{\partial x}{\partial t}\frac{\partial \boldsymbol{\ell}}{\partial x} + \frac{\partial y}{\partial t}\frac{\partial \boldsymbol{\ell}}{\partial y} + \frac{\partial z}{\partial t}\frac{\partial \boldsymbol{\ell}}{\partial z}. \tag{4-198}$$

Since $\boldsymbol{\ell}$ represents a displacement vector, its partial derivatives are simply the Cartesian basis vectors:

$$\frac{d\boldsymbol{\ell}}{dt} = \frac{\partial x}{\partial t}\hat{\mathbf{i}} + \frac{\partial y}{\partial t}\hat{\mathbf{j}} + \frac{\partial z}{\partial t}\hat{\mathbf{k}}. \tag{4-199}$$

Evaluating the dot product gives

$$\nabla\Phi \cdot \frac{d\boldsymbol{\ell}}{dt} = \frac{\partial\Phi}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial\Phi}{\partial y}\frac{\partial y}{\partial t} + \frac{\partial\Phi}{\partial z}\frac{\partial z}{\partial t}. \tag{4-200}$$

The right-hand side can be simplified by applying the multivariable chain rule in reverse:

$$\frac{d\Phi}{dt} = \frac{\partial\Phi}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial\Phi}{\partial y}\frac{\partial y}{\partial t} + \frac{\partial\Phi}{\partial z}\frac{\partial z}{\partial t} = \nabla\Phi \cdot \frac{d\boldsymbol{\ell}}{dt}. \tag{4-201}$$

Therefore, we can rewrite the integral as

$$-\int_a^b \frac{d\Phi}{dt} dt. \tag{4-202}$$

The dt can be cancelled and the integral can be carried out trivially:

$$-\int_a^b d\Phi = \Phi(a) - \Phi(b). \tag{4-203}$$

An important point is that the scalar function Φ is not unique. Suppose instead we let

$$\Phi = \Phi_0 + C \tag{4-204}$$

where C is an arbitrary scalar constant. Because the gradient of a constant is the same, we get the same result, i.e.,

$$\mathbf{F} = \nabla\Phi = \nabla(\Phi_0 + C) = \nabla\Phi_0. \tag{4-205}$$

While the potential function is not the potential energy, it does have the same feature (which is the source of much confusion) that where the potential equals zero is entirely arbitrary and defined as a matter of convenience. Regardless of where the zero point is defined, the physics will work out entirely the same.

Before continuing, it may seems a bit odd that a single scalar function Φ can encode enough information to describe the three components of a vector field $\mathbf{F} = -\nabla\Phi$. Note that, however, the F_x, F_y, and F_z components are not independent, because $\nabla \times \mathbf{F} = \mathbf{0}$. In other words,

$$\begin{aligned}
\frac{\partial F_z}{\partial y} &= \frac{\partial F_y}{\partial z}, \\
\frac{\partial F_z}{\partial x} &= \frac{\partial F_x}{\partial z}, \\
\frac{\partial F_y}{\partial x} &= \frac{\partial F_x}{\partial y}.
\end{aligned}$$

4.9.c Relation to Covector Fields

Recall that the line integral of the a conservative force field reduces to

$$\int_{P(a,b)} \mathbf{F} \cdot d\boldsymbol{\ell} = -\int_{P(a,b)} d\Phi = \Phi(a) - \Phi(b). \tag{4-206}$$

Recall that the differential operator d acting upon a scalar field Φ produces a covector field, which can be represented as a series of contour lines of equal potential. To illustrate, let us consider the electric potential from a point charge with charge Q at the origin in spherical coordinates:

$$\Phi = \frac{1}{4\pi\epsilon_0}\frac{Q}{r}. \tag{4-207}$$

Here the potential function is defined to be zero as $r \to \infty$, but we could add or subtract any arbitrary constant and get identical results.

The covector field $d\Phi$ can be computed using the multivariable chain rule

$$\begin{aligned} d\Phi &= \frac{\partial\Phi}{\partial r}dr + \frac{\partial\Phi}{\partial\theta}d\theta + \frac{\partial\Phi}{\partial\phi}d\phi \\ &= -\frac{1}{4\pi\epsilon_0}\frac{Q}{r^2}dr. \end{aligned} \tag{4-208}$$

This covector can be multiplied by a vector of coordinates and set equal to some constant a:

$$\begin{bmatrix} -\dfrac{1}{4\pi\epsilon_0}\dfrac{Q}{r^2} & 0 & 0 \end{bmatrix} \begin{bmatrix} r \\ \theta \\ \phi \end{bmatrix} = -\frac{1}{4\pi\epsilon_0}\frac{Q}{r} = a. \tag{4-209}$$

This gives a series of spherical surfaces of equal potential that become increasingly close together nearing the origin and spacing further apart the further away. Each spherical surface of equal potential also has an outward directed orientation giving the negative direction of increasing potential. A 2-D slice of these surfaces is taken along the origin to make a series of circles with arrows describing the orientation: this is illustrated in Fig. 4.10.

The electric potential is related to the electric field by minus the gradient of the potential

$$\mathbf{E} = -\nabla\Phi. \tag{4-210}$$

The force done on a particle of charge q by an electrostatic field is

$$\mathbf{F} = q\mathbf{E} = -q\nabla\Phi. \tag{4-211}$$

Therefore, the work done by an electrostatic field for the charged particle moving along path $P(a,b)$ is

$$W = -q\int_{P(a,b)} \nabla\Phi\cdot d\boldsymbol{\ell} = -q\int_a^b d\Phi = q\left(\Phi(a) - \Phi(b)\right). \tag{4-212}$$

In a similar manner to when a covector field acts upon a vector and produces a scalar, here the covector field $d\Phi$ acts upon each differential increment of path to produce some differential increment of work and the integral sums over the result to

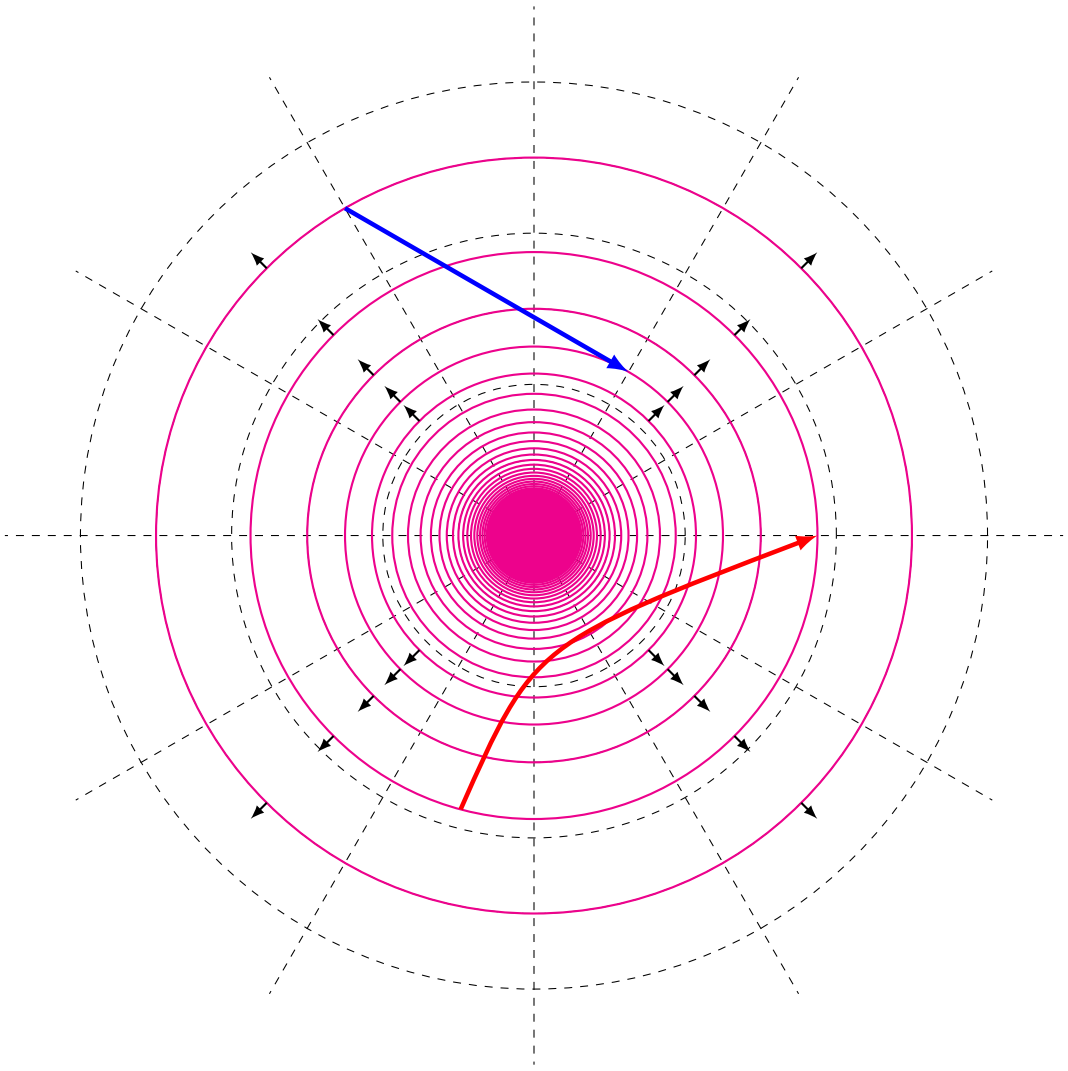

Figure 4.10: Illustration of contour curves for the electrostatic potential from a point charge at the origin and two different paths for another charged particle.

give the total work. In a similar manner, the work (per unit charge q) can be obtained by counting the number of covector contour lines that the path pierces considering the orientation of those contour lines.

The straight line path in blue displayed in Fig. 4.10 shows the work per unit charge q by counting the number of lines crossed is -3. If the two charges q and Q have the same sign, the charges will repel one another and the work done by will be negative since energy must be added by the charge q to reach the surface. Likewise, if the charges have opposite signs, the charge q will be attracted to the origin and the work will be positive.

The curved path in red displayed in Fig. 4.10 has a work per charge of zero since the positive and negative lines crossed are equal and the work cancels. In other words, the particle begins and ends at the same distance away from the origin. For a charge

q with the same sign as Q, the field does work against the particle (negative) initially going inward and then does work (positive) on the particle moving outward.

4.9.d Vector Potential Function

Physical vector field quantities for which the curl is nonzero cannot be described using the gradient of a scalar potential function. Vector fields that are divergence free or solenoidal (e.g., magnetic fields or fluids that are purely rotational) can be described by the curl of a vector potential $\mathbf{A}$. Mathematically,

$$\mathbf{F} = \nabla \times \mathbf{A}, \quad \text{if } \nabla \cdot \mathbf{F} = 0. \tag{4-213}$$

This result arises from the vector identity that the divergence of the curl of a vector field is zero.

Analogous to the scalar potential function, the vector potential function is unique to within a gradient of a scalar field. Suppose we write

$$\mathbf{A} = \mathbf{A}_0 + \nabla m. \tag{4-214}$$

Taking the curl gives

$$\mathbf{F} = \nabla \times \mathbf{A} = \nabla \times \mathbf{A}_0 + \nabla \times \nabla m = \nabla \times \mathbf{A}_0. \tag{4-215}$$

The term $\nabla \times \nabla m = \mathbf{0}$ since the curl of the gradient is always zero.

Unlike with the scalar potential, which is unique to within an additive constant, the addition of a function is not a trivial matter. While the choice of ∇m will not impact the final result, different choices may make the calculations significantly easier or more difficult. The selection of ∇m is called *gauge fixing* and is a challenge to enable theoretical calculations involving physical quantities. Also note that like the scalar potential function, the vector potential function is not by itself a physical quantity; rather, it is a useful mathematical quantity that can be used to obtain physical quantities.

4.9.e Example: Electric and Magnetic Potential Functions

Continuing to apply vector calculus to electromagnetism and Maxwell's equations, we first begin with Gauss' law of electric fields, which, in its differential form, states

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0}.$$

Expanding out the terms yields a first-order partial differential equation that can actually be difficult to solve. Perhaps unexpectedly, since it would result in a second-order partial differential equation, it is easier to solve this problem if we insert the electric potential. For this, we assume that the fields are static, i.e., not time varying, so that $\nabla \times \mathbf{E} = \mathbf{0}$ from Faraday's law. This gives

$$\nabla \cdot (-\nabla \Phi) = -\nabla^2 \Phi = \frac{\rho}{\epsilon_0}. \tag{4-216}$$

This equation is referred to as Poisson's equation. In the case where the charge density ρ is zero and the electric potential is specified at the boundaries, as is often the case, we have Laplace's equation. These second-order partial differential equations have mathematical properties that lend themselves to both analytical and numerical solution. Once we have Φ, we can easily determine the electric field $\mathbf{E}$ by taking $-\nabla\Phi$.

To show the magnetic potential, we begin with Ampere's law with a time-independent electric field

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J}.$$

Since the magnetic field $\mathbf{B}$ is always solenoidal because there are no point charges of magnetism, i.e., $\nabla \cdot \mathbf{B} = 0$, we can write $\mathbf{B} = \nabla \times \mathbf{A}$. Inserting this gives

$$\nabla \times (\nabla \times \mathbf{A}) = \mu_0 \mathbf{J}. \tag{4-217}$$

Now we can write the curl of the curl of a vector field in terms of the vector Laplacian:

$$\nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} = \mu_0 \mathbf{J}. \tag{4-218}$$

At this point, we can decide upon a gauge to simplify the expression. The common choice here is called the *Coulomb gauge*, for which we choose $\mathbf{A}$ such that

$$\nabla \cdot \mathbf{A} = 0. \tag{4-219}$$

This simplifies the equation for a static field as

$$\nabla^2 \mathbf{A} = -\mu_0 \mathbf{J}. \tag{4-220}$$

This equation is a vector form of Poisson's equation and is considerably easier to solve both analytically and numerically compared with Ampere's law. Once we have determined the vector potential, we can determine the magnetic field $\mathbf{B}$ as $\nabla \times \mathbf{A}$.

While here we assumed static electromagnetic fields, the results are not particularly difficult to generalize to the time-dependent equations of electrodynamics. The only major change is the electric field becomes more difficult to calculate, and requires the use of the vector potential $\mathbf{A}$. A comment about the Coulomb gauge for the electrodynamic equations is that the vector potential $\mathbf{A}$ is impacted instantaneously everywhere by changes to the charge or current distributions. This would seemingly violate special relativity, in that information cannot travel faster than the speed of light. This does not mean that the Coulomb gauge is incorrect in any way. Recall that the vector potential function is a purely mathematical quantity, not a physical quantity. Only when we take the curl of this vector potential do we get the magnetic field, which is a physical quantity. While we do not show this here, the magnetic field does preserve all of the elements of causality and is limited by the speed of light.

Chapter 5

Partial Differential Equations

Partial differential equations are fundamental to mathematical modeling in science and engineering. Much of the relevant physical phenomena that we study are described by partial differential equations, with their ordinary differential equation analogs being simplifications to one dimension. This chapter focuses on the analytical techniques for solving partial differential equations and leaves the theoretical considerations and the numerical techniques to another course (although the techniques encountered earlier in the text regarding ordinary differential equations are still applicable).

As with the ordinary differential equation chapter, this chapter on partial differential equations is split into two parts with the first focusing on first-order partial differential equations and the second focusing on second-order partial differential equations. For the first-order partial differential equations, we will learn about the method of characteristics for both the linear and the quasi-linear cases. For the latter, we will study the 1-D inviscid Burger's equation, which is the simplest model of shockwave formation. The second-order partial differential equations covered in this text are the 1-D heat equation, with a first derivative in time and a single second derivative in space, and the 2-D and 3-D Laplace equation in Cartesian and spherical coordinates.

5.1 First-Order Linear PDEs

The method of characteristics is useful for solving equations first-order partial differential equations of the form:

$$a(x,y)\frac{\partial u}{\partial x} + b(x,y)\frac{\partial u}{\partial y} = q(x,y,u). \tag{5-1}$$

Here a and b are coefficients that could be functions of y. The function q could depend upon $u(x,y)$ in addition to x and y. If its dependency upon u is constant or linear in u, the differential equation is linear. If q is a nonlinear function of u, the

partial differential equation is referred to as semi-linear. In the next section, we will discuss a particular case where we allow the coefficients a and b to depend upon u (but not on its derivatives). We call this set of equations quasilinear, and are relevant for describing shock formation, which is necessary for applications involving inertial confinement fusion.

5.1.a Method of Characteristics

The technique for solving first-order linear partial differential equations is called the method of characteristics. The goal with method of characteristics is to reduce a first-order partial differential equation into a system of ordinary differential equations that can hopefully be solved. In this section we will study its application to linear first-order partial differential equations; however, the method is quite versatile and can be applied to fairly general first-order partial differential equations.

To begin, we define an equation describing a 2-D surface in 3-D space

$$f = u(x, y) - u = 0. \tag{5-2}$$

This definition may seem odd at first. Here $u(x, y)$ is the function for a given x and y and u is the value taken when the function $u(x, y)$ is evaluated at those points. Now we find a normal vector to the surface by taking the gradient where the z coordinate is renamed the u coordinate. The normal vector to the surface is therefore

$$\nabla f = \frac{\partial u}{\partial x}\hat{\mathbf{i}} + \frac{\partial u}{\partial y}\hat{\mathbf{j}} - \hat{\mathbf{k}}. \tag{5-3}$$

Now let us take a vector

$$\mathbf{\Omega} = a(x, y, u)\hat{\mathbf{i}} + b(x, y, u)\hat{\mathbf{j}} + q(x, y, u)\hat{\mathbf{k}}. \tag{5-4}$$

Here we put a u dependence upon a and b to show that the method will work for that case. If we then take the directional derivative along $\mathbf{\Omega}$ we get

$$\mathbf{\Omega} \cdot \nabla f = a(x, y, u)\frac{\partial u}{\partial x} + b(x, y, u)\frac{\partial u}{\partial y} - q(x, y, u) = 0. \tag{5-5}$$

The directional derivative is zero because the result is the original partial differential equation. Therefore, the vector $\mathbf{\Omega}$ describes a direction for which the surface is constant. We call this a *characteristic curve*.

We specify a curve parameterized by a variable s, which has a differential length vector as

$$\frac{d\boldsymbol{\ell}}{ds} = \frac{dx}{ds}\hat{\mathbf{i}} + \frac{dy}{ds}\hat{\mathbf{j}} + \frac{du}{ds}\hat{\mathbf{k}}. \tag{5-6}$$

Since this differential length vector runs tangent to the curve $\mathbf{\Omega}$, the cross product between the them must be zero:

$$\mathbf{\Omega} \times \frac{d\boldsymbol{\ell}}{ds} = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ a & b & q \\ \dfrac{dx}{ds} & \dfrac{dy}{ds} & \dfrac{du}{ds} \end{vmatrix} = \mathbf{0}. \tag{5-7}$$

Since each vector component must be zero, this gives the system of equations:

$$b\frac{du}{ds} - q\frac{dy}{ds} = 0, \tag{5-8a}$$
$$q\frac{dx}{ds} - a\frac{du}{ds} = 0, \tag{5-8b}$$
$$a\frac{dy}{ds} - b\frac{dx}{ds} = 0. \tag{5-8c}$$

These equations are satisfied when

$$\frac{dx}{ds} = a(x, y, u), \tag{5-9a}$$
$$\frac{dy}{ds} = b(x, y, u), \tag{5-9b}$$
$$\frac{du}{ds} = q(x, y, u). \tag{5-9c}$$

This gives a system of linear equations that must be solved. Let's illustrate this with a few examples.

5.1.b Example: Uniform Transport

Let y take the role of a time variable t. Suppose we have a differential equation that reads

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0, \quad u(x, 0) = f(x). \tag{5-10}$$

We will see this equation describes motion at constant speed c. The initial condition gives some value of a function of x specified at time t.

The system of ordinary differential equations from method of characteristics for this problem is

$$\frac{dx}{ds} = c, \tag{5-11a}$$
$$\frac{dt}{ds} = 1, \tag{5-11b}$$
$$\frac{du}{ds} = 0. \tag{5-11c}$$

To solve these differential equations, we will have some arbitrary constants. We can choose these in terms of s based on our initial conditions such that

$$x(0) = \alpha, \tag{5-12a}$$
$$t(0) = 0, \tag{5-12b}$$
$$u(0) = f(\alpha). \tag{5-12c}$$

Here α is some constant value that we will eliminate using the initial condition.

Solving the differential equations is very simple. These yield

$$x = cs + \alpha, \tag{5-13a}$$
$$t = s, \tag{5-13b}$$
$$u = f(\alpha). \tag{5-13c}$$

From the equation for x we can solve for α to get

$$\alpha = x - cs = x - ct. \tag{5-14}$$

Inserting this into the equation for u gives the solution

$$u(x,t) = f(x - ct). \tag{5-15}$$

This differential equation takes the initial function $f(x)$ and moves it with a fixed speed c through time. This is depicted in Fig. 5.1. In other words, this describes uniform translation and is the simplest first-order partial differential equation. The characteristic curves, i.e., where u is constant, are straight lines with slope $1/c$. Fig. 5.2 plots these lines in the x-t plane. Here we can see the values of α denote where the characteristic lines hit the x axis.

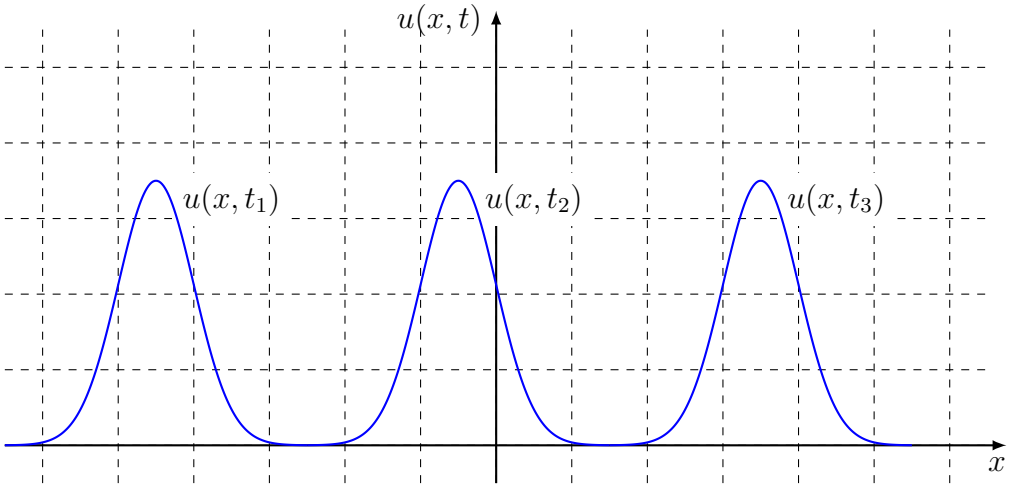

Figure 5.1: Illustration of an example solution of uniform transport at a few snapshots in time. The initial function translates along the x direction with a rate given by c.

5.1.c Example: Transport with Absorption

We can solve a slightly more difficult version of the same problem by including an absorption term on the right-hand side with constant absorption coefficient λ and initial condition $u(x,0) = f(x)$:

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = -\lambda u(x,t), \quad u(x,0) = f(x). \tag{5-16}$$

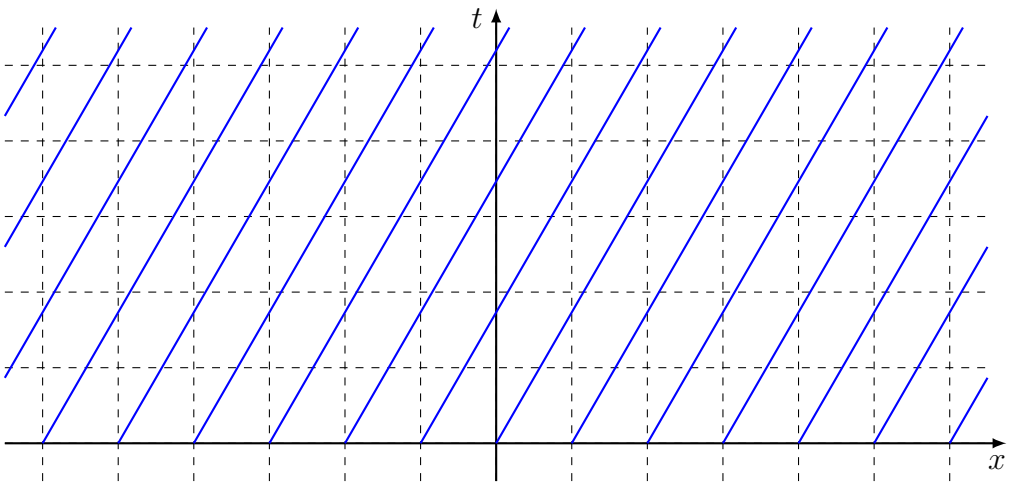

Figure 5.2: Illustration of characteristic lines for the uniform transport problem.

The system of ordinary differential equations from method of characteristics for this problem is

$$\frac{dx}{ds} = c, \tag{5-17a}$$
$$\frac{dt}{ds} = 1, \tag{5-17b}$$
$$\frac{du}{ds} = -\lambda u. \tag{5-17c}$$

with the initial conditions

$$x(0) = \alpha, \tag{5-18a}$$
$$t(0) = 0, \tag{5-18b}$$
$$u(0) = f(\alpha). \tag{5-18c}$$

The solution for the x and t are identical to before

$$x = cs + \alpha, \tag{5-19a}$$
$$t = s. \tag{5-19b}$$

The equation for u is separable and leads to an exponential with a constant

$$u = ke^{-\lambda s} = ke^{-\lambda t}.$$

Inserting the initial condition at $t = 0$ gives the solution for the constant

$$f(\alpha) = ke^{-\lambda(0)} = k.$$

Therefore,

$$u(x,t) = f(x - ct)e^{-\lambda t}. \tag{5-19c}$$

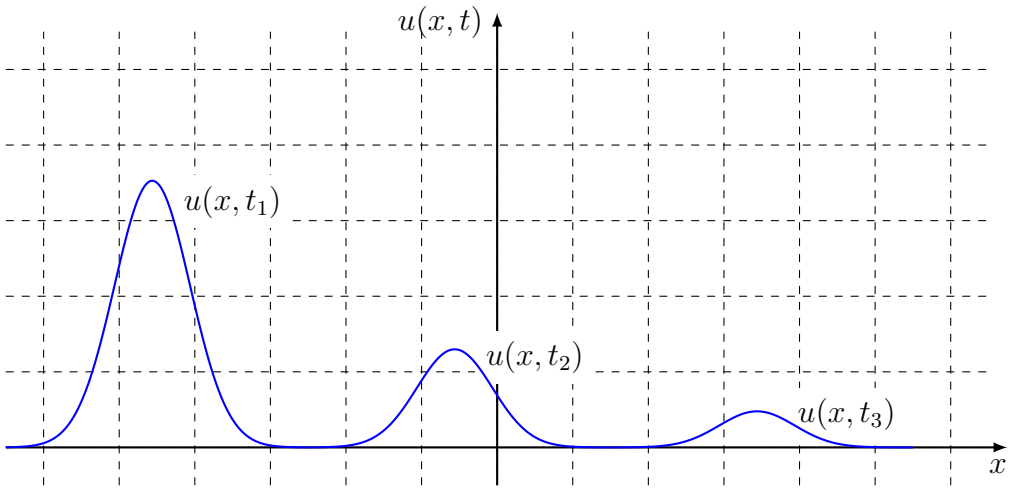

Figure 5.3: Illustration of an example solution of transport with absorption at a few snapshots in time. The initial function translates along the x direction with a rate given by c and the magnitude of the function decreases with increasing time.

The solution is very similar to what we had in the uniform transport case, except we how have an exponential decay term. This describes the translation of the initial condition with a speed c that decreases in magnitude exponentially. This is depicted in Fig. 5.3. This mathematical model is used to describe moving radioactive particles or how a pulse of radiation propagates through matter.

5.1.d Example: Photon Emission from a Moving Point Source

We have some flexibility when it comes to the initial condition. Suppose now we have a radioactive object that begins at $x = 0$ and moves with a constant speed v in the positive x direction that emits photons that move in the positive x direction with the speed of light c (we assume $v < c$). The photons are absorbed with a coefficient λ. The mathematical description of this problem is the same as what we had before, except that the initial condition is different. The mathematical model becomes

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = -\lambda u(x,t), \quad u(vt,t) = u_0, \quad t \geq 0 \tag{5-20}$$

The ordinary differential equations are identical,

$$\frac{dx}{ds} = c, \tag{5-21a}$$

$$\frac{dt}{ds} = 1, \tag{5-21b}$$

$$\frac{du}{ds} = -\lambda u. \tag{5-21c}$$

but we now have

$$x(0) = v\alpha, \tag{5-22a}$$

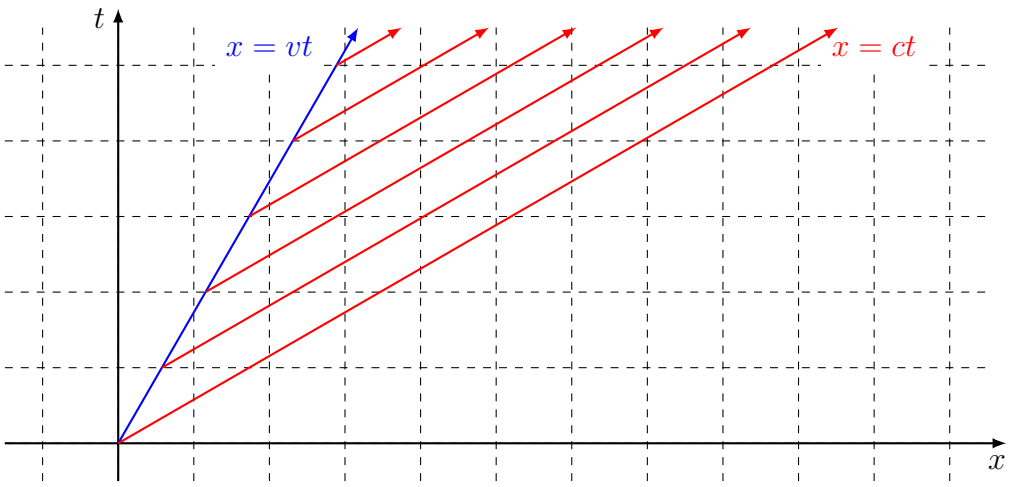

Figure 5.4: Illustration of photon emission problem. The radioactive source travel along the blue line and the photons travel along the red lines.

$$t(0) = \alpha, \tag{5-22b}$$
$$u(0) = u_0, \tag{5-22c}$$

for the initial conditions. Solving these equations gives

$$x = cs + v\alpha, \tag{5-23a}$$
$$t = s + \alpha, \tag{5-23b}$$
$$u = u_0 e^{-\lambda s}. \tag{5-23c}$$

The difference now being that the time coordinate t is a function of α. Eliminating s from the equation for t and then using the equation for x to solve for α gives

$$\alpha = \frac{ct - x}{c - v}. \tag{5-24}$$

Plugging in $s = t - \alpha$ into the equation for u gives the solution

$$u(x,t) = u_0 \exp\left[-\lambda\left(t - \frac{ct - x}{c - v}\right)\right], \quad vt \leq x \leq ct \tag{5-25}$$

To show the range solution where is valid based we study the characteristics and lines of photon emission in the x-t plane. These are shown in Fig. 5.4. Since $t \geq 0$ and the source starts at $x = 0$, we can assert that the position coordinate in the x-t plane cannot exceed ct, as it would require a particle to be emitted before $t = 0$. Likewise, the position coordinate cannot be less that vt since that denotes the point photons are emitted (remember we required that they travel in the positive x direction.

5.2 First-Order Quasi-Linear PDEs

A particular class of first-order partial differential equations that is particularly relevant to inertial confinement fusion is the quasi-linear case. These can be used to

describe the flow of inviscid fluids and permit the formation of shockwaves. The first-order quasi-linear partial differential equation has the form

$$a(x,y,u)\frac{\partial u}{\partial x} + b(x,y,u)\frac{\partial u}{\partial y} = q(x,y,u). \tag{5-26}$$

This differs from the linear case in that the coefficients a and b can now depend upon the function $u(x,t)$. The distinction between a quasi-linear and a fully nonlinear partial differential equation is that we do not allow the coefficients to depend upon the partial derivatives. Quasi-linear first-order partial differential equations, like their simpler linear kin, can be solved using the method of characteristics. In this section, we will focus explicitly on an equation called Burger's equation, which describes the flow of an inviscid fluid in one dimension with a unit speed and no external forces.

5.2.a Burger's Equation

Burger's equation has the form

$$\frac{\partial u}{\partial t} + \frac{1}{2}\frac{\partial}{\partial x}(u^2) = 0, \quad u(x,0) = f(x). \tag{5-27}$$

While at first glance, this appears nonlinear, the partial derivative in x can be expanded to obtain

$$\frac{\partial u}{\partial t} + u(x,t)\frac{\partial u}{\partial x} = 0, \quad u(x,0) = f(x). \tag{5-28}$$

In the form of quasi-linear partial differential equations (where we let $y = t$), the coefficients are

$$a = u, \tag{5-29a}$$
$$b = 1, \tag{5-29b}$$
$$q = 0. \tag{5-29c}$$

Burger's equation describes a system where the speed that the function travels in the x direction is equal to the magnitude of the function itself. This allows for parts of the function that are higher to "catch up" or "run away" from other parts of the solution.

Applying the method of characteristics we obtain the following system of ordinary differential equations:

$$\frac{dx}{ds} = u, \tag{5-30a}$$
$$\frac{dt}{ds} = 1, \tag{5-30b}$$
$$\frac{du}{ds} = 0, \tag{5-30c}$$

which have the initial conditions

$$x(0) = \alpha, \tag{5-31a}$$
$$t(0) = 0, \tag{5-31b}$$
$$u(0) = f(\alpha). \tag{5-31c}$$

Solving for t and applying the initial condition gives

$$t = s. \tag{5-32a}$$

Solving for u yields a constant value. After applying the initial condition we get

$$u = f(\alpha). \tag{5-32b}$$

Finally, solving for x gives

$$x = f(\alpha)s + \alpha = f(\alpha)t + \alpha. \tag{5-32c}$$

Solving for α gives

$$\alpha = x - f(\alpha)t = x - ut. \tag{5-33}$$

Therefore, the solution becomes the implicit function

$$u(x,t) = f(x - ut). \tag{5-34}$$

This solution can lead to some interesting issues. To illustrate, let us consider the example where the initial condition is given by the following piecewise ramp function

$$f(x) = \begin{cases} 1, & x < 0 \\ 1 - x, & 0 \leq x < 1 \\ 0, & x \geq 1 \end{cases}. \tag{5-35}$$

Using this initial condition, we can construct a set of characteristic lines in the x-t plane. If we parameterize the lines based on α we may use $x = f(\alpha)t + \alpha$ and insert the argument from the initial condition for $f(\alpha)$ replacing x with α. This gives the following functional form for the characteristics:

$$x(t) = \begin{cases} t + \alpha, & \alpha < 0 \\ (1 - \alpha)t + \alpha, & 0 \leq \alpha < 1 \\ \alpha, & \alpha \geq 1 \end{cases}. \tag{5-36}$$

The characteristic lines are plotted in Fig. 5.5. In the region $0 \leq \alpha < 1$, the lines exhibit a feature that the all intersect at the point $(1, 1)$ indicating that the the function appears to be converging at the time $t = 1$.

To obtain the solutions, we note that again $u = f(x - ut)$ and that the solution is constant along the characteristic curves. We develop solutions for three different

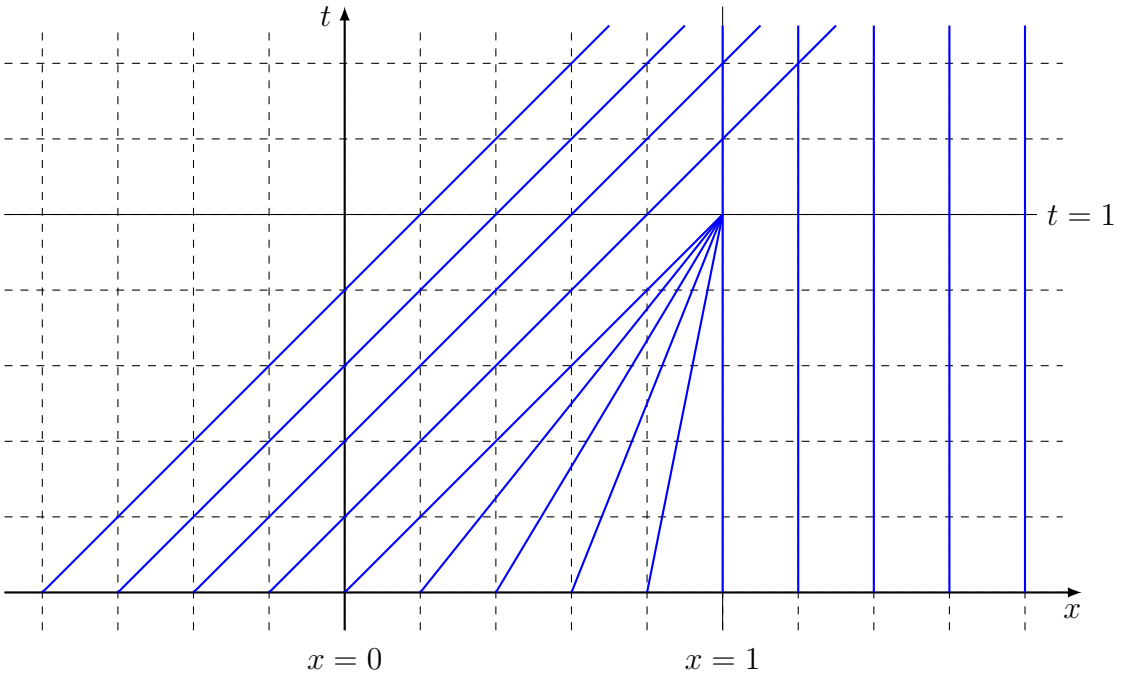

Figure 5.5: Characteristics for Burger's equation with a ramp function initial condition.

cases for each range of α on these characteristic curves. For the case where $\alpha < 0$, we have $f(\alpha) = 1$. Inserting this into the equation gives

$$u(x,t) = f(1) = 1, \quad \alpha < 0,$$

or a constant solution. Putting this in terms of x and t, this corresponds to the range where

$$0 > \alpha = x - f(\alpha)t = x - t$$

or

$$x < t.$$

Similarly, for the next range we have

$$f(x - ut) = 1 - (x - ut) = u, \quad 0 \leq \alpha < 1.$$

Solving for u gives

$$u = \frac{1 - x}{1 - t}.$$

The range is found by

$$0 \leq \alpha = x - f(\alpha)t = x - (1 - \alpha)t < 1.$$

Solving for α gives

$$\alpha = \frac{x-t}{1-t}.$$

For the left side of the range where $\alpha \geq 0$,

$$\frac{x-t}{1-t} \geq 0,$$

which gives

$$x \geq t.$$

For the right side of the range where $\alpha < 1$,

$$\frac{x-t}{1-t} < 1,$$

we get

$$x < 1.$$

Finally,

$$f(x-ut) = 0 = u, \quad \alpha > 1.$$

This corresponds to the range $x \geq 1$.

Pulling this altogether, the solution of Burger's equation is

$$u(x,t) = \begin{cases} 1, & x < t \\ \dfrac{1-x}{1-t}, & t \leq x < 1 \\ 0, & x \geq 1 \end{cases}. \tag{5-37}$$

The function $u(x,t)$ is well defined until we reach the convergence point of the characteristics at $t = 1$. At this point, the solution approaches a problem point where the slope goes towards infinity and the function becomes discontinuous. Strictly speaking, a discontinuous function cannot satisfy a differential equation. Instead we will demand the differential equation be satisfied under some integral. This leads to a concept called a weak formulation, which will not be discussed here. What we need to fully define the solution, is to obtain the behavior for $t \geq 1$. To do this, we will need to derive a jump condition.

