Matrix Analysis, CAAM 335, Spring 2012

Steven J Cox

Preface

Bellman has called matrix theory 'the arithmetic of higher mathematics.' Under the influence of Bellman and Kalman engineers and scientists have found in matrix theory a language for representing and analyzing multivariable systems. Our goal in these notes is to demonstrate the role of matrices in the modeling of physical systems and the power of matrix theory in the analysis and synthesis of such systems.

Beginning with modeling of structures in static equilibrium we focus on the linear nature of the relationship between relevant state variables and express these relationships as simple matrix–vector products. For example, the voltage drops across the resistors in a network are linear combinations of the potentials at each end of each resistor. Similarly, the current through each resistor is assumed to be a linear function of the voltage drop across it. And, finally, at equilibrium, a linear combination (in minus out) of the currents must vanish at every node in the network. In short, the vector of currents is a linear transformation of the vector of voltage drops which is itself a linear transformation of the vector of potentials. A linear transformation of n numbers into m numbers is accomplished by multiplying the vector of n numbers by an m-by-n matrix. Once we have learned to spot the ubiquitous matrix–vector product we move on to the analysis of the resulting linear systems of equations. We accomplish this by stretching your knowledge of three–dimensional space. That is, we ask what does it mean that the m–by–n matrix X transforms $\mathbf{R}^n$ (real n–dimensional space) into $\mathbf{R}^m$? We shall *visualize* this transformation by splitting both $\mathbf{R}^n$ and $\mathbf{R}^m$ each into two smaller spaces between which the given X behaves in very manageable ways. An understanding of this splitting of the ambient spaces into the so called *four fundamental subspaces* of X permits one to answer virtually every question that may arise in the study of structures in static equilibrium.

In the second half of the notes we argue that matrix methods are equally effective in the modeling and analysis of dynamical systems. Although our modeling methodology adapts easily to dynamical problems we shall see, with respect to analysis, that rather than splitting the ambient spaces we shall be better served by splitting X itself. The process is analogous to decomposing a complicated signal into a sum of simple harmonics oscillating at the natural frequencies of the structure under investigation. For we shall see that (most) matrices may be written as weighted sums of matrices of very special type. The weights are the eigenvalues, or natural frequencies, of the matrix while the component matrices are projections composed from simple products of eigenvectors. Our approach to the eigendecomposition of matrices requires a brief exposure to the beautiful field of Complex Variables. This foray has the added benefit of permitting us a more careful study of the Laplace Transform, another fundamental tool in the study of dynamical systems.

Contents

1. Matrix Methods for Electrical Systems

1.1. Nerve Cables and the Strang Quartet

We wish to confirm, by example, the prefatory claim that matrix algebra is a useful means of organizing (stating and solving) multivariable problems. In our first such example we investigate the response of a neuron to a constant current stimulus. Ideally, a neuron is simply a cylinder of radius a and length ℓ that conducts electricity both along its length and across its lateral membrane. Though we shall, in subsequent chapters, delve more deeply into the biophysics, here, in our first outing, we stick to its purely resistive properties. Theses are expressed via two quantities: ρ_i, the resistivity, in Ωcm, of the cytoplasm that fills the cell, and ρ_m, the resistivity in Ωcm^2 of the cell's lateral membrane.

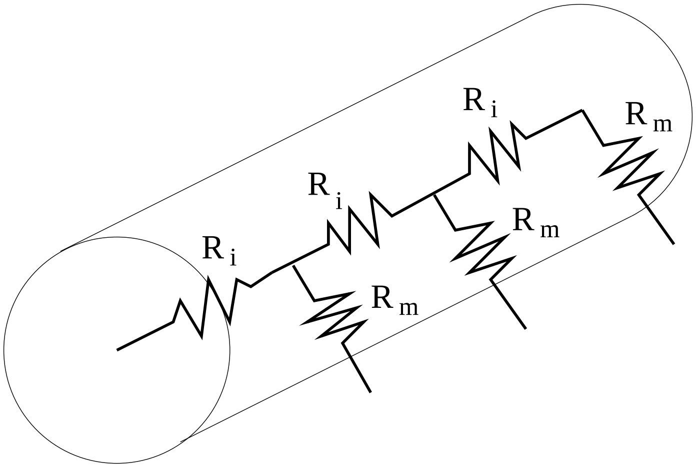


Figure 1.1. A 3 compartment model of a neuron.

Although current surely varies from point to point along the neuron it is hoped that these variations are regular enough to be captured by a multicompartment model. By that we mean that we choose a number N and divide the neuron into N segments each of length ℓ/N. Denoting a segment's axial resistance by

$$R_i = \frac{\rho_i \ell/N}{\pi a^2}$$

and membrane resistance by

$$R_m = \frac{\rho_m}{2\pi a \ell/N}$$

we arrive at the lumped circuit model of Figure 1.1. For a neuron in culture we may assume a constant extracellular potential, e.g., zero. We accomplish this by connecting and grounding the extracellular nodes, see Figure 1.2.

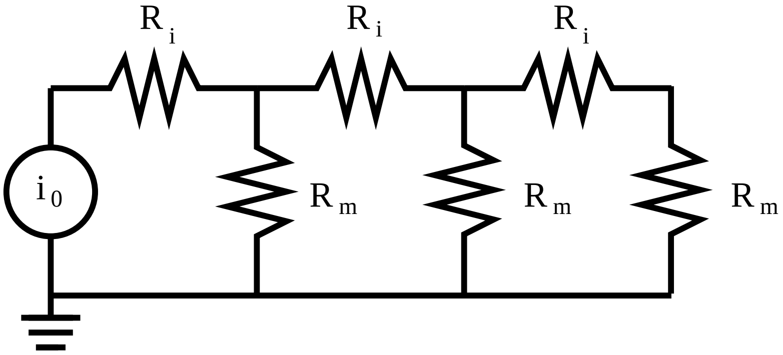


Figure 1.2. A rudimentary neuronal circuit model.

This figure also incorporates the exogenous disturbance, a current stimulus between ground and the left end of the neuron. Our immediate goal is to compute the resulting currents through each resistor and the potential at each of the nodes. Our long–range goal is to provide a modeling methodology that can be used across the engineering and science disciplines. As an aid to computing the desired quantities we give them names. With respect to Figure 1.3 we label the vector of potentials

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \quad \text{and vector of currents} \quad y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix}.$$

We have also (arbitrarily) assigned directions to the currents as a graphical aid in the consistent application of the basic circuit laws.

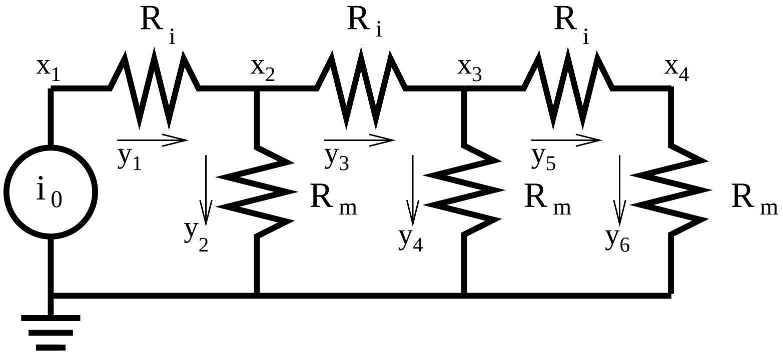


Figure 1.3 The fully dressed circuit model.

We incorporate the circuit laws in a modeling methodology that takes the form of a *Strang Quartet* after **?**,

(S1) Express the voltage drops via $e = -Ax$.
(S2) Express Ohm's Law via $y = Ge$.
(S3) Express Kirchhoff's Current Law via $A^T y = -f$.
(S4) Combine the above into $A^T GAx = f$.

The A in (S1) is the node–edge adjacency matrix – it encodes the network's connectivity. The G in (S2) is the diagonal matrix of edge conductances – it encodes the physics of the network. The f in (S3) is the vector of current sources – it encodes the network's stimuli. The culminating $A^T GA$

in (S4) is the symmetric matrix whose inverse, when applied to f, reveals the vector of potentials, x. In order to make these ideas our own we must work many, many examples.