5.2.b Jump Condition for Shocks

Before proceeding, we need to first define the notion of a conservation law. Differential equations are usually used to describe some physical process. In this case, Burger's equation is often used to describe the motion of inviscid fluids, and we can think of

the conserved quantity as the amount of mass or momentum in some volume. Let us define the conservation law as the integral of $u(x,t)$ over some range x_1 to x_2:

$$m = \int_{x_1}^{x_2} u(x,t)dx. \tag{5-38}$$

Here m denotes some conserved quantity. If we take the time rate of change of m, we can relate it to the difference between the inflow and outflow rates across x_1 and x_2 respectively. This gives

$$\frac{d}{dt}\int_{x_1}^{x_2} u(x,t)dx = \phi(u(x_1,t)) - \phi(u(x_2,t)). \tag{5-39}$$

Here $\phi(u)$ represents the flow rate of the conserved quantity across the boundary. Now, if we divide both sides by $h = x_2 - x_1$ we get

$$\frac{1}{x_2 - x_1}\frac{d}{dt}\int_{x_1}^{x_2} u(x,t)dx = \frac{\phi(u(x_1,t)) - \phi(u(x_2,t))}{x_2 - x_1}. \tag{5-40}$$

Next, if we take the limit is the interval $x_2 - x_1 \to 0$. The right-hand side is simply the definition of the partial derivative. We can see this if we rewrite the range to be centered around some fixed midpoint value we call x:

$$x_1 = x - \frac{h}{2}, \tag{5-41a}$$

$$x_2 = x + \frac{h}{2}. \tag{5-41b}$$

The right-hand side becomes

$$\frac{\partial u}{\partial x} = \lim_{h\to 0}\frac{\phi(u(x+h/2,t)) - \phi(u(x-h/2,t))}{h}. \tag{5-42}$$

The left-hand side becomes

$$\frac{d}{dt}\left(\lim_{h\to 0}\frac{1}{h}\int_{x-h/2}^{x+h/2} u(x,t)dx\right). \tag{5-43}$$

When taking the limit as $h \to 0$, we have the case where the integral goes to zero and the denominator h also goes to zero. To take this limit, we employ L'Hopital's rule.

$$\lim_{h\to 0}\frac{f(h)}{g(h)} = \frac{f'(h)}{g'(h)}. \tag{5-44}$$

To perform the derivative of the integral with respect to h, we note that the limits of integration depend upon h. This requires the use of the Leibniz rule:

$$\frac{d}{dt}\int_{a(t)}^{b(t)} u(x,t)dx = u(b(t),t)\frac{db}{dt} - u(a(t),t)\frac{da}{dt} + \int_{a(t)}^{b(t)}\frac{\partial u}{\partial t}dx. \tag{5-45}$$

The derivative becomes:

$$\frac{d}{dh}\left(\int_{x-h/2}^{x+h/2} u(x,t)dx\right) = u(x+h/2,t)\left(\frac{1}{2}\right) - u(x-h/2,t)\left(-\frac{1}{2}\right)$$
$$= \frac{1}{2}\left[u(x+h/2,t) + u(x-h/2,t)\right]. \tag{5-46}$$

The last partial derivative with respect to h vanishes because u does not depend upon h. The derivative of h with respect to h is simply 1. So taking the limit gives

$$\frac{d}{dt}\left[\frac{1}{2}\left(u(x,t)+u(x,t)\right)\right] = \frac{d}{dt}u(x,t) = \frac{\partial u}{\partial t}. \tag{5-47}$$

The derivative becomes the partial derivative since we are taking it with respect to some fixed value of x. Putting this together get a partial differential equation:

$$\frac{\partial u}{\partial t} = -\frac{\partial}{\partial x}(\phi(u)). \tag{5-48}$$

If we define

$$\phi(u) = \frac{1}{2}u^2(x,t) \tag{5-49}$$

we arrive at Burger's equation where $\phi(u)$ represents the outflow rate of a physical quantity.

Now let us return to our original conservation equation. Suppose we have $x_1 < x_s(t) < x_2$ where $x_s(t)$ is the time-dependent location of a shock where, based on our solution to Burger's equation, we define

$$u(x,t) = \begin{cases} u^-(x,t) & x < x_s(t) \\ u^+(x,t) & x > x_s(t) \end{cases} \tag{5-50}$$

We can write the integral term as the sum of two sub-integrals on each side of the shock

$$\frac{d}{dt}\left(\int_{x_1}^{x_s(t)} u^-(x,t)dx + \int_{x_s(t)}^{x_2} u^+(x,t)dx\right) = \phi(u^-(x_1,t)) - \phi(u^+(x_2,t)). \tag{5-51}$$

Now, to evaluate the total derivative with respect to an integral, where the limits of integration are again functions of the variable being differentiated, so we use the Leibniz rule. Taking the derivative gives

$$u^-(x_s(t),t)\frac{dx_s}{dt} - u^-(x_1,t)\frac{dx_1}{dt} + \int_{x_1}^{x_s(t)}\frac{\partial u^-}{\partial t}dx$$
$$+ u^+(x_2,t)\frac{dx_2}{dt} - u^+(x_s(t),t)\frac{dx_s}{dt} + \int_{x_s(t)}^{x_2}\frac{\partial u^+}{\partial t}dx$$
$$= \phi(u^-(x_1,t)) - \phi(u^+(x_2,t)). \tag{5-52}$$

Note that x_1 and x_2 are fixed in time, so those two terms can cancel, leaving:

$$u^-(x_s(t),t)\frac{dx_s}{dt} - u^+(x_s(t),t)\frac{dx_s}{dt} + \int_{x_1}^{x_s(t)} \frac{\partial u^-}{\partial t}dx + \int_{x_s(t)}^{x_2} \frac{\partial u^+}{\partial t}dx = \phi(u^-(x_1,t)) - \phi(u^+(x_2,t)). \tag{5-53}$$

Now we take the one-sided limits as $x_1 \to x_s(t)^-$ and $x_2 \to x_s(t)^+$. This causes the integral terms to vanish, leaving us with a solution we can solve for the time derivative of $x_s(t)$. This gives the *Rankine-Hugoniot jump condition*:

$$\frac{dx_s}{dt} = \frac{\phi(u^-(x_s(t),t)) - \phi(u^+(x_s(t),t))}{u^-(x_s(t),t) - u^+(x_s(t),t)}. \tag{5-54}$$

The Jump condition relates the speed of the shock to what occurs on each side of the shock.

Applying this to Burger's equation where $\phi(u) = u^2/2$, we get the jump condition is

$$\frac{dx_s}{dt} = \frac{1}{2}\frac{(u^-(x_s,t))^2 - (u^+(x_s,t))^2}{u^-(x_s,t) - u^+(x_s,t)}. \tag{5-55}$$

Based on our solution, we have the limiting case for the left side of the shock gives $u^- = 1$ and the limiting case on the right side is $u^+ = 0$. Plugging in these values gives

$$\frac{dx}{dt} = \frac{1}{2}. \tag{5-56}$$

Based on the characteristics, the problem point occurs at $(1,1)$, which gives us an "initial condition" for the jump condition. Integrating and applying the initial condition gives

$$x(t) = \frac{t+1}{2}. \tag{5-57}$$

Therefore, the solution for the problem for $t \geq 1$ is

$$u(x,t) = \begin{cases} 1, & x < \dfrac{t+1}{2} \\ 0, & x > \dfrac{t+1}{2} \end{cases}, \tag{5-58}$$

with the solution we obtained previously,

$$u(x,t) = \begin{cases} 1, & x < t \\ \dfrac{1-x}{1-t}, & t \leq x < 1 \\ 0, & x \geq 1 \end{cases},$$

valid for $t < 1$. This solution is such that for $0 < t < 1$, the solution moves toward forming a shock, which occurs at $t = 1$. The shock then propagates to the right at a speed of $1/2$. This solution is shown in Fig. 5.6. We must also incorporate this into our characteristics. A revised plot for the characteristics that includes the jump condition is given in Fig. 5.7.

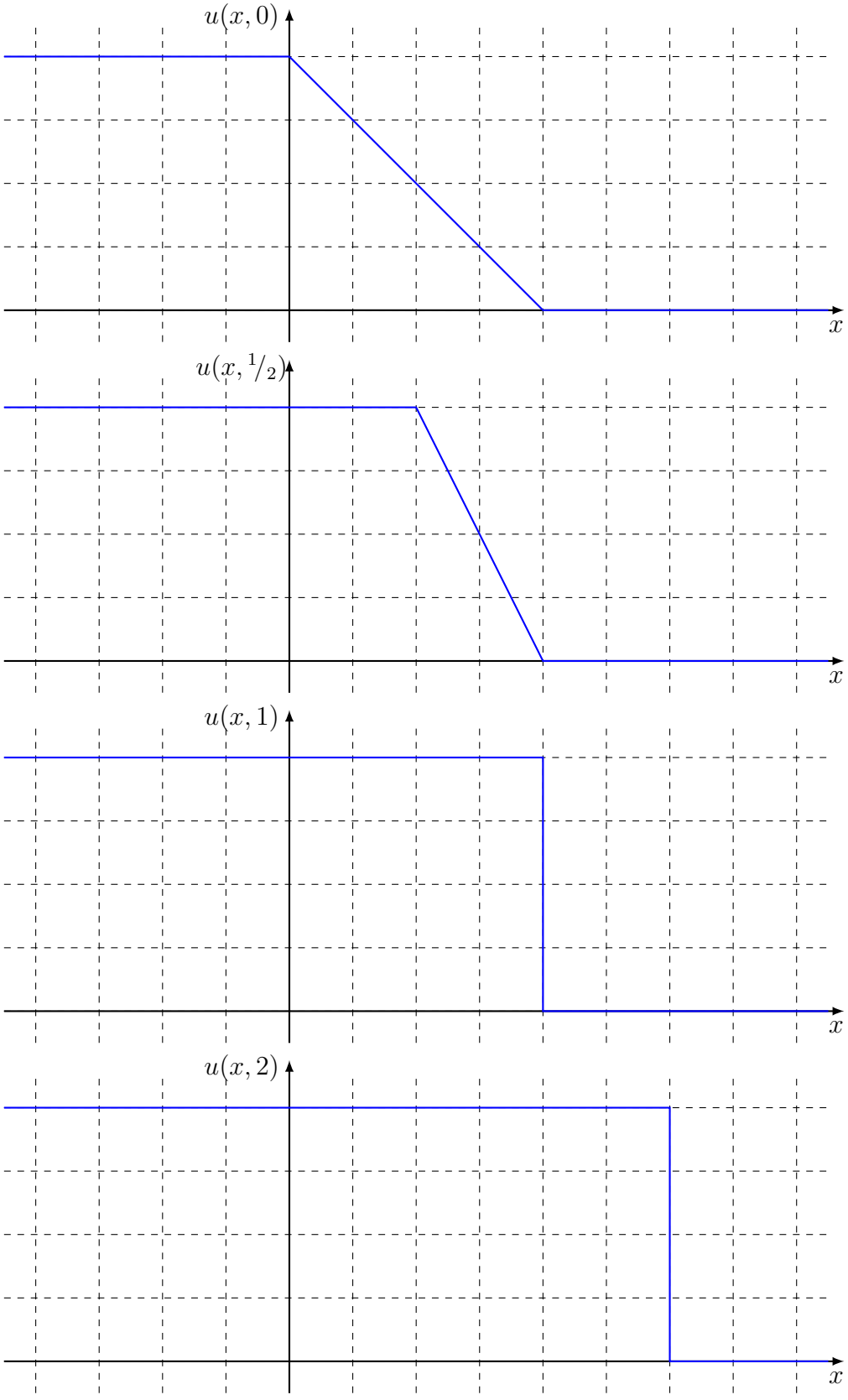

Figure 5.6: Illustration of a solution to Burger's equation with an initial ramp function.

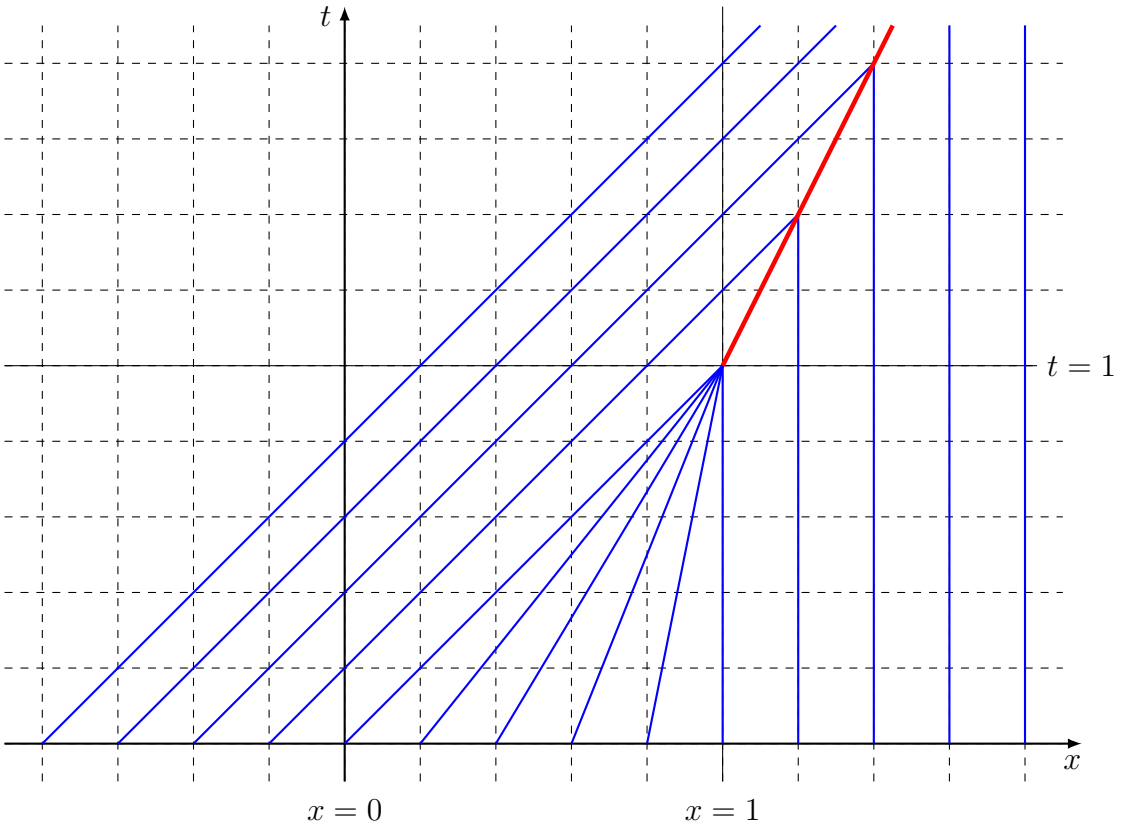

Figure 5.7: Characteristics for Burger's equation with a ramp function initial condition including the Rankine-Hugoniot jump condition after the shock formation. The shock front is denoted by a thick red line.

5.2.c Entropy Condition and Rarefactions

It is also possible to have cases where we have multiple valid solutions from quasi-linear partial differential equations that satisfy the jump conditions. In these cases, the mathematical model does not have sufficient information to uniquely determine the physics and we need to impose additional constraints based upon physics considerations. We illustrate this case for Burger's equation with the following initial condition:

$$f(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} . \tag{5-59}$$

This initial condition has a discontinuity at $x = 0$. A key feature here is that initial condition is greater to the right and smaller to the left. In these cases, we may run into the issue of multiple valid solutions.

Again, the characteristic curves along x parameterized by α are given by

$$x(\alpha) = f(\alpha)t + \alpha.$$

This implies

$$x(\alpha) = \begin{cases} \alpha, & \alpha < 0 \\ t + \alpha, & \alpha \geq 0 \end{cases} . \tag{5-60}$$

These characteristic curves produce the solution

$$u(x,t) = \begin{cases} 0, & x < 0 \\ ?, & 0 \geq x < t \\ 1, & x > t \end{cases} . \tag{5-61}$$

Here we have characteristic curves, and therefore solutions, for the region $x < 0$ and $x > t$, but no information about $0 < x < t$. In other words, the range between $x = 0$ and the edge of the shock is undefined for $t > 0$. The differential equation alone does not provide enough information.

We must therefore devise solutions that both satisfy both the differential equation and the Rankine-Hugoniot jump conditions. A possible solution that does this is to have the shock move to the right at the speed of $1/2$ with a value of $u = 1$ ahead of the shock and $u = 0$ behind it. We call this possible solution

$$u(x,t) \stackrel{?}{=} u_1(x,t) = \begin{cases} 0, & x < \dfrac{t}{2} \\ 1, & x \geq \dfrac{t}{2} \end{cases} . \tag{5-62}$$

Another possible solution continuously connects the points between $x = 0$ and $x = t$. We call this solution

$$u(x,t) \stackrel{?}{=} u_2(x,t) = \frac{x}{t}. \tag{5-63}$$

This also satisfies the differential equation and for $t \to 0$ limits to the initial and jump conditions. Both candidate solutions are plotting alongside the initial condition in Fig. 5.8.

Resolving which one of these two solutions most matches reality invokes the second law of thermodynamics, which states that in an isolated system the entropy increases. (Note that if we have external forces, the entropy could decrease locally so long as the net increase in entropy from generating the external force is positive.) The entropy condition whether a shock is allowed to propagate relates the shock speed to the derivative of the inflow/outflow rates:

$$\phi'(u^-) > \frac{dx_s}{dt} > \phi'(u^+). \tag{5-64}$$

We only admit shocks that follow this condition. If the curve is not a discontinuous shock, then we cannot apply this conditions.

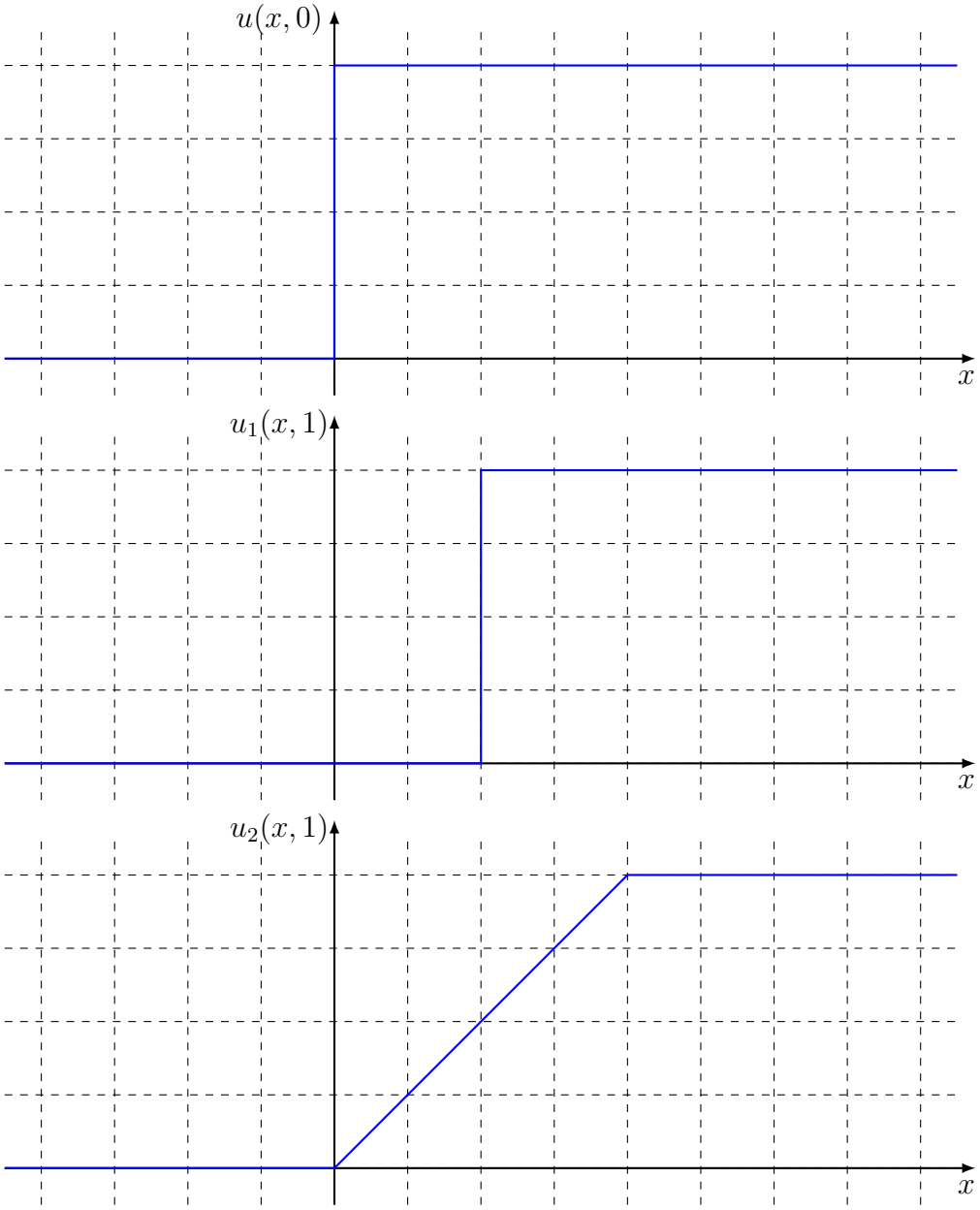

Figure 5.8: Illustration of the initial condition and two possible solutions of Burger's equation at $t = 1$.

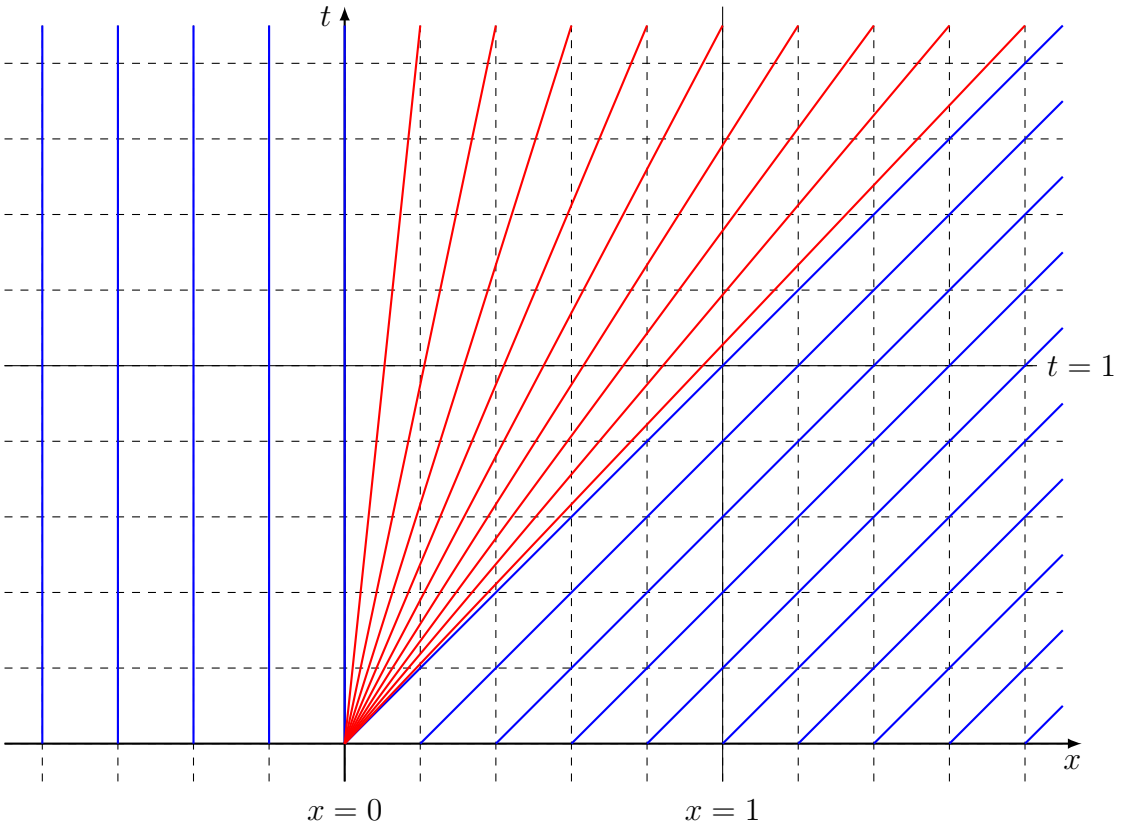

Figure 5.9: Characteristics for Burger's equation a step initial condition including a rarefaction solution (red) that satisfies the entropy condition.

For our case, the shock speed is 1/2 and

$$\phi'(u) = \frac{d}{du}\left(\frac{u^2}{2}\right) = u. \tag{5-65}$$

For candidate solution 1, we have a traveling shock with $u^- = 0$, which is obviously not greater than 1/2, and neither is $u^+ = 1$ less than 1/2. Therefore, we can reject this solution because it does not increase entropy. Candidate solution 2 is not a shock, rather it is a continuous solution. Since we have rejected the shock, it is reasonable to accept $u_2(x,t)$. Therefore, our solution becomes

$$u(x,t) = \begin{cases} 0, & x < 0 \\ \dfrac{x}{t}, & 0 \geq x < t \\ 1, & x > t \end{cases} . \tag{5-66}$$

The characteristic lines for the solution are plotted in Fig. 5.9. The solution $u_2 = x/t$ is called a *rarefaction wave*, which means the density of the quantity decreases as the solution propagates. From the plot, we see characteristic lines spreading out in a

fanlike pattern. This may be interpreted as for each point within the rarefaction for $t > 0$, we can trace the solution back in time to the point $x = 0, t = 0$.

Before concluding the discussion on shocks and rarefactions, one point to bring up that is that we have an overall conservation of area under the curve at all times. In our examples, the area is infinite so this is not apparent; however, these are contrived cases. In real-world problems, we have a finite system. As there is no absorption or growth term, we expect the total amount of the quantity to be constant, and just redistributing in space. We saw the similar case with the uniform transport problem where the quantity underwent simple translation.

5.3 Heat Equation in 1-D

The heat equation is a partial differential equation that was first proposed to study the time-dependent transfer of thermal energy. Using this equation, we are able to obtain the temperature distribution through some object given some prescribed initial condition and boundary conditions. While this may seem limited in scope, it turns out that the partial differential equations used to describe many other physical phenomena are either identical or similar in form to the heat equation. This includes mass transport through materials, neutron diffusion, and the potential functions of electromagnetism. The heat equation is sometimes (more precisely) referred to as the diffusion equation, but the former usage is more common because of its historical development.

The heat equation has the generic 3-D form

$$\frac{\partial u}{\partial t} - \alpha \nabla^2 u(\boldsymbol{x}, t) = 0. \tag{5-67}$$

The parameter α is often referred to as the diffusion coefficient. The negative sign on the Laplacian has a physical interpretation related to the entropy condition we discussed for the shock problem. Specifically, the negative sign leads to an overall tendency of a physical quantity to spread out uniformly, which is consistent with the second law of thermodynamics.

In this section we will study the case of one spatial dimension. In Cartesian coordinates the 1-D heat equation has the form,

$$\frac{\partial u}{\partial t} - \alpha \frac{\partial^2 u}{\partial x^2} = 0, \quad 0 \leq x \leq L, \quad t \geq 0. \tag{5-68}$$

Since we have three total derivatives: one in time and two in space, we need to specify three conditions on the solution. The first of these specifies an initial condition

$$u(x, 0) = f(x), \quad 0 \leq x \leq L. \tag{5-69}$$

Here the solution is allowed to take any function at time $t = 0$. As with the ordinary differential equations, we must also specify boundary conditions at $x = 0$ and $x = L$. Options for this are the same as for second-order ordinary differential equations. The

first, and simplest, are the Dirichlet boundary conditions, which specify the solution at the boundaries. The second is the Neumann boundary condition, which specifies the derivative in the direction normal to the boundary. The final condition is the Robin boundary condition, which is a combination of these two.

In cylindrical coordinates, the 1-D heat equation is obtained by expanding out the Laplacian and keeping the radial term

$$\frac{\partial u}{\partial t} - \frac{\alpha}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) = 0, \quad 0 \leq x \leq R, \quad t \geq 0. \tag{5-70}$$

In spherical coordinates the 1-D heat equation is

$$\frac{\partial u}{\partial t} - \frac{\alpha}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right) = 0, \quad 0 \leq x \leq R, \quad t \geq 0. \tag{5-71}$$

The heat equation also has a couple variants. The first adds an inhomogeneous term, which could be, for example, the generation of heat within the domain. We can also add an absorption term involving $u(x,t)$, which is common when we are studying mass or neutron diffusion where materials can be removed through chemical or nuclear reactions respectively. Finally, we can add a drift term involving the spatial partial derivative, which causes the solution to transport in time, identical to what we say when studying the first-order partial differential equations.

5.3.a Fourier Series Expansions

Unfortunately, solutions for the heat equation do not often have a closed form, and we need to express the solution as an infinite series with a particular choice of orthogonal basis functions. Orthogonal functions satisfy the property that if you have two functions $f_n(x)$ and $f_m(x)$ on some domain, then if we take the inner product of those two functions only terms where $n = m$ are nonzero. In terms of functions, the inner product is the integral over the entire domain. This is stated as

$$\int_{-\infty}^{\infty} f_n(x) f_m(x) dx = c_n \delta_{mn}. \tag{5-72}$$

Here c_n is some constant and δ_{mn} is the Kronecker delta function, which is

$$\delta_{mn} = \begin{cases} 1 & m = n, \\ 0 & m \neq n \end{cases}. \tag{5-73}$$

In Cartesian coordinates, the choice of orthogonal basis function are the trigonometric functions where we define the domain to be in such a way as to include full periods of that function. Let us assume the domain $-L \leq x \leq L$. (This does not lose generality, as we can always do a coordinate transformation to scale and shift to obtain any domain we like.) The trigonometric functions satisfy an orthogonality property. For the case of sine functions where $n > 0$ or $m > 0$ exclusively (not both equal to zero),

$$\int_{-L}^{L} \sin\left(\frac{m\pi x}{L}\right) \sin\left(\frac{n\pi x}{L}\right) dx = L\delta_{nm}. \tag{5-74a}$$

Similarly the cosine functions satisfy

$$\int_{-L}^{L} \cos\left(\frac{m\pi x}{L}\right) \cos\left(\frac{n\pi x}{L}\right) dx = L\delta_{nm}. \tag{5-74b}$$

Furthermore any product of sine and cosine functions integrated over the domain

$$\int_{-L}^{L} \sin\left(\frac{m\pi x}{L}\right) \cos\left(\frac{n\pi x}{L}\right) dx = 0. \tag{5-74c}$$

Note that if both $n = m = 0$ then we obviously get

$$\int_{-L}^{L} dx = 2L. \tag{5-74d}$$

The fundamental idea of the Fourier series is that any continuous function on the domain can be expressed using an infinite sum of trigonometric functions

$$f(x) = c_0 + \sum_{n=1}^{\infty} a_n \sin\left(\frac{n\pi x}{L}\right) + \sum_{n=1}^{\infty} b_n \cos\left(\frac{n\pi x}{L}\right), \quad -L < x < L. \tag{5-75}$$

Here a_n and b_n are coefficients that are used to fit the specific functions. These can be found by taking integrals. First, let us integrate the function over the domain

$$\int_{-L}^{L} f(x)dx = \int_{-L}^{L} c_0 dx + \sum_{n=1}^{\infty} a_n \int_{-L}^{L} \sin\left(\frac{n\pi x}{L}\right) dx + \sum_{n=1}^{\infty} b_n \int_{-L}^{L} \cos\left(\frac{n\pi x}{L}\right) dx. \tag{5-76}$$

Because of orthogonality, all of the terms involving sines and cosines vanish leaving us with $2Lc_0$. Therefore, the coefficient is

$$c_0 = \frac{1}{2L} \int_{-L}^{L} f(x)dx. \tag{5-77}$$

The a_n coefficients for $n > 0$ can be found by multiplying by the sine function with period n and integrating over the domain:

$$\begin{aligned} \int_{-L}^{L} \sin\left(\frac{n\pi x}{L}\right) f(x)dx = &\int_{-L}^{L} c_0 \sin\left(\frac{n\pi x}{L}\right) dx \\ &+ \sum_{n=1}^{\infty} a_n \int_{-L}^{L} \sin\left(\frac{n\pi x}{L}\right) \sin\left(\frac{n\pi x}{L}\right) dx \\ &+ \sum_{n=1}^{\infty} b_n \int_{-L}^{L} \sin\left(\frac{n\pi x}{L}\right) \cos\left(\frac{n\pi x}{L}\right) dx. \end{aligned} \tag{5-78}$$

Using the orthogonality, only the a_n term survives, giving

$$a_n = \frac{1}{L} \int_{-L}^{L} \sin\left(\frac{n\pi x}{L}\right) f(x)dx, \quad n > 0. \tag{5-79}$$

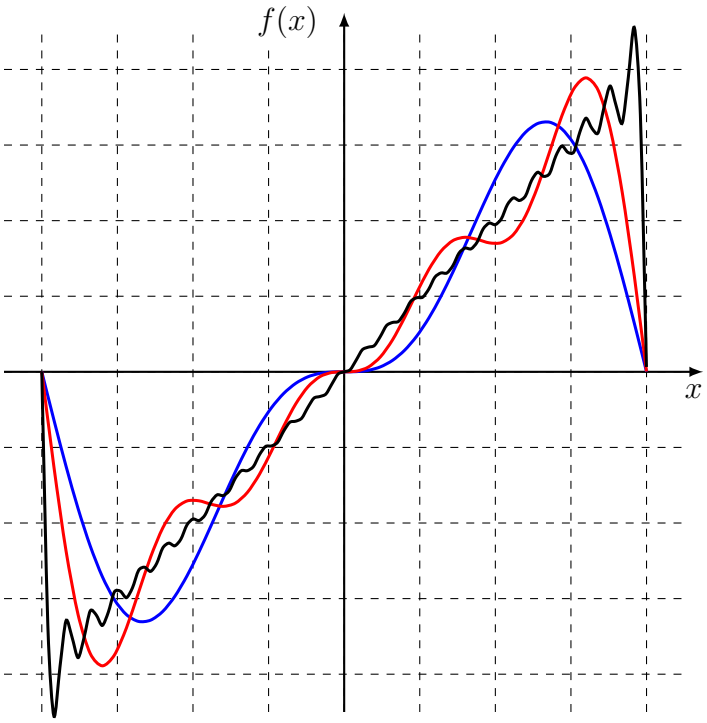

Figure 5.10: Illustration of a Fourier series expansion of $f(x) = x, -1 \leq x \leq 1$ for $N = 2$ terms (blue), $N = 4$ terms (red), and $N = 24$ terms (black).

The same process can be done using the cosine function. The result of this is

$$b_n = \frac{1}{L}\int_{-L}^{L} \cos\left(\frac{n\pi x}{L}\right) f(x)dx, \quad n > 0. \tag{5-80}$$

Before proceeding to try to apply this to partial differential equation, let us do a few examples. The first case we will consider is the linear function

$$f(x) = \frac{x}{L}, \quad -L \leq x \leq L. \tag{5-81}$$

Applying the equations to estimate coefficients, we get

$$c_0 = \frac{1}{2L}\int_{-L}^{L} \frac{x}{L}dx = 0, \tag{5-82a}$$

$$a_n = \frac{1}{L}\int_{-L}^{L} \sin\left(\frac{n\pi x}{L}\right)\frac{x}{L}dx = -\frac{2}{n\pi}\cos(n\pi), \quad n > 0, \tag{5-82b}$$

$$b_n = \frac{1}{L}\int_{-L}^{L} \cos\left(\frac{n\pi x}{L}\right)\frac{x}{L}dx = 0, \quad n > 0. \tag{5-82c}$$

We note that we can write the cosine term in the a_n coefficients as

$$\cos(n\pi) = (-1)^n.$$

Therefore, the coefficient becomes

$$a_n = (-1)^{n+1}\frac{2}{n\pi}. \tag{5-83}$$

The Fourier series expansion is therefore

$$\begin{aligned}\frac{x}{L} &= \sum_{n=1}^{\infty}(-1)^{n+1}\frac{2}{n\pi}\sin\left(\frac{n\pi x}{L}\right) \\ &= \frac{2}{\pi}\sin\left(\frac{\pi x}{L}\right) - \frac{1}{\pi}\sin\left(\frac{2\pi x}{L}\right) + \frac{2}{3\pi}\sin\left(\frac{3\pi x}{L}\right) - \dots\end{aligned} \tag{5-84}$$

An example of this is shown for $f(x) = x, -1 \leq x \leq 1$ in Fig. 5.10. The blue curve gives the case for 2 terms in the Fourier series, which does not do very well at approximating the function. The red curve shows the case with 4 terms. While still not a good approximation, it does seem to be getting closer. Finally, the black curve gives the case with 24 terms in the expansion. While have numerous oscillations, this curve does follow the line $f(x) = 1$.

Next, consider the example

$$f(x) = \frac{1}{L}\cosh(x), \quad -L \leq x \leq L. \tag{5-85}$$

Inserting $f(x)$ into the expressions for the coefficients, we get

$$c_0 = \frac{1}{2L}\int_{-L}^{L}\frac{\cosh(x)}{L}dx = \frac{\sinh(L)}{L^2}, \tag{5-86a}$$

$$a_n = \frac{1}{L}\int_{-L}^{L}\sin\left(\frac{n\pi x}{L}\right)\frac{\cosh(x)}{L}dx = 0, \quad n > 0, \tag{5-86b}$$

$$b_n = \frac{1}{L}\int_{-L}^{L}\cos\left(\frac{n\pi x}{L}\right)\frac{\cosh(x)}{L}dx = (-1)^n\frac{2\sinh(L)}{L^2 + n^2\pi^2}, \quad n > 0. \tag{5-86c}$$

The expansion is therefore

$$\frac{1}{L}\cosh(x) = \frac{\sinh(L)}{L^2} + \sum_{n=1}^{\infty}(-1)^n\frac{2\sinh(L)}{L^2 + n^2\pi^2}\cos\left(\frac{n\pi x}{L}\right).$$

This is plotted in Fig. 5.11. The blue and red curves give the approximation of the true function (plotted in black) for $N = 2$ and $N = 4$ terms respectively. It turns out that unlike $f(x) = x$, not very many terms are required to accurately describe the hyperbolic trigonometric functions.

It may seem odd to write out such simple functions in terms of infinite sums of trigonometric functions; however, these were merely simple examples to illustrate the concept of Fourier series expansions. Recall that when it comes to solving second-order partial differential equations, it is often the case that the solutions have no

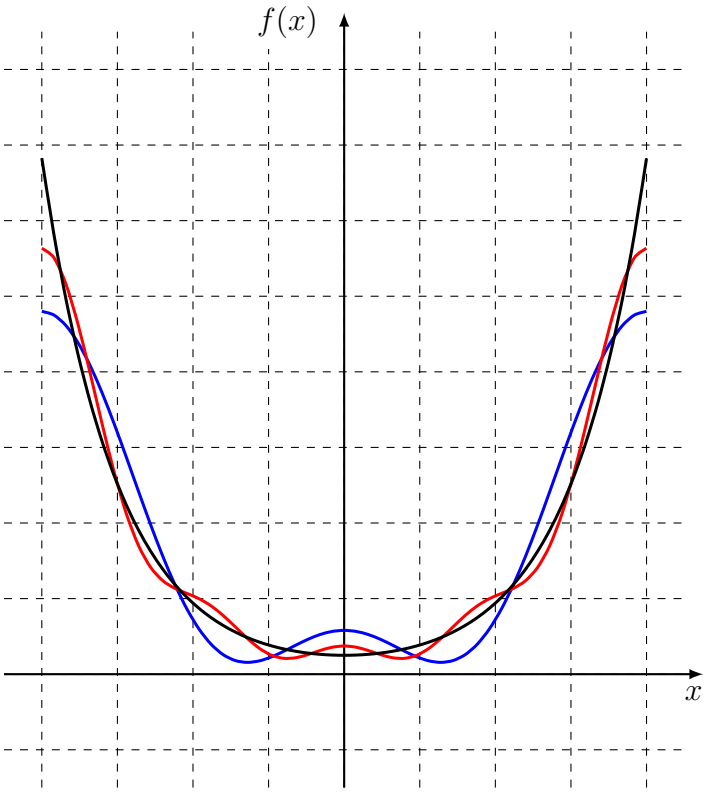

Figure 5.11: Illustration of a Fourier series expansion of $f(x) = \cosh(x)/4, -4 \leq x \leq 4$ for $N = 2$ terms (blue) and $N = 4$ terms (red) with the exact function (black).

closed form in terms of standard functions and we have no choice but to write those solutions as infinite sums of standard functinos.

Before concluding this discussion, it is worth mentioning that in each of our examples one of either the sine or cosine terms survived with the other vanishing. In first example, $f(x) = x/L$, we have an odd function. Since sine is also an odd function, those terms survive. Conversely, cosine is an even function and those, along with the constant term, vanished. In the second example, we were given an even function, $f(x) = \cosh(x)/L$ and the opposite occurs: the sine terms vanish, and the cosine plus constant terms survive.

It is often the case that we define our coordinate system for the differential equation so that one of the two sets of terms vanish. To illustrate, if we have the domain $0 \leq x \leq L$, we can then choose whether to represent a representation based on whether the function is even or odd, which is usually dictated by the boundary conditions.