1.2. Example 1

With respect to the circuit of figure 1.3, in accordance with step (S1), we express the six potentials differences (always tail minus head)

$$\begin{aligned}
e_1 &= x_1 - x_2 \\
e_2 &= x_2 \\
e_3 &= x_2 - x_3 \\
e_4 &= x_3 \\
e_5 &= x_3 - x_4 \\
e_6 &= x_4
\end{aligned}$$

Such long, tedious lists cry out for matrix representation, to wit

$$e = -Ax \quad \text{where} \quad A = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

Step (S2), Ohm's law, states that the current along an edge is equal to the potential drop across the edge divided by the resistance of the edge. In our case,

$$y_j = e_j/R_i, \; j = 1,3,5 \quad \text{and} \quad y_j = e_j/R_m, \; j = 2,4,6$$

or, in matrix notation,

$$y = Ge$$

where

$$G = \begin{pmatrix} 1/R_i & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/R_m & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/R_i & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/R_m & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/R_i & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/R_m \end{pmatrix}$$

Step (S3), Kirchhoff's Current Law, states that the sum of the currents into each node must be zero. In our case

$$\begin{aligned}
i_0 - y_1 &= 0 \\
y_1 - y_2 - y_3 &= 0 \\
y_3 - y_4 - y_5 &= 0 \\
y_5 - y_6 &= 0
\end{aligned}$$

or, in matrix terms

$$By = -f$$

where

$$B = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix} \quad \text{and} \quad f = \begin{pmatrix} i_0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

Turning back the page we recognize in B the **transpose** of A. Calling it such, we recall our main steps

$$e = -Ax, \quad y = Ge, \quad \text{and} \quad A^T y = -f.$$

On substitution of the first two into the third we arrive, in accordance with (S4), at

$$\boxed{A^T G A x = f.} \tag{1.1}$$

This is a linear system of four simultaneous equations for the 4 unknown potentials, x_1 through x_4. As you may know, the system (1.1) may have either 1, 0, or infinitely many solutions, depending on f and $A^T GA$. We shall devote chapters 3 and 4 to a careful analysis of the previous sentence. For now, we simply invoke the MATLAB `backslash` command and arrive at the response depicted in Figure 1.4.

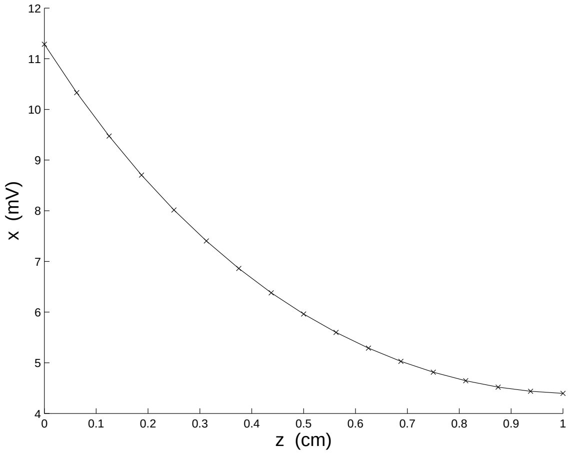


Figure 1.4. Results of a 16 compartment simulation. `cab1.m`.

Once the structure of the constituents in the fundamental system (1.1) is determined it is an easy matter to implement it, as we have done in `cab1.m`, for an arbitrary number of compartments. In Figure 1.4 we see that the stimulus at the neuron's left end produces a depolarization there that then attenuates with distance from the site of stimulation.

1.3. Example 2

We have seen in the previous section how a current source may produce a potential difference across a neuron's membrane. We note that, even in the absence of electrical stimuli, there is always a difference in potential between the inside and outside of a living cell. In fact, this difference is one of the biologist's definition of 'living.' Life is maintained by the fact that the neuron's interior is rich (relative to the cell's exterior) in potassium ions and poor in sodium and chloride ions. These concentration differences establish a resting potential difference, E_m, across the cell's lateral membrane. The modified circuit diagram is given in Figure 1.5.

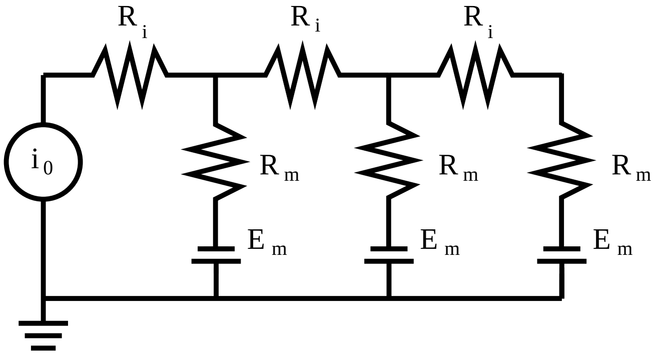


Figure 1.5 Circuit model with batteries associated with the rest potential.

The convention is that the potential difference across the battery is E_m. As the bottom terminal of each battery is grounded it follows that the potential at the top of each battery is E_m. Revisiting steps (S1–4) of the Strang Quartet we note that in (S1) the even numbered voltage drops are now

$$e_2 = x_2 - E_m, \quad e_4 = x_3 - E_m \quad \text{and} \quad e_6 = x_4 - E_m.$$

We accommodate such things by generalizing (S1) to

(S1') Express the voltage drops as $e = b - Ax$ where b is the vector that encodes the batteries.

No changes are necessary for (S2) and (S3). The final step now reads,

(S4') Combine (S1'), (S2) and (S3) to produce

$$\boxed{A^T GAx = A^T Gb + f.} \tag{1.2}$$

This is **the general** form for a resistor network driven by current sources and batteries.

Returning to Figure 1.5 we note that

$$b = -E_m[0\ 1\ 0\ 1\ 0\ 1]^T \quad \text{and} \quad A^T Gb = (E_m/R_m)[0\ 1\ 1\ 1]^T.$$

To build and solve (1.2) requires only minor changes to our old code. The new program is called `cab2.m` and results of its use are indicated in Figure 1.6.

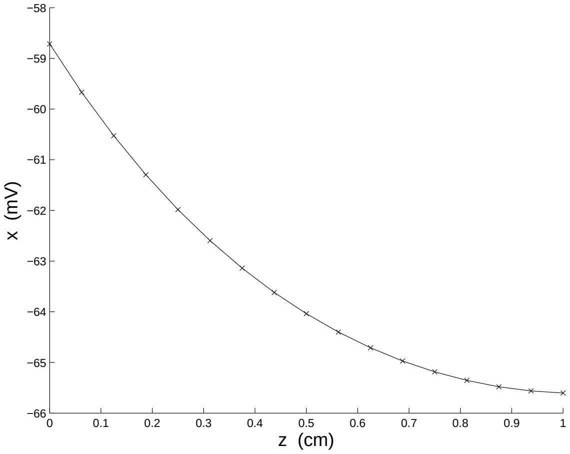


Figure 1.6. Results of a 16 compartment simulation with batteries, $E_m = -70\ mV$. `cab2.m`

1.4. Exercises

1. In order to refresh your matrix-vector multiply skills please calculate, by hand, the product A^TGA in the 3 compartment case and write out the 4 equations in (1.1). The second equation should read
$$(-x_1 + 2x_2 - x_3)/R_i + x_2/R_m = 0. \tag{1.3}$$

2. We began our discussion with the 'hope' that a multicompartment model could indeed adequately capture the neuron's true potential and current profiles. In order to check this one should run `cab1.m` with increasing values of N until one can no longer detect changes in the computed potentials.

   (a) Please run `cab1.m` with $N = 8, 16, 32$ and 64. Plot all of the potentials on the **same** (use `hold`) graph, using different line types for each. (You may wish to alter `cab1.m` so that it accepts N as an argument).

   Let us now interpret this convergence. The main observation is that the difference equation, (1.3), approaches a differential equation. We can see this by noting that
$$dz \equiv \ell/N$$
   acts as a spatial 'step' size and that x_k, the potential at $(k-1)dz$, is approximately the value of the true potential at $(k-1)dz$. In a slight abuse of notation, we denote the latter
$$x((k-1)dz).$$
   Applying these conventions to (1.3) and recalling the definitions of R_i and R_m we see (1.3) become
$$\frac{\pi a^2}{\rho_i}\frac{-x(0) + 2x(dz) - x(2dz)}{dz} + \frac{2\pi a dz}{\rho_m}x(dz) = 0,$$
   or, after multiplying through by $\rho_m/(\pi a dz)$,
$$\frac{a\rho_m}{\rho_i}\frac{-x(0) + 2x(dz) - x(2dz)}{dz^2} + 2x(dz) = 0.$$
   We note that a similar equation holds at each node (save the ends) and that as $N \to \infty$ and therefore $dz \to 0$ we arrive at
$$\frac{d^2x(z)}{dz^2} - \frac{2\rho_i}{a\rho_m}x(z) = 0. \tag{1.4}$$
   (b) With $\mu \equiv 2\rho_i/(a\rho_m)$ show that
$$x(z) = \alpha\sinh(\sqrt{\mu}z) + \beta\cosh(\sqrt{\mu}z) \tag{1.5}$$
   satisfies (1.4) regardless of α and β.

   We shall determine α and β by paying attention to the ends of the neuron. At the near end we find
$$\frac{\pi a^2}{\rho_i}\frac{x(0) - x(dz)}{dz} = i_0,$$
   which, as $dz \to 0$ becomes
$$\frac{dx(0)}{dz} = -\frac{\rho_i i_0}{\pi a^2}. \tag{1.6}$$

At the far end, we interpret the condition that no axial current may leave the last node to mean

$$\frac{dx(\ell)}{dz} = 0. \tag{1.7}$$

(c) Substitute (1.5) into (1.6) and (1.7) and solve for α and β and write out the final $x(z)$.

(d) Substitute into x the ℓ, a, ρ_i and ρ_m values used in `cab1.m`, plot the resulting function (using, e.g., `ezplot`) and compare this to the plot achieved in part (a).

2. Matrix Methods for Mechanical Systems

2.1. Elastic Fibers and the Strang Quartet

We connect 3 masses (nodes) with four springs (fibers) between two immobile walls, as in Figure 2.1, and apply forces at the masses and measure the associated displacement.

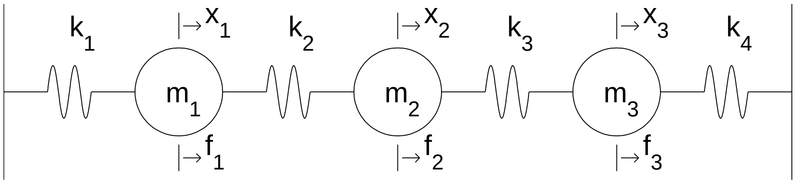


Figure 2.1. A fiber chain.

We suppose that a horizontal force, f_j, is applied to each m_j, and produces a horizontal displacement x_j, with the sign convention that rightward means positive. The bars at the ends of the figure indicate rigid supports incapable of movement. The k_j denote the respective spring stiffnesses. Regarding units, we measure f_j in Newtons (N) and x_j in meters (m) and so stiffness, k_j, is measured in (N/m). In fact each stiffness is a parameter composed of both 'material' and 'geometric' quantities. In particular,

$$k_j = \frac{Y_j a_j}{L_j} \tag{2.1}$$

where Y_j is the fiber's Young's modulus (N/m^2), a_j is the fiber's cross-sectional area (m^2) and L_j is the fiber's (reference) length (m).

The analog of potential difference is here elongation. If e_j denotes the elongation of the jth spring then naturally,

$$e_1 = x_1, \quad e_2 = x_2 - x_1, \quad e_3 = x_3 - x_2, \quad \text{and} \quad e_4 = -x_3,$$

or, in matrix terms,

$$e = Ax \quad \text{where} \quad A = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix}.$$

We note that e_j is positive when the spring is stretched and negative when compressed. The analog of Ohm's Law is here Hooke's Law: the restoring force in a spring is proportional to its elongation. We call this constant of proportionality the stiffness, k_j, of the spring, and denote the restoring force by y_j. Hooke's Law then reads, $y_j = k_j e_j$, or, in matrix terms

$$y = Ke \quad \text{where} \quad K = \begin{pmatrix} k_1 & 0 & 0 & 0 \\ 0 & k_2 & 0 & 0 \\ 0 & 0 & k_3 & 0 \\ 0 & 0 & 0 & k_4 \end{pmatrix}.$$

The analog of Kirchhoff's Current Law is here typically called 'force balance.' More precisely, equilibrium is synonymous with the fact that the net force acting on each mass must vanish. In symbols,

$$y_1 - y_2 - f_1 = 0, \quad y_2 - y_3 - f_2 = 0, \quad \text{and} \quad y_3 - y_4 - f_3 = 0,$$

or, in matrix terms

$$By = f \quad \text{where} \quad f = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

As is the previous section we recognize in B the transpose of A. Gathering our three important steps

$$\begin{aligned} e &= Ax \\ y &= Ke \\ A^T y &= f \end{aligned}$$

we arrive, via direct substitution, at an equation for x. Namely

$$A^T y = f \Rightarrow A^T K e = f \Rightarrow \boxed{A^T K A x = f.}$$

Assembling $A^T K A$ we arrive at the final system

$$\begin{pmatrix} k_1 + k_2 & -k_2 & 0 \\ -k_2 & k_2 + k_3 & -k_3 \\ 0 & -k_3 & k_3 + k_4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix}. \tag{2.2}$$

Although MATLAB solves such systems with ease our aim here is to develop a deeper understanding of Gaussian Elimination and so we proceed by hand. This aim is motivated by a number of important considerations. First, not all linear systems have unique solutions. A careful look at Gaussian Elimination will provide the general framework for not only classifying those systems that possess unique solutions but also for providing detailed diagnoses of those systems that lack solutions or possess too many.

In Gaussian Elimination one first uses linear combinations of preceding rows to eliminate nonzeros below the main diagonal and then solves the resulting triangular system via back–substitution. To firm up our understanding let us take up the case where each $k_j = 1$ and so (2.2) takes the form

$$\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} \tag{2.3}$$

We eliminate the $(2, 1)$ (row 2, column 1) element by implementing

$$\text{new row } 2 = \text{old row } 2 + \frac{1}{2}\text{row } 1, \tag{2.4}$$

bringing

$$\begin{pmatrix} 2 & -1 & 0 \\ 0 & 3/2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 + f_1/2 \\ f_3 \end{pmatrix}$$

We eliminate the current $(3, 2)$ element by implementing

$$\text{new row } 3 = \text{old row } 3 + \frac{2}{3}\text{row } 2, \tag{2.5}$$

bringing the upper–triangular system

$$Ux = g, \tag{2.6}$$

or, more precisely,

$$\begin{pmatrix} 2 & -1 & 0 \\ 0 & 3/2 & -1 \\ 0 & 0 & 4/3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 + f_1/2 \\ f_3 + 2f_2/3 + f_1/3 \end{pmatrix} \tag{2.7}$$

One now simply reads off

$$x_3 = (f_1 + 2f_2 + 3f_3)/4.$$

This in turn permits the solution of the second equation

$$x_2 = 2(x_3 + f_2 + f_1/2)/3 = (f_1 + 2f_2 + f_3)/2,$$

and, in turn,

$$x_1 = (x_2 + f_1)/2 = (3f_1 + 2f_2 + f_3)/4.$$

One must say that Gaussian Elimination has succeeded here. For, regardless of the actual elements of f we have produced an x for which $A^T KAx = f$.

Although Gaussian Elimination, remains the most efficient means for solving systems of the form $Sx = f$ it pays, at times, to consider alternate means. At the algebraic level, suppose that there exists a matrix that 'undoes' multiplication by S in the sense that multiplication by 2^{-1} undoes multiplication by 2. The matrix analog of $2^{-1}2 = 1$ is

$$S^{-1}S = I$$

where I denotes the identity matrix (all zeros except the ones on the diagonal). We call S^{-1} "the inverse of S" or "S inverse" for short. Its value stems from watching what happens when it is applied to each side of $Sx = f$. Namely,

$$Sx = f \Rightarrow S^{-1}Sx = S^{-1}f \Rightarrow Ix = S^{-1}f \Rightarrow x = S^{-1}f.$$

Hence, to solve $Sx = f$ for x it suffices to multiply f by the inverse of S. Let us now consider how one goes about computing S^{-1}. In general this takes a little more than twice the work of Gaussian Elimination, for we interpret

$$SS^{-1} = I$$

as n (the size of S) applications of Gaussian elimination, with f running through n columns of the identity matrix. The bundling of these n applications into one is known as the Gauss-Jordan method. Let us demonstrate it on the S appearing in (2.3). We first augment S with I.

$$\begin{pmatrix} 2 & -1 & 0 & | & 1 & 0 & 0 \\ -1 & 2 & -1 & | & 0 & 1 & 0 \\ 0 & -1 & 2 & | & 0 & 0 & 1 \end{pmatrix}$$

We then eliminate down, being careful to address each of the 3 f vectors. This produces

$$\begin{pmatrix} 2 & -1 & 0 & | & 1 & 0 & 0 \\ 0 & 3/2 & -1 & | & 1/2 & 1 & 0 \\ 0 & 0 & 4/3 & | & 1/3 & 2/3 & 1 \end{pmatrix}$$

Now, rather than simple back–substitution we instead eliminate up. Eliminating first the $(2,3)$ element we find

$$\begin{pmatrix} 2 & -1 & 0 & | & 1 & 0 & 0 \\ 0 & 3/2 & 0 & | & 3/4 & 3/2 & 3/4 \\ 0 & 0 & 4/3 & | & 1/3 & 2/3 & 1 \end{pmatrix}$$

Now eliminating the $(1,2)$ element we achieve

$$\begin{pmatrix} 2 & 0 & 0 & | & 3/2 & 1 & 1/2 \\ 0 & 3/2 & 0 & | & 3/4 & 3/2 & 3/4 \\ 0 & 0 & 4/3 & | & 1/3 & 2/3 & 1 \end{pmatrix}$$

In the final step we scale each row in order that the matrix on the left takes on the form of the identity. This requires that we multiply row 1 by 1/2, row 2 by 3/2 and row 3 by 3/4, with the result

$$\begin{pmatrix} 1 & 0 & 0 & | & 3/4 & 1/2 & 1/4 \\ 0 & 1 & 0 & | & 1/2 & 1 & 1/2 \\ 0 & 0 & 1 & | & 1/4 & 1/2 & 3/4 \end{pmatrix}.$$

Now in this transformation of S into I we have, *ipso facto*, transformed I to S^{-1}, i.e., the matrix that appears on the right upon applying the method of Gauss–Jordan is the inverse of the matrix that began on the left. In this case,

$$S^{-1} = \begin{pmatrix} 3/4 & 1/2 & 1/4 \\ 1/2 & 1 & 1/2 \\ 1/4 & 1/2 & 3/4 \end{pmatrix}.$$

One should check that $S^{-1}f$ indeed coincides with the x computed above.