For the case where $f(x)$ is odd, i.e., where $f(-x) = -f(x)$ should we temporarily extend the domain to negative values of x, i.e., $-L \leq x \leq 0$, then we all the cosine and constant (even) terms vanish. We can then note that for odd $f(x)$ that

$$\int_{-L}^{0} \sin\left(\frac{n\pi x}{L}\right) f(x)dx = \int_{0}^{L} \sin\left(\frac{n\pi x}{L}\right) f(x)dx. \tag{5-87}$$

This result is because $f(-x)$ and $\sin(-x)$ are $-f(x)$ and $-\sin(x)$ respectively and the two negatives cancel to make a positive integral. Therefore,

$$\begin{aligned} a_n &= \frac{1}{L}\int_{-L}^{L} \sin\left(\frac{n\pi x}{L}\right) f(x)dx \\ &= \frac{1}{L}\int_{-L}^{0} \sin\left(\frac{n\pi x}{L}\right) f(x)dx + \frac{1}{L}\int_{0}^{L} \sin\left(\frac{n\pi x}{L}\right) f(x)dx \\ &= \frac{2}{L}\int_{0}^{L} \sin\left(\frac{n\pi x}{L}\right) f(x)dx. \end{aligned} \tag{5-88}$$

This gives the Fourier expansion of an odd function $f(x)$ on the domain $0 \leq x \leq L$:

$$f(x) = \sum_{n=1}^{\infty} \left[\frac{2}{L}\int_{0}^{L} \sin\left(\frac{n\pi x'}{L}\right) f(x')dx'\right] \sin\left(\frac{n\pi x}{L}\right) \tag{5-89}$$

We can perform a similar analysis for the case where $f(x)$ is even, $f(-x) = f(x)$. The sine terms vanish leaving us with the constant plus cosine terms. This gives

$$c_0 = \frac{1}{2L}\int_{-L}^{L} f(x)dx = \frac{1}{L}\int_{0}^{L} f(x)dx, \tag{5-90}$$

and

$$b_n = \frac{1}{L}\int_{-L}^{L} \cos\left(\frac{n\pi x}{L}\right) f(x)dx = \frac{2}{L}\int_{0}^{L} \cos\left(\frac{n\pi x}{L}\right) f(x)dx. \tag{5-91}$$

Therefore, our Fourier expansion for even $f(x)$ on the domain $0 \leq x \leq L$ is

$$f(x) = \frac{1}{L}\int_{0}^{L} f(x')dx' + \sum_{n=1}^{\infty} \left[\frac{2}{L}\int_{0}^{L} \cos\left(\frac{n\pi x'}{L}\right) f(x')dx'\right] \cos\left(\frac{n\pi x}{L}\right). \tag{5-92}$$

5.3.b Separation of Variables

The solution technique that we can use to solve many linear second-order partial differential equations is called separation of variables. The technique involves guessing (hoping) that the solution to the partial differential equation can be described as a product of functions that each depend on only one of the independent variables. Given this guess, we insert the form of the solution into the partial differential equation and apply the initial and boundary conditions to check if we can obtain a solution consistent with this assumption. If this is indeed this case, then, since the partial differential equation is linear, we are then guaranteed to have the unique solution.

Given the 1-D heat equation in Cartesian coordinates, we guess that the solution can be written in the form of

$$u(x,t) = X(x)T(t). \tag{5-93}$$

Inserting this into the 1-D heat equation

$$\frac{1}{\alpha}\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0,$$

gives

$$\frac{X(x)}{\alpha}\frac{\partial}{\partial t}(T(t)) - T(t)\frac{\partial^2}{\partial x^2}(X(x)) = 0. \tag{5-94}$$

Note that we moved the α term to belong to the time term. This is not required; however, it will yield a more convenient form for the solution. Now we divide both sides of the equation by the solution $u = XT$ and rearrange to obtain

$$\frac{1}{\alpha T(t)}\frac{\partial}{\partial t}(T(t)) = \frac{1}{X(x)}\frac{\partial^2}{\partial x^2}(X(x)). \tag{5-95}$$

The terms on left-hand side are strictly functions of t and the terms on the right-hand side are strictly functions of x. The only way for the left- and right-hand sides of the equation to be equal is if they are both equal to constants:

$$\frac{1}{\alpha T(t)}\frac{dT}{dt} = C_1, \tag{5-96a}$$

$$\frac{1}{X(x)}\frac{d^2X}{dx^2} = C_2. \tag{5-96b}$$

These are now written as ordinary differential equations since they only depend upon a single variable. We can show that the constants must be equal by inserting the solution back into the equation.

$$C_1 = C_2. \tag{5-97}$$

The constant could be positive, negative, or even zero in some special cases. Depending on what happens, we will get different forms for the solution. To resolve this further, we need to evaluate the boundary conditions for a specific problem. We now proceed to analyze such a problem.

5.3.c Example: Transient Heat Conduction with Symmetric BCs

Let us consider the problem of heat conduction on the domain $0 \leq x \leq L$:

$$\frac{\partial u}{\partial t} - \alpha\nabla^2 u(\boldsymbol{x}, t) = 0, \quad u(x, 0) = f(x), \quad u(0, t) = u(L, t) = 0. \tag{5-98}$$

Here u describes the temperature of the system, α is the thermal diffusivity (which is the heat conductivity divided by the product of the density and specific heat capacity), an initial arbitrary temperature distribution $f(x)$ is prescribed, and the temperature is held constant at zero at the boundaries for all times.

Applying separation of variables with $u(x,t) = T(t)X(x)$, we get the result

$$\frac{1}{\alpha T(t)}\frac{dT}{dt} = C,$$
$$\frac{1}{X(x)}\frac{d^2X}{dx^2} = C.$$

Now we have to decide upon the whether C is positive, negative, or zero. For the case where C is zero, the differential equation in X becomes

$$\frac{d^2X}{dx^2} = 0. \tag{5-100}$$

This would yield solutions that is a line

$$X(x) = Ax + B. \tag{5-101}$$

The boundary conditions require that the solution go to zero on both ends, so the only solution that would satisfy this would be for the solution to be zero everywhere. This is a trivial solution and not of practical interest, so we reject this case. The case where $C > 0$, we have

$$\frac{d^2X}{dx^2} - CX(x) = 0. \tag{5-102}$$

This would yield exponentials or hyperbolic trigonometric functions:

$$X(x) = A\sinh(\sqrt{C}x) + B\cosh(\sqrt{C}x). \tag{5-103}$$

There are no sets of coefficients we could choose to force the solution go to zero at $x = 0$ and $x = L$ except for $A = 0$ and $B = 0$, which is again the trivial solution. We again, reject $C > 0$. Finally, we try a negative value of C. For convenience we define

$$C = -k^2 \tag{5-104}$$

which gives the differential equation

$$\frac{d^2X}{dx^2} + k^2X(x) = 0. \tag{5-105}$$

This gives the solution

$$X(x) = A\sin(kx) + B\cos(kx). \tag{5-106}$$

Applying our boundary condition at $X(0) = 0$ we have

$$X(0) = B = 0. \tag{5-107}$$

Next, applying the boundary condition at $X(L) =$ gives

$$X(L) = A\sin(kL) = 0. \tag{5-108}$$

Either $A = 0$, which is again the trivial solution, or

$$k = \frac{n\pi}{L}, \quad n = 1, 2, 3, \ldots \tag{5-109}$$

(The negative values of n would yield identical results with the sign flipped, so we leave those out, as they would be redundant.) This implies that there are infinitely many possible solutions to $X(x)$ that satisfy the differential equation and the boundary conditions. Since the equation is linear, we can take the sum of all of the solutions

$$X(x) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi x}{L}\right). \tag{5-110}$$

Note that by virtue of the boundary conditions $X(x)$ is an odd function if we were to temporarily extend the domain into the negative range.

Now, returning to the time term we have

$$\frac{1}{T}\frac{dT}{dt} = -\alpha k_n^2. \tag{5-111}$$

Here we denoted the constant as k_n to indicate that there are infinitely many values of k that solve the equation. Solving this gives

$$T(t) = Ke^{-\alpha k_n^2 t}. \tag{5-112}$$

where K is an arbitrary constant.

Multiplying these together, we have

$$u(x,t) = X(x)T(t) = \sum_{n=0}^{\infty} A_n \sin\left(\frac{n\pi x}{L}\right) Ke^{-\alpha k_n^2 t}.$$

Combining the constants and writing the solution explicitly gives the result:

$$u(x,t) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi x}{L}\right) \exp\left[-\alpha\left(\frac{n\pi}{L}\right)^2 t\right]. \tag{5-113}$$

Now there remains the matter of finding the constants A_n. This is done by applying the initial condition. We have

$$u(x,0) = f(x) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi x}{L}\right). \tag{5-114}$$

We can solve for the constants by multiplying by sine functions with different periods, integrating from $0 \le x \le L$.

$$\int_0^L \sin\left(\frac{m\pi x}{L}\right) f(x)dx = \sum_{n=1}^{\infty} A_n \int_0^L \sin\left(\frac{n\pi x}{L}\right) \sin\left(\frac{m\pi x}{L}\right) dx. \tag{5-115}$$

Applying orthogonality we get

$$\int_0^L \sin\left(\frac{m\pi x}{L}\right) f(x)dx = \sum_{n=1}^{\infty} A_n \frac{L}{2}\delta_{nm}. \tag{5-116}$$

The Kronecker delta ensures only the terms with $n = m$ are nonzero. Therefore, we can solve for the coefficient as

$$A_n = \frac{2}{L}\int_0^L \sin\left(\frac{n\pi x}{L}\right) f(x)dx. \tag{5-117}$$

Note that this is equivalent to the Fourier expansion of $f(x)$ for an odd function on the domain of $0 \leq x \leq L$.

Therefore, once we are given $f(x)$, we can obtain the coefficients and solution in terms of Fourier expansions. Unless $f(x)$ happens to just so happen to match some combination of sine waves itself, then the best we can do is represent the solution as an infinite sum of sine functions.

Let us illustrate this with an example. The domain is $L = 4$ and $\alpha = 0.1$ and the initial condition is

$$f(x) = \begin{cases} 4, & 2 \leq x \leq 3 \\ 0, & \text{otherwise} \end{cases}. \tag{5-118}$$

The Fourier expansion for these coefficients are

$$\begin{aligned} A_n &= 2\int_2^3 \sin\left(\frac{n\pi x}{4}\right) dx \\ &= \frac{8}{n\pi}\left[\cos\left(\frac{n\pi}{2}\right) - \cos\left(\frac{3n\pi}{4}\right)\right]. \end{aligned} \tag{5-119}$$

Because the function is discontinuous, we will require a very large number of coefficients to accurately represent the solution. The computations were done using a computer with $N = 1000$ coefficients. Fig. 5.12 provides plots of the solution at early times and later times. The early times show the initial condition spreading out in a diffusive process. Eventually, the shape of the solution begins to be significantly impacted by the boundary conditions. At later times, the solution evolves into a single exponentially decaying sine wave. This is because of the

$$\exp\left[-\alpha\left(\frac{n\pi}{L}\right)^2 t\right]$$

term. For large t the argument of the exponential becomes large making the exponential term small. Because the exponent is proportional to n^2, as t gets large, the terms for $n = 2, 3, \ldots$ make the exponent increasingly large, making the overall exponential vanishingly small. For late times the $n = 1$ term dominates the others and we are left with a single sine wave.

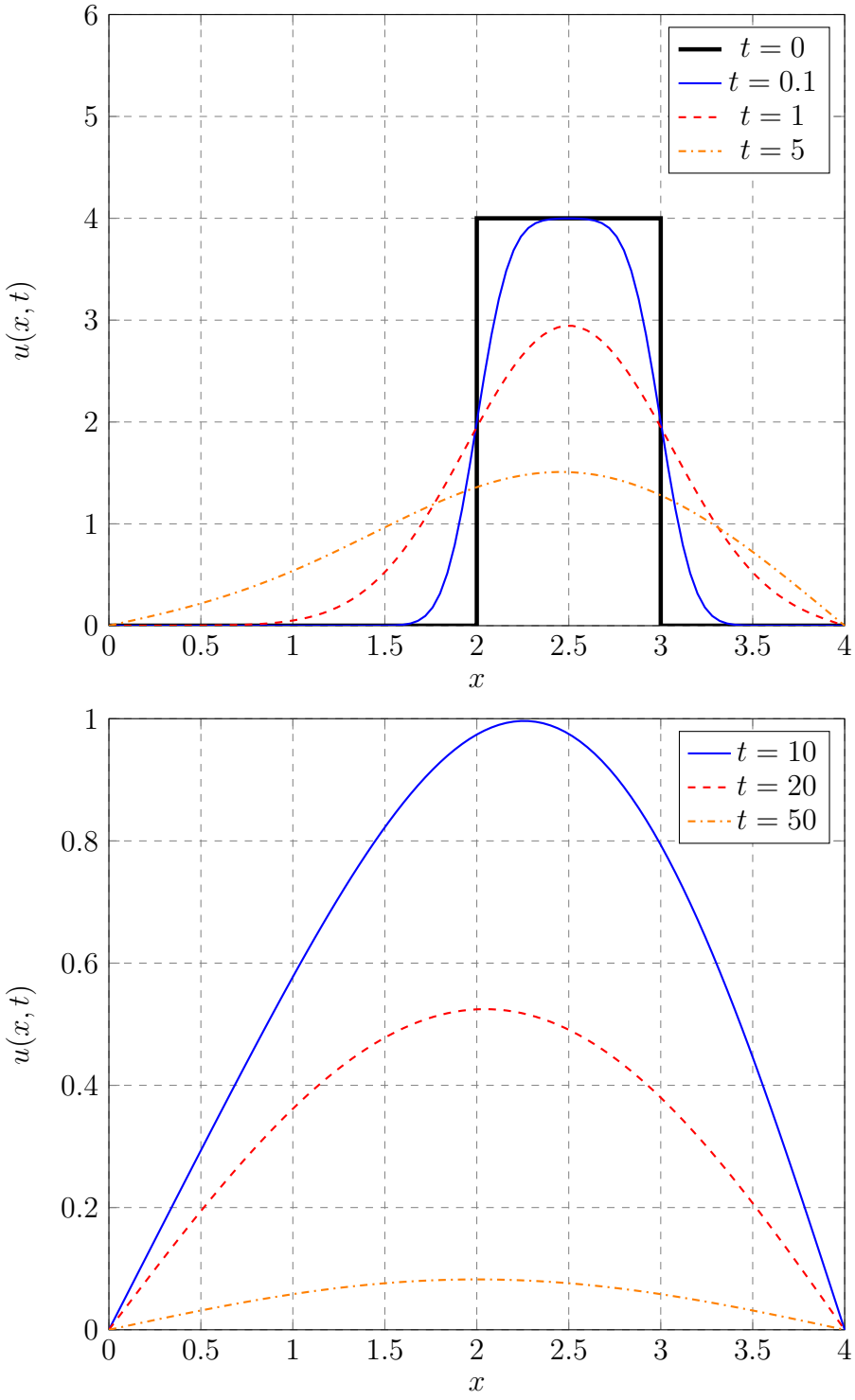

Figure 5.12: Plot of transient heat conduction for (top) early time and (bottom) late times.

5.3.d Example: Transient 1-D Neutron Diffusion and Criticality

The time-dependent neutron diffusion equation is

$$\frac{1}{v}\frac{\partial u}{\partial t} - D\frac{\partial^2 u}{\partial x^2} + \Sigma_a u(x,t) = \nu\Sigma_f u(x,t), \quad u(x,0) = f(x), \quad u(0,t) = u(L,t) = 0. \tag{5-120}$$

Here u represents the path-length density of neutrons, v is the neutron speed, D is the neutron diffusion coefficient, Σ_a is the absorption coefficient (macroscopic cross section), ν is the mean number of neutrons produced from fission, and Σ_f is the fission coefficient. We specify an initial condition for the neutrons $f(x)$ and require that the neutron distribution goes to zero on the edges. In practice, we have more sophisticated boundary conditions, but these complicate matters significantly and are beyond the scope of this section.

Before proceeding, let us first consider the difference of the neutron diffusion equation from the heat conduction equation. These differences are primarily the addition of absorption (loss) and fission (gain) terms. In the heat conduction problem we had no mechanism for thermal energy to be created or destroyed (e.g., converted into or from chemical energy through endothermic or exothermic chemical reactions).

We first rearrange the equation by dividing by the diffusion coefficient and move the absorption term to the right-hand side

$$\frac{1}{vD}\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = \frac{\nu\Sigma_f - \Sigma_a}{D}u(x,t). \tag{5-121}$$

As with the transient heat conduction problem, we assume the solution is separable of the form

$$u(x,t) = X(x)T(t), \tag{5-122}$$

insert this into the differential equation, and divide by u to obtain

$$\frac{1}{vDT}\frac{dT}{dt} - \frac{1}{X}\frac{d^2X}{dx^2} = \frac{\nu\Sigma_f - \Sigma_a}{D}. \tag{5-123}$$

We now have an equation where the right-hand side is equal to a constant and the time and space terms are only dependent on the time and space variables. Therefore, we can assert that the solutions are constants. As with the heat conduction, we need to analyze the possibilities. We will skip this step here and provide the one that will work and make the next steps easiest. We state that

$$\frac{1}{vDT}\frac{dT}{dt} = \frac{\nu\Sigma_f - \Sigma_a}{D} - B^2, \tag{5-124a}$$

$$\frac{1}{X}\frac{d^2X}{dx^2} = -B^2. \tag{5-124b}$$

Here B^2 is some constant for which we will find valid solutions. (This coefficient is referred to as the buckling, as it follows similar mathematics to structure mechanics, but otherwise has no physical connection.)

As for the heat conduction problem, the spatial function X depends upon the trigonometric functions

$$X(x) = A\sin(Bx) + C\cos(Bx). \tag{5-125}$$

Applying the boundary conditions $X(0) = X(L) = 0$ yields the same results as before:

$$C = 0, \tag{5-126a}$$

$$A = \frac{n\pi}{L}, \quad n = 1, 2, 3, \ldots \tag{5-126b}$$

Therefore, we have obtained infinitely many solutions to the X equation and can take a linear combination as with the heat conduction problem:

$$X(x) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi x}{L}\right). \tag{5-127}$$

Again, each term in the summation has a different constant coefficient.

The time term is obtained in a similar process, but now as a more complicated form

$$T(t) = K\exp\left[\left(\nu\Sigma_f - \Sigma_a - \left(\frac{n\pi}{L}\right)^2 D\right) vt\right]. \tag{5-128}$$

The values in the exponential will have a significant impact on the time-evolution and criticality of the system.

Multiplying the space and time terms together and combining constants gives the solution for 1-D transient neutron diffusion

$$u(x,t) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi x}{L}\right) \exp\left[\left(\nu\Sigma_f - \Sigma_a - \left(\frac{n\pi}{L}\right)^2 D\right) vt\right]. \tag{5-129}$$

The major difference with the heat conduction equation is in the exponential term for the time behavior. Namely, we have the term

$$\nu\Sigma_f - \Sigma_a - \left(\frac{n\pi}{L}\right)^2 D$$

which gives the production coefficient for fission, the loss coefficient for absorption, and a third term that is a loss term proportional to the neutron diffusion coefficient divided by the length of the reactor squared. The last term describes the physical phenomenon of leakage of neutrons from the system. Note that the least negative of these terms has $n = 1$ and, as we saw with the heat conduction problem, will be the term that survives the longest. We can state that if

$$\nu\Sigma_f = \Sigma_a + \left(\frac{\pi}{L}\right)^2 D$$

then the largest exponential term will be zero with all others negative. This is referred to as a critical reactor and implies that after long times the neutron distribution will reach a steady state sine wave distribution. The case where

$$\nu\Sigma_f < \Sigma_a + \left(\frac{\pi}{L}\right)^2 D$$

corresponds to a subcritical reactor. This means that the neutron population will reach a sine wave that is decaying exponentially in time, which is what we saw for the heat conduction problem. Finally, we have the case where

$$\nu\Sigma_f > \Sigma_a + \left(\frac{\pi}{L}\right)^2 D.$$

This is the supercritical configuration, which corresponds to a long time solution that is a sine wave that grows exponentially with time.

5.3.e Superposition

The two previous examples had zero boundary conditions and no inhomogeneous term. This simplified the analysis significantly. Now suppose instead we have boundary conditions that are not zero on both sides, but nonzero and potentially asymmetric. In these cases, we may not be able to easily find solutions that simultaneously satisfy all the conditions. Fortunately, because the heat equation is linear, we can apply the principle of superposition. The idea is rather than solving the entire problem at once, to break it up into simpler problems, combine the results, and then find values of the coefficients. We have seen similar analysis before in finding homogeneous and particular solutions of ordinary differential equations.

For the transient heat equation, we may split the solution into two parts:

$$u(x,t) = v(x,t) + s(x) \tag{5-130}$$

Here $v(x,t)$ represents the transient solution and $s(x)$ represents the steady-state solution. The mathematical justification for this is that because there is no growth term, eventually the solution $u(x,t)$ will reach some steady state value for large t. The process is to solve for the steady-state solution $s(x)$ separately, use that solution to form a new problem for the transient part $v(x,t)$, solve the transient part, and then combine the solutions to obtain $u(x,t)$.

5.3.f Example: Transient Heat Conduction with Asymmetric BCs

Let us illustrate this idea with the following transient heat conduction problem:

$$\frac{\partial u}{\partial t} - \alpha\frac{\partial^2 u}{\partial x^2} = 0, \quad u(x,0) = f(x), \quad u(0,t) = 0, \quad u(L,t) = u_r. \tag{5-131}$$

The difference now being the boundary condition at $x = L$, which is now some value u_r. We will now apply the superposition principle and write the solution as a linear combination of a transient part and steady state part as described before:

$$u(x,t) = v(x,t) + s(x).$$

We know that $s(x)$ must satisfy the partial differential equation and the boundary conditions because it is the limiting case as $t \to \infty$. Inserting the steady-state solution $s(x)$ into the differential equation yields a simple ordinary differential equation:

$$-\alpha \frac{d^2 s}{dx^2} = 0, \quad s(0) = 0, \quad s(L) = u_r. \tag{5-132}$$

Solving this equation gives the coefficients

$$s(x) = A_0 x + B_0. \tag{5-133}$$

Applying the boundary conditions gives

$$A_0 = \frac{u_r}{L}, \tag{5-134a}$$

$$B_0 = 0. \tag{5-134b}$$

Therefore the steady-state solution is

$$s(x) = \frac{u_r x}{L}. \tag{5-135}$$

Now we have to set up a problem for the transient solution $v(x,t)$. Let us take the original differential equation and expand out $u(x,t)$:

$$\frac{\partial}{\partial t}(v(x,t) + s(x)) - \alpha \frac{\partial^2}{\partial x^2}(v(x,t) + s(x)) = 0. \tag{5-136}$$

Since $s(x)$ is not a function of time, the partial derivative of $s(x)$ with respect to time goes to zero. From the solution we just obtained, we see that the second partial derivative of $s(x)$ with respect to x is also zero, therefore $v(x,t)$ satisfies the same differential equation

$$\frac{\partial v}{\partial t} - \alpha \frac{\partial^2 v}{\partial x^2} = 0, \tag{5-137}$$

subject to initial and boundary conditions that we need to obtain. To do this, let us insert the expanded form of $u(x,t)$ into the initial condition

$$\begin{aligned} u(x,0) &= v(x,0) + s(x) = f(x), \\ v(x,0) &= f(x) - s(x). \end{aligned} \tag{5-138}$$

Therefore, the initial condition for the transient solution is the initial condition for the solution with the steady-state solution subtracted off. Doing the same thing for the $x = 0$ boundary condition gives

$$\begin{aligned} u(0,t) &= v(0,t) + s(0) = 0, \\ v(0,t) &= 0. \end{aligned} \tag{5-139}$$

And likewise for the $x = L$ boundary condition

$$\begin{aligned} u(L,t) &= v(L,t) + s(L) = u_r, \\ v(L,t) + u_r &= u_r, \\ v(L,t) &= 0. \end{aligned} \tag{5-140}$$

Therefore, the problem for the transient case becomes

$$\frac{\partial v}{\partial t} - \alpha \frac{\partial^2 v}{\partial x^2} = 0, \quad v(x,0) = f(x) - s(x), \quad v(0,t) = 0, \quad v(L,t) = 0. \tag{5-141}$$

This is essentially identical to what we had before except that the initial condition is a different function. From our solution of the symmetric transient heat conduction we know that

$$v(x,t) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi x}{L}\right) \exp\left[-\alpha \left(\frac{n\pi}{L}\right)^2 t\right] \tag{5-142}$$

with the coefficients as

$$A_n = \frac{2}{L} \int_0^L \sin\left(\frac{n\pi x}{L}\right) \left(f(x) - \frac{u_r}{L} x\right) dx;$$

which we can expand out to be

$$A_n = (-1)^n \frac{2u_r}{n\pi} + \frac{2}{L} \int_0^L \sin\left(\frac{n\pi x}{L}\right) f(x) dx. \tag{5-143}$$

Therefore, the solution to the asymmetric transient heat conduction problem is

$$u(x,t) = \frac{u_r x}{L} + \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi x}{L}\right) \exp\left[-\alpha \left(\frac{n\pi}{L}\right)^2 t\right]. \tag{5-144}$$

Let us now try an example where the initial condition satisfies $f(x,0) = 0$, a flat temperature, with $L = 4$ and $\alpha = 0.1$. Instantaneously at $t = 0$, the temperature of the right boundary is brought to $u_r = 4$. From inserting numbers, we know that the steady-state solution will be a straight line with $u = 0$ and $x = 0$ and $u = 4$ at $x = 4$.

A few snapshots of the solution in time are shown in Fig. 5.13. At $t = 1$, the thermal energy begins to diffuse into the medium and raise the temperature near the right-boundary. The left boundary is largely unaffected, as there has not yet been sufficient time for the thermal energy to propagate through the entire domain. At $t = 10$, the thermal energy has reached the left boundary, but the shape of the temperature field still has curvature as the temperature has not yet equilibrated. By $t = 100$, the system has almost reached its equilibrium state where the temperature field is given by a straight line connecting the two boundaries.

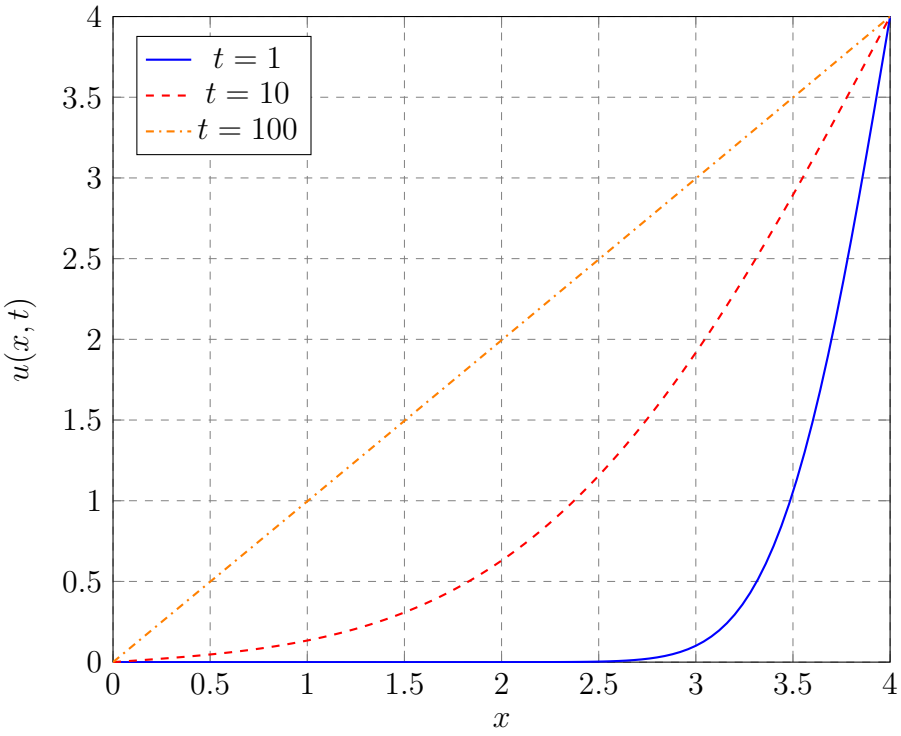

Figure 5.13: Plot of transient heat conduction for the case with different temperatures specified at the domain boundaries.

5.3.g Example: Transient Heat Conduction with Constant Source

The superposition principle also allows us to solve cases where we have an inhomogeneous term. Consider the problem

$$\frac{\partial u}{\partial t} - \alpha \frac{\partial^2 u}{\partial x^2} = Q, \quad u(x,0) = f(x), \quad u(0,t) = u(L,t) = 0. \tag{5-145}$$

Here Q represents a constant heat generation rate. As before, we apply superposition and write the solution as a transient plus steady-state solution:

$$u(x,t) = v(x,t) + s(x).$$

Inserting the steady state solution into the differential equation gives the ordinary differential equation

$$-\alpha \frac{d^2 s}{dx^2} = Q, \quad u(0,t) = u(L,t) = 0. \tag{5-146}$$

Solving this and applying the boundary conditions yields the steady-state solution

$$s(x) = \frac{Q}{2\alpha} x(L - x). \tag{5-147}$$

Now we insert the expanded form of $u(x,t)$ into the original differential equation

$$\frac{\partial v}{\partial t} - \alpha \frac{\partial^2 v}{\partial x^2} - \alpha \frac{\partial^2}{\partial x^2}\left[\frac{Q}{2\alpha}x(L-x)\right] = Q. \tag{5-148}$$

Since

$$-\alpha \frac{\partial^2}{\partial x^2}\left[\frac{Q}{2\alpha}x(L-x)\right] = Q$$

this will cancel out the inhomogeneous term to yield a homogeneous problem. We perform the same analysis as with the case with asymmetric boundary conditions to fully describe the problem for the transient part

$$\frac{\partial v}{\partial t} - \alpha \frac{\partial^2 v}{\partial x^2} = 0, \quad v(x,0) = f(x) - s(x), \quad v(0,t) = 0, \quad v(L,t) = 0, \tag{5-149}$$

which is identical to the previous problem except that $s(x)$ has a different form. The solution to the transient problem is

$$v(x,t) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi x}{L}\right) \exp\left[-\alpha\left(\frac{n\pi}{L}\right)^2 t\right]. \tag{5-150}$$

Adding on the steady-state solution gives the full solution

$$u(x,t) = \frac{Q}{2\alpha}x(L-x) + \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi x}{L}\right) \exp\left[-\alpha\left(\frac{n\pi}{L}\right)^2 t\right] \tag{5-151}$$

where the coefficients are

$$\begin{aligned} A_n &= \frac{2}{L}\int_0^L \sin\left(\frac{n\pi x}{L}\right)\left(f(x) - \frac{Q}{2\alpha}x(L-x)\right)dx \\ &= \frac{2QL^2}{\alpha n^3\pi^3}\left[(-1)^n - 1\right] + \frac{2}{L}\int_0^L \sin\left(\frac{n\pi x}{L}\right) f(x)dx. \end{aligned} \tag{5-152}$$

Consider an example where $L = 4$, $Q = 1$, $\alpha = 0.1$, and the initial condition

$$f(x) = \begin{cases} 10, & 2 < x < 4 \\ 0, & \text{otherwise} \end{cases}. \tag{5-153}$$

From plugging in numbers, we know that the equilibrium solution will be parabolic with a maximum value of $u = 20$ at the center $x = 2$ and going to zero at the boundaries.

Snapshots of the solution at various times are given in Fig. 5.14. At early times, the initial temperature distribution begins to spread out because of thermal conduction. Additionally, the temperature field rises as the constant heat source adds thermal energy to the system. At late times, the shape of the temperature distribution takes a parabolic shape and slowing rises to its steady state distribution.

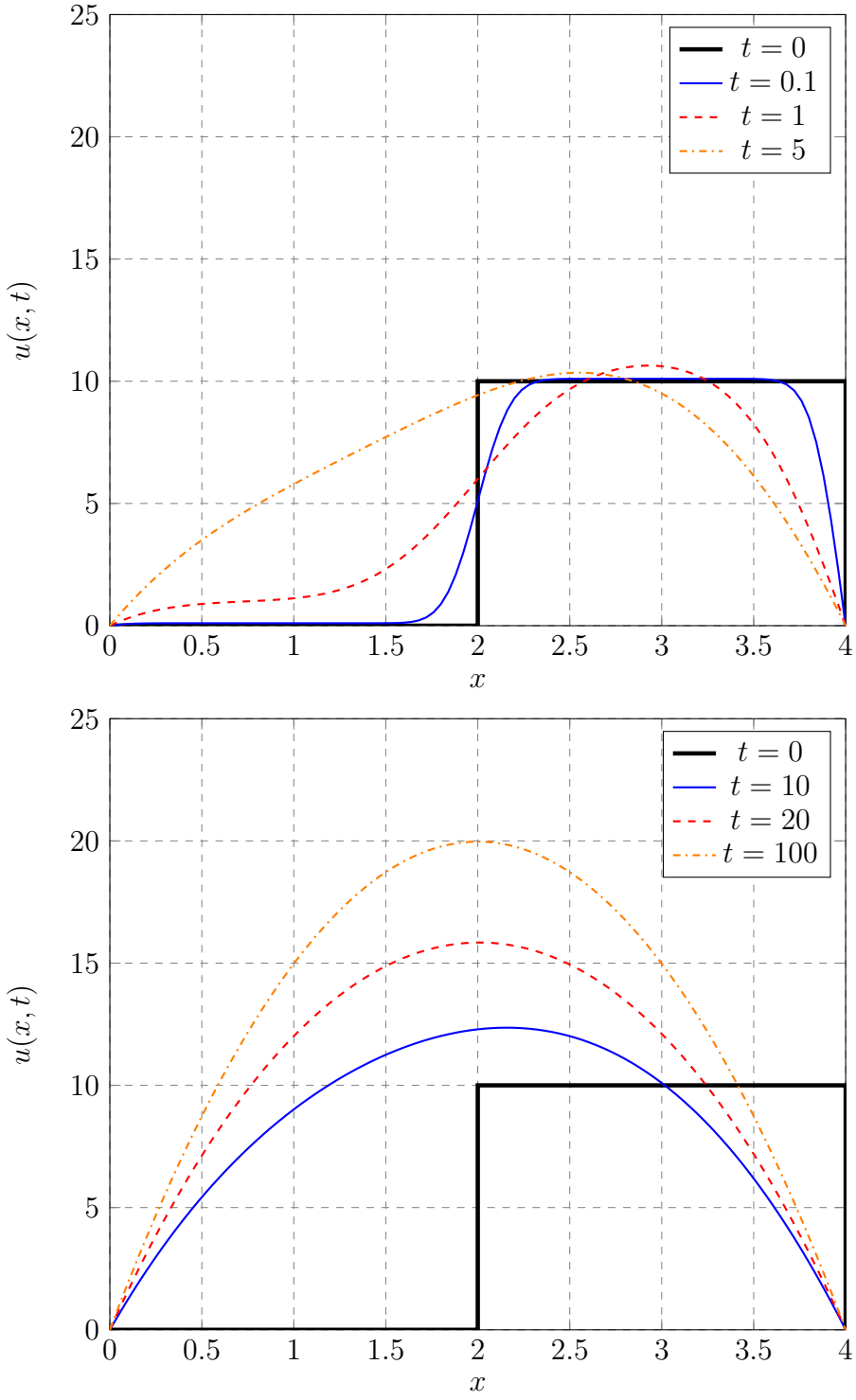

Figure 5.14: Plot of transient heat conduction with a constant source and an initial condition for (top) early time and (bottom) late times.

5.4 Laplace Equation

The Laplace equation is given by

$$\nabla^2 u = 0 \tag{5-154}$$

with the appropriate domain specified and boundary conditions given. Unlike the heat equation, this equation will involve only second derivatives. These notes will first cover a couple cases in electrostatics and heat transfer (conduction) in 2-D and 3-D Cartesian geometry. This will then be extended to spherical geometry for azimuthally symmetric boundary conditions.

5.4.a Example: Electric Field in an Infinite Square Duct

The electric field in a steady-state problem is given by the gradient of a potential function:

$$\mathbf{E} = -\nabla V. \tag{5-155}$$

Taking the divergence of the electric field and applying Gauss' law gives

$$\nabla \cdot \mathbf{E} = -\nabla^2 V = \frac{\rho}{\epsilon_0}. \tag{5-156}$$

Now suppose we have a square duct with dimensions $0 \leq x \leq L$, $0 \leq y \leq L$ that is infinite and uniform in the z direction. Within this duct, we will assume that there is no free electric charge. Finally, we will set one of the faces, $y = L$ to have an electric potential V_0 and the other faces set to zero. This problem is illustrated in Fig. 5.15.

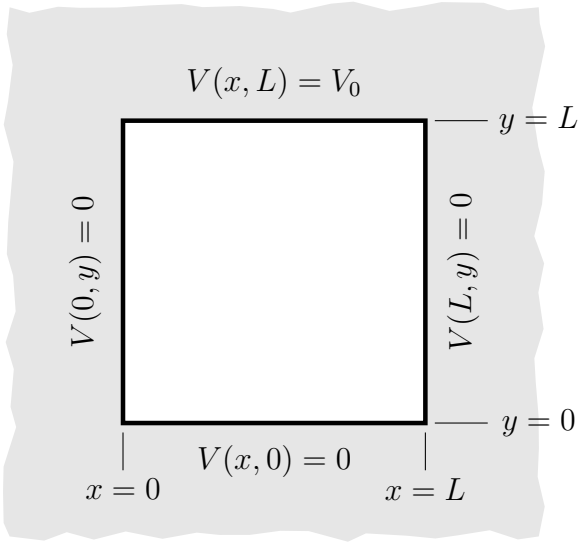

Figure 5.15: Illustration of a slice through an infinite square duct with prescribed electric potentials.

The problem statement for the divergence of the electric field can be written as the Laplace equation for the electric potential function:

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = 0, \quad V(0,y) = V(L,y) = V(x,0) = 0, \quad V(x,L) = V_0 \tag{5-157}$$

on the domain $0 \leq x \leq L$, $0 \leq y \leq L$. We expect the derivative of the potential in z to be zero because the prescribed potential function is uniform in the z direction.

To begin, we assume the solution is separable in x and y:

$$V(x,y) = X(x)Y(y). \tag{5-158}$$

Inserting the proposed separable solution and dividing by $V(x,y) = X(x)Y(y)$ gives

$$\frac{1}{X}\frac{\partial^2 X}{\partial x^2} + \frac{1}{Y}\frac{\partial^2 Y}{\partial y^2} = 0. \tag{5-159}$$

Similar to the 1-D heat equation, we have a function of x plus a function of y is equal to zero. The difference being that each equation is now the same sign, which will be important for the form of the solution. The only way for this to be true is if both terms are equal to a constant:

$$\frac{1}{X}\frac{\partial^2 X}{\partial x^2} = K_1 \tag{5-160a}$$

$$\frac{1}{Y}\frac{\partial^2 Y}{\partial y^2} = K_2. \tag{5-160b}$$

From the partial differential equation we know that

$$K_1 = -K_2, \tag{5-161}$$

or that the constants are equal and opposite to one another. At this point, we must assess whether the constant C_1 is positive, negative, or zero by trying each of the cases.

For the case where $K_1 = K_2 = 0$, we end up with

$$\frac{1}{X}\frac{\partial^2 X}{\partial x^2} = 0 \tag{5-162a}$$

$$\frac{1}{Y}\frac{\partial^2 Y}{\partial y^2} = 0, \tag{5-162b}$$

which have the solutions

$$X(x) = Ax + B \tag{5-163a}$$

$$Y(y) = Cy + D. \tag{5-163b}$$

Based on the boundary conditions in the x coordinate, we have that $V(0,y) = V(L,y) = 0$. This gives the equations for $X(x)$ as

$$X(0) = B = 0 \tag{5-164a}$$

$$X(L) = AL + B = 0. \tag{5-164b}$$

From the first of these, we end up with $B = 0$. The second gives $A = 0$. This would imply that $X(x) = 0$ everywhere, meaning $V(x, y) = 0$ everywhere, which is not a solution that can satisfy the boundary condition $V(x, L) = V_0$. Because this would lead to an inconsistency, we reject the case where $K_1 = K_2 = 0$.

For the case where $K_1 > 0$, we set

$$\begin{aligned} K_1 &= k^2, \\ K_2 &= -k^2. \end{aligned}$$

It follows that the solutions are

$$X(x) = A\sinh(kx) + B\cosh(kx) \tag{5-165a}$$
$$Y(y) = C\sin(ky) + D\cos(ky). \tag{5-165b}$$

From the boundary conditions $V(0, y) = V(L, y) = 0$, we have

$$\begin{aligned} X(0) &= B = 0, \\ X(L) &= A\sinh(kL) = 0. \end{aligned}$$

The only way for

$$\sinh(kL) = 0$$

is for $k = 0$. (This contrasts with the trigonometric functions, which are periodic.) The other case is where $A = 0$. In both of these cases, this would again imply $X(x) = 0$ everywhere, or $V(x, y) = 0$ everywhere. This is, as before, inconsistent with the boundary condition $V(x, L) = V_0$ and therefore the case with $K_1 > 0$ must be rejected.

We therefore conclude that $K_1 < 0$. Before proceeding, it is worth noting that in our examples we ruled out the other two cases; however, as we say for the heat conduction problem, we did end up with parts of the solution that implicitly satisfied the $C_1 = 0$ case and led to the linear solution—we did this by separating out the transient from the steady state solutions, but the same result is true nonetheless. It is therefore necessary to consider all possible cases and include them in the solution.