Not all matrices possess inverses. Those that do are called **invertible** or **nonsingular**. For example

$$\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$$

is **singular**.

Some matrices can be inverted by inspection. An important class of such matrices is in fact latent in the process of Gaussian Elimination itself. To begin, we build the elimination matrix that enacts the elementary row operation spelled out in (2.4),

$$E_1 = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Do you 'see' that this matrix (when applied from the left to S) leaves rows 1 and 3 unsullied but adds half of row one to two? This ought to be 'undone' by simply subtracting half of row 1 from row two, i.e., by application of

$$E_1^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Please confirm that $E_1^{-1}E_1$ is indeed I. Similarly, the matrix analogs of (2.5) and its undoing are

$$E_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 2/3 & 1 \end{pmatrix} \quad \text{and} \quad E_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2/3 & 1 \end{pmatrix}$$

Again, please confirm that $E_2 E_2^{-1} = I$. Now we may express the reduction of S to U (recall (2.6)) as

$$E_2 E_1 S = U$$

and the subsequent reconstitution by

$$S = LU, \quad \text{where} \quad L = E_1^{-1} E_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -1/2 & 1 & 0 \\ 0 & -2/3 & 1 \end{pmatrix}$$

One speaks of this representation as the **LU decomposition** of S. We have just observed that the inverse of a product is the product of the inverses in reverse order. Do you agree that $S^{-1} = U^{-1}L^{-1}$? And what do you think of the statement $S^{-1} = A^{-1}K^{-1}(A^T)^{-1}$?

LU decomposition is the preferred method of solution for the large linear systems that occur in practice. The decomposition is implemented in Matlab as

$$[\texttt{L U}] = \texttt{lu}(\texttt{S});$$

and in fact lies at the heart of Matlab 's blackslash command. To diagram its use, we write $Sx = f$ as $LUx = f$ and recognize that the latter is nothing more than a pair of triangular problems:

$$Lc = f \quad \text{and} \quad Ux = c,$$

that may be solved by forward and backward substitution respectively. This representation achieves its greatest advantage when one is asked to solve $Sx = f$ over a large class of f vectors. For example, if we wish to steadily increase the force, f_2, on mass 2, and track the resulting displacement we would be well served by

```
[L,U] = lu(S);
f = [1 1 1];
for j=1:100,
    f(2) = f(2) + j/100;
    x = U \ (L \ f);
    plot(x,'o')
end
```

You are correct in pointing out that we could have also just precomputed the inverse of S and then sequentially applied it in our for loop. The use of the inverse is, in general, considerably more costly in terms of both memory and operation counts. The exercises will give you a chance to see this for yourself.

2.2. A Small Planar Network

We move from uni-axial to biaxial elastic nets by first considering the swing in Figure 2.2.

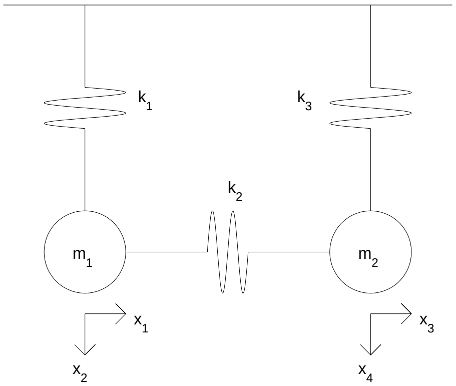


Figure 2.2. A simple swing.

We denote by x_1 and x_2 the respective horizontal and vertical displacements of m_1 (positive is right and down). Similarly, f_1 and f_2 will denote the associated components of force. The corresponding displacements and forces at m_2 will be denoted by $x_3,\ x_4$ and $f_3,\ f_4$. In computing the elongations of the three springs we shall make reference to their unstretched lengths, L_1, L_2, and L_3.

Now, if spring 1 connects $(0, -L_1)$ to $(0, 0)$ when at rest and $(0, -L_1)$ to (x_1, x_2) when stretched then its elongation is simply

$$e_1 = \sqrt{x_1^2 + (x_2 + L_1)^2} - L_1. \tag{2.8}$$

The price one pays for moving to higher dimensions is that lengths are now expressed in terms of square roots. The upshot is that the elongations are not linear combinations of the end displacements as they were in the uni-axial case. If we presume however that the loads and stiffnesses are matched in the sense that the displacements are small compared with the original lengths then we may effectively ignore the nonlinear contribution in (2.8). In order to make this precise we need only recall the Taylor development of $\sqrt{1+t}$ about $t = 0$, i.e.,

$$\sqrt{1+t} = 1 + t/2 + O(t^2)$$

where the latter term signifies the remainder. With regard to e_1 this allows

$$\begin{aligned} e_1 &= \sqrt{x_1^2 + x_2^2 + 2x_2L_1 + L_1^2} - L_1 \\ &= L_1\sqrt{1 + (x_1^2 + x_2^2)/L_1^2 + 2x_2/L_1} - L_1 \\ &= L_1 + (x_1^2 + x_2^2)/(2L_1) + x_2 + L_1O(((x_1^2 + x_2^2)/L_1^2 + 2x_2/L_1)^2) - L_1 \\ &= x_2 + (x_1^2 + x_2^2)/(2L_1) + L_1O(((x_1^2 + x_2^2)/L_1^2 + 2x_2/L_1)^2). \end{aligned}$$

If we now assume that

$$(x_1^2 + x_2^2)/(2L_1) \quad \text{is small compared to} \quad x_2 \tag{2.9}$$

then, as the O term is even smaller, we may neglect all but the first terms in the above and so arrive at

$$e_1 = x_2.$$

To take a concrete example, if L_1 is one meter and x_1 and x_2 are each one centimeter than x_2 is one hundred times $(x_1^2 + x_2^2)/(2L_1)$.

With regard to the second spring, arguing as above, its elongation is (approximately) its stretch along its initial direction. As its initial direction is horizontal, its elongation is just the difference of the respective horizontal end displacements, namely,

$$e_2 = x_3 - x_1.$$

Finally, the elongation of the third spring is (approximately) the difference of its respective vertical end displacements, i.e.,

$$e_3 = x_4.$$

We encode these three elongations in

$$e = Ax \quad \text{where} \quad A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Hooke's law is an elemental piece of physics and is not perturbed by our leap from uni-axial to biaxial structures. The upshot is that the restoring force in each spring is still proportional to its elongation, i.e., $y_j = k_j e_j$ where k_j is the stiffness of the jth spring. In matrix terms,

$$y = Ke \quad \text{where} \quad K = \begin{pmatrix} k_1 & 0 & 0 \\ 0 & k_2 & 0 \\ 0 & 0 & k_3 \end{pmatrix}.$$

Balancing horizontal and vertical forces at m_1 brings

$$-y_2 - f_1 = 0 \quad \text{and} \quad y_1 - f_2 = 0,$$

while balancing horizontal and vertical forces at m_2 brings

$$y_2 - f_3 = 0 \quad \text{and} \quad y_3 - f_4 = 0.$$

We assemble these into

$$By = f \quad \text{where} \quad B = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and recognize, as expected, that B is nothing more than A^T. Putting the pieces together, we find that x must satisfy $Sx = f$ where

$$S = A^T K A = \begin{pmatrix} k_2 & 0 & -k_2 & 0 \\ 0 & k_1 & 0 & 0 \\ -k_2 & 0 & k_2 & 0 \\ 0 & 0 & 0 & k_3 \end{pmatrix}.$$

Applying one step of Gaussian Elimination brings

$$\begin{pmatrix} k_2 & 0 & -k_2 & 0 \\ 0 & k_1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & k_3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_1 + f_3 \\ f_4 \end{pmatrix}$$

and back substitution delivers

$$\begin{aligned} x_4 &= f_4/k_3, \\ 0 &= f_1 + f_3, \\ x_2 &= f_2/k_1, \\ x_1 - x_3 &= f_1/k_2. \end{aligned}$$

The second of these is remarkable in that it contains no components of x. Instead, it provides a condition on f. In mechanical terms, it states that there can be no equilibrium unless the horizontal forces on the two masses are equal and opposite. Of course one could have observed this directly from the layout of the truss. In modern, three–dimensional structures with thousands of members meant to shelter or convey humans one should not however be satisfied with the 'visual' integrity of the structure. In particular, one desires a detailed description of all loads that can, and, especially, all loads that can not, be equilibrated by the proposed truss. In algebraic terms, given a matrix S one desires a characterization of (1) all those f for which $Sx = f$ possesses a solution and (2) all those f for which $Sx = f$ does not possess a solution. We provide such a characterization in Chapter 3 in our discussion of the *column space* of a matrix.

Supposing now that $f_1 + f_3 = 0$ we note that although the system above is consistent it still fails to uniquely determine the four components of x. In particular, it specifies only the difference between x_1 and x_3. As a result both

$$x = \begin{pmatrix} f_1/k_2 \\ f_2/k_1 \\ 0 \\ f_4/k_3 \end{pmatrix} \quad \text{and} \quad x = \begin{pmatrix} 0 \\ f_2/k_1 \\ -f_1/k_2 \\ f_4/k_3 \end{pmatrix}$$

satisfy $Sx = f$. In fact, one may add to either an arbitrary multiple of

$$z \equiv \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \tag{2.10}$$

and still have a solution of $Sx = f$. Searching for the source of this lack of uniqueness we observe some redundancies in the columns of S. In particular, the third is simply the opposite of the first. As S is simply A^TKA we recognize that the original fault lies with A, where again, the first and third columns are opposites. These redundancies are encoded in z in the sense that

$$Az = 0.$$

Interpreting this in mechanical terms, we view z as a displacement and Az as the resulting elongation. In $Az = 0$ we see a nonzero displacement producing zero elongation. One says in this case that the truss deforms without doing any work and speaks of z as an 'unstable mode.' Again, this mode could have been observed by a simple glance at Figure 2.2. Such is not the case for more complex structures and so the engineer seeks a systematic means by which *all* unstable modes may be identified. We shall see in Chapter 3 that these modes are captured by the *null space* of A.

From $Sz = 0$ one easily deduces that S is singular. More precisely, if S^{-1} were to exist then $S^{-1}Sz$ would equal $S^{-1}0$, i.e., $z = 0$, contrary to (2.10). As a result, Matlab will fail to solve $Sx = f$ even when f is a force that the truss can equilibrate. One way out is to use the pseudo–inverse, as we shall see below.

2.3. A Large Planar Network

We close with the (scalable) example of the larger planar net in Figure 2.3. Elastic fibers, numbered 1 – 20, meet at nodes, numbered 1 – 9. We limit our observation to the motion of the nodes by denoting the horizontal and vertical displacements of node j by x_{2j-1} and x_{2j} respectively. Retaining the convention that down and right are positive we note that the elongation of fiber 1 is

$$e_1 = x_2 - x_8$$

while that of fiber 3 is

$$e_3 = x_3 - x_1.$$

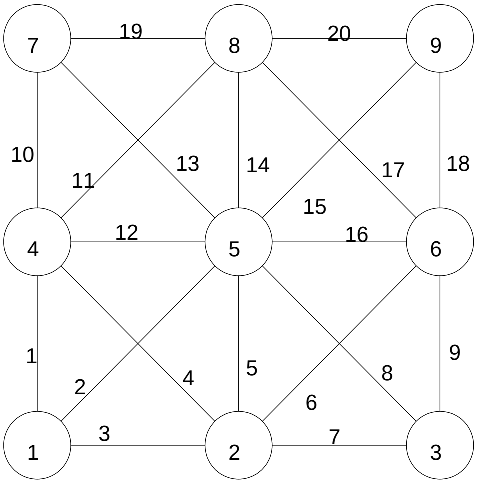


Figure 2.3. A crude tissue model.

As fibers 2 and 4 are neither vertical nor horizontal their elongations, in terms of nodal displacements, are not so easy to read off. This is more a nuisance than an obstacle however, for recalling our earlier discussion, the elongation is approximately just the stretch along its undeformed axis. With respect to fiber 2, as it makes the angle $-\pi/4$ with respect to the positive horizontal axis, we find

$$e_2 = (x_9 - x_1)\cos(-\pi/4) + (x_{10} - x_2)\sin(-\pi/4) = (x_9 - x_1 + x_2 - x_{10})/\sqrt{2}.$$

Similarly, as fiber 4 makes the angle $-3\pi/4$ with respect to the positive horizontal axis, its elongation is

$$e_4 = (x_7 - x_3)\cos(-3\pi/4) + (x_8 - x_4)\sin(-3\pi/4) = (x_3 - x_7 + x_4 - x_8)/\sqrt{2}.$$

These are both direct applications of the general formula

$$e_j = (x_{2n-1} - x_{2m-1})\cos(\theta_j) + (x_{2n} - x_{2m})\sin(\theta_j) \tag{2.11}$$

for fiber j, as depicted in the figure below, connecting node m to node n and making the angle θ_j with the positive horizontal axis when node m is assumed to lie at the point $(0, 0)$. The reader should check that our expressions for e_1 and e_3 indeed conform to this general formula and that e_2 and e_4 agree with ones intuition. For example, visual inspection of the specimen suggests that fiber 2 can not be supposed to stretch (i.e., have positive e_2) unless $x_9 > x_1$ and/or $x_2 > x_{10}$. Does this jibe with (2.11)?

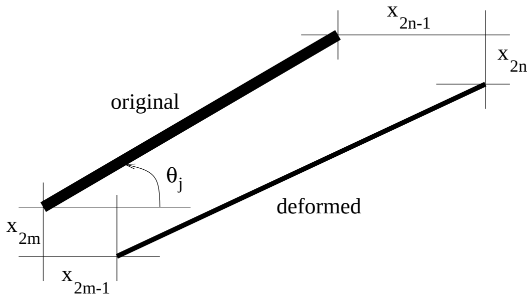


Figure 2.4. Elongation of a generic bar, see (2.11).

Applying (2.11) to each of the remaining fibers we arrive at $e = Ax$ where A is 20-by-18, one row for each fiber, and one column for each degree of freedom. For systems of such size with such a well defined structure one naturally hopes to automate the construction. We have done just that in the accompanying M-file and diary. The M-file begins with a matrix of raw data that anyone with a protractor could have keyed in directly from Figure 2.3. More precisely, the data matrix has a row for each fiber and each row consists of the starting and ending node numbers and the angle the fiber makes with the positive horizontal axis. This data is precisely what (2.11) requires in order to know which columns of A receive the proper cos or sin. The final A matrix is displayed in the diary.

The next two steps are now familiar. If K denotes the diagonal matrix of fiber stiffnesses and f denotes the vector of nodal forces then $y = Ke$ and $A^Ty = f$ and so one must solve $Sx = f$ where $S = A^TKA$. In this case there is an entire three–dimensional class of z for which $Az = 0$ and therefore $Sz = 0$. The three indicates that there are three independent unstable modes of the specimen, e.g., two translations and a rotation. As a result S is singular and `x = S\f` in Matlab will get us nowhere. The way out is to recognize that S has $18 - 3 = 15$ stable modes and that if we restrict S to 'act' only in these directions then it 'should' be invertible. We will begin to make these notions precise in Chapter 4 on the Fundamental Theorem of Linear Algebra.

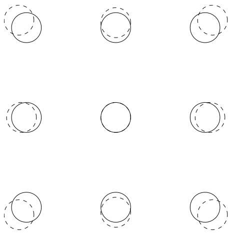


Figure 2.5. The solid(dashed) circles correspond to the nodal positions before(after) the application of the traction force, f.

For now let us note that every matrix possesses such a **pseudo-inverse** and that it may be computed in Matlab via the `pinv` command. On supposing the fiber stiffnesses to each be one and the edge traction to be of the form

$$f = [-1\ 1\ 0\ 1\ 1\ 1\ \ -1\ 0\ 0\ 0\ 1\ 0\ \ -1\ \ -1\ 0\ \ -1\ 1\ \ -1]^T,$$

we arrive at x via `x=pinv(S)*f` and refer to Figure 2.5 for its graphical representation.

2.4. Exercises

1. With regard to Figure 2.1, (i) Derive the A and K matrices resulting from the removal of the fourth spring (but not the third mass) and assemble $S = A^TKA$.

(ii) Compute S^{-1}, by hand via Gauss–Jordan, and compute L and U where $S = LU$ by hand via the composition of elimination matrices and their inverses. Assume throughout that with $k_1 = k_2 = k_3 = k$,

(iii) Use the result of (ii) with the load $f = [0\ 0\ F]^T$ to solve $Sx = f$ by hand two ways, i.e., $x = S^{-1}f$ and $Lc = f$ and $Ux = c$.

2. With regard to Figure 2.2

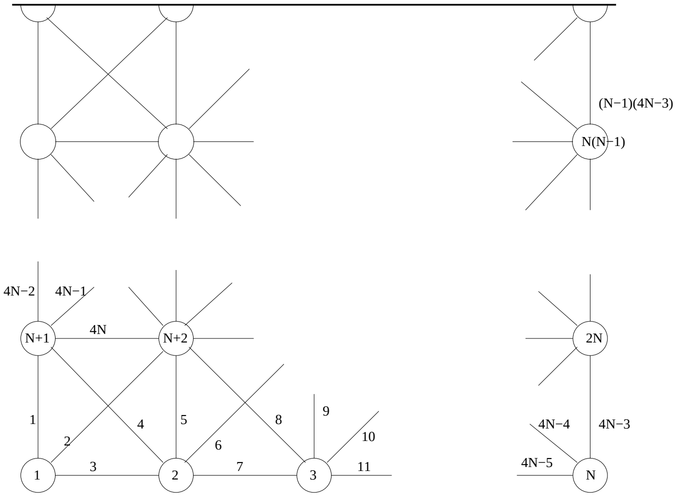
   (i) Derive the A and K matrices resulting from the addition of a fourth (diagonal) fiber that runs from the top of fiber one to the second mass and assemble $S = A^T KA$.