Proceeding with $K_1 < 0$, we define

$$\begin{aligned} K_1 &= k^2, \\ K_2 &= -k^2, \end{aligned}$$

which leads to the solutions

$$X(x) = A\sin(kx) + B\cos(kx) \tag{5-167a}$$
$$Y(y) = C\sinh(ky) + D\cosh(ky). \tag{5-167b}$$

Now as before, we apply the boundary conditions $V(0, y) = V(L, y) = 0$. For the first boundary condition at $x = 0$ we have

$$X(0) = B = 0. \tag{5-168a}$$

For the second boundary condition at $x = L$ we have

$$X(L) = A\sin(kL) = 0. \tag{5-168b}$$

This can be true when $A = 0$, which yields the trivial solution or where kL is an integer multiple of π. Therefore, as with the heat equation, we have

$$k_n = \frac{n\pi}{L}, \quad n = 1, 2, 3\ldots \tag{5-169}$$

Therefore, for $C_1 < 0$, we do have meaningful solutions to the $X(x)$ equation. In fact, we have infinitely many solutions, so we include all of them as with the heat equation:

$$X(x) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi x}{L}\right). \tag{5-170}$$

The electric potential is therefore

$$V(x,y) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi x}{L}\right)\left[C\sinh\left(\frac{n\pi y}{L}\right) + D\cosh\left(\frac{n\pi y}{L}\right)\right]. \tag{5-171}$$

Absorbing the constants we have

$$V(x,y) = \sum_{n=1}^{\infty} \left[C_n \sinh\left(\frac{n\pi y}{L}\right) + D_n\cosh\left(\frac{n\pi y}{L}\right)\right]\sin\left(\frac{n\pi x}{L}\right). \tag{5-172}$$

Next, we must resolve the other two boundary conditions to solve for the remaining sets of constants. First, let us use $V(x,0) = 0$. This gives

$$V(x,0) = \sum_{n=1}^{\infty} D_n \sin\left(\frac{n\pi x}{L}\right) = 0. \tag{5-173}$$

The only way for this to be true is if all $D_n = 0$. Next, applying the $V(x,L) = V_0$ boundary condition gives

$$V(x,L) = \sum_{n=1}^{\infty} C_n \sinh(n\pi)\sin\left(\frac{n\pi x}{L}\right) = V_0. \tag{5-174}$$

Note that $C_n\sinh(n\pi)$ is just a constant and that this has the expression of a Fourier series expansion. Therefore, this constant may be expressed as

$$C_n\sinh(n\pi) = \frac{2}{L}\int_0^L \sin\left(\frac{n\pi x}{L}\right)V_0 dx, \quad n = 1, 2, 3\ldots \tag{5-175}$$

Carrying out the integral gives

$$C_n\sinh(n\pi) = \frac{2V_0}{n\pi}\left(1 - \cos(n\pi)\right), \quad n = 1, 2, 3\ldots \tag{5-176}$$

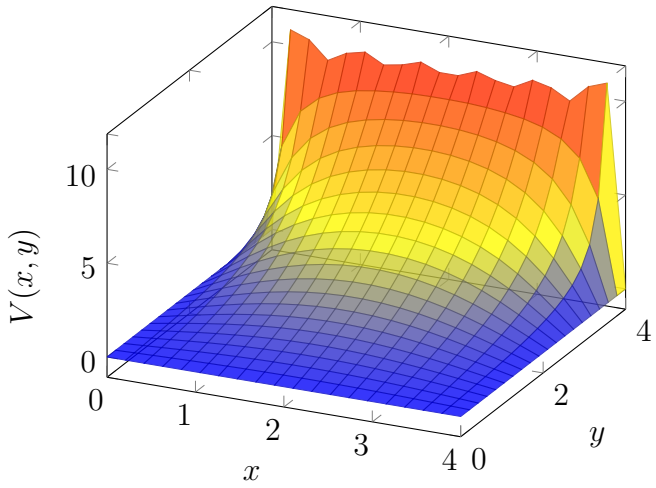

Figure 5.16: Approximate solution for the electric potential of a 2-D slice within the square duct.

However, recall that for integer n

$$\cos(n\pi) = (-1)^n, \quad n = 1, 2, 3 \ldots$$

The term $1 - \cos(n\pi)$ is either 2 when n is odd or 0 when n is even. Therefore, we can write the constant as

$$C_n = \begin{cases} \dfrac{4V_0}{n\pi \sinh(n\pi)}, & n \text{ odd} \\ 0, & n \text{ even} \end{cases}, \quad n = 1, 2, 3 \ldots \tag{5-177}$$

Therefore, the solution for the electric potential is

$$V(x,y) = \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{4V_0}{n\pi \sinh(n\pi)} \sin\left(\frac{n\pi x}{L}\right) \sinh\left(\frac{n\pi y}{L}\right). \tag{5-178}$$

Figure 5.16 shows an approximate solution with $N = 25$ terms in the expansion for $L = 4$ and $V_0 = 10$. As we can see, there are some spurious oscillations on the (x, L) surface, however, overall it does line up with what we expect for the solution. Adding more terms would improve the smoothness of the plot.

The electric field is again $\mathbf{E} = -\nabla V$, therefore by taking the gradient we have

$$\mathbf{E}(x,y) = -\sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{4V_0}{L \sinh(n\pi)} \left[\cos\left(\frac{n\pi x}{L}\right) \sinh\left(\frac{n\pi y}{L}\right) \hat{\mathbf{i}} + \sin\left(\frac{n\pi x}{L}\right) \cosh\left(\frac{n\pi y}{L}\right) \hat{\mathbf{j}}\right]. \tag{5-179}$$

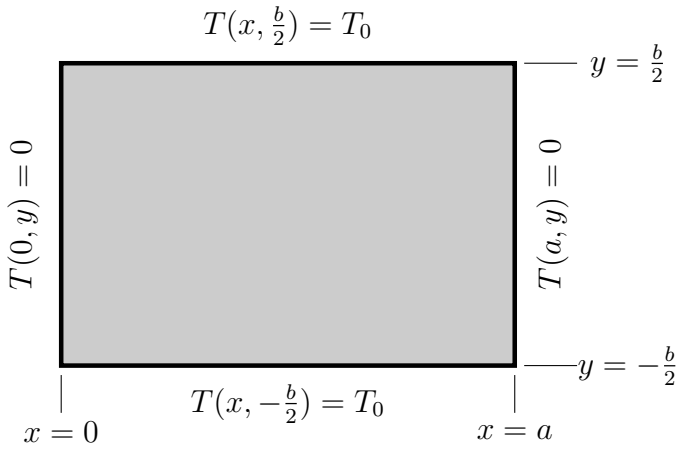

Figure 5.17: Illustration of the first example for heat conduction on a 2-D rectangular plate.

5.4.b Example: Heat Conduction on a Rectangular Plate

Suppose we have a thin rectangular plate of thicknesses a and b in the x and y directions respectively. The plate is held at constant temperature zero on two of the left and right sides and T_0 at the top and bottom sides. This is depicted in Fig. 5.17 and will be the first of two examples in this section.

We wish to obtain the temperature distribution $T(x, y)$ within the rectangular plate. To solve this problem, we define the origin in a specific way to simplify the boundary conditions using $0 \leq x \leq a$, and $-b/2 \leq y \leq b/2$ where

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0, \quad T(0, y) = T(a, y) = 0, \quad T(x, -\tfrac{b}{2}) = T(x, \tfrac{b}{2}) = T_0. \tag{5-180}$$

As with the electric potential problem, we assume a separable solution $T(x, y) = X(x)Y(y)$ and plug it into the differential equation to obtain

$$\frac{1}{X}\frac{\partial^2 X}{\partial x^2} + \frac{1}{Y}\frac{\partial^2 Y}{\partial y^2} = 0.$$

The solutions have each term as equal and opposite constants K_1 and K_2. As before, we would attempt to find solutions satisfying the different boundary conditions that give the sign of the constants. We can reject $K_1 = 0$ and $K_1 > 0$ using the same arguments as with the electric potential problem, so they will not be repeated here. Using the $K_1 < 0$ case, everything is identical until we reach the point

$$T(x, y) = \sum_{n=1}^{\infty} \left[C_n \sinh\left(\frac{n\pi y}{a}\right) + D_n \cosh\left(\frac{n\pi y}{a}\right) \right] \sin\left(\frac{n\pi x}{a}\right). \tag{5-181}$$

Now we must consider the boundary conditions $T(x, -b/2) = T(x, b/2) = T_0$. This gives a symmetric solution in y. Since the hyperbolic trigonometric functions are not

periodic, only the hyperbolic cosine terms may satisfy this symmetry condition. We can therefore reject the odd functions in y and set $C_n = 0$. This gives

$$T(x,y) = \sum_{n=1}^{\infty} D_n \sin\left(\frac{n\pi x}{a}\right) \cosh\left(\frac{n\pi y}{a}\right). \tag{5-182}$$

Inserting either of the $T(x,-\frac{b}{2}) = T(x,\frac{b}{2}) = T_0$ gives

$$T(x,\pm\tfrac{b}{2}) = \sum_{n=1}^{\infty} D_n \cosh\left(\frac{n\pi b}{2a}\right) \sin\left(\frac{n\pi x}{a}\right) = T_0. \tag{5-183}$$

As before, we recognize the terms on the left of the summation as constants that are coefficients in a Fourier series expansion. This gives the relationship

$$D_n \cosh\left(\frac{n\pi b}{2a}\right) = \frac{2}{a}\int_0^a \sin\left(\frac{n\pi x}{a}\right) T_0\, dx, \quad n = 1,2,3\ldots \tag{5-184}$$

which has the result

$$D_n = \begin{cases} \dfrac{4T_0}{n\pi \cosh(\frac{n\pi b}{2a})}, & n \text{ odd} \\ 0, & n \text{ even} \end{cases}, \quad n = 1,2,3\ldots \tag{5-185}$$

This gives the solution for the temperature field as

$$T(x,y) = \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{4T_0}{n\pi \cosh(\frac{n\pi b}{2a})} \sin\left(\frac{n\pi x}{a}\right) \cosh\left(\frac{n\pi y}{a}\right). \tag{5-186}$$

This problem was largely identical to the previous problem except that the boundary condition was nonzero on the bottom surface. Because of the symmetry with the top surface, this motivated choosing the origin for y to be at the center of the plate. This made the boundary conditions cause the hyperbolic sine term to vanish and yielded a solution that was largely identical to the other case except that the hyperbolic cosine term is used.

Let us now consider another 2-D heat conduction problem with different boundary conditions

$$\begin{aligned} &\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0, \\ &T(0,y) = T(a,y) = 0, \quad T(x,-\tfrac{b}{2}) = T_0, \quad T(x,\tfrac{b}{2}) = 2T_0. \end{aligned} \tag{5-187}$$

This problem is illustrated in Fig. 5.18.

To solve this problem, we apply the principle of superposition and write the temperature field as the sum of two temperature fields

$$T(x,y) = T_1(x,y) + T_2(x,y). \tag{5-188}$$

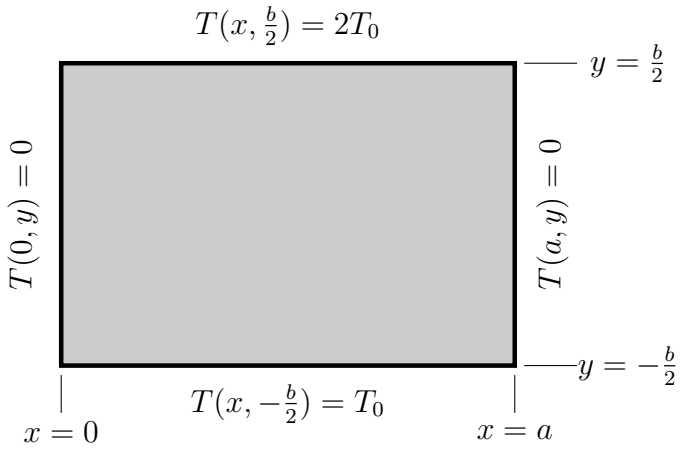

Figure 5.18: Illustration of the second example for heat conduction on a 2-D rectangular plate.

We define the temperature field $T_1(x, y)$ to satisfy the problem we just solved

$$\frac{\partial^2 T_1}{\partial x^2} + \frac{\partial^2 T_1}{\partial y^2} = 0, \quad T_1(0, y) = T_1(a, y) = 0, \quad T_1(x, -\tfrac{b}{2}) = T_1(x, \tfrac{b}{2}) = T_0. \tag{5-189}$$

It follows from what was done previously that the solution is

$$T_1(x, y) = \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{4T_0}{n\pi \cosh(\frac{n\pi b}{2a})} \sin\left(\frac{n\pi x}{a}\right) \cosh\left(\frac{n\pi y}{a}\right). \tag{5-190}$$

Now we build the problem for $T_2(x, y)$ by revisiting the differential equation

$$\begin{aligned} &\frac{\partial^2}{\partial x^2}(T_1 + T_2) + \frac{\partial^2}{\partial y^2}(T_1 + T_2) = 0, \\ &\frac{\partial^2 T_1}{\partial x^2} + \frac{\partial^2 T_1}{\partial y^2} + \frac{\partial^2 T_2}{\partial x^2} + \frac{\partial^2 T_2}{\partial y^2} = 0. \end{aligned}$$

However, we know that the first two terms on the left-hand side are zero because of how the problem for $T_1(x, y)$ is defined. This gives the differential equation

$$\frac{\partial^2 T_2}{\partial x^2} + \frac{\partial^2 T_2}{\partial y^2} = 0. \tag{5-191}$$

Now we must resolve the boundary conditions for $T_2(x, y)$. For the boundary condition at $(0, y)$ we have

$$\begin{aligned} T(0, y) &= T_1(0, y) + T_2(0, y) = 0 + T_2(0, y) = 0 \\ T_2(0, y) &= 0. \end{aligned} \tag{5-192a}$$

It follows from the same line of reasoning that

$$T_2(a, y) = 0. \tag{5-192b}$$

For the boundary condition $(x, -\frac{b}{2})$ we have

$$\begin{aligned} T(x, -\tfrac{b}{2}) &= T_1(x, -\tfrac{b}{2}) + T_2(x, -\tfrac{b}{2}) = T_0 + T_2(x, -\tfrac{b}{2}) = T_0 \\ T_2(x, \tfrac{b}{2}) &= 0. \end{aligned} \tag{5-192c}$$

And for the boundary condition $(x, \frac{b}{2})$ we have

$$\begin{aligned} T(x, \tfrac{b}{2}) &= T_1(x, \tfrac{b}{2}) + T_2(x, \tfrac{b}{2}) = T_0 + T_2(x, \tfrac{b}{2}) = 2T_0 \\ T_2(x, \tfrac{b}{2}) &= T_0. \end{aligned} \tag{5-192d}$$

This gives a problem that is effectively identical to the 2-D electrostatic potential problem except for how the coordinate system is defined. To make this problem identical, we define a new coordinate system using the translation:

$$u = x, \tag{5-193a}$$

$$v = y + \frac{b}{2}. \tag{5-193b}$$

This leads to the problem

$$\frac{\partial^2 T_2}{\partial u^2} + \frac{\partial^2 T_2}{\partial v^2} = 0, \quad T_2(0, v) = T_2(a, v) = T_2(u, 0) = 0, \quad T_2(u, b) = 1. \tag{5-194}$$

From our electrostatics problem, we obtain the equivalent solution:

$$T_2(u, v) = \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{4T_0}{n\pi \sinh(\frac{n\pi b}{a})} \sin\left(\frac{n\pi u}{a}\right) \sinh\left(\frac{n\pi v}{a}\right). \tag{5-195}$$

Substituting in values of x and y gives

$$T_2(x, y) = \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{4T_0}{n\pi \sinh(\frac{n\pi b}{a})} \sin\left(\frac{n\pi x}{a}\right) \sinh\left(\frac{n\pi(y + \frac{b}{2})}{a}\right). \tag{5-196}$$

Applying superposition, we can arrive at our combined solution for the temperature field

$$T(x, y) = \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \frac{4T_0}{n\pi} \sin\left(\frac{n\pi x}{a}\right) \left[\frac{\cosh\left(n\pi y/a\right)}{\cosh(n\pi b/(2a))} + \frac{\sinh\left(n\pi(y + \frac{b}{2})/a\right)}{\sinh(n\pi b/a)}\right]. \tag{5-197}$$

An approximation solution using $N = 15$ is plotted in Fig. 5.19 using the values of $a = 3$, $b = 2$, and $T_0 = 100$. As with the 2-D electrostatics problem, the solution exhibits spurious oscillations that would diminish as N becomes very large. The solution does, however, follow the general trends of satisfying the boundary conditions with zero on the sides, and $T = 100$ and $T = 200$ on the other two ends.

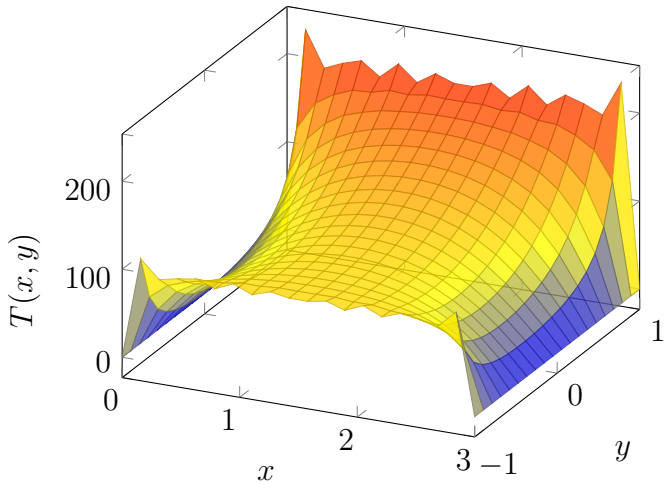

Figure 5.19: Approximate solution for the temperature field on a rectangular plate for the second example problem.

5.4.c Example: Electric Field in a Semi-infinite Rectangular Duct

Now consider the case where we have a square 3-D duct with $0 \leq x \leq a$, $0 \leq y \leq b$, defined over the right half-space in the z direction such that $0 \leq z < \infty$. We set the electric potential to zero on all faces except for the $z = 0$ face, which we set to some constant V_0. The equation for the electrostatic potential is

$$\begin{aligned} &\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = 0, \\ &V(0,y,z) = V(a,y,z) = V(x,0,z) = V(x,b,z) = 0, \\ &V(x,y,0) = V_0, \quad V(x,y,z) \to 0, z \to \infty. \end{aligned} \tag{5-198}$$

Since the electric potential function should be similar to the adjacent faces, we have the potential going to zero as we get far away from the $z = 0$ face.

To solve this problem, we apply separation of variables

$$V(x,y,z) = X(x)Y(y)Z(z). \tag{5-199}$$

Inserting this into the differential equation and dividing by $V = XYZ$ we get

$$\frac{1}{X}\frac{\partial^2 X}{\partial x^2} + \frac{1}{Y}\frac{\partial^2 Y}{\partial y^2} + \frac{1}{Z}\frac{\partial^2 Z}{\partial z^2} = 0. \tag{5-200}$$

Analogous to the 2-D case, we have a function of x plus a function of y plus a function of z is equal to zero. This can only be true if each of the functions are equal to constants:

$$\frac{1}{X}\frac{\partial^2 X}{\partial x^2} = K_1, \tag{5-201a}$$

$$\frac{1}{Y}\frac{\partial^2 Y}{\partial y^2} = K_2, \tag{5-201b}$$

$$\frac{1}{Z}\frac{\partial^2 Z}{\partial z^2} = K_3. \tag{5-201c}$$

We now must determine the signs of the constants. This laborious step is omitted here, but from prior experience with the 2-D cases, we can surmise that $K_1 < 0$ and $K_2 < 0$. Given that

$$K_1 + K_2 + K_3 = 0,$$

we know that $K_3 > 0$. We define

$$\frac{1}{X}\frac{\partial^2 X}{\partial x^2} = -k^2, \tag{5-202a}$$

$$\frac{1}{Y}\frac{\partial^2 Y}{\partial y^2} = -\ell^2; \tag{5-202b}$$

this gives the following solutions

$$X(x) = A\sin(kx) + B\cos(kx), \tag{5-203a}$$

$$Y(y) = C\sin(\ell y) + D\cos(\ell y). \tag{5-203b}$$

It follows for the z equation that

$$\frac{1}{Z}\frac{\partial^2 Z}{\partial z^2} = k^2 + \ell^2, \tag{5-204}$$

which gives the solution

$$Z(z) = Fe^{\sqrt{k^2+\ell^2}z} + Ge^{-\sqrt{k^2+\ell^2}z}. \tag{5-205}$$

Here we use exponentials rather than hyperbolic trigonometric functions because of the boundary condition that as $z \to \infty$, the potential goes to zero. This implies that $F = 0$, giving the result

$$Z(z) = Ge^{-\sqrt{k^2+\ell^2}z}. \tag{5-206a}$$

By applying the boundary conditions at $x = 0$ and $y = 0$, we can show that the coefficients $B = 0$ and $D = 0$ respectively. Giving us

$$X(x) = A\sin(kx), \tag{5-206b}$$

$$Y(y) = C\sin(\ell y). \tag{5-206c}$$

Using the boundary conditions at $x = a$ and $y = b$, gives the familiar result that k and ℓ must be integer multiples of π. That is

$$k_n = \frac{n\pi}{a}, \quad n = 1, 2, 3\ldots \tag{5-207a}$$

$$\ell_n = \frac{m\pi}{b}, \quad m = 1, 2, 3 \ldots \tag{5-207b}$$

Therefore, we have infinitely many values of k and ℓ that yield the solution and we can take a linear combination of those solutions

$$X(x) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi x}{a}\right), \tag{5-208a}$$

$$Y(y) = \sum_{m=1}^{\infty} C_m \sin\left(\frac{m\pi y}{b}\right). \tag{5-208b}$$

The solution for the electric potential is therefore

$$V(x,y,z) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi x}{a}\right) \sum_{m=1}^{\infty} C_m \sin\left(\frac{m\pi y}{b}\right) G \exp\left[-\sqrt{\left(\frac{n\pi}{a}\right)^2 + \left(\frac{m\pi}{b}\right)^2}\, z\right].$$

Combining the constants and rearranging gives

$$V(x,y,z) = \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} A_{n,m} \sin\left(\frac{n\pi x}{a}\right) \sin\left(\frac{m\pi y}{b}\right) \exp\left[-\sqrt{\left(\frac{n\pi}{a}\right)^2 + \left(\frac{m\pi}{b}\right)^2}\, z\right]. \tag{5-209}$$

Here $A_{n,m}$ is a combined constant that depends upon both indices. We have one remaining boundary condition to determine this constant, $V(x,y,0) = V_0$. Plugging this in gives

$$V(x,y,0) = \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} A_{n,m} \sin\left(\frac{n\pi x}{a}\right) \sin\left(\frac{m\pi y}{b}\right) = V_0. \tag{5-210}$$

We can obtain this constant by multiplying by products of sine functions in indices ν and μ and then integrating x from 0 to a and y form 0 to b:

$$\begin{aligned}&\sum_{n=1}^{\infty}\sum_{m=1}^{\infty} A_{n,m} \int_0^a \sin\left(\frac{\nu\pi x}{a}\right) \sin\left(\frac{n\pi x}{a}\right) dx \int_0^b \sin\left(\frac{\mu\pi y}{b}\right) \sin\left(\frac{m\pi y}{b}\right) dy \\ &= \int_0^a \int_0^b V_0 \sin\left(\frac{\nu\pi x}{a}\right) \sin\left(\frac{\mu\pi y}{b}\right) dydx.\end{aligned} \tag{5-211}$$

Using the orthogonality property, we know that all but the terms where $n = \nu$ and $m = \mu$ survive in the inner product. This gives

$$A_{n,m}\left(\frac{a}{2}\right)\left(\frac{b}{2}\right) = \int_0^a \int_0^b V_0 \sin\left(\frac{n\pi x}{a}\right) \sin\left(\frac{m\pi y}{b}\right) dydx. \tag{5-212}$$

Performing the integral and solving for the constant gives

$$A_{n,m} = \begin{cases} \dfrac{16V_0}{\pi^2 nm}, & \text{both } n \text{ and } m \text{ odd} \\ 0, & \text{either } n \text{ or } m \text{ even} \end{cases}. \tag{5-213}$$

The electric potential function is therefore

$$V(x,y,z) = \sum_{\substack{n=1 \\ n \text{ odd}}}^{\infty} \sum_{\substack{m=1 \\ m \text{ odd}}}^{\infty} \frac{16V_0}{\pi^2 nm} \sin\left(\frac{n\pi x}{a}\right) \sin\left(\frac{m\pi y}{b}\right) \exp\left[-\sqrt{\left(\frac{n\pi}{a}\right)^2 + \left(\frac{m\pi}{b}\right)^2}\, z\right]. \tag{5-214}$$

The electric field may be obtained by $\mathbf{E} = -\nabla V$.

5.4.d Spherical Coordinates and Legendre Polynomials

Previously we had considered solving the Laplace equation in Cartesian coordinates. We now turn our attention to spherical coordinates. (Cylindrical coordinates require special functions called Bessel functions, and are not covered.) The Laplace equation in spherical coordinates is

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial u}{\partial r}\right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial u}{\partial \theta}\right) + \frac{1}{r^2 \sin^2\theta}\frac{\partial^2 u}{\partial \phi^2} = 0 \tag{5-215}$$

with boundary conditions prescribed as appropriate.

This equation is quite complicated. We can simplify it significantly if we restrict ourselves to problems that are azimuthally symmetric. This allows us to remove the derivative in ϕ to yield the Laplace equation for spherical coordinates with azimuthally symmetry:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial u}{\partial r}\right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial u}{\partial \theta}\right) = 0. \tag{5-216}$$

As with the Cartesian case, we assume that the solution is separable and write

$$u(r,\theta) = R(r)\Theta(\theta). \tag{5-217}$$

Inserting this into the differential equation and dividing by $u = R\Theta$ gives the equation

$$\frac{1}{R}\frac{\partial}{\partial r}\left(r^2 \frac{\partial R}{\partial r}\right) + \frac{1}{\Theta \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial \Theta}{\partial \theta}\right) = 0. \tag{5-218}$$

Here the $1/r^2$ term has been multiplied away.

Since we have a function of r plus a function of θ equals zero, we know that each of the terms must equal to a constant K and that those constants are equal and opposite. As with Cartesian coordinates we must deduce the sign of K. The sign of K that produces sensible results is to assign the positive term to the radial R equation and the negative term to the polar Θ equation:

$$\frac{1}{R}\frac{\partial}{\partial r}\left(r^2 \frac{\partial R}{\partial r}\right) = K, \tag{5-219a}$$

$$\frac{1}{\Theta \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial \Theta}{\partial \theta}\right) = -K. \tag{5-219b}$$

Therefore, when we expand out these equations, we have

$$r^2\frac{\partial^2 R}{\partial r^2} + 2r\frac{\partial R}{\partial r} - KR(r) = 0, \tag{5-220a}$$

$$\frac{\partial^2 \Theta}{\partial \theta^2} + \frac{\cos\theta}{\sin\theta}\frac{\partial \Theta}{\partial \theta} + K\Theta(\theta) = 0. \tag{5-220b}$$

These equations are unwieldy, and their solutions are certainly not obvious. First, we will focus our attention on the Θ equation. It turns out (again, not obvious) that the equation only has a solution when

$$K = \ell(\ell+1), \quad \ell = 0, 1, 2\ldots \tag{5-221}$$

and the solutions correspond to a special set of polynomials we denote by $P_\ell(\cos\theta)$, where ℓ represents the highest order term in the polynomial. These polynomials are called the *Legendre polynomials*, and like the trigonometric functions form a complete an orthogonal basis. Let us now briefly turn our attention away from the spherical Laplace equation and discuss the Legendre polynomials.

Suppose we wish to construct a series of orthogonal polynomial functions defined on the domain $-1 \leq \mu \leq 1$. (If the domain is not -1 to 1, we can always transform the coordinates to make it as such.) The simplest polynomial is a constant equal to one, so we start by defining

$$P_0(\mu) = 1.$$

Now we wish to find a polynomial of degree $\ell = 1$, call it $P_1(\mu)$, such that

$$\int_{-1}^{1} P_1(\mu)P_0(\mu)d\mu = 0,$$
$$\int_{-1}^{1} P_1(\mu)P_1(\mu)d\mu = c_1 \neq 0.$$

To satisfy the first constraint, we require the polynomial be odd or be symmetric about the origin. A polynomial that meets this criterion is

$$P_1(\mu) = \mu.$$

We could apply a multiplicative constant and still meet this constraint, but it turns out that as we go further, having a multiplicative constant of one is the right choice. We then continue and attempt to find a quadratic polynomial $P_2(\mu)$ that now satisfies the constraints

$$\int_{-1}^{1} P_2(\mu)P_0(\mu)d\mu = 0,$$
$$\int_{-1}^{1} P_2(\mu)P_1(\mu)d\mu = 0,$$
$$\int_{-1}^{1} P_2(\mu)P_2(\mu)d\mu = c_2 \neq 0.$$

A polynomial that does this is

$$P_2(\mu) = \frac{1}{2}\left(3\mu^2 - 1\right).$$

We continue this process to construct a series of polynomials satisfying the orthogonality property

$$\int_{-1}^{1} P_\ell(\mu) P_m(\mu) d\mu = \frac{2}{2\ell+1}\delta_{\ell m}, \tag{5-222}$$

where here we have explicitly wrote out the constant that we will enforce.

Continuing this process, we can deduce the following definition of the Legendre polynomials:

$$P_0(\mu) = 1, \tag{5-223a}$$
$$P_1(\mu) = \mu, \tag{5-223b}$$
$$P_{\ell+1}(\mu) = \frac{(2\ell+1)\mu P_\ell(\mu) - \ell P_{\ell-1}(\mu)}{\ell+1}, \quad \ell \geq 1. \tag{5-223c}$$

The last of these is a recursion relationship that allows us to generate any order polynomial $\ell \geq 2$.

As we mentioned previously, the Legendre polynomials form a complete an orthogonal basis. Therefore, similar to Fourier expansions, we can take Legendre polynomial expansions of functions defined on the domain $-1 \leq \mu \leq 1$:

$$f(\mu) = \sum_{\ell=0}^{\infty} \left(\frac{2\ell+1}{2}\right) f_\ell P_\ell(\mu). \tag{5-224}$$

The expansion coefficient f_ℓ (sometimes called the Legendre moment) can be obtained from

$$f_\ell = \int_{-1}^{1} f(\mu) P_\ell(\mu) d\mu. \tag{5-225}$$

As an aside, the Legendre polynomials have many applications outside solving the Laplace equation. For example, they are frequently used in performing fast integration method called the Gauss-Legendre quadrature scheme. In fact, we can very rapidly integrate polynomials exactly on a computer by evaluating the function at roots of a certain order of the Legendre polynomials multiplying by special weighting factors and adding up the results. The Gauss-Legendre quadrature is one of most efficient tools for numerical integration available. Another application of Legendre polynomials and their expansions is in radiation interactions with matter. The change in direction that a photon or neutron experiences in a scattering event is random and often well described by an expansion in Legendre polynomials.

Returning to the Laplace equation, applying the Legendre polynomials we let $\mu = \cos\theta$ and we have that solution for $\Theta(\theta)$ is

$$\Theta(\theta) = C_\ell P_\ell(\cos\theta), \quad \ell = 0, 1, 2 \ldots \tag{5-226}$$

where C_ℓ is some arbitrary constant. Before moving onto the radial equation. It is worth noting that, being a second-order ordinary differential equation, we expect there to be two sets of solutions to $\Theta(\theta)$. Indeed there is, and this is called the *Legendre function of the second kind* given by $Q_\ell(\cos\theta)$. The first couple Legendre functions of the second kind in terms of μ is

$$Q_0(\mu) = \frac{1}{2}\ln\left(\frac{1+\mu}{1-\mu}\right), \tag{5-227a}$$

$$Q_1(\mu) = \frac{\mu}{2}\ln\left(\frac{1+\mu}{1-\mu}\right) - 1. \tag{5-227b}$$

The higher-order Legendre functions of the second kind satisfy the same recursion relationship as the Legendre polynomials. Note that the $Q_\ell(\cos\theta)$ functions blow up as θ approaches 0 to 2π. Therefore, the Legendre function can often be rejected because we normally require that the solution be finite, unless we have a case where the polar axis is not included in the problem domain for some reason. For completeness the solution to the polar equation is

$$\Theta(\theta) = C_\ell P_\ell(\cos\theta) + D_\ell Q_\ell(\cos\theta), \quad \ell = 0, 1, 2\ldots \tag{5-228}$$

even through the second term is seldom required.

Since $\mu = \cos\theta$ then it will be convenient to rewrite the equations for the Legendre polynomials in terms of θ. The orthogonality property becomes

$$\int_0^\pi P_\ell(\cos\theta)P_m(\cos\theta)\sin\theta d\theta = \frac{2}{2\ell+1}\delta_{\ell m}; \tag{5-229}$$

the Legendre polynomial expansion is

$$f(\cos\theta) = \sum_{\ell=0}^{\infty}\left(\frac{2\ell+1}{2}\right) f_\ell P_\ell(\cos\theta); \tag{5-230}$$

and the expansion coefficient is

$$f_\ell = \int_0^\pi f(\cos\theta)P_\ell(\cos\theta)\sin\theta d\theta. \tag{5-231}$$

Moving on to the radial equation, we have with the choice of constant that satisfies $\Theta(\theta)$:

$$r^2\frac{\partial^2 R}{\partial r^2} + 2r\frac{\partial R}{\partial r} - \ell(\ell+1)R(r) = 0. \tag{5-232}$$

The solution to this equation is also not obvious, but we can verify that

$$R(r) = A_\ell r^\ell + B_\ell\frac{1}{r^{\ell+1}}, \quad \ell = 0, 1, 2\ldots \tag{5-233}$$

does indeed satisfy the differential equation for $R(r)$.

Multiplying our solutions for $R(r)$ and $\Theta(\theta)$ together and combining constants, we have the solution for the Laplace equation in spherical coordinates with azimuthal symmetry as

$$u(r,\theta) = \sum_{\ell=0}^{\infty} \left(A_\ell r^\ell + B_\ell \frac{1}{r^{\ell+1}} \right) P_\ell(\cos\theta). \tag{5-234}$$

Since there are infinitely many integer values of ℓ that satisfy the equation, we take a linear combination of all possible solutions. We will then need to apply the boundary conditions to solve for the coefficients. Let us illustrate this with a few examples.

5.4.e Example: Heat Conduction in a Sphere

Suppose we have a sphere of radius a where the temperature is specified on the exterior of the sphere at $r = a$ as a function of θ. The Laplace equation can be used as before to solve the heat conduction, but this time in spherical coordinates:

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2 \frac{\partial T}{\partial r} \right) + \frac{1}{r^2 \sin\theta}\frac{\partial}{\partial \theta}\left(\sin\theta \frac{\partial T}{\partial \theta} \right) = 0, \quad T(a,\theta) = f(\theta), \quad 0 \le r \le a. \tag{5-235}$$

We will not repeat the separation of variables, as it is the same as what was just discussed, and simply write out the form of the solution:

$$T(r,\theta) = \sum_{\ell=0}^{\infty} \left(A_\ell r^\ell + B_\ell \frac{1}{r^{\ell+1}} \right) P_\ell(\cos\theta).$$

We must now apply the boundary conditions to resolve the coefficients.

First, we know the temperature must be finite everywhere and the domain includes $r = 0$. The terms in

$$B_\ell \frac{1}{r^{\ell+1}}, \quad \ell = 0, 1, 2 \ldots$$

always diverge as $r \to 0$. For this reason, we must exclude them and we do this by setting $B_\ell = 0$ for all values of ℓ.

Applying the boundary condition at $r = a$ gives

$$f(\theta) = \sum_{\ell=0}^{\infty} A_\ell a^\ell P_\ell(\cos\theta). \tag{5-236}$$

This is a a Legendre polynomial expansion if we let

$$A_\ell a^\ell = \left(\frac{2\ell+1}{2} \right) f_\ell \tag{5-237}$$

Applying the relationship for the Legendre coefficient and solving for the integration constant gives the result:

$$A_\ell = \left(\frac{2\ell+1}{2a^\ell} \right) \int_0^\pi f(\theta) P_\ell(\cos\theta) \sin\theta d\theta. \tag{5-238}$$

In general, we would need to carry out the integrals (often numerically) to find the coefficients. Fortunately, it is often the case where $f(\theta)$ has a form where the θ having only the dependence of $\cos^n\theta$ where n are non-negative integers. In this case, we can avoid doing the integrals and "pick out" the coefficients directly by starting with the highest order term and working down to the constant.

To illustrate, consider the boundary condition:

$$f(\theta) = 1 - \cos\theta + 3\cos^2\theta. \tag{5-239}$$

Here the highest-order term in $\cos\theta$ is 2. Therefore we try to write the equation in terms of the second-order Legendre polynomial

$$P_2(\cos\theta) = \frac{1}{2}\left(3\cos^2\theta - 1\right).$$

To do this, we multiply and divide the $\cos^2\theta$ term by 2:

$$f(\theta) = 1 - \cos\theta + 2\cdot\frac{1}{2}\left(3\cos^2\theta\right).$$

Then we add an subtract 1 inside the parentheses to get

$$f(\theta) = 1 - \cos\theta + 2\cdot\frac{1}{2}\left(3\cos^2\theta - 1\right) + 1.$$

Recognizing $P_2(\cos\theta)$ and regrouping terms gives

$$f(\theta) = 2 - \cos\theta + 2P_2(\cos\theta).$$

Noting that $P_0(\cos\theta) = 1$ and $P_1(\cos\theta) = \cos\theta$ we can write this as

$$f(\theta) = 2P_0(\cos\theta) - P_1(\cos\theta) + 2P_2(\cos\theta) = \sum_{\ell=0}^{\infty} A_\ell a^\ell P_\ell(\cos\theta). \tag{5-240}$$

Therefore, the coefficients are

$$A_0 = 2, \tag{5-241a}$$
$$A_1 = -\frac{1}{a}, \tag{5-241b}$$
$$A_2 = \frac{2}{a^2}, \tag{5-241c}$$
$$A_\ell = 0, \quad \ell > 2 \tag{5-241d}$$

The temperature within the sphere is therefore

$$T(r,\theta) = 2 - \frac{r}{a}\cos\theta + \frac{r^2}{a^2}\left(3\cos^2\theta - 1\right). \tag{5-242}$$

5.4.f Example: Fluid Velocity Around a Spinning Ball

Suppose we have a solid ball or radius a immersed in large vat containing a viscous fluid such that it is centered at the origin of a spherical coordinate system. Suppose the ball spins about the z axis such that the velocity of the ball at the surface has an azimuthal component

$$u_\phi(a,\theta) = a\omega\sin\theta. \tag{5-243}$$

Here ω is the angular velocity. The ball rotates with an azimuthal velocity that is faster at the equator than near the poles. We can assume the vat is large and fluid is stagnant so that faraway from the ball, the azimuthal velocity is zero, $u_\phi(\infty,\theta) = 0$.

If we assume that the fluid is sufficiently viscous and ω is sufficiently small so that we can neglect turbulence, then the azimuthal velocity is described by the Laplace equation in spherical coordinates. Because the rotation is azimuthally symmetric, the azimuthal component of the fluid velocity is azimuthally constant and only depends on the radial position and the polar angle with respect to the axis of rotation. The differential equation is therefore

$$\begin{aligned}&\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial u_\phi}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial u_\phi}{\partial\theta}\right) = 0,\\ &u_\phi(a,\theta) = a\omega\sin\theta, \quad u_\phi(\infty,\theta) = 0, \qquad r \geq a.\end{aligned} \tag{5-244}$$

As we have seen, the solution is

$$u_\phi(r,\theta) = \sum_{\ell=0}^{\infty}\left(A_\ell r^\ell + B_\ell\frac{1}{r^{\ell+1}}\right)P_\ell(\cos\theta). \tag{5-245}$$

This time, because the domain does not include the origin, we cannot eliminate the B_ℓ terms. On the other hand, what we can do is eliminate the A_ℓ terms, as this is the only way for the solution to go to zero for large r. Applying the boundary condition at the surface of the sphere gives

$$u_\phi(a,\theta) = \sum_{\ell=0}^{\infty} B_\ell\frac{1}{a^{\ell+1}}P_\ell(\cos\theta) = a\omega\sin\theta. \tag{5-246}$$

We recognize that this is a Legendre polynomial expansion if we equate

$$B_\ell\frac{1}{a^{\ell+1}} = \left(\frac{2\ell+1}{2}\right)f_\ell. \tag{5-247}$$

Solving for the integration constant and writing the equation in terms of the integral for the expansion coefficient gives

$$B_\ell = a^{\ell+2}\omega\left(\frac{2\ell+1}{2}\right)\int_0^\pi \sin\theta P_\ell(\cos\theta)\sin\theta d\theta,$$

At this point, let us check if we can equate terms as we did in the previous example. Unfortunately, by

$$\cos^2\theta + \sin^2\theta = 1$$

we would get

$$B_\ell = a^{\ell+2}\omega\left(\frac{2\ell+1}{2}\right)\int_0^\pi \sqrt{1-\cos\theta}P_\ell(\cos\theta)\sin\theta d\theta,$$

which does not depend upon only nonnegative integer powers of $\cos\theta$. Therefore, we will be unable to write the expansion exactly for at a finite number of terms. Rather, we will need to estimate the terms directly and truncate at a certain point. On the bright side, because $\sin^2\theta$ is an even function, we know that only the even values of the expansion will be nonzero. Evaluating the integral

$$B_\ell = a^{\ell+2}\omega\left(\frac{2\ell+1}{2}\right)\int_0^\pi \sin^2\theta P_\ell(\cos\theta)d\theta \tag{5-248}$$

for the first few even values of ℓ gives

$$B_0 = \frac{\pi a^2\omega}{4}, \tag{5-249a}$$

$$B_2 = -\frac{5\pi a^4\omega}{32}, \tag{5-249b}$$

$$B_4 = -\frac{9\pi a^6\omega}{256}, \tag{5-249c}$$

$$B_6 = -\frac{55\pi a^8\omega}{4096}. \tag{5-249d}$$

The remaining terms are all negative. Plugging in these values and factoring out some terms gives the following expression for the azimuthal velocity:

$$u_\phi(r,\theta) = \frac{\pi a^2\omega}{4r}\left[1 - \frac{5}{8}\left(\frac{a}{r}\right)^2 P_2(\cos\theta) - \frac{9}{64}\left(\frac{a}{r}\right)^4 P_4(\cos\theta) - \cdots\right]. \tag{5-250}$$

As more terms are included, the expression will become increasingly accurate. An important point to note is that the terms are ordered in a way that they become increasingly small as the radial coordinate r moves further from the surface of the sphere. Because only the even-ordered terms are nonzero, the leading terms diminish rapidly with r. This also means the further away from the sphere, the variation in the polar angle diminishes rapidly. As we move very far away from the sphere $r \gg a$, only the leading term is significant, or

$$u_\phi(r,\theta) \approx \frac{\pi a^2\omega}{4r}, \quad r \gg a,$$

which has no polar angle dependence at all.