   (ii) Compute S^{-1}, by hand via Gauss–Jordan, and compute L and U where $S = LU$ by hand via the composition of elimination matrices and their inverses. Assume throughout that with $k_1 = k_2 = k_3 = k_4 = k$.

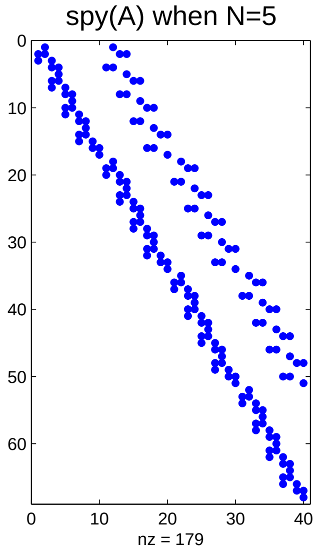
   (iii) Use the result of (ii) with the load $f = [0\ 0\ F\ 0]^T$ to solve $Sx = f$ by hand two ways, i.e., $x = S^{-1}f$ and $Lc = f$ and $Ux = c$.

3. Generalize figure 2.3 to the case of 16 nodes connected by 42 fibers. Introduce one stiff (say $k = 100$) fiber and show how to detect it by 'properly' choosing f. Submit your well-documented M-file as well as the plots, similar to Figure 2.5, from which you conclude the presence of a stiff fiber.

4. We generalize Figure 2.3 to permit ever finer meshes. In particular, with reference to the figure below we assume $N(N-1)$ nodes where the horizontal and vertical fibers each have length $1/N$ while the diagonal fibers have length $\sqrt{2}/N$. The top row of fibers is anchored to the ceiling.



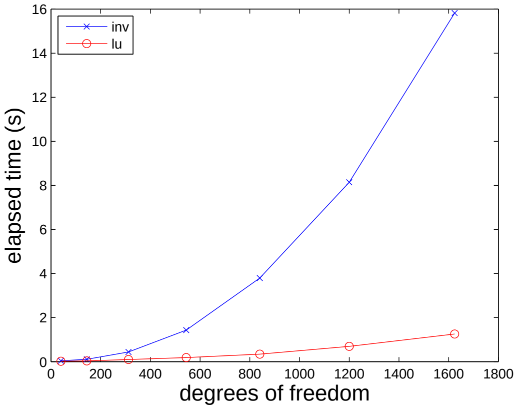
   (i) Write and test a Matlab function `S=bignet(N)` that accepts the **odd** number N and produces the stiffness matrix $S = A^T KA$. As a check on your work we offer a `spy` plot of A when $N = 5$. Your K matrix should reflect the fiber lengths as spelled out in (2.1). You may assume $Y_j a_j = 1$ for each fiber. The sparsity of A also produces a sparse S. In order to exploit this, please use `S=sparse(S)` as the final line in `bignet.m`.



(ii) Write and test a driver called `bigrun` that generates S for $N = 5:4:29$ and for each N solves $Sx = f$ two ways for 100 choices of f. In particular, f is a steady downward pull on the bottom set of nodes, with a continual increase on the pull at the center node. This can be done via `f=zeros(size(S,1),1); f(2:2:2*N) = 1e-3/N;`

```
for j=1:100,
f(N+1) = f(N+1) + 1e-4/N;
```

This construction should be repeated twice, with the code that closes §2.1 as your guide. In the first scenario, precompute S^{-1} via `inv` and then apply $x = S^{-1}f$ in the `j` loop. In the second scenario precompute L and U and then apply $x = U\backslash(L\backslash f)$ in the `j` loop. In both cases use `tic` and `toc` to time each for loop and so produce a graph of the form



Submit your well documented code, a spy plot of S when $N = 9$, and a time comparison like (will vary with memory and cpu) that shown above.

3. The Column and Null Spaces

3.1. The Column Space

We begin with the direct geometric interpretation of matrix–vector multiplication. Namely, the multiplication of the n-by-1 vector x by the m-by-n matrix S produces a linear combination of the columns of S. More precisely, if s_j denotes the jth column of S, then

$$Sx = [s_1 \; s_2 \; \cdots \; s_n] \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = x_1 s_1 + x_2 s_2 + \cdots + x_n s_n. \tag{3.1}$$

The picture I wish to place in your mind's eye is that Sx lies in the plane spanned by the columns of S. This plane occurs so frequently that we find it useful to distinguish it with a

Definition 3.1. The **column space** of the m-by-n matrix S is simply the span of its columns, i.e.,

$$\mathcal{R}(S) \equiv \{Sx : x \in \mathbb{R}^n\}.$$

This is a subset of $\mathbb{R}^m$. The letter $\mathcal{R}$ stands for range.

For example, let us recall the S matrix associated with Figure 2.2. Its column space is

$$\mathcal{R}(S) = \left\{ x_1 \begin{pmatrix} k_2 \\ 0 \\ -k_2 \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 0 \\ k_1 \\ 0 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} -k_2 \\ 0 \\ k_2 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} 0 \\ 0 \\ 0 \\ k_3 \end{pmatrix} : x \in \mathbb{R}^4 \right\}.$$

As the first and third columns are colinear we may write

$$\mathcal{R}(S) = \left\{ x_1 \begin{pmatrix} k_2 \\ 0 \\ -k_2 \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 0 \\ k_1 \\ 0 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} 0 \\ 0 \\ 0 \\ k_3 \end{pmatrix} : x \in \mathbb{R}^3 \right\}.$$

As the remaining three columns are linearly independent we may go no further. We 'recognize' then $\mathcal{R}(S)$ as a three dimensional subspace of $\mathbb{R}^4$. In order to use these ideas with any confidence we must establish careful definitions of subspace, independence, and dimension.

A subspace is a natural generalization of line and plane. Namely, it is any set that is closed under vector addition and scalar multiplication. More precisely,

Definition 3.2. A subset M of $\mathbb{R}^d$ is a **subspace** of $\mathbb{R}^d$ when
(1) $p + q \in M$ whenever $p \in M$ and $q \in M$, and
(2) $tp \in M$ whenever $p \in M$ and $t \in \mathbb{R}$.

Let us confirm now that $\mathcal{R}(S)$ is indeed a subspace. Regarding (1) if $p \in \mathcal{R}(S)$ and $q \in \mathcal{R}(S)$ then $p = Sx$ and $q = Sy$ for some x and y. Hence, $p+q = Sx+Sy = S(x+y)$, i.e., $(p+q) \in \mathcal{R}(S)$. With respect to (2), $tp = tSx = S(tx)$ so $tp \in \mathcal{R}(S)$.

This establishes that every column space is a subspace. The converse is also true. Every subspace is the column space of some matrix. To make sense of this we should more carefully explain what we mean by 'span'.

Definition 3.3. A collection of vectors $\{s_1, s_2, \ldots, s_n\}$ in a subspace M is said to **span** M when $M = \mathcal{R}(S)$ where $S = [s_1\ s_2\ \cdots\ s_n]$.

We shall be interested in how a subspace is 'situated' in its ambient space. We shall have occasion to speak of complementary subspaces and even the sum of two subspaces. Lets take care of the latter right now,

Definition 3.4. If M and Q are subspaces of the same ambient space, $\mathbb{R}^d$, we define their **direct sum**

$$M \oplus Q \equiv \{p + q : p \in M \text{ and } q \in Q\}$$

as the union of all possible sums of vectors from M and Q.

Do you see how $\mathbb{R}^3$ may be written as the direct sum of $\mathbb{R}^1$ and $\mathbb{R}^2$?

3.2. The Null Space

Definition 3.5. The **null space** of an m-by-n matrix S is the collection of those vectors in $\mathbb{R}^n$ that S maps to the zero vector in $\mathbb{R}^m$. More precisely,

$$\mathcal{N}(S) \equiv \{x \in \mathbb{R}^n : Sx = 0\}.$$

Let us confirm that $\mathcal{N}(S)$ is in fact a subspace. If both x and y lie in $\mathcal{N}(S)$ then $Sx = Sy = 0$ and so $S(x + y) = 0$. In addition, $S(tx) = tSx = 0$ for every $t \in \mathbb{R}$.

As an example we remark that the null space of the S matrix associated with Figure 2.2 is

$$\mathcal{N}(S) = \left\{ t \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} : t \in \mathbb{R} \right\},$$

a line in $\mathbb{R}^4$.

The null space answers the question of uniqueness of solutions to $Sx = f$. For, if $Sx = f$ and $Sy = f$ then $S(x - y) = Sx - Sy = f - f = 0$ and so $(x - y) \in \mathcal{N}(S)$. Hence, a solution to $Sx = f$ will be unique if, and only if, $\mathcal{N}(S) = \{0\}$.

Recalling (3.1) we note that if $x \in \mathcal{N}(S)$ and $x \neq 0$, say, e.g., $x_1 \neq 0$, then $Sx = 0$ takes the form

$$s_1 = -\sum_{j=2}^{n} \frac{x_j}{x_1} s_j.$$

That is, the first column of S may be expressed as a linear combination of the remaining columns of S. Hence, one may determine the (in)dependence of a set of vectors by examining the null space of the matrix whose columns are the vectors in question.

Definition 3.6. The vectors $\{s_1, s_2, \ldots, s_n\}$ are said to be **linearly independent** if $\mathcal{N}(S) = \{0\}$ where $S = [s_1\ s_2\ \cdots\ s_n]$.

As lines and planes are described as the set of linear combinations of one or two generators, so too subspaces are most conveniently described as the span of a few basis vectors.

Definition 3.7. A collection of vectors $\{s_1, s_2, \ldots, s_n\}$ in a subspace M is a **basis** for M when the matrix $S = [s_1\ s_2\ \cdots\ s_n]$ satisfies

(1) $M = \mathcal{R}(S)$, and
(2) $\mathcal{N}(S) = \{0\}$.

The first stipulates that the columns of S span M while the second requires the columns of S to be linearly independent.

3.3. A Blend of Theory and Example

Let us compute bases for the null and column spaces of the adjacency matrix associated with the ladder below

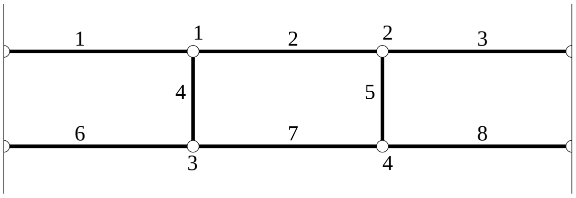


Figure 3.1. An unstable ladder?

The ladder has 8 bars and 4 nodes, so 8 degrees of freedom. Continuing to denote the horizontal and vertical displacements of node j by x_{2j-1} and x_{2j} we arrive at the A matrix

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \end{pmatrix}$$

To determine a basis for $\mathcal{R}(A)$ we must find a way to discard its dependent columns. A moment's reflection reveals that columns 2 and 6 are colinear, as are columns 4 and 8. We seek, of course, a more systematic means of uncovering these, and perhaps other less obvious, dependencies. Such dependencies are more easily discerned from the row reduced form

$$A_{\text{red}} = \texttt{rref(A)} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Recall that `rref` performs the elementary row operations necessary to eliminate all nonzeros below the diagonal. For those who can't stand to miss any of the action I recommend `rrefmovie`.

Each nonzero row of A_{red} is called a pivot row. The first nonzero in each row of A_{red} is called a pivot. Each column that contains a pivot is called a pivot column. On account of the staircase nature of A_{red} we find that there are as many pivot columns as there are pivot rows. In our example there are six of each and, again on account of the staircase nature, the pivot columns are **the** linearly independent columns of A_{red}. One now asks how this might help us distinguish the independent columns of A. For, although the rows of A_{red} are linear combinations of the rows of A no such thing is true with respect to the columns. The answer is: pay attention only to the indices of the pivot columns. In our example, columns $\{1, 2, 3, 4, 5, 7\}$ are the pivot columns. In general

Proposition 3.1. Suppose A is m-by-n. If columns $\{c_j : j = 1, \dots, r\}$ are the pivot columns of A_{red} then columns $\{c_j : j = 1, \dots, r\}$ of A constitute a basis for $\mathcal{R}(A)$.

Proof: Note that the pivot columns of A_{red} are, by construction, linearly independent. Suppose, however, that columns $\{c_j : j = 1, \dots, r\}$ of A are linearly dependent. In this case there exists a nonzero $x \in \mathbb{R}^n$ for which $Ax = 0$ and

$$x_k = 0, \quad k \notin \{c_j : j = 1, \dots, r\}. \tag{3.2}$$

Now $Ax = 0$ necessarily implies that $A_{\text{red}}x = 0$, contrary to the fact that columns $\{c_j : j = 1, \dots, r\}$ are the pivot columns of A_{red}. (The implication $Ax = 0 \Rightarrow A_{\text{red}}x = 0$ follows from the fact that we may read row reduction as a sequence of linear transformations of A. If we denote the product of these transformations by T then $TA = A_{\text{red}}$ and you see why $Ax = 0 \Rightarrow A_{\text{red}}x = 0$. The reverse implication follows from the fact that each of our row operations is reversible, or, in the language of the land, invertible.)

We now show that the span of columns $\{c_j : j = 1, \dots, r\}$ of A indeed coincides with $\mathcal{R}(A)$. This is obvious if $r = n$, i.e., if *all* of the columns are linearly independent. If $r < n$ there exists a $q \notin \{c_j : j = 1, \dots, r\}$. Looking back at A_{red} we note that its qth column is a linear combination of the pivot columns with indices not exceeding q. Hence, there exists an x satisfying (3.2) and $A_{\text{red}}x = 0$ and $x_q = 1$. This x then necessarily satisfies $Ax = 0$. This states that the qth column of A is a linear combination of columns $\{c_j : j = 1, \dots, r\}$ of A. End of Proof.

Let us now exhibit a basis for $\mathcal{N}(A)$. We exploit the already mentioned fact that $\mathcal{N}(A) = \mathcal{N}(A_{\text{red}})$. Regarding the latter, we partition the elements of x into so called **pivot** variables,

$$\{x_{c_j} : j = 1, \dots, r\}$$

and **free** variables

$$\{x_k : k \notin \{c_j : j = 1, \dots, r\}\}.$$

There are evidently $n - r$ free variables. For convenience, let us denote these in the future by

$$\{x_{c_j} : j = r + 1, \dots, n\}.$$

One solves $A_{\text{red}}x = 0$ by expressing each of the pivot variables in terms of the nonpivot, or free, variables. In the example above, x_1, x_2, x_3, x_4, x_5 and x_7 are pivot while x_6 and x_8 are free. Solving for the pivot in terms of the free we find

$$x_7 = 0,\ x_5 = 0,\ x_4 = x_8,\ x_3 = 0,\ x_2 = x_6,\ x_1 = 0,$$

or, written as a vector,

$$x = x_6 \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_8 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \tag{3.3}$$

where x_6 and x_8 are free. As x_6 and x_8 range over all real numbers the x above traces out a plane in $\mathbb{R}^8$. This plane is precisely the null space of A and (3.3) describes a generic element as the linear combination of two basis vectors. Compare this to what MATLAB returns when faced with `null(A,'r')`. Abstracting these calculations we arrive at

Proposition 3.2. Suppose that A is m-by-n with pivot indices $\{c_j : j = 1, \ldots, r\}$ and free indices $\{c_j : j = r + 1, \ldots, n\}$. A basis for $\mathcal{N}(A)$ may be constructed of $n - r$ vectors $\{z_1, z_2, \ldots, z_{n-r}\}$ where z_k, and only z_k, possesses a nonzero in its c_{r+k} component.

With respect to our ladder the free indices are $c_7 = 6$ and $c_8 = 8$. You still may be wondering what $\mathcal{R}(A)$ and $\mathcal{N}(A)$ tell us about the ladder that did not already know. Regarding $\mathcal{R}(A)$ the answer will come in the next chapter. The null space calculation however has revealed two independent motions against which the ladder does no work! Do you see that the two vectors in (3.3) encode rigid vertical motions of bars 4 and 5 respectively? As each of these lies in the null space of A the associated elongation is zero. Can you square this with the ladder as pictured in figure 3.1? I hope not, for vertical motion of bar 4 must 'stretch' bars 1,2,6 and 7. How does one resolve this (apparent) contradiction?

We close a few more examples. We compute bases for the column and null spaces of

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

Subtracting the first row from the second lands us at

$$A_{\text{red}} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}$$

hence both rows are pivot rows and columns 1 and 2 are pivot columns. Proposition 3.1 then informs us that the first two columns of A, namely

$$\left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\} \tag{3.4}$$

comprise a basis for $\mathcal{R}(A)$. In this case, $\mathcal{R}(A) = \mathbb{R}^2$.

Regarding $\mathcal{N}(A)$ we express each row of $A_{\text{red}}x = 0$ as the respective pivot variable in terms of the free. More precisely, x_1 and x_2 are pivot variables and x_3 is free and $A_{\text{red}}x = 0$ reads

$$\begin{aligned} x_1 + x_2 &= 0 \\ -x_2 + x_3 &= 0 \end{aligned}$$

Working from the bottom up we find

$$x_2 = x_3 \quad \text{and} \quad x_1 = -x_3$$

and hence every vector in the null space is of the form

$$x = x_3 \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}.$$

In other words

$$\mathcal{N}(A) = \left\{ x_3 \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} : x_3 \in \mathbb{R} \right\}$$

and

$$\begin{pmatrix}-1\\1\\1\end{pmatrix}$$

constitutes a basis for $\mathcal{N}(A)$.