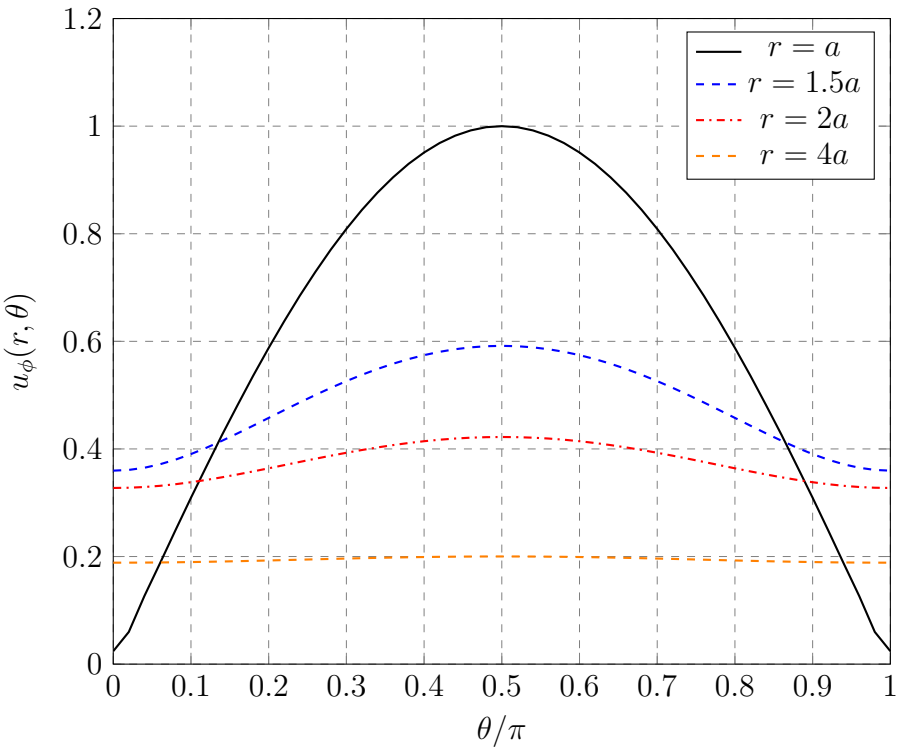

Figure 5.20: Plot of the polar angle dependence (units of θ/π to scale from 0 to 1) of the azimuthal fluid velocity for various radial coordinates.

To illustrate, we consider a problem with $a = 1$ and $\omega = 1$. Figure 5.20 shows the polar angle dependence of the azimuthal velocity $u_\phi(r, \theta)$ at various radial coordinates away from the sphere. A total of 20 terms were used to generate the results. At the surface of the sphere, we see a sinusoidal dependence upon the fluid velocity with maximal flow at the equator and zero at the poles. As we move further away, the diffusive process causes the momentum to mix and the fluid velocity rises near the poles and decreases near the equator with an overall decreasing fluid velocity the further away from the sphere. As we move further away from the surface of the sphere, the curve flattens out, as the viscous shear forces dissipate any of the gradients in the polar angle such that by $r = 4a$, there is little polar dependence remaining.

5.5 Finite Difference Schemes for PDEs

While the analytical techniques often provide physical insight into the problem, rarely are they adequate for practical engineering problems. Therefore, we must resort to applying numerical schemes. A plethora of numerical methods for solving PDEs have been and continue to be developed. This section covers the standard methods.

5.5.a Crank-Nicholson for 1-D Heat Equation

The 1-D heat equation with a reaction coefficient can be written in the following form:

$$\rho(x)\frac{\partial \phi}{\partial t} + \frac{\partial J}{\partial x} + \lambda(x)\phi(x,t) = Q(x,t). \tag{5-251}$$

Here the coefficient ρ can be interpreted as a type of inertia such that the larger the value, the slower the time variation of the quantity $\phi(x,t)$; J is the local flow rate given by the gradient,

$$J(x,t) = -D\frac{\partial \phi}{\partial x}, \tag{5-252}$$

where D is a diffusion coefficient; λ is a reaction coefficient; and Q is the internal source.

We must impose a spatial and temporal discretization upon the problem. As with ODEs, here we apply the cell-centered differencing scheme such that full integer indices i correspond to cell centers and half-integer indices correspond to cell edges. The temporal discretization follows a different convention where integer values n correspond to the solution at the beginning and end of the time steps. We use half-integer time indices to denote some average value for the time step. Uniform spatial properties are assumed within a cell and the properties are assumed to be constant in time.

The first step is to integrate the heat equation over a spatial cell having width $\Delta x_i = x_{i+1/2} - x_{i-1/2}$:

$$\begin{aligned}\rho_i \frac{\partial}{\partial t}\left(\int_{x_{i-1/2}}^{x_{i+1/2}} \phi(x,t)dx\right) + J(x_{i+1/2},t) - J(x_{i-1/2},t) \\ + \lambda \int_{x_{i-1/2}}^{x_{i+1/2}} \phi(x,t)dx = \int_{x_{i-1/2}}^{x_{i+1/2}} Q(x,t)dx.\end{aligned} \tag{5-253}$$

We then define the cell-averaged quantity as

$$\phi_i(t) = \frac{1}{\Delta x_i}\int_{x_{i-1/2}}^{x_{i+1/2}} \phi(x,t)dx. \tag{5-254}$$

Inserting this definition gives

$$\rho_i \Delta x_i \frac{\partial \phi_i}{\partial t} + J(x_{i+1/2},t) - J(x_{i-1/2},t) + \lambda_i \Delta x_i \phi_i(t) = \int_{x_{i-1/2}}^{x_{i+1/2}} Q(x,t)dx. \tag{5-255}$$

To eliminate the time derivative, integrate over the time step $\Delta t_n = t_{n+1} - t_n$:

$$\begin{aligned}&\rho_i \Delta x_i \left[\phi_i(t_{n+1}) - \phi_i(t_n)\right] + \int_{t_n}^{t_{n+1}} J(x_{i+1/2},t) - J(x_{i-1/2},t) + \lambda_i \Delta x_i \phi_i(t)dt \\ &= \int_{t_n}^{t_{n+1}} \int_{x_{i-1/2}}^{x_{i+1/2}} Q(x,t)dxdt.\end{aligned} \tag{5-256}$$

To clean this up, we introduce a few quantities. First, we have the cell-average value at time t_n as $\phi_i^n = \phi_i(t_n)$, where n is a superscript not a power. We then require average values for the quantity, flow rate, and internal source. These are

$$\phi_i^{n+1/2} = \frac{1}{\Delta t_n} \int_{t_n}^{t_{n+1}} \phi(x_{i+1/2}, t) dt, \tag{5-257a}$$

$$J_{i+1/2}^{n+1/2} = \frac{1}{\Delta t_n} \int_{t_n}^{t_{n+1}} J(x_{i+1/2}, t) dt, \tag{5-257b}$$

$$Q_i^{n+1/2} = \frac{1}{\Delta x_i} \frac{1}{\Delta t_n} \int_{t_n}^{t_{n+1}} \int_{x_{i-1/2}}^{x_{i+1/2}} Q(x_{i+1/2}, t) dx dt. \tag{5-257c}$$

Then,

$$\rho_i \Delta x_i \left(\phi_i^{n+1} - \phi_i^n \right) + \Delta t_n \left[J_{i+1/2}^{n+1/2} - J_{i-1/2}^{n+1/2} + \lambda_i \Delta x_i \phi_i^{n+1/2} \right] = Deltax_i \Delta t_n Q_i^{n+1/2}. \tag{5-258}$$

We need to relate the quantity and flow rates at the beginning and end of the time step with the average value. A second-order accurate scheme may be obtained by using improved Euler and assuming that the average values within a time step are the arithmetic mean of the values at the beginning and the end:

$$\phi_i^{n+1/2} = \frac{1}{2} \left(\phi_i^n + \phi_i^{n+1} \right), \tag{5-259a}$$

$$J_{i+1/2}^{n+1/2} = \frac{1}{2} \left(J_i^n + J_i^{n+1} \right). \tag{5-259b}$$

Inserting these approximations and moving all quantities with time step n (which is known information from either the initial condition or a previous calculation) to the right-hand side gives

$$\begin{aligned} &\rho_i \Delta x_i \phi_i^{n+1} + \frac{\Delta t_n}{2} \left[J_{i+1/2}^{n+1} - J_{i-1/2}^{n+1} + \lambda_i \Delta x_i \phi_i^{n+1} \right] \\ &= \rho_i \Delta x_i \phi_i^n - \frac{\Delta t_n}{2} \left[J_{i+1/2}^n - J_{i-1/2}^n + \lambda_i \Delta x_i \phi_i^n \right] + \Delta x_i \Delta t_n Q_i^{n+1/2}. \end{aligned} \tag{5-260}$$

Note that the internal source $Q_i^{n+1/2}$ is assumed to be known.

The next step is to approximate the flow rates J with a finite difference scheme. This process is detailed in Sec. 3.9.b and only the end results are quoted here.

For an interior element (not on the boundary), the flow rate on the edge of a cell is related to the cell-average quantities of the adjoining cells by

$$J_{i+1/2}^n = -\widetilde{D}_{i+1/2} \left(\phi_{i+1}^n - \phi_i^n \right), \tag{5-261}$$

where $\widetilde{D}$ is the edge-average diffusion coefficient:

$$\widetilde{D}_{i+1/2} = 2 \frac{(D_i/\Delta x_i)(D_{i+1}/\Delta x_{i+1})}{D_i/\Delta x_i + D_{i+1}/\Delta x_{i+1}}. \tag{5-262}$$

Inserting this relationship and arranging terms gives an equation for an interior element:

$$
\begin{aligned}
&-\frac{\Delta t_n}{2}\widetilde{D}_{i-1/2}\phi_{i-1}^{n+1}+\left[\frac{\Delta t_n}{2}\left(\widetilde{D}_{i-1/2}+\widetilde{D}_{i+1/2}\right)+(\lambda_i+\rho_i)\Delta x_i\right]\phi_i^{n+1}-\frac{\Delta t_n}{2}\widetilde{D}_{i+1/2}\phi_{i+1}^{n+1}\\
&=\frac{\Delta t_n}{2}\widetilde{D}_{i-1/2}\phi_{i-1}^{n}-\left[\frac{\Delta t_n}{2}\left(\widetilde{D}_{i-1/2}+\widetilde{D}_{i+1/2}\right)+(\lambda_i-\rho_i)\Delta x_i\right]\phi_i^{n}+\frac{\Delta t_n}{2}\widetilde{D}_{i+1/2}\phi_{i+1}^{n}\\
&+\Delta x_i\Delta t_n Q_i^{n+1/2}.
\end{aligned}
\tag{5-263}
$$

The equations for the boundary elements have a slightly different form. The generic form of the Robin boundary condition is written as

$$\alpha\phi+\beta J=\gamma$$

with α, β, and γ being coefficients that depend on the condition type. The equations for the left ($i=1$) and right ($i=N$) boundary elements are

$$
\begin{aligned}
&\left[\frac{\Delta t_n}{2}\left(\widetilde{D}_{1/2}\alpha_\ell+\widetilde{D}_{3/2}\right)+(\lambda_1+\rho_1)\Delta x_1\right]\phi_1^{n+1}-\frac{\Delta t_n}{2}\widetilde{D}_{3/2}\phi_2^{n+1}\\
&=-\left[\frac{\Delta t_n}{2}\left(\widetilde{D}_{1/2}\alpha_\ell+\widetilde{D}_{3/2}\right)+(\lambda_1-\rho_1)\Delta x_1\right]\phi_1^{n}+\frac{\Delta t_n}{2}\widetilde{D}_{3/2}\phi_2^{n}\\
&+\Delta x_1\Delta t_n Q_1^{n+1/2}+\Delta t_n\widetilde{D}_{1/2}\gamma_\ell.
\end{aligned}
\tag{5-264a}
$$

and

$$
\begin{aligned}
&-\frac{\Delta t_n}{2}\widetilde{D}_{N-1/2}\phi_{N-1}^{n+1}+\left[\frac{\Delta t_n}{2}\left(\widetilde{D}_{N-1/2}+\widetilde{D}_{N+1/2}\alpha_r\right)+(\lambda_N+\rho_N)\Delta x_N\right]\phi_N^{n+1}\\
&=\frac{\Delta t_n}{2}\widetilde{D}_{N-1/2}\phi_{N-1}^{n}-\left[\frac{\Delta t_n}{2}\left(\widetilde{D}_{N-1/2}+\widetilde{D}_{N+1/2}\alpha_r\right)+(\lambda_N-\rho_N)\Delta x_N\right]\phi_N^{n}\\
&+\Delta x_N\Delta t_n Q_N^{n+1/2}+\Delta t_n\widetilde{D}_{N+1/2}\gamma_r.
\end{aligned}
\tag{5-264b}
$$

respectively, where the ℓ and r subscripts denote left and right boundary coefficients and the respective edge-average diffusion coefficients are

$$\widetilde{D}_{1/2}=\frac{2D_1/\Delta_1}{\alpha_\ell+\beta_\ell(2D_1/\Delta_1)},\tag{5-265a}$$

$$\widetilde{D}_{N+1/2}=\frac{2D_N/\Delta_N}{\alpha_r-\beta_r(2D_N/\Delta_N)}.\tag{5-265b}$$

All terms on the right-hand sides of the difference equations are known either from the internal source, boundary condition, and the initial condition or previous time-step calculation. Inspecting the left-hand side, we observe that the coupling is only to the adjacent elements, meaning this system of equations can be formulated as a tridiagonal system. We have for the subdiagonal element ℓ_i, diagonal element d_i,

superdiagonal element u_i and right-hand side element r_i as follows. The left-boundary cell elements are

$$d_1 = \frac{\Delta t_n}{2}\left(\widetilde{D}_{1/2}\alpha_\ell + \widetilde{D}_{3/2}\right) + (\lambda_1 + \rho_1)\Delta x_1, \tag{5-266a}$$

$$u_1 = -\frac{\Delta t_n}{2}\widetilde{D}_{3/2}, \tag{5-266b}$$

$$r_1 = -\left[\frac{\Delta t_n}{2}\left(\widetilde{D}_{1/2}\alpha_\ell + \widetilde{D}_{3/2}\right) + (\lambda_1 - \rho_1)\Delta x_1\right]\phi_1^n + \frac{\Delta t_n}{2}\widetilde{D}_{3/2}\phi_2^n + \Delta x_1 \Delta t_n Q_1^{n+1/2} + \Delta t_n \widetilde{D}_{1/2}\gamma_\ell; \tag{5-266c}$$

the interior cell elements are

$$\ell_i = -\frac{\Delta t_n}{2}\widetilde{D}_{i-1/2}, \tag{5-266d}$$

$$d_i = \frac{\Delta t_n}{2}\left(\widetilde{D}_{i-1/2} + \widetilde{D}_{i+1/2}\right) + (\lambda_i + \rho_i)\Delta x_i, \tag{5-266e}$$

$$u_i = -\frac{\Delta t_n}{2}\widetilde{D}_{i+1/2}, \tag{5-266f}$$

$$r_i = \frac{\Delta t_n}{2}\widetilde{D}_{i-1/2}\phi_{i-1}^n - \left[\frac{\Delta t_n}{2}\left(\widetilde{D}_{i-1/2} + \widetilde{D}_{i+1/2}\right) + (\lambda_i - \rho_i)\Delta x_i\right]\phi_i^n + \frac{\Delta t_n}{2}\widetilde{D}_{i+1/2}\phi_{i+1}^n + \Delta x_i \Delta t_n Q_i^{n+1/2}, \tag{5-266g}$$

for $i = 2, \ldots, N-1$. The right-boundary cell elements are

$$\ell_N = -\frac{\Delta t_n}{2}\widetilde{D}_{N-1/2}, \tag{5-266h}$$

$$d_N = \frac{\Delta t_n}{2}\left(\widetilde{D}_{N-1/2} + \widetilde{D}_{N+1/2}\alpha_r\right) + (\lambda_N + \rho_N)\Delta x_N, \tag{5-266i}$$

$$r_N = \frac{\Delta t_n}{2}\widetilde{D}_{N-1/2}\phi_{N-1}^n - \left[\frac{\Delta t_n}{2}\left(\widetilde{D}_{N-1/2} + \widetilde{D}_{N+1/2}\alpha_r\right) + (\lambda_N - \rho_N)\Delta x_N\right]\phi_N^n + \Delta x_N \Delta t_n Q_N^{n+1/2} + \Delta t_n \widetilde{D}_{N+1/2}\gamma_r. \tag{5-266j}$$

Contrasting these terms with those given for the steady-state form of the reaction-diffusion equation in Sec. 3.9.b, we note that the structure is the same, but there are a few notable differences. The diffusion terms have an extra factor of $\Delta t_n/2$ that limits the amount of spreading out of the quantity within a time step. The diagonal elements contain an extra factor of $\rho\Delta x$, which arises from the time derivative. Finally, the right-hand side vector contains information about the initial condition or the field from the previous time step.

While computing the elements of the tridiagonal system is more complicated, solving for the quantity at the next time step is fundamentally the same. Performing an entire transient simulation involves repeating the tridiagonal solve for each time step.

5.5.b Finite Difference for 2-D Reaction-Diffusion Equation

The 2-D reaction-diffusion equation is a generalization of the Laplace equation. This has the form

$$-\nabla \cdot D(x,y)\nabla\phi(x,y) + \lambda(x,y)\phi(x,y) = q(x,y). \tag{5-267}$$

Here ϕ is some physical quantity of interest (e.g., the neutron path-length density), D is a diffusion coefficient, λ is a reaction coefficient, and q is an inhomogeneous source term. We define the flow rate vector field as

$$\mathbf{J}(x,y) = -D(x,y)\nabla\phi(x,y). \tag{5-268}$$

Substituting this into the reaction-diffusion equation yields the continuity equation:

$$\nabla \cdot \mathbf{J}(x,y) + \lambda(x,y)\phi(x,y) = q(x,y). \tag{5-269}$$

Expanding out the divergence in 2-D Cartesian coordinates gives

$$\frac{\partial J_x}{\partial x} + \frac{\partial J_y}{\partial y} + \lambda(x,y)\phi(x,y) = q(x,y). \tag{5-270}$$

Here J_x and J_y are the x and y components of the flow rate vector $\mathbf{J}$, respectively.

As with the 1-D case in Sec. 3.9.b, we define a spatial discretization. Now we apply a rectangular grid where the (i,j) cell is defined such that $x_{i-1/2} \leq x \leq x_{i+1/2}$ and $y_{j-1/2} \leq y \leq y_{j+1/2}$ with spacing

$$\Delta x_i = x_{i+1/2} - x_{i-1/2}, \tag{5-271a}$$
$$\Delta y_j = y_{j+1/2} - y_{j-1/2}, \tag{5-271b}$$

and the coordinate (x_i, y_j) is taken at the center of the cell. We also demand that all reaction and diffusion coefficients are constant within each cell:

$$\lambda(x,y) = \lambda_{i,j}, \tag{5-272a}$$
$$D(x,y) = D_{i,j}, \quad (x,y) \in (i,j). \tag{5-272b}$$

We now integrate this equation over the (i,j) cell and arrive at

$$\int_{y_{j-1/2}}^{y_{j+1/2}} J_x(x_{i+1/2},y) - J_x(x_{i-1/2},y)dy + \int_{x_{i-1/2}}^{x_{i+1/2}} J_y(x,y_{j+1/2}) - J_y(x,y_{j-1/2})dx$$
$$+ \lambda_{i,j}\int_{y_{j-1/2}}^{y_{j+1/2}}\int_{x_{i-1/2}}^{x_{i+1/2}} \phi(x,y)dxdy = \int_{y_{j-1/2}}^{y_{j+1/2}}\int_{x_{i-1/2}}^{x_{i+1/2}} q(x,y)dxdy. \tag{5-273}$$

We then make the following definitions for edge-averaged flow rates and cell-averaged quantities and sources:

$$J_{x,i\pm1/2,j} = \frac{1}{\Delta y_j}\int_{y_{j-1/2}}^{y_{j+1/2}} J_x(x_{i\pm1/2},y)dy, \tag{5-274a}$$

$$J_{y,i,j\pm1/2} = \frac{1}{\Delta x_j}\int_{x_{i-1/2}}^{y_{i+1/2}} J_y(x, y_{j\pm1/2})dx, \tag{5-274b}$$

$$\phi_{i,j} = \frac{1}{\Delta x_i \Delta y_j}\int_{y_{j-1/2}}^{y_{j+1/2}}\int_{x_{i-1/2}}^{x_{i+1/2}} \phi(x,y)dxdy, \tag{5-274c}$$

$$q_{i,j} = \frac{1}{\Delta x_i \Delta y_j}\int_{y_{j-1/2}}^{y_{j+1/2}}\int_{x_{i-1/2}}^{x_{i+1/2}} q(x,y)dxdy. \tag{5-274d}$$

We substitute in these definitions and divide by $\Delta x_i \Delta y_j$ to get

$$\frac{1}{\Delta x_i}\left(J_{x,i+1/2,j} - J_{x,i-1/2,j}\right) + \frac{1}{\Delta y_i}\left(J_{y,i,j+1/2} - J_{y,i,j-1/2}\right) + \lambda_{i,j}\phi_{i,j} = q_{i,j}. \tag{5-275}$$

Interior Cells

As with the 1-D case, we now must apply a finite difference approximation to relate the currents to the fluxes. First we consider the case of an interior cell, i.e. one where there are neighboring cells on all four sides.

We can relate the flow rate on the left edge of the cell to the cell-average quantities in the (i, j) and $(i-1, j)$ cells:

$$J_{x,i-1/2,j} = -D_{i,j}\frac{\phi_{i,j} - \phi_{i-1/2,j}}{\Delta x_i/2} \tag{5-276}$$

$$J_{x,i-1/2,j} = -D_{i-1,j}\frac{\phi_{i-1/2,j} - \phi_{i-1,j}}{\Delta x_{i-1}/2}. \tag{5-277}$$

We then equate these, solve for the quantity on the left edge $\phi_{i-1/2,j}$, and then back substitute to write the current in terms of the cell-average fluxes, eliminating the cell-edge quantity:

$$J_{x,i-1/2,j} = -\widetilde{D}_{i-1/2,j}\left(\phi_{i,j} - \phi_{i-1,j}\right), \tag{5-278}$$

where the edge-average diffusion coefficient is

$$\widetilde{D}_{i-1/2,j} = 2\frac{(D_{i-1,j}/\Delta x_{i-1})(D_{i,j}/\Delta x_i)}{(D_{i-1,j}/\Delta x_{i-1}) + (D_{i,j}/\Delta x_i)}. \tag{5-279}$$

We can repeat this process for the right, bottom, and top edges to arrive at similar expressions. To illustrate, the current on the top edge is

$$J_{y,i,j+1/2} = -\widetilde{D}_{i,j+1/2}\left(\phi_{i,j+1} - \phi_{i,j}\right), \tag{5-280}$$

with edge-average diffusion coefficient

$$\widetilde{D}_{i,j+1/2} = 2\frac{(D_{i,j}/\Delta y_j)(D_{i,j+1}/\Delta y_{j+1})}{(D_{i,j}/\Delta y_j) + (D_{i,j+1}/\Delta y_{j+1})}. \tag{5-281}$$

We then insert these edge flow rates into the balance equation and obtain

$$\begin{aligned}
&-\frac{\widetilde{D}_{i-1/2,j}}{\Delta x_i}\phi_{i-1,j} - \frac{\widetilde{D}_{i,j-1/2}}{\Delta y_j}\phi_{i,j-1} \\
&+\left[\frac{\widetilde{D}_{i-1/2,j}+\widetilde{D}_{i+1/2,j}}{\Delta x_i} + \frac{\widetilde{D}_{i,j-1/2}+\widetilde{D}_{i,j+1/2}}{\Delta y_j} + \lambda_{i,j}\right]\phi_{i,j} \\
&-\frac{\widetilde{D}_{i+1/2,j}}{\Delta x_i}\phi_{i+1,j} - \frac{\widetilde{D}_{i,j+1/2}}{\Delta y_j}\phi_{i,j+1} = q_{i,j}.
\end{aligned} \tag{5-282}$$

Contrasting this with the 1-D result, we see that there are additional terms coupling the top and bottom edges and the coefficient on $\phi_{i,j}$ has another term.

Boundary and Corner Cells

We now need to apply a similar analysis to obtain edge-average diffusion coefficients and balance relationships at the boundaries. There are a couple additional complications in 2-D versus the 1-D case. First, the boundary conditions on each side are, in general, functions of position. For example, we write the boundary condition on the left boundary at $x = x_{1/2}$ as

$$\alpha(x_{1/2}, y)\phi(x_{1/2}, y) + \beta(x_{1/2}, y)J(x_{1/2}, y) = \gamma(x_{1/2}, y). \tag{5-283}$$

The coefficients α, β, and γ are now functions of y. To handle this in the discretization scheme, we take the boundary conditions to be spatially uniform within each each cell-edge. For the left boundary this is then

$$\alpha_{1/2,j}\phi_{1/2,j} + \beta_{1/2,j}J_{1/2,j} = \gamma_{1/2,j}, \tag{5-284a}$$

where we introduce the cell subscripts as before. On the bottom, right, and top boundaries these are

$$\alpha_{i,1/2}\phi_{i,1/2} + \beta_{i,1/2}J_{i,1/2} = \gamma_{i,1/2}, \tag{5-284b}$$
$$\alpha_{N_x+1/2,j}\phi_{N_x+1/2,j} + \beta_{N_x+1/2,j}J_{N_x+1/2,j} = \gamma_{N_x+1/2,j}, \tag{5-284c}$$
$$\alpha_{i,N_y+1/2}\phi_{i,N_y+1/2} + \beta_{i,N_y+1/2}J_{i,N_y+1/2} = \gamma_{i,N_y+1/2}, \tag{5-284d}$$

respectively. The second complication in 2-D is that we now have to distinguish between boundary cells, where there is one boundary, and the four corner cells where there are two. This means we require special logic to handle all of the cases: one for interior cells, four for non-corner boundary cells, and another four for the corner cells.

For a cell on the left boundary, we have two equations, the finite difference approximation to the boundary,

$$J_{1/2,j} = -D_{1,j}\frac{\phi_{1,j} - \phi_{1/2,j}}{\Delta x_1/2}, \tag{5-285}$$

and Eq. (5-284a) that we can solve to eliminate the edge quantity $\phi_{1/2,j}$ to solve for the flow rate at the boundary as

$$J_{1/2,j} = -\widetilde{D}_{1/2,j}\left(\alpha_{1/2,j}\phi_{1,j} + \gamma_{1/2,j}\right), \tag{5-286a}$$

where the edge-averaged left boundary diffusion coefficient is

$$\widetilde{D}_{1/2,j} = \frac{2D_{1,j}/\Delta x_1}{\alpha_{1/2,j} + \beta_{1/2,j}(2D_{1,j}/\Delta x_1}. \tag{5-286b}$$

The proces can be repeated for the a cell on the bottom boundary at $y = y_{1/2}$. The flow rate is

$$J_{i,1/2} = -\widetilde{D}_{i,1/2}\left(\alpha_{i,1/2}\phi_{i,1} + \gamma_{i,1/2}\right), \tag{5-286c}$$

with edge-averaged diffusion coefficient of

$$\widetilde{D}_{i,1/2} = \frac{2D_{i,1}/\Delta y_1}{\alpha_{i,1/2} + \beta_{i,1/2}(2D_{i,1}/\Delta y_1)}. \tag{5-286d}$$

On the right boundary, $x = x_{N_x+1/2}$, we have

$$J_{N_x+1/2,j} = \widetilde{D}_{N_x+1/2,j}\left(\alpha_{N_x+1/2,j}\phi_{N_x,j} - \gamma_{N_x+1/2,j}\right), \tag{5-286e}$$

where the edge-averaged left boundary diffusion coefficient is

$$\widetilde{D}_{N_x+1/2,j} = \frac{2D_{N_x,j}/\Delta x_{N_x}}{\alpha_{N_x+1/2,j} - \beta_{N_x+1/2,j}(2D_{N_x,j}/\Delta x_{N_x})}. \tag{5-286f}$$

And finally on the top boundary, $y = y_{N_y+1/2}$:

$$J_{i,N_y+1/2} = \widetilde{D}_{i,N_y+1/2}\left(\alpha_{i,N_y+1/2}\phi_{i,N_y} - \gamma_{i,N_y+1/2}\right), \tag{5-286g}$$

with edge-averaged diffusion coefficient of

$$\widetilde{D}_{i,N_y+1/2} = \frac{2D_{i,N_y}/\Delta y_{N_y}}{\alpha_{i,N_y+1/2} - \beta_{i,N_y+1/2}(2D_{i,N_y}/\Delta y_{N_y})}. \tag{5-286h}$$

Given these results for the flow rates and the associated edge-averaged diffusion coefficients on the boundaries, we can insert them into the balance equations for the cells on the boundaries and corner cells, expressing them in terms of the cell-centered quantities ϕ. The results of these calculations are provided in Table 5.1.

Table 5.1: Cell-Balance Equations for 2-D Reaction Diffusion Equations

Interior	$-\frac{\widetilde{D}_{i-1/2,j}}{\Delta x_i}\phi_{i-1,j} - \frac{\widetilde{D}_{i,j-1/2}}{\Delta y_j}\phi_{i,j-1} + \left[\frac{\widetilde{D}_{i-1/2,j}+\widetilde{D}_{i+1/2,j}}{\Delta x_i} + \frac{\widetilde{D}_{i,j-1/2}+\widetilde{D}_{i,j+1/2}}{\Delta y_j} + \lambda_{i,j}\right]\phi_{i,j} - \frac{\widetilde{D}_{i+1/2,j}}{\Delta x_i}\phi_{i+1,j} - \frac{\widetilde{D}_{i,j+1/2}}{\Delta y_j}\phi_{i,j+1} = q_{i,j}$
Left	$-\frac{\widetilde{D}_{1,j-1/2}}{\Delta y_j}\phi_{1,j-1} + \left[\frac{\widetilde{D}_{1/2,j}\alpha_{1/2,j}+\widetilde{D}_{3/2,j}}{\Delta x_1} + \frac{\widetilde{D}_{1,j-1/2}+\widetilde{D}_{1,j+1/2}}{\Delta y_j} + \lambda_{1,j}\right]\phi_{1,j} - \frac{\widetilde{D}_{3/2,j}}{\Delta x_1}\phi_{2,j} - \frac{\widetilde{D}_{1,j+1/2}}{\Delta y_j}\phi_{1,j+1} = q_{1,j} + \frac{\widetilde{D}_{1/2,j}}{\Delta x_1}\gamma_{1/2,j}$
Bottom	$-\frac{\widetilde{D}_{i-1/2,1}}{\Delta x_i}\phi_{i-1,1} + \left[\frac{\widetilde{D}_{i-1/2,1}+\widetilde{D}_{i+1/2,1}}{\Delta x_i} + \frac{\widetilde{D}_{i,1/2}\alpha_{i,1/2}+\widetilde{D}_{i,3/2}}{\Delta y_1} + \lambda_{i,1}\right]\phi_{i,1} - \frac{\widetilde{D}_{i+1/2,1}}{\Delta x_i}\phi_{i+1,1} - \frac{\widetilde{D}_{i,3/2}}{\Delta y_1}\phi_{i,2} = q_{i,1} + \frac{\widetilde{D}_{i,1/2}}{\Delta y_1}\gamma_{i,1/2}$
Right	$-\frac{\widetilde{D}_{N_x-1/2,j}}{\Delta x_{N_x}}\phi_{N_x-1,j} - \frac{\widetilde{D}_{N_x,j-1/2}}{\Delta y_j}\phi_{N_x,j-1} + \left[\frac{\widetilde{D}_{N_x-1/2,j}+\widetilde{D}_{N_x+1/2,j}\alpha_{N_x+1/2,j}}{\Delta x_{N_x}} + \frac{\widetilde{D}_{N_x,j-1/2}+\widetilde{D}_{N_x,j+1/2}}{\Delta y_j} + \lambda_{N_x,j}\right]\phi_{N_x,j} - \frac{\widetilde{D}_{N_x,j+1/2}}{\Delta y_j}\phi_{N_x,j+1}$ $= q_{N_x,j} + \frac{\widetilde{D}_{N_x+1/2,j}}{\Delta x_{N_x}}\gamma_{N_x+1/2,j}$
Top	$-\frac{\widetilde{D}_{i-1/2,N_y}}{\Delta x_i}\phi_{i-1,N_y} - \frac{\widetilde{D}_{i,N_y-1/2}}{\Delta y_{N_y}}\phi_{i,N_y-1} + \left[\frac{\widetilde{D}_{i-1/2,N_y}+\widetilde{D}_{i+1/2,N_y}}{\Delta x_i} + \frac{\widetilde{D}_{i,N_y-1/2}+\widetilde{D}_{i,N_y+1/2}\alpha_{i,N_y+1/2}}{\Delta y_{N_y}} + \lambda_{i,N_y}\right]\phi_{i,N_y} - \frac{\widetilde{D}_{i+1/2,N_y}}{\Delta x_i}\phi_{i+1,N_y}$ $= q_{i,N_y} + \frac{\widetilde{D}_{i,N_y+1/2}}{\Delta y_{N_y}}\gamma_{i,N_y+1/2}$
Bottom-Left	$\left[\frac{\widetilde{D}_{1/2,1}\alpha_{1/2,1}+\widetilde{D}_{3/2,1}}{\Delta x_1} + \frac{\widetilde{D}_{1,1/2}\alpha_{1,1/2}+\widetilde{D}_{1,3/2}}{\Delta y_1} + \lambda_{1,1}\right]\phi_{1,1} - \frac{\widetilde{D}_{3/2,1}}{\Delta x_1}\phi_{2,1} - \frac{\widetilde{D}_{1,3/2}}{\Delta y_1}\phi_{1,2} = q_{1,1} + \frac{\widetilde{D}_{1/2,1}}{\Delta x_1}\gamma_{1/2,1} + \frac{\widetilde{D}_{1,1/2}}{\Delta y_1}\gamma_{1,1/2}$
Bottom-Right	$-\frac{\widetilde{D}_{N_x-1/2,1}}{\Delta x_{N_x}}\phi_{N_x-1,1} + \left[\frac{\widetilde{D}_{N_x-1/2,1}+\widetilde{D}_{N_x+1/2,1}\alpha_{N_x+1/2,1}}{\Delta x_{N_x}} + \frac{\widetilde{D}_{N_x,1/2}\alpha_{N_x,1/2}+\widetilde{D}_{N_x,3/2}}{\Delta y_1} + \lambda_{N_x,1}\right]\phi_{N_x,1} - \frac{\widetilde{D}_{N_x,3/2}}{\Delta y_1}\phi_{N_x,2}$ $= q_{N_x,1} + \frac{\widetilde{D}_{N_x+1/2,1}}{\Delta x_{N_x}}\gamma_{N_x+1/2,1} + \frac{\widetilde{D}_{N_x,1/2}}{\Delta y_1}\gamma_{N_x,1/2}$
Top-Left	$-\frac{\widetilde{D}_{1,N_y-1/2}}{\Delta y_{N_y}}\phi_{1,N_y-1} + \left[\frac{\widetilde{D}_{1/2,N_y}\alpha_{1/2,N_y}+\widetilde{D}_{3/2,N_y}}{\Delta x_1} + \frac{\widetilde{D}_{1,N_y-1/2}+\widetilde{D}_{1,N_y+1/2}\alpha_{1,N_y+1/2}}{\Delta y_{N_y}} + \lambda_{1,N_y}\right]\phi_{1,N_y} - \frac{\widetilde{D}_{3/2,N_y}}{\Delta x_1}\phi_{2,N_y}$ $= q_{1,N_y} + \frac{\widetilde{D}_{1/2,N_y}}{\Delta x_1}\gamma_{1/2,N_y} + \frac{\widetilde{D}_{1,N_y+1/2}}{\Delta y_{N_y}}\gamma_{1,N_y+1/2}$
Top-Right	$-\frac{\widetilde{D}_{N_x-1/2,N_y}}{\Delta x_{N_x}}\phi_{N_x-1,N_y} - \frac{\widetilde{D}_{N_x,N_y-1/2}}{\Delta y_{N_y}}\phi_{N_x,N_y-1} + \left[\frac{\widetilde{D}_{N_x-1/2,N_y}+\widetilde{D}_{N_x+1/2,N_y}\alpha_{N_x+1/2,N_y}}{\Delta x_{N_x}} + \frac{\widetilde{D}_{N_x,N_y-1/2}+\widetilde{D}_{N_x,N_y+1/2}\alpha_{N_x,N_y+1/2}}{\Delta y_{N_y}} + \lambda_{N_x,N_y}\right]\phi_{N_x,N_y}$ $= q_{i,N_y} + \frac{\widetilde{D}_{N_x+1/2,N_y}}{\Delta x_{N_x}}\gamma_{N_x+1/2,N_y} + \frac{\widetilde{D}_{N_x,N_y+1/2}}{\Delta y_{N_y}}\gamma_{N_x,N_y+1/2}$

Gauss-Seidel Iteration Scheme and Successive Overrelaxation

The linear system given by Eq. (5-282) for the interior cells and the associated results for the boundary and corner points do not, unlike for the 1-D case, form a tridiagonal system. Rather, it is a banded tridiagonal system. Unfortunately, the simplistic tridiagonal solver cannot be used. Instead, we employ an iterative solution scheme such as Gauss-Seidel. A couple comments. One question is whether convergence can be guaranteed. Recall that if the matrix $\mathbf{A}$ is diagonally dominant, than the algorithm will converge. We can show that for this case, diagonal dominance is guaranteed. The second point is that, as we will see, we do not actually need to represent the matrices explicitly. This is because the second-derivative operator introduces coupling only between one point and its (four) adjacent points. On account of this, the matrix is sparse and has a pattern. Therefore, we can devise a scheme that only requires representing the non-zero coefficients in vectors corresponding to their relative orientation with respect to the equation being solved.

To apply this to the 2-D neutron diffusion equation, can express Eq. (5-282) compactly as

$$C_m\phi_m + L_m\phi_{m-1} + R_m\phi_{m+1} + B_m\phi_{m-N} + A_m\phi_{m+N} = r_m, \quad m = i + jN_x. \tag{5-287}$$

Here m represents a flattened one-dimensional index,

$$m = i + (j-1)N_x, \quad i = 1, \ldots, N_x, \quad j = 1, \ldots, N_y, \quad m = 1, \ldots N_xN_y, \tag{5-288}$$

and the coefficients are as follows: C_m on the center element or cell, which is given by the correspoinding term in square brackets in Table 5.1. For an interior element, this is

$$C_m = \frac{\widetilde{D}_{i-1/2,j} + \widetilde{D}_{i+1/2,j}}{\Delta x_i} + \frac{\widetilde{D}_{i,j-1/2} + \widetilde{D}_{i,j+1/2}}{\Delta y_j} + \lambda_{i,j}, \tag{5-289a}$$

and the boundary and corner cells follow from the table. The remaining coefficients have the same form regardless as to whether the cell is interior or not. We have: L_m on the cell to the left,

$$L_m = -\frac{\widetilde{D}_{i-1/2,j}}{\Delta x_i}; \tag{5-289b}$$

R_m on the cell to the right,

$$R_m = -\frac{\widetilde{D}_{i+1/2,j}}{\Delta x_i}; \tag{5-289c}$$

B_m for the cell below,

$$B_m = -\frac{\widetilde{D}_{i,j-1/2}}{\Delta y_j}; \tag{5-289d}$$

A_m for the cell above,

$$A_m = -\frac{\widetilde{D}_{i,j+1/2}}{\Delta y_j}. \tag{5-289e}$$

And finally, r_m is the right-hand side or inhomogeneous source, which is the right-hand side of Table 5.1. For an interior cell this is

$$r_m = q_{i,j}, \tag{5-289f}$$

and we apply the respective forms for the boundary and corner cells.

Once the equations have been written down, we then need to proceed with solving them. This can be represented as as matrix algebra $\mathbf{Ax} = \mathbf{b}$ problem using the flattened index m. The ordering of the index is in such a manner that it gives the matrix a regular structure. This matrix can be organized into a banded-tridiagonal structure. An example of this for the grid 6×5 grid is as follows:

$$\left[\begin{array}{cccccc|cccccc|cccccc|cccccc|cccccc}
C_1 & R_1 & 0 & 0 & 0 & 0 & A_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
L_2 & C_2 & R_2 & 0 & 0 & 0 & 0 & A_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & L_3 & C_3 & R_3 & 0 & 0 & 0 & 0 & A_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & L_4 & C_4 & R_4 & 0 & 0 & 0 & 0 & A_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & L_5 & C_5 & R_5 & 0 & 0 & 0 & 0 & A_5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & L_6 & C_6 & 0 & 0 & 0 & 0 & 0 & A_6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
B_7 & 0 & 0 & 0 & 0 & 0 & C_7 & R_7 & 0 & 0 & 0 & 0 & A_7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & B_8 & 0 & 0 & 0 & 0 & L_8 & C_8 & R_8 & 0 & 0 & 0 & 0 & A_8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & B_9 & 0 & 0 & 0 & 0 & L_9 & C_9 & R_9 & 0 & 0 & 0 & 0 & A_9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & B_{10} & 0 & 0 & 0 & 0 & L_{10} & C_{10} & R_{10} & 0 & 0 & 0 & 0 & A_{10} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & B_{11} & 0 & 0 & 0 & 0 & L_{11} & C_{11} & R_{11} & 0 & 0 & 0 & 0 & A_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & B_{12} & 0 & 0 & 0 & 0 & L_{12} & C_{12} & 0 & 0 & 0 & 0 & 0 & A_{12} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & B_{13} & 0 & 0 & 0 & 0 & 0 & C_{13} & R_{13} & 0 & 0 & 0 & 0 & A_{13} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & B_{14} & 0 & 0 & 0 & 0 & L_{14} & C_{14} & R_{14} & 0 & 0 & 0 & 0 & A_{14} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & B_{15} & 0 & 0 & 0 & 0 & L_{15} & C_{15} & R_{15} & 0 & 0 & 0 & 0 & A_{15} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & B_{16} & 0 & 0 & 0 & 0 & L_{16} & C_{16} & R_{16} & 0 & 0 & 0 & 0 & A_{16} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & B_{17} & 0 & 0 & 0 & 0 & L_{17} & C_{17} & R_{17} & 0 & 0 & 0 & 0 & A_{17} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & B_{18} & 0 & 0 & 0 & 0 & L_{18} & C_{18} & 0 & 0 & 0 & 0 & 0 & A_{18} & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & B_{19} & 0 & 0 & 0 & 0 & 0 & C_{19} & R_{19} & 0 & 0 & 0 & 0 & A_{19} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & B_{20} & 0 & 0 & 0 & 0 & L_{20} & C_{20} & R_{20} & 0 & 0 & 0 & 0 & A_{20} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & B_{21} & 0 & 0 & 0 & 0 & L_{21} & C_{21} & R_{21} & 0 & 0 & 0 & 0 & A_{21} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & B_{22} & 0 & 0 & 0 & 0 & L_{22} & C_{22} & R_{22} & 0 & 0 & 0 & 0 & A_{22} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & B_{23} & 0 & 0 & 0 & 0 & L_{23} & C_{23} & R_{23} & 0 & 0 & 0 & 0 & A_{23} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & B_{24} & 0 & 0 & 0 & 0 & L_{24} & C_{24} & 0 & 0 & 0 & 0 & 0 & A_{24} \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & B_{25} & 0 & 0 & 0 & 0 & 0 & C_{25} & R_{25} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & B_{26} & 0 & 0 & 0 & 0 & L_{26} & C_{26} & R_{26} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & B_{27} & 0 & 0 & 0 & 0 & L_{27} & C_{27} & R_{27} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & B_{28} & 0 & 0 & 0 & 0 & L_{28} & C_{28} & R_{28} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & B_{29} & 0 & 0 & 0 & 0 & L_{29} & C_{29} & R_{29} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & B_{30} & 0 & 0 & 0 & 0 & L_{30} & C_{30}
\end{array}\right].$$

Partitions are included in this matrix to illustrate where the problem edges reside and to highlight features in the matrix structure. First, it should be apparent that most elements in this matrix are zero, i.e., the matrix is sparse. Therefore, we do not need to explicitly represent it on the computer (which can be prohibitive as the size of the grid becomes large).

Starting the lower-left corner $m = 1$ or $\phi_{1,1}$, we solve for the corresponding value of the field there. For the corner point, the value can be solved as

$$\phi_1^{(n+1)} = \frac{1}{C_1}\left(r_1 - R_1\phi_2^{(n)} - A_1\phi_{1+N_x}^{(n)}\right). \tag{5-290}$$

Here we introduced a superscript in parentheses to denote an iteration index and use the flattened index. Next, we move to the right and solve for the adjacent value $\phi_2 = \phi_{2,0}$, using the new value of $\phi_{1,1}$ we just solved for:

$$\phi_2^{(n+1)} = \frac{1}{C_2}\left(r_2 - L_2\phi_1^{(n+1)} - R_2\phi_3^{(n)} - A_2\phi_{2+N_x}^{(n)}\right). \tag{5-291}$$

Again, the iteration index of $(n+1)$ on ϕ_1 is because we are using the value we just computed. We then continue the process, moving to the right until we reach the end of the first row. For a generic element on this row with flattened index $m = i$, we have

$$\phi_i^{(n+1)} = \frac{1}{C_i}\left(r_i - L_i\phi_{i-1}^{(n+1)} - R_i\phi_{i+1}^{(n)} - A_i\phi_{i+N_x}^{(n)}\right), \tag{5-292}$$

and at the lower-right corner we have

$$\phi_{N_x}^{(n+1)} = \frac{1}{C_{N_x}}\left(r_{N_x} - L_{N_x}\phi_{N_x-1}^{(n+1)} - A_{N_x}\phi_{2N_x}^{(n)}\right). \tag{5-293}$$

Next, we move onto the second row and solve for each field value from left to right and then over the rows from bottom to top, always using the most up-to-date information. For a generic interior element, we have

$$\phi_m^{(n+1)} = \frac{1}{C_m}\left(r_m - L_m\phi_{m-1}^{(n+1)} - B_m\phi_{m-N_x}^{(n+1)} - R_m\phi_{m+1}^{(n)} - A_m\phi_{m+N_x}^{(n)}\right). \tag{5-294}$$

Note that the element to the left and below have already been solved for on this iteration, so we use those values in the equation.

After we solve for the element at the top-right corner $\phi_{N_xN_y} = \phi_{N_x,N_y}$, we check whether the new field values are within some tolerance ϵ of the previous solution. This is done using the following norm:

$$\frac{\sum_{m=1}^{N_xN_y}|\phi_m^{(n+1)} - \phi_m^{(n)}|}{\sum_{m=1}^{N_xN_y}|\phi_m^{(n+1)}|} < \epsilon. \tag{5-295}$$

If the tolerance is satisfied, the calculation finishes and the current solution is output. If it is not, the process is repeated for more iterations until convergence is achieved.

The implementation of this is actually (perhaps surprisingly) not too difficult. The coefficient vectors C, L, R, B, A along with the inhomogeneous right-hand side source term can be precomputed. We then begin with the top row of the matrix and solve for a new value using the appropriate equation, computing the contribution to the error, and then overwriting the old value with that new value. Next, we proceed down the rows, repeating the process and using the most up-to-date information. The tricky part is handling the edges gracefully so as not to access invalid array elements. These can be handled either with if-then statements or by sizing vectors larger than they need to be and storing zeroes in those elements.

The convergence rate for Gauss-Seidel can be slow for large linear systems. Accelerating this convergence becomes especially important because we can have multiple energy groups (typically not too many for diffusion, thankfully) and, more significantly, an outer iteration loop to converge the fission source. The most common acceleration method is called *successive over-relaxation*, which is a kind of extrapolation.

We introduce an over-relaxation parameter ω that is typically on the domain $1 \leq \omega < 2$. The value of $\omega = 1$ reduces to Gauss-Seidel and $\omega \geq 2$ can be numerically unstable. Note that sometimes we let $\omega < 1$ to under-relax numerical schemes to assist with numerical stability.

Anyway, we return to the $\mathbf{Ax} = \mathbf{b}$ problem. We split out the matrix $\mathbf{A}$ as before and then multiply by a factor ω. Moving the upper triangular part to the right-hand side we get

$$(\omega\mathbf{L} + \omega\mathbf{D})\,\mathbf{x} = \omega\mathbf{b} - \omega\mathbf{U}\mathbf{x} \tag{5-296}$$

Next we add $(1-\omega)\mathbf{D}\mathbf{x}$ to both sides of the equation. After using the fact that $\omega + (1-\omega) = 1$ and applying iteration indices we have

$$[\omega\mathbf{L} + \mathbf{D}]\,\mathbf{x}^{(n+1)} = \omega\mathbf{b} + [(1-\omega)\mathbf{D} - \omega\mathbf{U}]\,\mathbf{x}^{(n)}. \tag{5-297}$$

We can then do as before and write out each term given that the system only couples with its adjacent grid points. To illustrate, the updated flux for the generic interior grid point is

$$\phi_m^{(n+1)} = (1-\omega)\phi_m^{(n)} + \frac{\omega}{C_m}\left(r_m - L_m\phi_{m-1}^{(n+1)} - B_m\phi_{m-N_x}^{(n+1)} - R_m\phi_{m+1}^{(n)} - A_m\phi_{m+N_x}^{(n)}\right). \tag{5-298}$$

The other expression follow directly. By quick observation, the application of successive over-relaxation requires only a trivial modification to the equations used in the loops. This is a major reason why it is such popular acceleration scheme.

One question is the choice of ω. There is a unique, optimal value of ω. While there are some analyses that can be done under simple conditions to determine this, in practice this is often chosen based on numerical experimentation for the class of problems. If ω is chosen close to the optimal value, then the number of iterations required for convergence is significantly reduced.

Chapter 6

Probability

In this chapter, we turn our attention to the ideas of probability and randomness. This is important because in many physical systems we do have true randomness because of the laws of quantum mechanics. Individual particle interactions, for example can only be described probabilistically in the sense that when and what interacts a particle undergoes is inherently random. Furthermore, we often apply the ideas of probability to systems that are essentially deterministic, but appear random because of uncertainties resulting from our limited information of a particular system. For example, the failure time of a machine could be predicted if we fully understood the physics of every component of that machine at all points in time. Because this is not practical, we often use a mathematical model involving randomness to analyze failure rates and overall consequences of such systems. Furthermore, the ideas of probability theory are used strongly in many algorithms related to data analytics or machine learning in that they use large amounts of data to make inferences about particular systems with many unknowns.

6.1 Basic Concepts

6.1.a Interpretations of Probability

Before going into the discussion of the mathematics of probability theory, let's pause and very briefly discuss the meaning of probability. There are two major philosophical interpretations of the concept of probability.

The first is the *frequentist* interpretation. The frequentist interpretations views probability, as the name implies, as measuring the frequency of a given random event having a particular outcome. For example, suppose we measure the number of radioactive decays from a particular sample in a given time window. Because radioactive decay is inherently a random process, and each time we perform the experiment we expect to get a different result. If we repeat the experiment numerous times, we can make inferences about how likely it is to have a particular number or range of counts

in a given time interval.

The second major philosophical interpretation of probability is the *Bayesian* interpretation. The Bayesian approach to probability views the concept as measuring the degree of belief or logical support for a given hypothesis. For example, suppose we wish to decide whether we should perform an expensive and time consuming preventative maintenance procedure in a nuclear power plant before the next scheduled outage. We could assign a relative probability based on our degree of believe whether equipment would fail and therefore maintenance would be required. We could assign this prior probability, based on, for example, past performance data at this or other nuclear power plants or even a subjective guess. Suppose we then collect information from engineers performing inspections on various components throughout the plant and obtain data from various sensors throughout the plant that could indicate potentially off-normal conditions. Based on the results of these inspections and sensor data we can refine our degree of belief as to whether the unplanned outage for maintenance is required.

Both interpretations of probability are valid and useful in their own contexts. There is come conflict in terms of how to make inference and decision making, but largely we will ignore these issues in this text and focus on the general concepts.

6.1.b Random Events

To begin our discussion, we need to define the concept of an event. An event is a set of possible outcomes that have a defined probability. Let A be a particular outcome and $P(A)$ be the probability that outcome A occurs. A few examples of events and outcomes are as follows:

- Rolling two dice and getting a particular total;
- A piece of equipment in a laboratory fails before a replacement arrives;
- A sensor reading in a nuclear power plant is accurate within a particular range;
- An engineer has performed a calculation without any mistakes.

Probabilities must be assigned so that for any event:

$$0 \leq P(A) \leq 1. \tag{6-1}$$

An event also has a complementary event $\overline{A}$, which is the oppose of the event. If the event is complimentary, then we have

$$P(A) + P(\overline{A}) = 1. \tag{6-2}$$

We can also have multiple events A and B that may or may not be related. We define that the probability that both A and B occurs as $P(A, B)$. If the events are disjoint or mutually exclusive, we say that $P(A, B) = 0$. Alternatively, we can state

that the probability of either A or B occurring is $P(A) + P(B)$. In general, we can state that

$$P(A \cup B) = P(A) + P(B) - P(A, B). \tag{6-3}$$

In other words, the probability that either A or B occurs is the probability that A occurs (regardless of B) plus the probability that B occurs (regardless of A) minus the probability that both A and B occur.

If the two events are unrelated, we call them independent and can state that

$$P(A, B) = P(A)P(B), \text{ only if events } A \text{ and } B \text{ are independent.} \tag{6-4}$$

6.1.c Conditional Probability and Bayes' Theorem

It is often the case we know a particular event B occurred and we wish to know the probability whether a different event A will occur given that information. We denote this as the conditional probability $P(A|B)$. Examples of this include:

- If the first of two die is a 4, what is the probability that the total of two die is at least 6?
- If a piece of equipment has failed in the first year of operation, what is the probability that an otherwise identical piece of equipment will fail in its first year?
- If two sensors report a reading indicating the condition of a system is normal and a third reports an off-normal condition, what is the probability the system is in a normal condition?

If A and B are independent of each other then $P(A|B) = P(A)$, since the outcome B has no impact on how likely A is to occur.

We can also infer the result that

$$P(A) = P(A|B)P(B) + P(A|\overline{B})P(\overline{B}). \tag{6-5}$$

This states the probability of event A occurring is equal to the probability of A conditional on event B and A conditional on B not occurring. We can generalize this to the law of total probability. If B_i are disjoint events then:

$$P(A) = \sum_i P(A|B_i)P(B_i). \tag{6-6}$$

We can relate the joint probability to conditional probability as follows:

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A). \tag{6-7}$$

From this we can obtain the result known as Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{6-8}$$

6.1.d Example: Predicting Pump Failure Based on a Sensor

As an example of Bayes' theorem, let us consider the event A that a pump in a nuclear power plant is not operating normally, indicated by a degraded flow, and requiring maintenance. Based on operating experience, we know that this brand of pump operates correctly 98% of the time. Therefore, with no additional information, the best we can do is assert that at any given time, there is a 2% chance that the pump will not be functioning correctly.

Now suppose we have a pressure sensor attached to the pump that can indicate that the pump is failing through measuring degradation in the flow. Let B be the event that the sensor indicates the flow is not normal. Unfortunately, the sensor itself is a piece of equipment that can become uncalibrated and report inaccurate readings, and this must be taken into account in the analysis. From historical operating experience, the sensor will successfully detect degraded flows 99.5% of the time. Because the sensor can become miscalibrated with time, it will successfully indicate normal flows only 95% of the time. This implies that the sensor will fail to report a degraded flow 0.5% of the time and misreport a normal flow as degraded 5% of the time. Let event B be the sensor indicating an off-normal flow. Suppose we wish to find the probability that the pump is failing.

From the information, we know that the probability that the pump is indeed failing is 2%, or

$$P(A) = 0.02.$$

The probability that the sensor indicates an off-normal condition regardless of the condition of the pump can be determined from the provided conditional probabilities. We have

$$\begin{aligned} P(B) &= P(B|A)P(A) + P(B|\overline{A})P(\overline{A}) \\ &= 0.995 \times 0.02 + 0.05 \times 0.98 \\ &= 0.0199 + 0.049 = 0.0689. \end{aligned}$$

This equation states the probability that the sensor will indicate a degraded flow is the probability of actually detecting the degraded flow plus the probability of misreporting the normal flow as degraded. Note that this result suggests that more often than not, the sensor indicating an off-normal flow is because of it becoming miscalibrated and not because of a problem with the flow itself.

Using Bayes' theorem, we can calculate the probability that the pump is indeed failing if the sensor indicates an off-normal flow:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0.995 \times 0.02}{0.0689} = 0.288.$$

Therefore, the sensor indicating that the flow is off-normal only gives us a 28.8% chance that the pump is actually in need of maintenance. In other words, roughly 7 out of 10 times, we would make the wrong decision as to repair the pump.

Since the impact of the former decision is simply lost revenue from conducting an unnecessary repair, it may be worthwhile to analyze the false negative rate of failing to detect a failing pump. For this we require the probability that the sensor indicates normal flow:

$$P(\overline{B}) = 1 - 0.0689 = 0.9311.$$

Using Bayes' theorem we can calculate the probability that the pump is actually failing given that the sensor reports normal conditions. This is

$$P(A|\overline{B}) = \frac{P(\overline{B}|A)P(A)}{P(\overline{B})} = \frac{0.005 \times 0.02}{0.9311} = 1.07 \times 10^{-4}.$$

This indicates that the chance of the sensor reporting a normal reading and the pump being in need of maintenance is quite low.

6.1.e Random Variables

A random variable is a variable whose values depend on some random process. We usually denote a random variable using a capital letter such as X and indicate a particular value that the random variable can take as the lowercase version x. Therefore, we say that random variable X takes on value x with probability $P(X = x)$. Underlying every random variable is a probability distribution function that gives the likelihoods of the random variable X taking on a specific value.

Random variables are either discrete or continuous (or possibly a mixture of discrete and continuous, but we will ignore this case). The equations and methods for analyzing each case are very similar, but have some key differences. In the next section we will analyze discrete random variables and then we will analyze continuous random variables in the section that follows. In these will introduce the concepts of the probability mass/density function and the cumulative distribution function.

6.2 Discrete Random Variables

As the name implies, discrete random variables are permitted to take on a set of possibly infinite discrete values. Examples of discrete random variables are:

- The result of a coin flip where heads is given a value of 1 and tails is given a value of 0;
- The result of the role of a die, which is the set $\{1, 2, 3, 4, 5, 6\}$;
- The number of unplanned nuclear power plant outages in the US in a given year;
- The number of counts measured from a radioactive source in a minute;

- The energy of a gamma photon emitted during a radioactive decay (the nuclei have discrete energy levels);

- The type of reaction that occurs in a neutron-nucleus interaction: scattering, absorption, fission, etc.

A discrete random variable is characterized by a probability mass function or cumulative distribution function. In the context of discrete random variables, both of these are probabilities describing the same information in different ways. For continuous random variables, these quantities are fundamentally different, but we will get into that later.

Note that the outcomes of discrete random variables need not be numbers. A common case in nuclear engineering is determining the particular type of interaction that occurs, e.g., scattering, absorption, or fission. We could (and often do) enumerate these reactions with numbers such as 1, 2, 3, ...; however, the numbering is arbitrary and merely useful for mathematical bookkeeping.

In this section, we will introduce the concept of a probability mass function and its partner the cumulative distribution function. We will then provide a couple examples.

6.2.a Probability Mass Function

The probability mass function gives the probability that a random variable X takes on a particular value x. This is denoted with a lowercase function variable f with a subscript being the random variable as

$$f_X(x) = P(X = x). \tag{6-9}$$

The probability mass function must satisfy the property that if we sum up the probabilities of all possible events, we get one

$$\sum_{x_i} f_X(x_i) = 1. \tag{6-10}$$

All valid probability mass functions are normalized in this way.

A simple example of a probability mass function is that of the case where we have a six-sided die where all sides have an equal probability; the probability mass function is as follows:

$$f_X(x) = \begin{cases} \frac{1}{6}, & x \in \{1,2,3,4,5,6\} \\ 0, & \text{otherwise} \end{cases}. \tag{6-11}$$

In other words, the probability is $\frac{1}{6}$ for all values between 1 and 6 inclusive and zero for any other value.

6.2.b Cumulative Distribution Function

The cumulative distribution function gives the probability that the of some random variable X is x or less. This is denoted by a capital F with a subscript being the random variable as

$$F_X(x) = P(X \leq x) = \sum_{x_i \leq x} f_X(x_i). \tag{6-12}$$

The cumulative distribution function is a monotonically increasing function that satisfies:

$$\begin{aligned} F_X(-\infty) &= 0, \\ F_X(\infty) &= 1. \end{aligned}$$

An important result is that if we want to know the probability that a random variable X is between values a exclusive and b inclusive; we can write

$$P(a < x \leq b) = P(X \leq b) - P(X \leq a) = F_X(b) - F_X(a). \tag{6-13}$$

We can use this result to get to the probability mass function. Suppose we know the cumulative distribution function and want to know the probability mass function at each point x_i numbered in order of $i = 1, 2, \ldots$. This relationship is

$$f_X(x_i) = P(X = x_i) = P(X \leq x_i) - P(X \leq x_{i-1}) = F_X(x_i) - F_X(x_{i-1}). \tag{6-14}$$

6.2.c Example: Sum of Two Six-Sided Dice

Let us build the probability mass and cumulative distribution functions for the sum of two six-sided dice. The probability of a single die landing on any side from 1 to 6 is $\frac{1}{6}$. The dice are rolled independently of each other so that the result of one has no impact on the other. Since the number of combinations is relatively limited, 36, we can "brute force" the probability mass function by enumerating the events. This is done in the table.

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

To get the probability mass function, we simply count the frequency of a particular

result. This gives

$$f_X(x) = \begin{cases} \frac{1}{36}, & x = 2, 12 \\ \frac{1}{18}, & x = 3, 11 \\ \frac{1}{12}, & x = 4, 10 \\ \frac{1}{9}, & x = 5, 9 \\ \frac{5}{36}, & x = 6, 8 \\ \frac{1}{6}, & x = 7 \\ 0, & \text{otherwise} \end{cases} . \tag{6-15}$$

The probability mass function is displayed in Fig. 6.1. This plot has spikes with a height corresponding to the probability at the discrete integer values of the valid dice rolls. The probability of having a dice roll is zero for any non-integer value or integers outside the range of 2 to 12.

The cumulative distribution function is

$$F_X(x) = \begin{cases} 0, & x < 2 \\ \frac{1}{36}, & 2 \leq x < 3 \\ \frac{1}{12}, & 3 \leq x < 4 \\ \frac{1}{6}, & 4 \leq x < 5 \\ \frac{5}{18}, & 5 \leq x < 6 \\ \frac{5}{12}, & 6 \leq x < 7 \\ \frac{7}{12}, & 7 \leq x < 8 \\ \frac{13}{18}, & 8 \leq x < 9 \\ \frac{5}{6}, & 9 \leq x < 10 \\ \frac{11}{12}, & 10 \leq x < 11 \\ \frac{35}{36}, & 11 \leq x < 12 \\ 1, & x \geq 12 \end{cases} . \tag{6-16}$$

The cumulative distribution function is also displayed in Fig. 6.1. The plot forms a monotonically increasing set of piecewise constant values where the open circles denote that the particular value is excluded and the closed circles denote that the value is included. Note that the cumulative distribution is defined for all real numbers, not just the discrete integers and has a number between zero and one at all points. Jumps occur in the function upon reaching the integer values corresponding to the valid dice rolls. The height of those jumps is equal to the probability of getting that particular value.

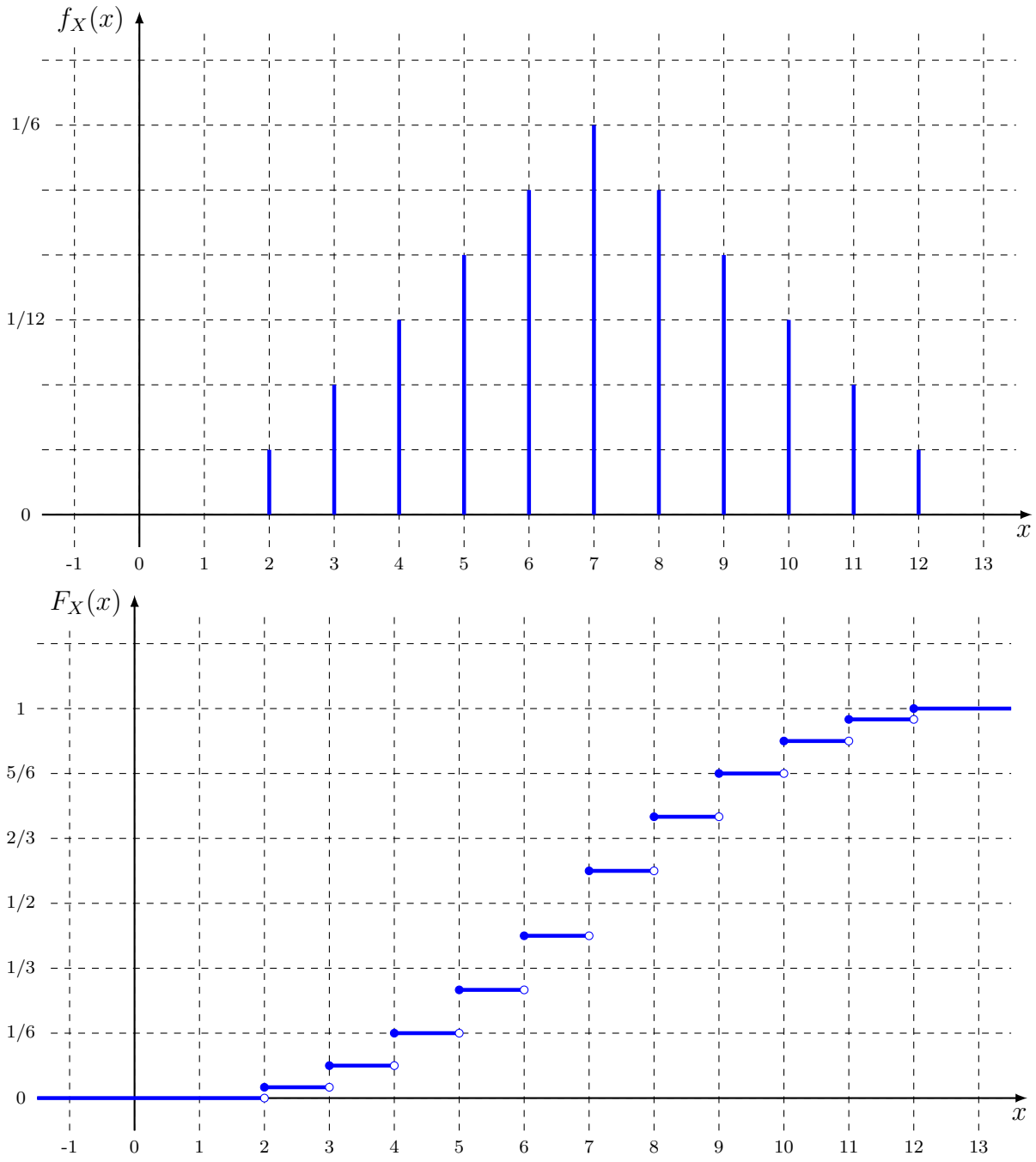

Figure 6.1: Probability mass and cumulative distribution functions for the sum of two six-sided dice.

6.2.d Example: Nuclear Reaction Probabilities

When two particles interact, the outcome of that interaction is inherently random because of the laws of quantum mechanics. Here we will consider the case of a neutron interacting the a fissile nucleus such as ^{235}U. In this case there are three different types

of nuclear reactions that may occur: elastic scattering, radiative capture (n,γ), and nuclear fission.

The total likelihood of interaction is governed by the microscopic total cross section σ_t (which is a strong function of the neutron kinetic energy, but we will ignore this for the time being). Each individual reaction has a cross section: elastic scattering is σ_s; radiative capture is σ_γ, and fission is σ_f. The reaction probabilities are given by the ratio of the reaction cross section divided by the total cross section. For example, the probability of fission $p_f = \sigma_f/\sigma_t$.

Suppose we have $\sigma_t = 4$ b, $\sigma_s = 2.5$ b, $\sigma_c = 0.5$ b, and $\sigma_f = 1$ b. the reaction probabilities are $p_s = \frac{5}{8}$, $p_c = \frac{1}{8}$, and $p_f = \frac{1}{4}$. If we number these as 1 being scattering, 2 being radiative capture, and 3 being nuclear fission, we can compute the probability mass function and cumulative distribution function. The probability mass function is

$$f_X(x) = \begin{cases} \frac{5}{8}, & x = 1, \text{elastic scattering} \\ \frac{1}{8}, & x = 2, \text{radiative capture} \\ \frac{1}{4}, & x = 3, \text{nuclear fission} \\ 0, & \text{otherwise} \end{cases}, \tag{6-17}$$

and the cumulative distribution function is

$$F_X(x) = \begin{cases} 0, & x < 1 \\ \frac{5}{8}, & 1 \leq x < 2 \\ \frac{3}{4}, & 2 \leq x < 3 \\ 1, & x \geq 3 \end{cases}. \tag{6-18}$$

6.3 Continuous Random Variables

In the previous section, we introduced the concept of random variables for the case that they can take on discrete values. We can also talk about random variables that can take on a continuum of values. The support may be finite or infinite. Examples of continuous random variables are

- The time for pump in a nuclear power plant to fail from its installation date;
- The distance a photon travels before undergoing an interaction with an electron in a solid;
- The kinetic energy of a neutron emitted from fission;
- The thickness of a component in given uncertainties in the manufacturing process (dimensional tolerances).

In this section, we will introduce the probability density function and the cumulative distribution function for continuous random variables. These concepts are similar, but do differ from the discrete case and warrant separate discussion. We will then give a few examples and discuss common continuous random distributions.

6.3.a Probability Density Function

The probability density function is the analog of the probability mass function of discrete random variables, but for continuous random variables. The major difference is that the probability density function does not return a probability, but rather a probability density, which is probability per whatever unit the random variable has, e.g., probability per unit length, probability per unit time, etc.

The probability that a random variable X with have a result between x and $x+dx$ is given as

$$f_X(x)dx = P(x \leq X \leq x + dx). \tag{6-19}$$

To get a probability, we take the area under the curve of a probability density function:

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx. \tag{6-20}$$

The probability density function satisfies the normalization property that

$$\int_{-\infty}^{\infty} f_X(x)dx = 1. \tag{6-21}$$

The probability density function satisfies the property that it is never negative:

$$f_X(x) \geq 0. \tag{6-22}$$

Since the probability density function is not a probability, it is not restricted to be less than or equal to one. On the contrary, it is often the case that a probability density function has portions that are greater than one, so long as the area under the curve of any region is always less than or equal to one.

6.3.b Cumulative Distribution Function

There is also a version of the cumulative distribution function for continuous random variables. Similar to discrete random variables, the cumulative distribution is the probability that a continuous random variable X is at most a value x:

$$F_X(x) = P(X \leq x), \tag{6-23}$$

which may be obtained from the integral

$$F_X(x) = \int_{-\infty}^{x} f_X(x')dx'. \tag{6-24}$$

The cumulative distribution function is a monotonically increasing function again satisfying:

$$F_X(-\infty) = 0,$$
$$F_X(\infty) = 1.$$

Because the cumulative distribution function is a continuous function, we can get the corresponding probability density function from

$$f_X(x) = \frac{dF_X}{dx}. \tag{6-25}$$

6.3.c Example: Radioactive Decay

An important application of continuous random variables in nuclear engineering is radioactive decay, which follows an exponential decay law. The probability density function for the exponential decay law is

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t}, & t \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{6-26}$$

where λ is the decay constant, which has units of inverse time. We can show that the exponential distribution is indeed normalized:

$$\int_{-\infty}^{\infty} f_T(t)dt = \int_{-\infty}^{0} 0dt + \int_{0}^{\infty} \lambda e^{-\lambda t}dt = 0 + \lambda\frac{1}{\lambda} = 1. \tag{6-27}$$

If we want to know the probability a radioisotope decays by a certain time, we can calculate the cumulative distribution function:

$$F_T(t) = \int_{-\infty}^{t} f_T(t')dt' = \int_{0}^{t} \lambda e^{-\lambda t'}dt' = 1 - e^{-\lambda t}, \quad t \geq 0. \tag{6-28}$$

If $t < 0$, the cumulative distribution function is zero.

6.3.d Example: Piecewise-Linear Function

It is common to represent probability density functions of more complicated functions as piecewise-constant or piecewise-linear functions. An example of this is as follows:

$$f_X(x) = \begin{cases} C(1+x), & -1 \leq x \leq 0 \\ C(1-\frac{x}{2}), & 0 \leq x \leq 2 \\ 0, & \text{otherwise} \end{cases}. \tag{6-29}$$

Here C is a normalization constant we need to determine. To find this, we integrate the probability density function from $-\infty$ to ∞, set the result equal to 1, and solve for the constant. This proceeds as

$$\begin{aligned} \int_{-\infty}^{\infty} f_X(x) &= \int_{-1}^{0} C(1+x)dx + \int_{0}^{2} C(1-\tfrac{x}{2})dx = 1 \\ &= \frac{3}{2}C = 1 \end{aligned} \tag{6-30}$$

Therefore, the constant is $C = \frac{2}{3}$ and we can rewrite the probability density function as

$$f_X(x) = \begin{cases} \frac{2}{3}(1+x), & -1 \leq x \leq 0 \\ \frac{2}{3}(1-\frac{x}{2}), & 0 < x \leq 2 \\ 0, & \text{otherwise} \end{cases}. \tag{6-31}$$

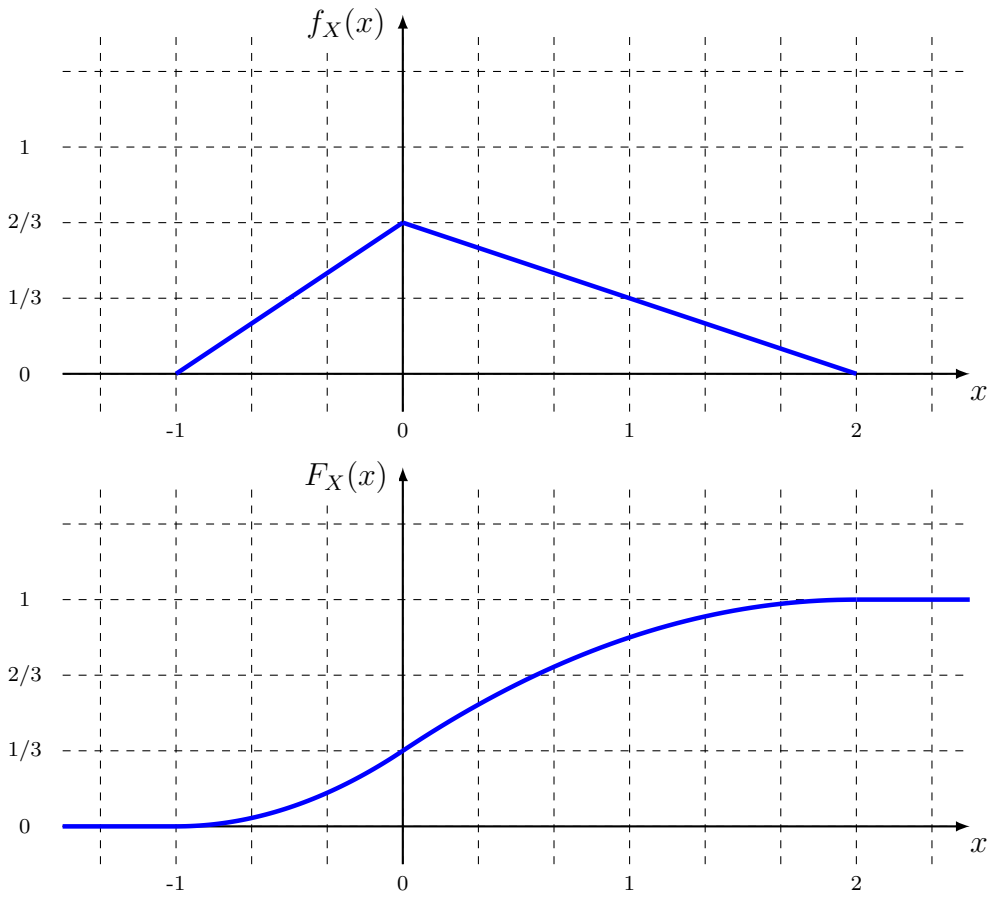

Figure 6.2: Probability density and cumulative distribution functions for the piecewise-linear function.

This probability density function is plotted in Fig. 6.2.

To obtain the cumulative distribution function, we consider each range. For the range with $x < -1$, the cumulative distribution function is zero. For the range between $-1 \leq x \leq 0$, we evaluate

$$F_X(x) = \int_{-1}^{x} \tfrac{2}{3}(1 + x')dx' = \frac{1}{3}(1 + 2x + x^2). \tag{6-32}$$

For the range between $0 < x \leq 2$, we must integrate the previous region up to 0 and the following region up to x:

$$\begin{aligned} F_X(x) &= \int_{-1}^{0} \tfrac{2}{3}(1 + x')dx' + \int_{-1}^{x} \tfrac{2}{3}(1 - \tfrac{x'}{2})dx' \\ &= \tfrac{1}{3}(1 + 2x - \tfrac{1}{2}x^2). \end{aligned} \tag{6-33}$$

We can verify that $F_X(2) = 1$, and since the probability density function is zero for

$x > 2$, the cumulative density function will remain one. Pulling this together we have

$$F_X(x) = \begin{cases} 0, & x < 2 \\ \frac{1}{3}(1 + 2x + x^2), & -1 \leq x \leq 0 \\ \frac{1}{3}(1 + 2x - \frac{1}{2}x^2), & 0 < x \leq 2 \\ 1, & x > 2 \end{cases} . \tag{6-34}$$

This is also plotted in Fig. 6.2.

6.4 Multivariate Distributions

In the previous sections we introduced discrete and continuous random variables in the context of a single random variable. It is common to have multiple random variables simultaneous that may or may not be independent. Most of the following discussion will focus on the case of two random variables, but the concepts are generalizable to any number.

6.4.a Probability Mass/Density Functions

We can generalize the probability mass function and probability density function to multiple variables in a relatively straightforward manner. The joint probability mass function for two random variables X and Y describes the joint probability that $X = x$ and $Y = y$. This is

$$f_{X,Y}(x, y) = P(X = x, Y = y). \tag{6-35}$$

If we are given the joint probability mass function want to compute the probability mass function of a single random variable X or Y, we sum over the other variable(s):

$$f_X(x) = \sum_y f_{X,Y}(x, y), \tag{6-36a}$$

$$f_Y(x) = \sum_x f_{X,Y}(x, y). \tag{6-36b}$$

If we have more than two, we sum over all the variables that we do not want information about. The single variable probability mass functions derived from a joint probability mass function are called the marginal probability mass function.

The probability density function for the two variable case is defined so that the probability that random variable X is between x and $x + dx$ and Y is between y and $y + dy$ is

$$f_{X,Y}(x, y)dxdy = P(x \leq X \leq x + dx, y \leq Y \leq y + dy). \tag{6-37}$$

Similar to the single variable case, the units of the probability density function are probability per the units of random variable X per the units of random variable Y. In a similar manner to the discrete case, we can compute the marginal probability density functions in X and Y alone by integrating out the variables that are not of interest:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy, \tag{6-38a}$$

$$f_Y(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx. \tag{6-38b}$$

The final point to bring up in this section is the case where random variables X and Y are independent. In this case we may write the joint probability mass or density function as the product of the marginals:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \quad \text{only if } X \text{ and } Y \text{ are independent.} \tag{6-39}$$

6.4.b Cumulative Distribution Functions

It may not be surprising that we can also generalize the cumulative distribution functions to multiple variables. In both cases this states that random variables X and Y are simultaneously at most x and y respectively. For the discrete random variable case, the cumulative distribution function for random variables X and Y is

$$F_{X,Y}(x,y) = \sum_{x_i<x}\sum_{y_i<y} f_{X,Y}(x_i,y_i) = P(X \leq x, Y \leq Y). \tag{6-40}$$

The cumulative distribution for the continuous case is

$$F_{X,Y}(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(x',y')dx'dy' = P(X \leq x, Y \leq Y). \tag{6-41}$$

6.4.c Conditional Distribution Functions

Earlier on we discussed the idea of conditional probability, which is the idea that the probability of a particular event outcome is influenced by the outcome of a different, but related random event. The notion of conditional probability also can be applied in the context of probability mass/density functions and their cumulative distribution function counterparts.

The conditional probability mass function for a discrete random variable is the probability random variable $X = x$ given that we know another random variable $Y = y$. This is written as

$$f_{X|Y}(x,y) = P(X = x|Y = y). \tag{6-42}$$

The cumulative distribution function is the probability that $X \leq x$ given that $Y = y$ and is computed by as follows:

$$F_{X|Y}(x,y) = \sum_{x_i<x} f_{X|Y}(x_i) = P(X \leq x|Y = y). \tag{6-43}$$

The conditional probability density function is defined similarly as before. The probability that random variable X is between x and $x + dx$ given that random variable $Y = y$ is

$$f_{X|Y}(x, y)dx = P(X = x|Y = y), \tag{6-44}$$

and the cumulative distribution function is

$$F_{X|Y}(x, y) = \int_{-\infty}^{x} f_{X|Y}(x')dx' = P(X \leq x|Y = y). \tag{6-45}$$

We can relate the conditional probability mass or density functions to the joint mass or density functions for either the discrete or continuous case:

$$f_{X,Y}(x, y) = f_{Y|X}(x, y)f_Y(x) = f_{X|Y}(x, y)f_Y(y). \tag{6-46}$$

6.4.d Example: Discrete Binary Distribution

Suppose we have two random variables X and Y that both take on the value of zero or one. The probability mass function is given by the function

$$f_{X,Y}(x, y) = C\frac{1 + 3x}{1 + xy}, \quad x \in \{0, 1\}, \quad y \in \{0, 1\}. \tag{6-47}$$

We wish to find the marginal probability mass functions.

First, we normalize the distribution function by taking

$$\begin{aligned}\sum_{x=0}^{1}\sum_{y=0}^{1} C\frac{1 + 2x}{1 + y} &= C\left(\frac{1 + 3(0)}{1 + (0)(0)} + \frac{1 + 3(0)}{1 + (0)(1)} + \frac{1 + 3(1)}{1 + (1)(0)} + \frac{1 + 3(1)}{1 + (1)(1)}\right) \\ &= C\left(1 + 1 + 4 + 2\right) = 8C = 1.\end{aligned} \tag{6-48}$$

Therefore $C = \frac{1}{8}$. Rewriting the probability mass function gives

$$f_{X,Y}(x, y) = \frac{1 + 3x}{8(1 + xy)}, \quad x \in \{0, 1\}, \quad y \in \{0, 1\}. \tag{6-49}$$

To find the marginal probability mass functions in x and y, we sum over y and x respectively. The marginal probability mass function in x is

$$\begin{aligned}f_X(x) = \sum_{y=0}^{1} f_{X,Y}(x, y) &= \frac{1 + 3x}{8(1 + 0)} + \frac{1 + 2x}{8(1 + x)} \\ &= \frac{(1 + 3x)(1 + x)}{8(1 + x)} + \frac{1 + 3x}{8(1 + x)} \\ &= \frac{3x^2 + 7x + 2}{8(1 + x)}.\end{aligned} \tag{6-50}$$

The marginal probability mass function in y is

$$\begin{aligned} f_Y(y) = \sum_{x=0}^{1} f_{X,Y}(x,y) &= \frac{1}{8(1+0)} + \frac{1+2}{8(1+y)} \\ &= \frac{1}{8}\left(1 + \frac{3}{1+y}\right) \\ &= \frac{4+y}{8(1+y)}. \end{aligned} \tag{6-51}$$

6.5 Random Variable Operators

In this chapter we will discuss the common operators used upon random variables, which are used to extract information about a random variable and its distribution. Here we will cover expectation (the mean value), variance (square of the standard deviation), and covariance (related to correlation).

6.5.a Expectation

The most important and commonly used operator for random variables is the expectation operator $E(\cdot)$. The expectation operator is often used in the definitions of other operators. The simplest form is to take the expectation of a random variable $E(X)$. The expectation of the discrete and continuous random variables are, respectively,

$$E(X) = \sum_x x f_X(x), \tag{6-52a}$$

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx. \tag{6-52b}$$

The expectation operator on a random variable produces a number that corresponds to its mean value and we therefore often colloquially call $E(X)$ as the average of random variable X. The mean describes the central tendency of the distribution, i.e., where is the probability of the distribution centered about. The concept is similar to the "center of mass" from physics.

An important point is that expectation does not necessarily mean the most probable value—although this is sometimes the case, it is often that it does not. For example, let us consider the simple example of a fair coin where we assign heads a value of 1 and tails a value of 0. Since the coin is fair, we have an equal probability of getting heads or tails. The expectation in this case is

$$E(X) = \sum_x x f_X(x) = (0)(\tfrac{1}{2}) + (1)(\tfrac{1}{2}) = \tfrac{1}{2}.$$

The expectation or mean value is $\frac{1}{2}$, which does not correspond to either heads or tails. The takeaway again is that expectation measures centrality of a distribution and not maximal likeliness.

The expectation operator is a linear operator and satisfies the following properties:

$$E(c) = c, \tag{6-53}$$
$$E(aX) = aE(X), \tag{6-54}$$
$$E(X + Y) = E(X) + E(Y). \tag{6-55}$$

Here a and c are non-random constants.

Additionally, the result of the expectation operator is related to the sample mean. If we collect N samples x_i from random variable X, we can estimate the mean using the sample mean

$$E(X) \approx \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{6-56}$$

The approximation becomes exact in the limit as $N \to \infty$, provided that the expectation exists. There are certain pathological cases (see the Cauchy distribution) where there is no mean value. In this case, the sample mean will never converge to any value.

We can extend the expactation operator to take what we call moments of a random variable. For example, the second moment is

$$E(X^2) = \sum_x x^2 f_X(x), \tag{6-57a}$$
$$E(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)dx, \tag{6-57b}$$

for discrete and continuous random variables respectively. The second moment is useful in defining the variance operator, which is covered in the next section.

We can further generalize the expectation to consider any function of a random variable. We define

$$E(g(X)) = \sum_x g(x) f_X(x), \tag{6-58a}$$
$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)dx, \tag{6-58b}$$

again for the respective discrete and continuous cases.

We can generalize the concept to multiple variables by stating that the expectation operator always returns a number. If we have a bivariate joint random distribution for X and Y, the expectation of the random variable X is

$$E(X) = \sum_x \sum_y x f_{X,Y}(x, y) = \sum_x x f_X(x), \tag{6-59a}$$
$$E(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x, y)dydx = \int_{-\infty}^{\infty} x f_X(x)dx. \tag{6-59b}$$

Furthermore, we take the expectation with any function of random variables $g(X, Y)$ using

$$E(g(X,Y)) = \sum_x \sum_y g(x,y) f_{X,Y}(x,y), \tag{6-60a}$$

$$E(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y) dy dx. \tag{6-60b}$$

We can get an estimate from the sample mean for this case if we have N samples of x_i and y_i using

$$E(g(X,Y)) \approx \frac{1}{N} \sum_{i=1}^{N} g(x_i, y_i), \tag{6-61}$$

where the approximation becomes exact in the limit as $N \to \infty$ provided the expectation exists.

A final important property is independence related to the term $E(XY)$. If X and Y are independent, then we can show that $E(XY) = E(X)E(Y)$. This is

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{X,Y}(x,y) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_X(x) f_Y(y) \\ &= \left(\int_{-\infty}^{\infty} x f_X(x) dx \right) \left(\int_{-\infty}^{\infty} y f_Y(y) dy \right) \\ &= E(X)E(Y), \quad \text{if } X \text{ and } Y \text{ are independent.} \end{aligned} \tag{6-62}$$

6.5.b Variance and Standard Deviation

The variance is the second most commonly encountered operator for random variables. The variance operator is defined as

$$\text{Var}(X) = E((X - E(X))^2). \tag{6-63}$$

We can expand this out into a more convenient form for calculations:

$$\begin{aligned} \text{Var}(X) &= E(X^2 - 2XE(X) + E(X)^2) \\ &= E(X^2) - 2E(XE(X)) + E(E(X)^2) \\ &= E(X^2) - 2E(X)E(X) + E(X)^2 \\ &= E(X^2) - E(X)^2. \end{aligned} \tag{6-64}$$

Using the variance we can compute the standard deviation

$$\sigma = \sqrt{\text{Var}(X)}. \tag{6-65}$$

The standard deviation is a measure of how much the distribution deviates from its mean and can be thought of as how "spread out" the probability is. The variance is a strictly positive quantity such that the larger the variance, the more dispersed a the probability of a distribution is relative to its mean.

The variance operator is nonlinear because of the squared term and does not satisfy the linearity properties. However, we can derive another relationship. Suppose we have independent random variables X and Y. We then can calculate

$$\begin{aligned}
\text{Var}(aX+bY) &= E((aX+bY)^2) - E(aX+bY)^2 \\
&= E(a^2X^2 + 2abXY + b^2Y^2) - E(aX+bY)E(aX+bY) \\
&= a^2E(X^2) + 2abE(XY) + b^2E(Y^2) \\
&\quad - [aE(X) + bE(Y)]\,[aE(X) + bE(Y)] \\
&= a^2E(X^2) + 2abE(X)E(Y) + b^2E(Y^2) \\
&\quad - a^2E(X)^2 - b^2E(X)^2 - 2abE(X)E(Y) \\
&= a^2\left[E(X^2) - E(X)^2\right] + b^2\left[E(Y^2) - E(Y)^2\right] \\
&= a^2\text{Var}(X) + b^2\text{Var}(Y).
\end{aligned} \tag{6-66}$$

6.5.c Covariance and Correlation

We can also define the covariance between two random variables X and Y. The covariance measures how much two random variables change with respect to one another about their respective means. It has the definition of

$$\text{Cov}(X,Y) = E[(X - E(X))(Y - E(Y))]. \tag{6-67}$$

This can be expanded out to give the more convenient form for calculation of

$$\text{Cov}(X,Y) = E(XY) - E(X)E(Y). \tag{6-68}$$

From the definition, we can see that the covariance of X with itself is simply the variance,

$$\text{Cov}(X,X) = E[(X - E(X))(X - E(X))] = \text{Var}(X). \tag{6-69}$$

The covariance is also symmetric:

$$\text{Cov}(X,Y) = \text{Cov}(Y,X) \tag{6-70}$$

If we have multiple random variables, we can define a covariance matrix that contains information about how each random variable varies with respect to the others

$$\boldsymbol{\Sigma} = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1,X_2) & \cdots & \text{Cov}(X_1,X_N) \\ \text{Cov}(X_2,X_1) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2,X_N) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_N,X_1) & \text{Cov}(X_N,X_2) & \cdots & \text{Var}(X_N) \end{bmatrix}. \tag{6-71}$$

This covariance matrix is useful when we discuss the propagation of uncertainties.

Unlike the variance, the covariance may be either positive, zero, or negative. To get an interpretation, let us introduce the Pearson correlation coefficient

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}. \tag{6-72}$$

This is the covariance of random variables X and Y divided by the product of their standard deviations. The Pearson correlation coefficient is a scaled covariance that ranges from -1 to 1 and measures linear dependence. Positive correlation implies that if we take a sample of random variables X and Y and if our sample x is high (or low) with respect to its mean, then a corresponding sampled value y will also be high (or low) with respect to its mean. Likewise, if the correlation is negative a low sample of x tends to mean a high sample of y.

If the correlation coefficient is zero, then the two random variables are said to be uncorrelated or linearly independent. A common misconception is that a zero correlation or linear independence implies that two random variables are actually independent. The correlation coefficient only measures the linear component of the dependency between two random variables. If, for example, two random variables are only dependent in a purely quadratic nature, the correlation would be zero even though the two random variables are statistically dependent upon each other.

6.6 Discrete Distributions

In this section, we will discuss several common distributions for discrete random variables that are commonly encountered in science and engineering applications.

6.6.a Bernoulli Distribution

The simplest type of discrete random variable is a Bernoulli distribution. The distribution describes a "coin flip". We have a success with probability p that we assign a value of 1, and a failure with probability $1-p$ that we assign a value of 0. This has the probability mass function of

$$f_X(x) = \begin{cases} 1-p, & x = 0 \\ p, & x = 1 \\ 0, & \text{otherwise} \end{cases}. \tag{6-73}$$

The cumulative distribution function is

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1-p, & 0 \le x < 1 \\ 1, & x \ge 1 \end{cases}. \tag{6-74}$$

6.6.b Discrete Uniform Distribution

Another common distribution is the discrete uniform distribution, which represents the n-sided die where each face is equiprobable. Let us enumerate the faces $x = 1, 2, \ldots, n$. The probability mass function is

$$f_X(x) = \begin{cases} \frac{1}{n}, & x = 1, 2, \ldots, n \\ 0, & \text{otherwise} \end{cases} . \tag{6-75}$$

The cumulative distribution function is

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{n}, & 0 \leq x < 1 \\ \frac{2}{n}, & 1 \leq x < 2 \\ \vdots & \\ \frac{n-1}{n}, & n-1 \leq x < n \\ 1, & x \geq n \end{cases} . \tag{6-76}$$

6.6.c Binomial Distribution

The binomial distribution describes the number of successful independent trials n of a Bernoulli distribution with probability of success p. The probability mass function is

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, \ldots, n \tag{6-77}$$

and zero otherwise. The binomial coefficient may be written in terms of the factorial:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \tag{6-78}$$

The binomial coefficient gives the number of unique combinations of x members from a set of n elements. Unfortunately, the cumulative distribution function does not have a form in terms of standard functions and is simply:

$$F_X(x) = \sum_{k=1}^{x} \binom{n}{k} p^k (1-p)^{n-k}. \tag{6-79}$$

6.6.d Example: Determining Number of Experimental Trials

Suppose we are designing a high-energy density physics experiment that requires collecting data from a laser facility. To collect enough data, we need to ensure we have at least 10 successful shots on the laser facility. We know from historical data of the facility that the success rate of any individual shot is 90%. We desire a probability

of 95% of having at least 10 successful shots. As part of planning, we need to allocate resources to have a number of shots that gives us this level of confidence that we will meet our objective. Because we have limited resources to purchase time at the laser facility, we want to have as few shots as possible to achieve our desired confidence level.

The appropriate distribution for analyzing this problem is the binomial distribution with $p = 0.9$. To get the number, we know we need at least 10 shots total, so let's start by finding the probability that all 10 shots are successful:

$$\begin{aligned} P(X \geq 10; n = 10) &= f_X(10) \\ &= \binom{10}{10} 0.9^{10}(1 - 0.9)^{10-10} \\ &= \frac{10!}{10! \times 0!} \times 0.9^{10} \times 0.1^0 = 0.349. \end{aligned}$$

Clearly, just scheduling ten shots would be quite risky and is well below our desired confidence level.

Let's try 11 shots. The probability of having at least 10 successful shots out of 11 is

$$\begin{aligned} P(X \geq 10; n = 11) &= P(X = 10) + P(X = 11) \\ &= f_X(10) + f_X(11) \\ &= \binom{11}{10} 0.9^{10}(1 - 0.9)^{11-10} + \binom{11}{11} 0.9^1 1(1 - 0.9)^{11-11} \\ &= 0.3835 + 0.3138 = 0.6974. \end{aligned} \tag{6-80}$$

Clearly doing 11 shots improves our chances significantly, but this is still well below our desired confidence level. Let's try 12:

$$\begin{aligned} P(X \geq 10; n = 12) &= P(X = 10) + P(X = 11) + P(X = 12) \\ &= f_X(10) + f_X(11) + f_X(12) \\ &= 0.2301 + 0.3766 + 0.2824 = 0.8891. \end{aligned} \tag{6-81}$$

Getting better, but not quite there yet. Let's schedule a lucky 13 shots:

$$\begin{aligned} P(X \geq 10; n = 13) &= P(X = 10) + P(X = 11) + P(X = 12) + P(X = 13) \\ &= f_X(10) + f_X(11) + f_X(12) + f_X(13) \\ &= 0.0997 + 0.2448 + 0.3672 + 0.2542 = 0.9658. \end{aligned} \tag{6-82}$$

Therefore 13 shots achieves our stated objective of having a confidence of at least 95% of having at least 10 successes.

6.6.e Geometric Distribution

The geometric distribution can be used to describe the number of successes x before a failure if the failure probability is p. For example, we can use the distribution to

model the case where we wish to know the number of times a piece of equipment with failure probability p will operate before requiring maintenance. The probability mass function for the geometric distribution is

$$f_X(x) = p(1-p)^x, \quad x \geq 0, \tag{6-83}$$

and zero otherwise. The probability mass function gives the exact number of successes before failure. The cumulative distribution function is

$$F_X(x) = 1 - (1-p)^{x+1}, \quad x \geq 0. \tag{6-84}$$

This is we will have at most x successes before failure. Note that unlike the other distributions we have encountered, the geometric distribution has an infinite support. In other words, when we use the geometric distribution we do not restrict the number of successes that can occur before a failure.

6.6.f Example: Machine Failure Probability

Suppose the failure probability for a piece of equipment is 0.1%. If the machine is used everyday, we wish to know the probability that the machine will operator for at least a year, which we take to have 365 days. We let random variable X be the number of days the machine operates successfully before failure. We have

$$\begin{aligned} P(X \geq 365) &= 1 - P(X \leq 364) \\ &= 1 - F_X(364) \\ &= 1 - \left[1 - (1-0.001)^{364+1}\right] = 0.999^{365} = 0.694. \end{aligned} \tag{6-85}$$

6.6.g Poisson Distribution

The Poisson distribution can be used to describe the number of radioactive decays during a time t if the sample has a decay constant λ, which is the expected number of counts per unit time. The probability mass function of the Poisson distribution is

$$f_X(x) = \frac{(\lambda t)^x}{x!} e^{-\lambda t}, \quad x \geq 0, \tag{6-86}$$

and zero otherwise. Unfortunately, like the binomial distribution, the Poisson distribution does not have a cumulative distribution function in a closed form.

6.6.h Example: Detecting Radioactive Contamination

Suppose we have soil that we suspect may be contaminated with a radioactive isotope that does not occur in the environment naturally. The minimum activity of concern is 1 μBq/g $= 10^{-6}$. If we take a 10 g sample and count the radioactivity for 48 hours and get no counts indicating the specific radioisotope, what is the probability that we got no counts by chance alone assuming we have perfect detection equipment?

To find this we use the Poisson distribution at $x = 0$. The total activity of the 10 g sample that would be of minimum concern is 10^{-5} Bq. We therefore have

$$f_X(0) = \frac{(1.0 \times 10^{-6} \times 10 \times 3600 \times 48)^0}{0!} e^{-(1.0\times10^{-6}\times10\times3600\times48)} = 0.178. \tag{6-87}$$

This says that even counting for two entire days, we have an 17.8% chance that a sample with 1 μBq/g of activity did not emit any decays and our test would produce a false negative. Since this probability is still rather high, it would be advisable to count for a longer period of time or, if practical, to take a larger sample to establish or rule out contamination.

6.7 Continuous Distributions

6.7.a Uniform Distribution

The simplest of the continuous distributions is the uniform distribution, which is constant between $a \leq x \leq b$ and zero elsewhere. The probability density function for the uniform distribution is

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases} . \tag{6-88}$$

The cumulative distribution function is

$$F_X(x) = \begin{cases} 0, & x < b \\ \dfrac{x-a}{b-a}, & a \leq x \leq b \\ 1, & x > b \end{cases} . \tag{6-89}$$

6.7.b Exponential Distribution

A common distribution that arises in applications such as radioactive decay and the interaction of radiation with matter is the exponential distribution. The probability density function is

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases} , \tag{6-90}$$

where λ is a rate parameter, which, in the context of radioactive decay corresponds to the expected frequency of the decay. The cumulative distribution function is

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases} . \tag{6-91}$$

6.7.c Normal Distribution

Perhaps the most common distribution in all of probability is the normal or Gaussian distribution. The collective behavior of large populations, whether they be velocities of gas molecules, lifetimes of equipment, or grades for students in a class, tend to settle toward normally distributed behaviors. More precisely, the sum of numerous independent and identically distributed random events tends toward a normal distribution by way of the Central Limit Theorem.

The normal distribution contains information about the mean μ and standard deviation σ, which give the centrality and width of the distribution. The normal distribution has the probability density function

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]. \tag{6-92}$$

The cumulative distribution function is

$$F_X(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right]. \tag{6-93}$$

The function $\text{erf}(\cdot)$ is a special function that is, perhaps inappropriately, named the *error function*. The error function is available in most mathematical software and is determined by numerically tabulating the following integral:

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{-\infty}^{x} e^{-t^2} dt. \tag{6-94}$$

6.7.d Multivariate Normal Distribution

Perhaps the most common multivariate random distribution is the multivariate normal distribution. Here we have a set of random variables $\{X_1, X_2, \ldots, X_N\}$. These random variables have individual means described by a column vector

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{bmatrix}, \tag{6-95}$$

which describes the centrality of each random variable in each independent coordinate direction. The spread of each random variable is given by a covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1N}\sigma_1\sigma_N \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2N}\sigma_2\sigma_N \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1N}\sigma_1\sigma_N & \rho_{2N}\sigma_2\sigma_N & \cdots & \sigma_N^2 \end{bmatrix}. \tag{6-96}$$

The diagonal terms represent the variance of each element σ_i^2 and the off-diagonal terms are the covariance of each element involving the correlation coefficient ρ_{ij}. Since

the covariance is a symmetric operator, the covariance matrix is also symmetric. Given these the multivariate normal has a probability density function of

$$f_{X_1,X_2,\ldots,X_N}(x_1,x_2,\cdots,x_N)=\frac{1}{\sqrt{(2\pi)^N\det(\boldsymbol{\Sigma})}}\exp\left[-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^\top\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right]. \tag{6-97}$$

A simpler and common case encountered is the bivariate normal distribution involving means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and the correlation coefficient ρ. For this case, we can expand out Eq. (6-97) to obtain

$$\begin{aligned}f_{X_1,X_2}(x_1,x_2)&=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\\&\times\exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}-\frac{2(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right)\right].\end{aligned} \tag{6-98}$$

6.7.e Log-Normal Distribution

An important continuous random distribution found in many engineering applications is the log-normal distribution, which occurs when the logarithm of a random variable is normally distributed. Just as the normal distribution arises when we have the sum of a large number of independent and identically distributed events, the log-normal distribution arises when we have the product of random variables that are strictly positive. The advantage of the log-normal distribution is that the result is strictly positive and tends to arise when modeling phenomenon such as failure times of complex equipment.

The log-normal distribution has the probability density function of

$$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma x}\exp\left[-\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right],\quad x\geq 0, \tag{6-99}$$

and zero for $x<0$, and the cumulative distribution function

$$F_X(x)=\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{\ln(x)-\mu}{\sqrt{2}\sigma}\right)\right],\quad x\geq 0, \tag{6-100}$$

and also zero for $x<0$.

6.8 Fundamental Theorems of Probability

There are two fundamental results from probability theory that are fundamental to the analysis of sequences of random variables. These are the law of large numbers and the central limit theorem. It is these two results that allow us to make conclusions about averages and standard deviations. This is especially important when analyzing results of experiments where we take repeated measurements, which have inherent randomness.

6.8.a Law of Large Numbers

The law of large numbers states that if we take a sum of identically-distributed (possibly dependent) random variables, the sample mean will converge to the expectation of the random variable given that the expectation exists. In other words, if we have random variables $X_1, X_2, \ldots, X_N$ that all have the same underlying random distribution, the law of large numbers states

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} X_i = E(X). \tag{6-101}$$

The law of large numbers tells us that even under very modest restrictions (finite expectation) then the sample mean will converge to the expectation.

6.8.b Central Limit Theorem

To understand the convergence rate or how to form confidence intervals of how well the sample mean estimates the expectation, we require the use of the central limit theorem. The central limit theorem states that if we have a sequence of random variables that are: (1) identically-distributed, (2) independent, (3) have a finite expectation, and (4) has a variance that is finite; then, we can assert that as the number of random variables N becomes large, then the (unknown) expectation of the random variable is itself randomly distributed about its sample mean as a normal distribution with a variance given by the sample variance of the sample mean.

To unpack this statement, the normal distribution again is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

The sample mean that we use is given by

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{6-102}$$

where x_i is the result of each sample. Now we need to calculate the sample variance of the sample mean. This is

$$\begin{aligned}
\mathrm{Var}(\overline{X}) &= \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}(X_i) \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}(X) \\
&= \frac{1}{N}\mathrm{Var}(X) = \frac{1}{N}s_x^2.
\end{aligned} \tag{6-103}$$

Here s_x^2 is the sample variance measured from the population. The population sample variance can be estimated by

$$s_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2. \tag{6-104}$$

The factor of $N-1$ as opposed to N is a result of the fact that the we are using the same set of samples to estimate the sample mean and the sample variance, and to create an unbiased estimate for the sample variance, we must remove one sample. Therefore, the variance of the sample mean is

$$\sigma^2 = \frac{1}{N}s_x^2. \tag{6-105}$$

The implication of this is that our estimate of the expectation becomes better as we take more samples, however, the rate that the standard deviation of the sample mean, which is a measure of our uncertainty, decreases, but does so slowly at the rate of $1/\sqrt{N}$.

6.9 Transformations of Random Variables

We often need to transform random variables from one coordinate system to another. Just as it is often convenient to work science or engineering problems in different coordinate systems, we do the same with random variables and probability distributions. Furthermore, we often have combinations of random variables, which themselves are transformations. We discuss both cases in this section. While we discussion discrete random variables, here we focus most of our attention upon continuous random variables, since they have some subtleties that need to be addressed.

6.9.a Univariate Transformations

For the discrete random variable case, we can equate the probability mass or cumulative distribution functions. We say that if we have discrete random variable Z related by some coordinate transformation upon X, we have for both the probability mass and cumulative distribution function.

$$f_Z(z) = f_X(x), \tag{6-106}$$
$$F_Z(z) = F_X(x), \tag{6-107}$$

The first of these results only applies to discrete random variables, whereas the relationship cumulative distribution functions are the same regardless of whether the random variable is discrete or continuous.

To begin the discussion of continuous random variables we introduce the principle that the amount of probability in some differential unit of measurement is the same regardless of the coordinate system. Therefore, we can only equate probabilities and

not probability densities, which have units of probability per some unit coordinate. Recall that the probability density is defined in terms of a probability in some differential dx about some x. Therefore, it stands to reason that we equate the probabilities as

$$f_X(x)|dx| = f_Z(z)|dz|.$$

Note that we have placed absolute values around the differentials by convention. Coordinate transformations are not guaranteed to preserve the sign, and in these cases the integration limits need to be reversed so as to always integrate from a smaller to a larger number. Rather than attempting to track the directionality of all dimensions over all coordinate transformations, the convention of including of the absolute value ensures that we do not have to keep track of this information. Therefore, if we know the probability density function for X, we can find the probability density function for the transformed variable Z using

$$f_Z(z) = f_X(x)\left|\frac{dx}{dz}\right|. \tag{6-108}$$

As mentioned previously, the cumulative distribution function for the continuous random variable under a coordinate transformation is the same as the discrete case. This is because the cumulative distribution function is a probability.

6.9.b Example: Neutron Lethargy

An important application of random variable transformations are in neutron scattering. The most important neutron scattering interaction in analyzing the behavior of a nuclear reactor is the elastic scattering process, which governs the process of neutron moderation. The elastic scattering process begins with a neutron of kinetic energy E scattering off a nucleus and emerging after the collision with a random kinetic energy x, which is given by a uniform distribution:

$$f(E' \to E) = \begin{cases} \dfrac{1}{1-\alpha}\dfrac{1}{E'}, & \alpha E' \leq E \leq E' \\ 0, & \text{otherwise} \end{cases}. \tag{6-109}$$

Here α is a scattering parameter

$$\alpha = \frac{A-1}{A+1} \tag{6-110}$$

with A as the ratio of the target mass to the neutron mass. (Please excuse the change in notation here; but it is consistent to how it is used in reactor analysis.) Note that this assumes the thermal motion of the background atoms is negligible. A significantly more complicated analysis needs to be performed when the neutron kinetic energy is on the same order as the kinetic energy of the atoms in the material.

We can apply this to probability density function to understand the moderation of neutrons in a nuclear fission reactor by calculating the mean loss in kinetic energy

$$\begin{aligned}
-\overline{\Delta E} &= -\int_{\alpha E'}^{E'} (E - E') \frac{1}{1-\alpha} \frac{1}{E'} dE \\
&= -\frac{1}{1-\alpha} \frac{1}{E'} \int_{\alpha E'}^{E'} (E - E') dE \\
&= -\frac{1}{1-\alpha} \frac{1}{E'} \left(\frac{(E')^2 - \alpha^2 (E')^2}{2} - E'(E' - \alpha E') \right) \\
&= \frac{1}{1-\alpha} \left(\frac{\alpha^2 - 2\alpha + 1}{2} \right) E' \\
&= \left(\frac{1-\alpha}{2} \right) E'. \qquad (6\text{-}111)
\end{aligned}$$

The mean energy loss in a single collision, unfortunately, depends upon the initial energy. This makes calculating the average change in energy from two, three, or an arbitrary number of collisions very difficult.

Since the energy loss is multiplicative, i.e., the expected energy loss is proportional to the incident energy, we can get around this by introducing a transformed dimensionless energy coordinate called neutron lethargy

$$u = -\ln \left(\frac{E}{E_0} \right). \qquad (6\text{-}112)$$

Here E_0 denotes the maximum plausible neutron kinetic energy in the system, which is usually assigned a value of 10 or 20 MeV. The lethargy coordinate is such that when a neutron has the maximum plausible kinetic energy, it has zero lethargy, and when a neutron has no energy, it has infinite lethargy.

To find the probability density function for the change in lethargy we use the relationship

$$f(u' \to u) = f(E' \to E) \left| \frac{dE}{du} \right|. \qquad (6\text{-}113)$$

Here u' is the lethargy corresponding to the neutron's kinetic energy prior to the collision.

To find the transformation, we solve for the energy in terms of lethargy

$$E = E_0 e^{-u}. \qquad (6\text{-}114)$$

Taking the derivative yields

$$\frac{dE}{du} = -E_0 e^{-u} = -E.$$

Therefore, we have

$$f(u' \to u) = E f(E' \to E).$$

Before writing out the distribution, we must compute the limits in terms of lethargy. If no energy is lost, then zero lethargy is gained and the initial and final lethargies are equal. We can also show that the maximum change in lethargy is $\ln(1/\alpha)$, therefore, the probability density function for the lethargy change becomes

$$f(u' \to u) = \begin{cases} \dfrac{e^{-(u-u')}}{1-\alpha}, & u' \leq u \leq u' + \ln(1/\alpha) \\ 0, & \text{otherwise} \end{cases} . \tag{6-115}$$

We can now calculate the mean lethargy gain per collision. This is

$$\begin{aligned} \xi = \overline{\Delta u} &= \int_{u'}^{u'+\ln(1/\alpha)} (u-u') \frac{e^{-(u-u')}}{1-\alpha} du \\ &= 1 + \frac{\alpha \ln \alpha}{1-\alpha}. \end{aligned} \tag{6-116}$$

While this result appears much more complicated, it has the attractive property that it does not depend upon the incident lethargy. In other words, the mean change in lethargy is independent of the lethargy the neutron had prior to the collision. This means the mean lethargy gain in two collisions is 2ξ, for three collisions, this is 3ξ, and so on. Therefore, if we want to understand the mean number of collisions N to gain a lethargy ΔU, corresponding to the amount of energy a typical fission neutron needs to lose to become a thermal neutron, we simply have

$$N = \frac{\Delta U}{\xi}. \tag{6-117}$$

This expression is frequently used to analyze the ability of moderators to slow down neutrons in nuclear fission reactors.

6.9.c Multivariate Transformations

It is sometimes the case where we have possibly correlated random variables X and Y, and we wish to rewrite them as random variables U and V after a coordinate transformation. To calculate the new probability density function, we make use of the determinant of the Jacobian matrix. For the two variable case this is

$$f_{U,V}(u,v) = f_{X,Y}(x,y) \left| \frac{\partial(x,y)}{\partial(u,v)} \right|, \tag{6-118}$$

where again the Jacobian determinant is

$$\frac{\partial(x,y)}{\partial(u,v)} = \begin{vmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \end{vmatrix}. \tag{6-119}$$

The generalization to three or more random variables, say N involves taking an $N \times N$ determinant.

6.9.d Sums of Random Variables

It is common that we wish take the sum of two or more random variables. The probability distributions of a sum of random variables is much more complicated than taking the sum of random distributions—one would end up with invalid distributions. To do this properly, we use transformation rules.

Let us suppose a random variable Z is the sum of two other random variables X and Y, which may be dependent, i.e., $Z = X + Y$. For the discrete random variable case, the joint probability mass function is $f_{X,Y}(x, y)$. If we solve for one of the variables,$y = z - x$ or $x = z - y$, we can write

$$f_{X,Z}(x, z) = f_{X,Y}(x, z - x) = f_{X,Y}(z - y, y).$$

To find the probability mass function for random variable Z we compute the marginal probability mass function by summing over the other variable:

$$f_Z(z) = \sum_x f_{X,Y}(x, z - x) = \sum_y f_{X,Y}(z - y, y), \tag{6-120}$$

taking whichever sum is most convenient.

For the continuous random variable case, we have to be more careful when we transform random variables, since we are working with probability densities as opposed to probabilities. As with the discrete case, we let $Z = X + Y$ and then solve for one of the variables and relate

$$f_{X,Z}(x, z)|dz| = f_{X,Y}(x, z - x)|dy| = f_{X,Y}(z - y, y)|dx|.$$

We can then divide by the differential $|dz|$ and then integrate over the other variable to obtain the marginal density function in z. This is

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z - x) \left|\frac{dy}{dz}\right| dx = \int_{-\infty}^{\infty} f_{X,Y}(z - y, y) \left|\frac{dx}{dz}\right| dy. \tag{6-121}$$

The derivation can be extended to include the difference of two random variables as well. This will be shown for a continuous random variable in the example in the next section.

6.9.e Example: Probability for Time Between Outages

Suppose we have two nuclear fission reactors operating. The times between unplanned outages are given by random variables X and Y and are exponentially distributed as follows:

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0, \tag{6-122a}$$
$$f_Y(y) = \mu e^{-\mu y}, \quad y \geq 0, \tag{6-122b}$$

and zero for $x < 0$ or $y < 0$. We assume the two nuclear reactors operate completely independently of each other so that the operating state of one does not impact the

other. Suppose the time to bring a reactor up to operation is a fixed constant time τ. We wish to find the probability that the two nuclear reactors will be simultaneously under repair.

To solve this problem, we define a random variable $Z = X - Y$, the difference between the two outage times. For two reactors to be under repair at the same time, we have

$$P(-\tau \leq Z \leq \tau),$$

with the positive and negative cases depending on which of the two reactors enter an unplanned outage first. To obtain this probability, we obtain the probability density function and then the cumulative density function.

Since $z = x - y$, we solve for $x = z + y$ and write the joint probability density function:

$$\begin{aligned} f_{Z,Y}(z,y) &= f_{X,Y}(z+y,y)\left|\frac{dx}{dz}\right| \\ &= (\lambda e^{-\lambda(z+y)})(\mu e^{-\mu y})|1| \\ &= \lambda\mu e^{-\lambda z}e^{-(\lambda+\mu)y}. \end{aligned} \tag{6-123}$$

We then integrate over y to obtain the marginal density function for z; however, we must be careful with the limits of integration. It is often useful to make a plot to show the domain of the joint density function. This is given in Fig. 6.3 with y and z being on the horizontal and vertical axes respectively. To illustrate, let's try a few numbers. When $y = 0$, z can be any number greater than or equal to zero, since $x \geq 0$. When $y = 1$, z has a lower bound of -1 when $x = 0$ and is otherwise unbounded in positive z. Therefore, we conclude that z is limited by the line $z = -y$. Therefore, $y \geq 0$ and $z \geq -y$.

Since we are integrating over y and the lower limit changes depending on whether $z < 0$ or $z \geq 0$, we break up the integral into two parts. For $z < 0$, the blue shaded are in Fig. 6.3, the integral for the marginal density function is

$$f_Z(z) = \int_{-z}^{\infty} \lambda\mu e^{-\lambda z}e^{-(\lambda+\mu)y}dy = \frac{\lambda\mu}{\lambda+\mu}e^{\mu z}, \quad z < 0. \tag{6-124}$$

For $z \geq 0$, the red shaded are in Fig. 6.3, we have

$$f_Z(z) = \int_{0}^{\infty} \lambda\mu e^{-\lambda z}e^{-(\lambda+\mu)y}dy = \frac{\lambda\mu}{\lambda+\mu}e^{-\lambda z}, \quad z \geq 0. \tag{6-125}$$

Pulling these two together, the probability density function is therefore given by the piecewise function

$$f_Z(z) = \begin{cases} \dfrac{\lambda\mu}{\lambda+\mu}e^{\mu z}, & z < 0 \\ \dfrac{\lambda\mu}{\lambda+\mu}e^{-\lambda z}, & z \geq 0 \end{cases}. \tag{6-126}$$

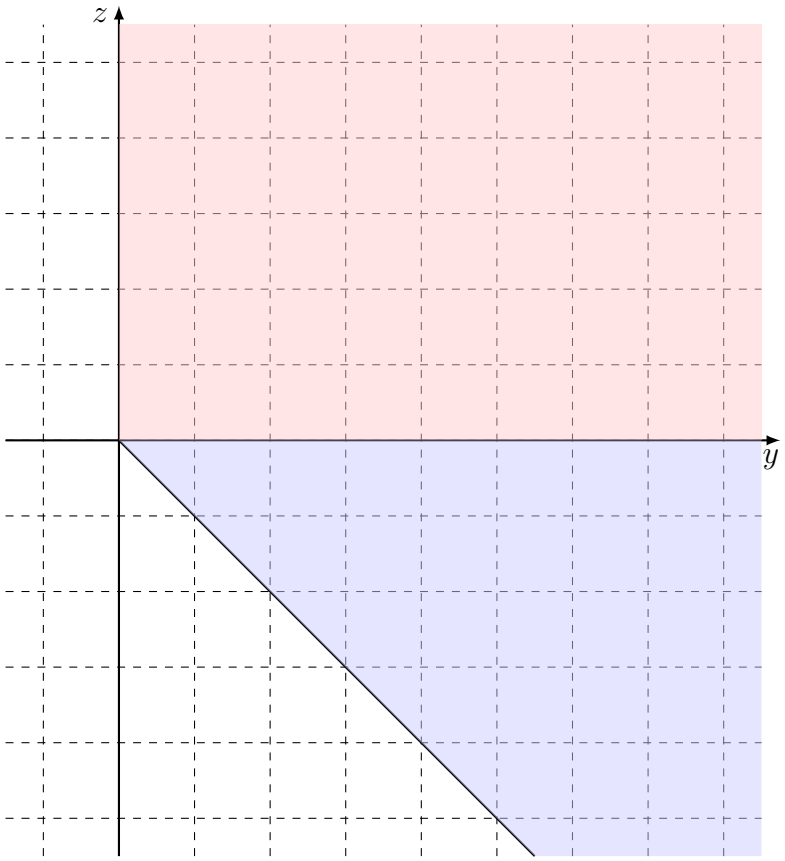

Figure 6.3: Domain of integration for the joint density function for the difference of two exponential distributions.

To find our desired probability, we integrate this over z from $-\tau$ to τ:

$$\begin{aligned} P(-\tau \leq Z \leq \tau) &= \int_{-\tau}^{0} \frac{\lambda\mu}{\lambda+\mu} e^{\mu z} dz + \int_{0}^{\tau} \frac{\lambda\mu}{\lambda+\mu} e^{-\lambda z} dz \\ &= \frac{\lambda}{\lambda+\mu}(1 - e^{-\mu\tau}) + \frac{\mu}{\lambda+\mu}(1 - e^{-\lambda\tau}) \\ &= 1 - \frac{\lambda e^{-\mu\tau} + \mu e^{-\lambda\tau}}{\lambda+\mu}. \end{aligned} \tag{6-127}$$

6.9.f Products of Random Variables

Another common operation for random variables is to take their product. Let $Z = XY$ where X and Y are (possibly dependent) random variables having joint probability density function $f_{X,Y}(x, y)$. As we did with the sum of random variables, we perform a transformation on either X or Y. If we perform the let $y = z/x$, we have

$$\frac{dy}{dz} = -\frac{1}{x};$$

therefore the joint density function for X and Z becomes

$$f_{X,Z}(x,z) = f_{X,Y}(x,y)\left|\frac{dy}{dz}\right| = f_{X,Y}(x,z/x)\frac{1}{|x|} \tag{6-128}$$

The marginal density function in the product Z may be obtained by integrating the joint density function over x:

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x,z/x)\frac{dx}{|x|}. \tag{6-129}$$

6.10 Error Propagation

In performing experiments and collecting data, we often take the raw data and perform analysis by evaluating functions $g(X_1, X_2, \ldots)$. Unfortunately, there is always inherent randomness because of factors that are outside the control of the experimenters. We often do not have a detailed knowledge of how these variations in data from experimental results are distributed. Alternatively, we may have an understanding of the random distributions, but it is too difficult to calculate an exact random distribution for an arbitrary function $g(X_1, X_2, \ldots)$. For these cases, we approximate the mean and variance using error propagation formulas, which are based on Taylor series approximations. This section discusses the first-order Taylor series approximation and derives the error propagation formulas.

6.10.a First-Order Taylor Approximation

To perform this analysis, let us consider the case of two random variables X and Y with a known measured sample means $\overline{x}$ and $\overline{y}$ respectively, sample variances s_x^2 and s_y^2 respectively, and correlation coefficient ρ_{xy}. The analysis will extend to any number of random variables, but using two simplifies the process. Now we introduce a new random variable $Z = g(X,Y)$ where we assume g is a differentiable function that allows us to take a Taylor series expansion.

To begin, we perform a first-order Taylor series expansion of the random variable Z about the sample means of X and Y, $\overline{x}$ and $\overline{y}$ respectively:

$$z = g(x,y) \approx g(\overline{x},\overline{y}) + (x-\overline{x})\frac{\partial g}{\partial x} + (y-\overline{y})\frac{\partial g}{\partial y}. \tag{6-130}$$

Here the partial derivatives are taken to be evaluated at $(\overline{x},\overline{y})$. We then take the expectation of random variable Z:

$$E(Z) = \overline{z} \approx E(g(\overline{x},\overline{y})) + \frac{\partial g}{\partial x}(E(X)-\overline{x}) + \frac{\partial g}{\partial y}(E(Y)-\overline{y}). \tag{6-131}$$

The first term is a non-random variable, so the expectation is itself. Our best estimate of the expectations are the sample means, $E(X) = \overline{x}$ and $E(Y) = \overline{y}$. Therefore, the second and third terms vanish, leaving us with the simple result:

$$E(Z) = \overline{z} \approx g(\overline{x},\overline{y}). \tag{6-132}$$

To compute the variance, let us take the difference $Z - E(Z)$,

$$Z - E(Z) = z - \overline{z} = g(\overline{x}, \overline{y}) + (x - \overline{x})\frac{\partial g}{\partial x} + (y - \overline{y})\frac{\partial g}{\partial y} - g(\overline{x}, \overline{y}) \tag{6-133}$$

The first and fourth terms cancel. Squaring the remaining two terms gives

$$\begin{aligned}(Z - E(Z))^2 &= (z - \overline{z})^2 \\ &= (x - \overline{x})^2 \left(\frac{\partial g}{\partial x}\right)^2 + (y - \overline{y})^2 \left(\frac{\partial g}{\partial y}\right)^2 \\ &\quad + 2(x - \overline{x})(y - \overline{y}) \left(\frac{\partial g}{\partial x}\right)\left(\frac{\partial g}{\partial y}\right).\end{aligned} \tag{6-134}$$

Taking the expectation gives

$$\begin{aligned}E[(Z - E(Z))^2] = s_z^2 = E[(x - \overline{x})^2] \left(\frac{\partial g}{\partial x}\right)^2 + E[(y - \overline{y})^2] \left(\frac{\partial g}{\partial y}\right)^2 \\ + 2E[(x - \overline{x})(y - \overline{y})] \left(\frac{\partial g}{\partial x}\right)\left(\frac{\partial g}{\partial y}\right).\end{aligned}$$

For the first and second terms, we recognize the expectations as the variances of random variables X and Y respectively, for which our best estimates are the respective sample variances s_x^2 and s_y^2. The third term is the covariance, which can be put in terms of the correlation coefficient as $\rho_{xy} s_x s_y$. This gives the first-order error propagation formula for two random variables:

$$s_z^2 = s_x^2 \left(\frac{\partial g}{\partial x}\right)^2 + s_y^2 \left(\frac{\partial g}{\partial y}\right)^2 + 2\rho_{xy} s_x s_y \left(\frac{\partial g}{\partial x}\right)\left(\frac{\partial g}{\partial y}\right). \tag{6-135}$$

This result can be generalized to any number of random variables:

$$s_z^2 = \sum_{i=1}^{N} s_{x_i}^2 \left(\frac{\partial g}{\partial x_i}\right)^2 + 2\sum_{j>i}^{N}\sum_{i=1}^{N} \rho_{ij} s_{x_i} s_{x_j} \left(\frac{\partial g}{\partial x_i}\right)\left(\frac{\partial g}{\partial x_j}\right). \tag{6-136}$$

Equivalently if we have the sample covariance matrix $\mathbf{\Sigma}$ and the partial derivatives encoded in column vector $\mathbf{d}$ we can write this in matrix-vector form:

$$s_z^2 = \mathbf{d}^\top \mathbf{\Sigma} \mathbf{d}. \tag{6-137}$$

This is often referred to as the *sandwich rule* because it encloses the covariance matrix between two vectors of partial derivatives.

The first-order Taylor series is usually sufficient for most engineering applications with one major assumption. This assumption is that the function g is sufficiently flat in the range of a few standard deviations that we can approximate it as a linear function. If g is a rapidly varying function, a special case where the mean is at or near a critical point (where the derivatives would evaluate to zero), or if the variance is too large to make this assumption, then the first-order approximation is not particularly good and we may need to consider higher-order Taylor series expansions. The cost of doing this leads to very complicated systems of equations involving higher-order moments and correlations of the random variables, which may not be easy to measure in a statistically significant manner.

6.10.b Example: Attenuation of a Beam

The equation for the attenuation of a beam without scattering is given as

$$I(x) = I_0 e^{-N\sigma x}. \tag{6-138}$$

This mathematical model is valid for a beam of low-energy (< 100 keV) photons on a high atomic-mass target such as lead or tungsten. Here I_0 is the initial intensity of the beam at $x = 0$, N is the atomic density, σ is the microscopic cross section, and x is the depth into the target.

We wish to understand the relative uncertainty, the ratio of the standard deviation to the expectation, of the beam intensity for various depths. To do this, we can apply the first-order Taylor series approximation to obtain the variance of the intensity. Here we assume the beam intensity I_0, atomic density N, and microscopic cross section σ are uncertain parameters that are independent of each other. From the first-order Taylor series we can find the estimated variance of the intensity,

$$s_I^2 = s_{I_0}^2 \left(\frac{\partial I}{\partial I_0}\right)^2 + s_N^2 \left(\frac{\partial I}{\partial N}\right)^2 + s_\sigma^2 \left(\frac{\partial I}{\partial \sigma}\right)^2. \tag{6-139}$$

The partial derivatives are

$$\frac{\partial I}{\partial I_0} = e^{-N\sigma x}, \tag{6-140a}$$

$$\frac{\partial I}{\partial N} = I_0 \sigma x e^{-N\sigma x}, \tag{6-140b}$$

$$\frac{\partial I}{\partial \sigma} = I_0 N x e^{-N\sigma x}. \tag{6-140c}$$

Inserting these into the equation and factoring out the exponential, we have the variance of

$$s_I^2(x) = \left[s_{I_0}^2 + (I_0 \sigma x)^2 s_N^2 + (I_0 N x)^2 s_\sigma^2\right] e^{-2N\sigma x}. \tag{6-141}$$

Taking the square root gives the standard deviation. If we then divide by $I(x)$ we then have the relative uncertainty as

$$R_I(x) = \frac{s_I(x)}{I(x)} = \frac{\sqrt{\left[s_{I_0}^2 + (I_0 \sigma x)^2 s_N^2 + (I_0 N x)^2 s_\sigma^2\right]}}{I_0}. \tag{6-142}$$

If we are given the following information:

$$\begin{aligned} I_0 &= 2 \times 10^8 \pm 1 \times 10^7 \text{ photons/s}, \\ N &= 0.05 \pm 0.001 \text{ atoms/b/cm}, \\ \sigma &= 2 \pm 0.2 \text{ b}, \end{aligned}$$

then we can calculate

$$R_I(x) = \frac{\sqrt{\left[(1 \times 10^{14}) + (1.6 \times 10^{11})x^2 + (4 \times 10^{12})x^2\right]}}{I_0}.$$

Here each term corresponds to (from left to right) the contribution from the uncertainties in the initial beam intensity, the atom density, and nuclear cross section. We see that the term for uncertainty in the beam intensity has the largest magnitude, but the other two have a factor of x^2, which means the effect of the uncertainties in the atom density and nuclear cross section become more important with increasing depth. Of these two, the uncertainty in the nuclear cross section is more important.

6.10.c Linear Systems of Normally-Distributed Variables

It is often the case that we have random variables that follow a normal distribution (or at least approximately so), which is often a consequence of the central limit theorem. We can show (with a good deal of math that is omitted) that the linear combination of normally distributed random variables is also normally distributed. This allows us to perform error propagation on these random variables without having to resort to performing convolution integrals or Taylor-series approximations.

Suppose we have the linear system of random variables:

$$\begin{aligned} Y_1 &= a_{1,1}X_1 + a_{1,2}X_2 + \cdots + a_{1,N}X_N, \\ Y_2 &= a_{2,1}X_1 + a_{2,2}X_2 + \cdots + a_{1,N}X_N, \\ &\vdots \\ Y_N &= a_{N,1}X_1 + a_{N,2}X_2 + \cdots + a_{N,N}X_N. \end{aligned} \tag{6-143}$$

The random variables are all normally distributed with different mean values given by the column vector $\boldsymbol{\mu}_X$ and the covariance matrix $\boldsymbol{\Sigma}_X$. If the coefficients $a_{i,j}$ are stored in matrix $\mathbf{A}$ then the mean value of the Y random variables are simply

$$\boldsymbol{\mu}_Y = \mathbf{A}\boldsymbol{\mu}_X. \tag{6-144}$$

The covariance matrix for the Y random variables is given by

$$\boldsymbol{\Sigma}_Y = \mathbf{A}^\top \boldsymbol{\Sigma}_X \mathbf{A}. \tag{6-145}$$

6.11 Random Sampling

In many applications the probability distributions are too complicated to analytically or even numerically evaluate on a computer. In these cases we perform random sampling of the underlying random or stochastic process. If we create numerous realizations of the stochastic process, we can then make inferences about its behavior. This is generally called the Monte Carlo method. Before we can discuss this approach, we need to develop techniques for sampling random variates from non-uniform distributions.

All major computer languages have the ability to generate random numbers from a uniform distribution. Often these random numbers are scaled to be in the range from 0 to 1 (with 1 often excluded). Using these sampled uniform random variables,

called uniform variates with the symbol ξ_i (or simply ξ when we only are using a single uniform random variate), we are able to derive methods to find random variates from other distributions. The methods by which a computer generates uniform random variates are called pseudorandom number generation. The sequence of uniform variates are pseudorandom because they are usually generated using a deterministic algorithm; however, if we are given a sequence of these uniform variates, they will essentially appear uniformly random. This means that the sequence has no preference for any particular values in the range of 0 to 1 and there are no discernible patterns, i.e., if we do not know the underlying algorithm, simply knowing one random variate provides no information about the next random variate.

The process of generating random variates of non-uniform distributions is generally called random sampling. Textbooks have been written on this topic alone, so these notes can only cover the most basic techniques for generating random samples. Here we will discuss the methods of inversion and rejection sampling.

6.11.a Direct Inversion Sampling

One approach for generating non-uniform random variates is called the direct inversion sampling method or simply the inversion method.

The inversion method proceeds by finding the cumulative distribution function of the random variable we wish to sample, call this $F_X(x)$. The cumulative distribution function, again, is a probability that a random variable X is less than or equal to some value x. Since the cumulative distribution function is a probability, we know that it ranges from 0 to 1. We therefore equate the cumulative distribution function with a uniform variate ξ and then solve for x. Formally, this is expressed as:

$$x = F_X^{-1}(\xi). \tag{6-146}$$

To illustrate this, let us suppose that we wish to sample random variates of an exponential distribution with rate parameter λ. The reason for doing this may be to find the random time that a radioactive isotope undergoes decay or to find the random position that a particle undergoes an interaction with the background material. The cumulative density function is

$$F_X(x) = 1 - e^{-\lambda x} = \xi. \tag{6-147}$$

If we solve for x, we have

$$x = -\frac{1}{\lambda}\ln(1-\xi).$$

We can make one more simplification and note that one minus a uniform random variable from 0 to 1 is still a uniform random variable from 0 to 1. Therefore, we can obtain a random sample from an exponential distribution using

$$x = -\frac{1}{\lambda}\ln\xi. \tag{6-148}$$

A slightly more complicated example is the truncated exponential distribution with the probability density function

$$f_X(x) = \begin{cases} \dfrac{\lambda e^{-\lambda x}}{1 - e^{-\lambda a}}, & 0 \leq x \leq a \\ 0, & \text{otherwise} \end{cases} . \tag{6-149}$$

The cumulative distribution function is found by integrating from 0 to x, which gives

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \dfrac{1 - e^{-\lambda x}}{1 - e^{-\lambda a}}, & 0 \leq x \leq a \\ 1, & x > a \end{cases} . \tag{6-150}$$

The range from $0 \leq x \leq a$ is the only portion of interest, so we set that equal to a uniform random variate,

$$\frac{1 - e^{-\lambda x}}{1 - e^{-\lambda a}} = \xi, \tag{6-151}$$

and solve for x, giving

$$x = -\frac{1}{\lambda} \ln \left[1 - \xi(1 - e^{-\lambda a}) \right] . \tag{6-152}$$

This equation is slightly more complicated than the previous form, largely because an extra factor of $1 - e^{-\lambda a}$, which was the one over the normalization constant.

The direct inversion sampling method is relatively straightforward in principle: obtain the cumulative distribution function, set it equal to a uniform variate, and solve for x. Unfortunately, only a limited number of random distributions have cumulative distribution functions that are analytically invertible. A common random distribution with a non-invertible cumulative distribution function is the normal distribution. For these distributions, we need to use other techniques.

6.11.b Rejection Sampling

If we have a random variable with a cumulative distribution function that cannot be inverted directly, a common approach to find random variates is to use the rejection sampling technique.

The basic idea with rejection sampling is as follows: Suppose we wish to sample a random variable X with probability density function $f_X(x)$. We take another distribution with a probability density function we know how to sample with probability density function $g_X(x)$ scaled by some constant $c > 1$ so that $cg_X(x) \geq f_X(x)$ for all x in the domain of random variable X. We then sample a candidate value of $\hat{x}$ from $g_X(x)$ using, e.g., direct inversion. Next we decide whether to accept the candidate value $\hat{x}$. To do this, we sample a value $\hat{y}$ from a uniform distribution ranging from 0 to $cg_X(\hat{x})$. If the sampled value $\hat{y} \leq f_X(\hat{x})$ we then accept $\hat{x} = x$. If it is not, we

```
1. do until accept
2.   sample x_hat from g(x)
3.   xi = uniform random variate
4.   y  = xi * c * g(xhat)
5.   if y_hat <= f(x_hat)
6.      accept x = x_hat
7.   else
8.      repeat loop
9. return x
```

Figure 6.4: Pseudocode for rejection sampling of $f_X(x)$ using $g_X(x)$.

repeat the entire process until we are successful. Pseudocode for this algorithm is given in Fig. 6.4.

One drawback with rejection sampling is that we may require numerous trials from sampling from $g_X(x)$ to obtain a valid sample of $f_X(x)$. Indeed, if $g_X(x)$ is not chosen carefully, we can end up with cases where we may spend a significant amount of time sampling $g_X(x)$ numerous times. We can calculate a sampling efficiency using the expectation:

$$\begin{aligned}\epsilon &= \int_{-\infty}^{\infty} \left[\frac{f_X(x)}{cg_X(x)}\right] g_X(x)dx \\ &= \frac{1}{c}\int_{-\infty}^{\infty} f_X(x)dx = \frac{1}{c}.\end{aligned} \tag{6-153}$$

Therefore, the sampling efficiency is inversely proportional to the scaling constant c and this shows that we should select a distribution for which c is smallest, i.e., a $g_X(x)$ that can tightly bound $f_X(x)$.

As an example, let us consider Rayleigh scattering of photons. In Rayleigh scattering, a photon interacts with an atom and elastically scatters, changing direction, but maintaining its energy. The cosine of the angle between the incoming and outgoing photon velocity vectors is random because of the laws of quantum mechanics. The probability density function for the cosine of the scattering angle μ is

$$f_\mu(\mu) = \frac{3}{8}(1+\mu^2), \quad -1 \leq \mu \leq 1. \tag{6-154}$$

This probability density function is plotted on the left plot of Fig. 6.5. It turns out this distribution yields an invertible cumulative density function, but the inversion process yields a very complicated expression. A better approach would be to apply rejection sampling.

To begin, we first note the Rayleigh scattering probability density function is symmetric about the y axis. Therefore, we can sample the right half of the distribution,

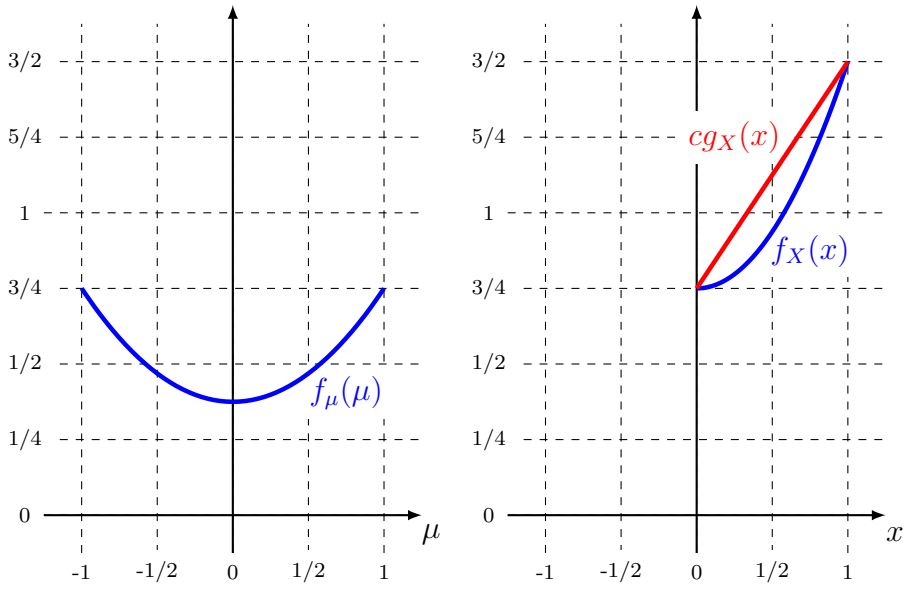

Figure 6.5: Probability density function for scattering cosine μ of Rayleigh scattering (left) and illustration of rejection sampling of the half distribution using a linear bounding function (right).

and then with a probability of $\frac{1}{2}$ flip the sign. Since we are only going to sample half of the distribution, we scale $f(\mu)$ by a factor of 2 to get

$$f_X(x) = 2f_\mu(x = |\mu|) = \frac{3}{4}(1 + x^2), \quad 0 \leq x \leq 1. \tag{6-155}$$

Now, we have several choices to sample half of the parabola. One choice that is fairly straight-forward is to use a probability density function $g_X(x)$ as a line that connects the endpoints of the parabola from $f_X(x)$. This is

$$cg_X(x) = \frac{3}{4}(1 + x). \tag{6-156}$$

An illustration of $cg_X(x)$ with the probability density function $f_X(x)$ is displayed on the right plot in Fig. 6.5. To find the value of c and $g_X(x)$, we need a probability density function proportional to $1 + x$ from $0 \leq x \leq 1$. This can be found from

$$\int_0^1 A(1 + x)dx = 1, \tag{6-157}$$

where A is a normalization constant. If we carry out the integral, we can find that $A = \frac{2}{3}$ and have

$$g_X(x) = \frac{2}{3}(1 + x). \tag{6-158}$$

The constant is therefore $c = \frac{9}{8}$ implying the sampling efficiency is $\epsilon = \frac{8}{9} \approx 0.89$. Now, we seek to sample from $g_X(x)$ using direct inversion. To do this, we obtain the cumulative distribution function

$$G_X(x) = \int_0^x g_X(x')dx' = \frac{1}{3}(x^2 + 2x). \tag{6-159}$$

Setting this equal to a uniform variate ξ_1 yields the quadratic equation

$$x^2 + 2x - 3\xi_1 = 0. \tag{6-160}$$

Solving this and retaining the positive root gives our candidate value of

$$\hat{x} = \sqrt{3\xi_1 + 1} - 1. \tag{6-161}$$

Next we sample our value of $\hat{y}$ using a new uniform random variate ξ_2,

$$\hat{y} = \xi_2 c g_X(\hat{x}) = \frac{3\xi_2}{4}(1 + \hat{x}). \tag{6-162}$$

Now we check if we have a successful sample of $f_X(x)$ if $\hat{y} \leq f_X(\hat{x})$. This is

$$\xi_2(1 + \hat{x}) \leq 1 + \hat{x}^2. \tag{6-163}$$

If we satisfy this condition, we let $x = \hat{x}$; if not, we repeat the process. Finally, we need find μ using the symmetry applied earlier. With probability $\frac{1}{2}$, $\mu = x$, and also with probability $\frac{1}{2}$, $\mu = -x$. To do this, we draw another uniform variate ξ_3 and set μ as

$$\mu = \begin{cases} x, & 0 \leq \xi_3 < \frac{1}{2} \\ -x, & \frac{1}{2} \leq \xi_3 < 1 \end{cases}. \tag{6-164}$$

6.12 Monte Carlo Methods

The Monte Carlo method describes a broad class of methods that are used to solve a variety of problems using random numbers. The method was originally developed shortly after the conclusion of the Second World War during the waning days of the Manhattan Project at Los Alamos to model the transport of radiation, which is a random process. Since the method has been applied in just about every quantitative discipline ranging from particle physics, evolutionary biology, economics, political science, and more. In this section, we will discuss Monte Carlo integration and its application to the transport of radiation through matter.

6.12.a Monte Carlo Integration

It is often the case we encounter integrals that cannot be performed analytically. In these cases, we need to use numerical approximation schemes such as the trapezoid or Simpson's rule. The approach of breaking the integration domain into small intervals

works well for one-dimensional integrals. However, it is often the case we need to evaluate multi-dimensional integrals and it can be computationally prohibitive to perform the integration. For example, in radiation transport, we are required to perform six-dimensional integrals: three spatial, two directional, and one energy. In solid state physics, we need to integrate probability density functions for the positions of all particles in a the element, which can range in the hundreds of dimensions. The statistical models encountered in social science or economics can have a similar dimensionality.

For high-dimensional integrals, it is simply impractical to perform the integration directly using standard numerical techniques. Fortunately, the Monte Carlo method can often be employed to make evaluating these integrals tractable. In this section, we will cover the one-dimensional case.

The idea is very similar to rejection sampling. Suppose we wish to evaluate the integral

$$\int_a^b f(x)dx,$$

where $f(x)$ is some function satisfying $0 \leq f(x) < \infty$ in the domain $a \leq x \leq b$. If we can find a probability density function $g_X(x)$ on the domain $a \leq x \leq b$ and a constant $c > 0$ such that $cg_X(x) \geq f(x)$ in the domain, we can evaluate the integral using rejection sampling via the sampling efficiency. The sampling efficiency is the expected number of samples that are accepted and is

$$\epsilon = \int_a^b \left[\frac{f(x)}{cg_X(x)}\right] g_X(x)dx = \frac{1}{c}\int_a^b f(x)dx. \tag{6-165}$$

Therefore, the integral is

$$\int_a^b f(x)dx = c\epsilon. \tag{6-166}$$

The constant c is determined by bounding $f(x)$ with $cg_X(x)$ and the sampling efficiency ϵ is determined using rejection sampling.

As an example, suppose we wish to evaluate the integral

$$I = \int_0^1 (x-1)\ln\left[\cos\left(\frac{\pi x}{2}\right)\right] dx. \tag{6-167}$$

This integral does not have any result in terms of standard (non-special) functions. Since this is a single integral, we could perform this integral using numerical techniques without too much trouble and get the result of approximately 0.133434; however, we will also use Monte Carlo integration to illustrate the concept.

A plot of $f(x)$ is given by the blue curve in Fig. 6.6. From this curve, it appears that we could reasonably bound this with a linear function as we did in the previous rejection sampling example for Rayleigh scattering. The probability density function is

$$g_X(x) = 2x, \quad 0 \leq x \leq 1, \tag{6-168}$$

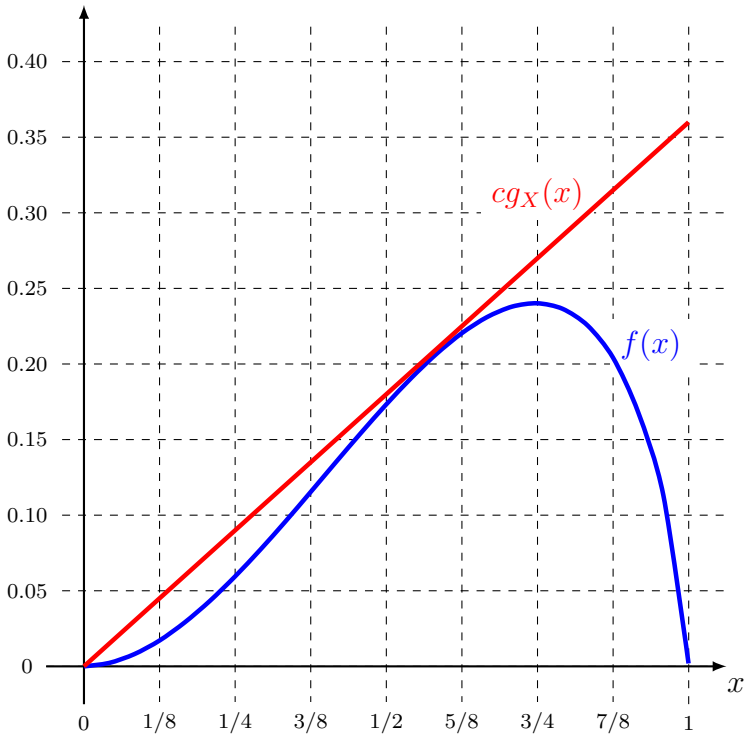

Figure 6.6: Example for Monte Carlo integration.

which may be sampled using direct inversion. The cumulative density function is

$$G_X(x) = x. \tag{6-169}$$

Setting this equal to a uniform variate and solving, gives the sampling scheme

$$x = \sqrt{\xi}. \tag{6-170}$$

Using graphical methods, we plot $g_X(x)$ scaled by a constant c and determine that $c = 0.18$ appears to allow us to bound $f(x)$ with $cg_X(x)$ tightly. For illustration, the function $cg_X(x)$ is plotted using the red line in Fig. 6.6.

Using the rejection method for N random samples from $g_X(x)$, we tabulate the sampling efficiency and multiply the result by $c = 0.18$ to get an approximate value of the integral. Table 6.1 gives a list of numerical results for various sample sizes ranging from 10^3 to 10^8 compared to the reference result obtained using numerical integration. The results converge as the sample size increases with a convergence rate of $1/\sqrt{N}$ by way of the central limit theorem.

6.12.b Application to Particle Transport

A common application of the Monte Carlo method is to simulate the transport of radiation particles. From these simulations, we can estimate quantities of interest that are describable through integrals of the function describing the average behavior of

Table 6.1: Monte Carlo Integration Results of $(x-1)\ln(\cos(\pi x/2))$ from 0 to 1

N	I
10^3	0.135
10^4	0.132912
10^5	0.133411
10^6	0.133398
10^7	0.133416
10^8	0.133443
Numerical	0.133434

the radiation field. The behavior of individual particles is inherently random because of the laws of quantum mechanics; however, the collective behavior of large numbers of particles tends toward a deterministic function as the number gets large. For many applications, we are only concerned with the average behavior of particle behavior, and this section will cover this case. The Monte Carlo method is powerful in that it can also inform us about random statistical fluctuations in the radiation field as well, but this topic will not be covered here. Additionally, here we will discuss the transport of neutral particles only (e.g., neutrons and photons).

To begin, we introduce the concept of a *particle history*. A particle history is a sequence of events spanning the birth of the particle to the removal of that particle and any other particles it may have created during that history (e.g., through nuclear fission). The Monte Carlo particle transport simulation performs numerous random particle histories. During these histories, information about physical quantities of interest is extracted and put into an accumulator to compute average values of physical quantities.

The first step in any particle history is to create the source particle. For this, we must be provided a source distribution $Q(\boldsymbol{x}, \hat{\boldsymbol{\Omega}}, E, t)$ where $\boldsymbol{x}$ is the position, $\hat{\boldsymbol{\Omega}}$ is a unit vector describing the particle direction, E is the kinetic energy of the particle, and t is the time that it is emitted. We then normalize $Q(\boldsymbol{x}, \hat{\boldsymbol{\Omega}}, E, t)$ and use random sampling techniques to find the random starting location of the particle.

After a particle is emitted, the particle will move until it interacts with a background atom/nucleus or transits into another region. The distance to particle interaction is proportional to the macroscopic total cross section or attenuation coefficient $\Sigma_t(E)$ in the current region. The distance to interaction or collision is given by an exponential distribution and found using direct inversion as

$$\ell_c = -\frac{1}{\Sigma_t(E)} \ln \xi. \tag{6-171}$$

We then calculate the distance that the particle would travel until it crosses into the adjacent region or edge of the problem, call this ℓ_s. The distance to surface crossing is a deterministic, or non-random, quantity given a starting position $\boldsymbol{x}$ and direction $\hat{\boldsymbol{\Omega}}$. Most often, we describe the surfaces in a Monte Carlo particle transport

equation using analytic equations. For example, a plane has the equation

$$S(x, y, z) = Ax + By + Cz - D = 0, \tag{6-172}$$

where A, B, C, and D are coefficients that determine the orientation of the plane. To find the distance we set each coordinate to the point at the surface point $x \to x+\Omega_x\ell_s$, $y \to y + \Omega_y\ell_s$, $z \to z + \Omega_z\ell_s$ and then solve for ℓ_s. Here (x, y, z) are the particle positions at the start of the trajectory and $(\Omega_x, \Omega_y, \Omega_z)$ are components of the unit direction vector of the particle. For the plane this is

$$A(x + \Omega_x\ell_s) + B(y + \Omega_y\ell_s) + C(z + \Omega_z\ell_s) - D = 0. \tag{6-173}$$

Solving this gives the distance to the plane as

$$\ell_s = \frac{D - Ax - By - Cz}{A\Omega_x + B\Omega_y + C\Omega_z}. \tag{6-174}$$

If we get a negative value for ℓ_s, then that means the particle is traveling away from the plane and we throw out the distance as not a possible intersection point.

Another common surface is a sphere at the origin, which has the equation

$$S(x, y, z) = x^2 + y^2 + z^2 - R^2 = 0, \tag{6-175}$$

with R being the radius. As with the plane, we get the equation for the distance ℓ_s as

$$S(x, y, z) = (x + \Omega_x\ell_s)^2 + (y + \Omega_y\ell_s)^2 + (z + \Omega_z\ell_s)^2 - R^2 = 0, \tag{6-176}$$

We can rewrite this as a quadratic equation for ℓ_s as

$$\ell_s^2 + 2(x\Omega_x + y\Omega_y + z\Omega_z)\ell_s + \left(x^2 + y^2 + z^2 - R^2\right) = 0. \tag{6-177}$$

We can then solve this quadratic equation to obtain its two roots. The value of ℓ_s is the smallest positive, real root. If we have two positive roots, then the particle is entering the surface and will intersect twice. We take the smallest, as it is the next intersection point. Finding a positive and negative root implies that the particle is inside the surface and the positive root gives the distance until it leaves, whereas the negative root is the distance in the opposite direction. Having two negative roots implies the particle is outside the sphere and moving away from it. Finally, complex roots imply that the particle will not enter the surface at all. In all these cases, except for the smallest positive root, we eliminate those roots from consideration.

To decide which surface the particle intersections, we calculate the distances to all valid surfaces and take the smallest positive distance. This gives us the distance to surface ℓ_s.

Next, given ℓ_c and ℓ_s we need to determine the next event for the particle by taking the smallest of the two. To summarize this, pseudocode for calculating the distance to the next event is given in Fig. 6.7.

```
1. dist to collision = -log(rand)/sigma_t
2. dist to surface   = infinity
3. loop over surfaces bounding region:
4.   candiate dist = solution to surface equation
5.   if candidate dist > 0:
6.     dist to surface = min( dist to surface, candidate dist )
7. dist to next event = min( dist to collision, dist to surface )
```

Figure 6.7: Pseudocode for finding the distance to the next event in Monte Carlo particle transport.

If ℓ_s is smaller, we advance the particle to the adjacent region by moving the particle a distance ℓ_s along its direction $\hat{\mathbf{\Omega}}$. If the particle reaches the edge of the problem, we stop that particle trajectory. If there is another adjacent region, we randomly sample a new distance to collision, calculate a new distance to intersecting surface, and determine the next event. If ℓ_c is smaller, we then advance the particle a distance ℓ_c along its direction $\hat{\mathbf{\Omega}}$ and then we randomly determine the result of the collision, which is a random process because of the laws of quantum mechanics.

To process the random collision, first we need to determine the atom or nuclide that the particle interacts with. Each of these atoms/nuclides has an index j and has its own total macroscopic cross section or interaction coefficient $\Sigma_t^j(E)$ (note that j is a superscript, not a power) such that the total interaction coefficient for the material is

$$\Sigma_t(E) = \sum_{j=1}^{J} \Sigma_t^j(E). \tag{6-178}$$

We can use these total interaction coefficients to construct a probability mass and cumulative distribution function for the random atom/nuclide. The probability that the particle interacts with a particular atom/nuclide j is

$$p_j = \frac{\Sigma_t^j(E)}{\Sigma_t(E)}. \tag{6-179}$$

To find this, we construct a cumulative distribution function, get a new uniform random variate ξ, and determine the range that it is within.

Once we have determined the atom or nuclide j, we have to determine the reaction type. The conditional probability that a specific type of reaction r with atom/nuclide j given that a collision with atom/nuclide j occurs is

$$p_r^j = \frac{\Sigma_r^j(E)}{\Sigma_t^j(E)}. \tag{6-180}$$

```
 1. r1 = rand
 2. s1 = 0
 3. loop j over all atoms/nuclides:
 4.   s1 += total xs for atom/nuclide j
 5.   if r1 < s1 * total xs for material:
 6.     select atom/nuclide j
 7.     exit loop
 8. r2 = rand
 9. s2 = 0
10. loop k over all reactions in atom/nuclide j:
11.   s2 += xs for reaction k of atom/nuclide j
12.   if r2 < s2 * total xs for atom/nuclide j:
13.     select reaction k
14.     exit loop
```

Figure 6.8: Pseudocode for sampling the atom/nuclide and corresponding reaction.

In the exact same manner as randomly sampling the atom or nuclide, we construct a cumulative distribution, draw another different uniform random variate, and find where it falls to determine the reaction. A detailed algorithm of sampling the atom/nuclide and the reaction is given in Fig. 6.8.

Depending on the particle type and the atom or nuclide, there are numerous possible types of reactions that may be possible. For neutrons, the common reactions are capture, elastic scattering, and fission. For photons, we have as common reactions as photoelectric absorption, Compton (incoherent elastic) scattering, and pair production. We then discuss the general idea for handling each of those cases.

For the capture or photoelectric absorption reactions, the neutron or photon respectively disappears. In the case of neutron capture, the resulting nucleus is usually in an excited state and the de-excitation leads to the emission of one or more photons. These photons are then stored and continue in the current history. Likewise, photoelectric effect can lead to the emission of fluorescence x-ray photons that may need to be followed. Additionally, the ejected electron can create lower-energy Bremsstrahlung photons. Depending on the level of fidelity desired, these secondary photons may or may not be tracked. In this section, we will ignore these secondaries.

In elastic scattering events, we are given a continuous probability density function for μ_0, which is the cosine of the angle between the incident and outgoing photon direction vectors. We sample a μ_0 from the probability density function $f_{\mu_0}(\mu_0)$ and then compute the outgoing direction vector using a 3-D rotation matrix. Going through this process is beyond the scope of these notes. For lower energy photons (< 100 keV) or lower energy neutrons scattering off of heavy targets (e.g., < 100 keV off of uranium), however, we may assume that scattering is isotropic or equiprobable

in all directions. In this case, we sample the polar cosine uniformly from -1 to 1 using

$$\mu_0 = 2\xi_1 - 1, \tag{6-181a}$$

and an azimuthal angle uniformly from 0 to 2π using

$$\gamma_0 = 2\pi\xi_2. \tag{6-181b}$$

The outgoing direction vector components are the spherical coordinates with unit radius written as

$$\Omega_x = \sqrt{1-\mu_0^2}\cos\gamma_0, \tag{6-182a}$$

$$\Omega_y = \sqrt{1-\mu_0^2}\sin\gamma_0, \tag{6-182b}$$

$$\Omega_z = \mu_0. \tag{6-182c}$$

The outgoing kinetic energy is determined from conservation of momentum and mechanical energy.

In pair production, the original photon is destroyed and two photons are emitted with a kinetic energy of $m_e c^2 = 0.511$ MeV. The direction $\hat{\mathbf{\Omega}}$ of one of the photons is sampled isotropically using the method described for isotropic elastic scattering. The second photon is given a direction in the opposite direction $-\hat{\mathbf{\Omega}}$ and is stored for later on in the history after the first photon is removed from the problem.

For nuclear fission, the original neutron is absorbed and we have a random number of neutrons emitted each isotropically and at random energies. For the number of neutrons emitted, e.g., the multiplicity, we rarely have a reliably-measured probability mass function for the multiplicity that we can sample. Rather, we are usually only provided the mean value $\overline{\nu}$, which is a non-integer quantity. In this case, the best we can do is sample the multiplicity in a way that preserves this average. There are infinitely many ways one could do this, but a simple approach uses the distribution

$$\nu = \lfloor \overline{\nu} + \xi \rfloor, \tag{6-183}$$

where $\lfloor \cdot \rfloor$ is the "floor" operator that rounds down to the nearest integer. For example if $\overline{\nu} = 2.6$, then with probability 0.4 we will have $\nu = 2$ neutrons emitted, and probability 0.6 we will have $\nu = 3$ neutrons emitted with the average coming out to 2.6. For each neutron, we independently sample a direction isotropically and an outgoing energy from some provided distribution $f(E)$ often given the symbol $\chi(E)$. All but one of these particles are stored in a bank for later in the same history with the last one continuing the history.

After the collision is processed, if we have a surviving particle, the history continues by sampling a new distance to collision and calculating a new distance to surface. The particle keeps going until it is absorbed or leaks out of the problem. At this point, we look to see if we have any stored particles in a bank of particles. If there are, the history continues using by taking one of these stored particles out of the bank

and executing the particle transport routines on it. Once the bank is fully depleted and no particles remain, the history terminates. The calculation continues with the next history until all the number of particle histories have been performed, at which point the calculation stops and results are provided to the user.

In terms of getting results, we keep track of accumulators called estimators or tallies that are "scored" during the particle histories. The most common physical quantities to be estimated are the particle scalar flux or path-length density or the particle current. The scalar flux or path-length density is used to calculate reaction rates, heating rates, biological dose, etc. As the name implies, the path-length density is the amount of path per unit volume the particles generate in a given history. For each region that we wish to know the scalar flux or path-length density, we get an estimate for that history i as

$$\phi_i = \frac{1}{V} \sum_{\text{tracks}k} \ell_k, \tag{6-184}$$

where V is the volume of the region, k is the index for the track in that region, and ℓ_k is the length of the track. This estimate can be multiplied by a response function $r(E)$ (e.g., cross section or flux to biological dose conversion factor) to get a reaction rate, dose, etc. by

$$R_i = \frac{1}{V} \sum_{\text{tracks } k} r_k(E)\ell_k, \tag{6-185}$$

where $r_k(E)$ is the response function for the kth track. The particle current is simply the rate that particles cross a surface, which is obtained by counting up the number of surface crossings. The particle current is used in in applications where we, for example, wish to understand the effectiveness of a radiation shield. The particle current for history i is

$$J_i = \sum_{\text{crossings } k} 1, \tag{6-186}$$

where k in this context denotes the kth time a particle has crossed the surface in this history. At the end of the history, these estimates are added to a running sum and the average over N histories is taken. For example, the scalar flux or path-length density is simply the sample mean

$$\phi = \frac{1}{N} \sum_{i=1}^{N} \phi_i, \tag{6-187}$$

and likewise for any other response.

To summarize all of this discussion, the pseudocode for the overall Monte Carlo particle transport algorithm is given in Fig. 6.9.

```
 1. loop i over number of histories:
 2.   draw starting particles from source distribution
 3.   insert source particles into back
 4.   while bank not empty:
 5.     pull particle out of bank
 6.     while particle alive:
 5.       sample random distance to collision dc
 6.       calculate distance to surface intersection ds
 7.       distance = min( dc, ds )
 8.       move particle by distance along path
 9.       compute scalar flux contribution along path
10.       if dc <= ds:
11.         select nuclide and reaction
12.         process selected reaction (possibly kill particle)
13.         put secondaries in particle bank
14.       else:
15.         compute particle current contribution
16.         determine next region
17.         kill particle if exited problem
18.   add scalar flux and current estimators to accumulator
19. divide accumulators by number of histories, write results
```

Figure 6.9: Pseudocode for the overall Monte Carlo particle transport.