Let us stretch this example a bit and see what happens. In particular, we append a new column and arrive at

$$B = \begin{pmatrix}1 & 1 & 0 & 2\\1 & 0 & 1 & 3\end{pmatrix}.$$

The column space of A was already the 'whole' space and so adding a column changes, with respect to $\mathcal{R}(A)$, nothing. That is, $\mathcal{R}(B) = \mathcal{R}(A)$ and (3.4) is a basis for $\mathcal{R}(B)$.

Regarding $\mathcal{N}(B)$ we again subtract the first row from the second,

$$B_{\text{red}} = \begin{pmatrix}1 & 1 & 0 & 2\\0 & -1 & 1 & 1\end{pmatrix}$$

and identify x_1 and x_2 as pivot variables and x_3 and x_4 as free. We see that $B_{\text{red}}x = 0$ means

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

or, equivalently,

$$x_2 = x_3 + x_4 \quad \text{and} \quad x_1 = -x_3 - 3x_4$$

and so

$$\mathcal{N}(B) = \left\{ x_3 \begin{pmatrix}-1\\1\\1\\0\end{pmatrix} + x_4 \begin{pmatrix}-3\\1\\0\\1\end{pmatrix} : x_3 \in \mathbb{R},\ x_4 \in \mathbb{R} \right\}$$

and

$$\left\{ \begin{pmatrix}-1\\1\\1\\0\end{pmatrix}, \begin{pmatrix}-3\\1\\0\\1\end{pmatrix} \right\}$$

constitutes a basis for $\mathcal{N}(B)$.

The number of pivots, r, of an m-by-n matrix A appears to be an important indicator. We shall refer to it from now on as the **rank** of A. Our canonical bases for $\mathcal{R}(A)$ and $\mathcal{N}(A)$ possess r and $n - r$ elements respectively. The number of elements in a basis for a subspace is typically called the **dimension** of the subspace.

3.4. Exercises

1. Which of the following subsets of $\mathbf{R}^3$ are actually subspaces? Check both conditions in definition 2 and show your work.

   (a) All vectors whose first component $x_1 = 0$.

   (b) All vectors whose first component $x_1 = 1$.

(c) All vectors whose first two components obey $x_1x_2 = 0$.

(d) The vector $(0, 0, 0)$.

(e) All linear combinations of the pair $(1, 1, 0)$ and $(2, 0, 1)$.

(f) All vectors for which $x_3 - x_2 + 3x_1 = 0$.

2. I encourage you to use `rref` and `null` for the following. (i) Add a diagonal crossbar between nodes 3 and 2 in Figure 3.1 and compute bases for the column and null spaces of the new adjacency matrix. As this crossbar fails to stabilize the ladder we shall add one more bar. (ii) To the 9 bar ladder of (i) add a diagonal cross bar between nodes 1 and the left end of bar 6. Compute bases for the column and null spaces of the new adjacency matrix.

3. We wish to show that $\mathcal{N}(A) = \mathcal{N}(A^TA)$ regardless of A.

   (i) We first take a concrete example. Report the findings of `null` when applied to A and A^TA for the A matrix associated with Figure 3.1.

   (ii) For arbitrary A show that $\mathcal{N}(A) \subset \mathcal{N}(A^TA)$, i.e., that if $Ax = 0$ then $A^TAx = 0$.

   (iii) For arbitrary A show that $\mathcal{N}(A^TA) \subset \mathcal{N}(A)$, i.e., that if $A^TAx = 0$ then $Ax = 0$. (Hint: if $A^TAx = 0$ then $x^TA^TAx = 0$ and says something about $\|Ax\|$, recall that $\|y\|^2 \equiv y^Ty$.)

4. Suppose that A is m-by-n and that $\mathcal{N}(A) = \mathbb{R}^n$. Argue that A must be the zero matrix.

5. Suppose that both $\{s_1, \ldots, s_n\}$ and $\{t_1, \ldots, t_m\}$ are both bases for the subspace M. Prove that $m = n$ and hence that our notion of dimension makes sense.

4. The Fundamental Theorem of Linear Algebra

The previous chapter, in a sense, only told half of the story. In particular, an m-by-n matrix A maps $\mathbb{R}^n$ into $\mathbb{R}^m$ and its null space lies in $\mathbb{R}^n$ and its column space lies in $\mathbb{R}^m$. Having seen examples where $\mathcal{R}(A)$ was a proper subspace of $\mathbb{R}^m$ one naturally asks about what is left out. Similarly, one wonders about the subspace of $\mathbb{R}^n$ that is complimentary to $\mathcal{N}(A)$. These questions are answered by the column space and null space of A^T.

4.1. The Row Space

As the columns of A^T are simply the rows of A we call $\mathcal{R}(A^T)$ the row space of A. More precisely

Definition 4.1. The **row space** of the m-by-n matrix A is simply the span of its rows, i.e.,

$$\mathcal{R}(A^T) \equiv \{A^T y : y \in \mathbb{R}^m\}.$$

This is a subspace of $\mathbb{R}^n$.

Regarding a basis for $\mathcal{R}(A^T)$ we recall that the rows of $A_{\text{red}} \equiv$`rref(A)` are merely linear combinations of the rows of A and hence

$$\mathcal{R}(A^T) = \mathcal{R}((A_{\text{red}})^T).$$

Recalling that pivot rows of A_{red} are linearly independent and that all remaining rows of A_{red} are zero leads us to

Proposition 4.1. Suppose A is m-by-n. The pivot rows of A_{red} constitute a basis for $\mathcal{R}(A^T)$.

As there are r pivot rows of A_{red} we find that the dimension of $\mathcal{R}(A^T)$ is r. Recalling Proposition 2.2 we find the dimensions of $\mathcal{N}(A)$ and $\mathcal{R}(A^T)$ to be complementary, i.e., they sum to the dimension of the ambient space, n. Much more in fact is true. Let us compute the dot product of an arbitrary element $x \in \mathcal{R}(A^T)$ and $z \in \mathcal{N}(A)$. As $x = A^T y$ for some y we find

$$x^T z = (A^T y)^T z = y^T A z = 0.$$

This states that every vector in $\mathcal{R}(A^T)$ is perpendicular to every vector in $\mathcal{N}(A)$.

Let us test this observation on the A matrix stemming from the unstable ladder of §3.4. Recall that

$$z_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad z_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

constitute a basis for $\mathcal{N}(A)$ while the pivot rows of $A_{\rm red}$ are

$$x_1 = \begin{pmatrix}1\\0\\0\\0\\0\\0\\0\\0\end{pmatrix}, \quad x_2 = \begin{pmatrix}0\\1\\0\\0\\0\\-1\\0\\0\end{pmatrix}, \quad x_3 = \begin{pmatrix}0\\0\\1\\0\\0\\0\\0\\0\end{pmatrix},$$

and

$$x_4 = \begin{pmatrix}0\\0\\0\\1\\0\\0\\0\\-1\end{pmatrix}, \quad x_5 = \begin{pmatrix}0\\0\\0\\0\\1\\0\\0\\0\end{pmatrix}, \quad x_6 = \begin{pmatrix}0\\0\\0\\0\\0\\0\\1\\0\end{pmatrix}.$$

Indeed, each z_j is perpendicular to each x_k. As a result,

$$\{z_1, z_2, x_1, x_2, x_3, x_4, x_5, x_6\}$$

comprises a set of 8 linearly independent vectors in $\mathbb{R}^8$. These vectors then necessarily span $\mathbb{R}^8$. For, if they did not, there would exist nine linearly independent vectors in $\mathbb{R}^8$! In general, we find

Fundamental Theorem of Linear Algebra (Preliminary). Suppose A is m-by-n and has rank r. The row space, $\mathcal{R}(A^T)$, and the null space, $\mathcal{N}(A)$, are respectively r and $n-r$ dimensional subspaces of $\mathbb{R}^n$. Each $x \in \mathbb{R}^n$ may be uniquely expressed in the form

$$x = x_R + x_N, \quad \text{where} \quad x_R \in \mathcal{R}(A^T) \quad \text{and} \quad x_N \in \mathcal{N}(A). \tag{4.1}$$

Recalling Definition 4 from Chapter 3 we may interpret (4.1) as

$$\mathbb{R}^n = \mathcal{R}(A^T) \oplus \mathcal{N}(A).$$

As the constituent subspaces have been shown to be orthogonal we speak of $\mathbb{R}^n$ as the **orthogonal direct sum** of $\mathcal{R}(A^T)$ and $\mathcal{N}(A)$.

4.2. The Left Null Space

The Fundamental Theorem will more than likely say that $\mathbb{R}^m = \mathcal{R}(A) \oplus \mathcal{N}(A^T)$. In fact, this is already in the preliminary version. To coax it out we realize that there was nothing special about the choice of letters used. Hence, if B is p-by-q then the preliminary version states that $\mathbb{R}^q = \mathcal{R}(B^T) \oplus \mathcal{N}(B)$. As a result, letting $B = A^T$, $p = n$ and $q = m$, we find indeed $\mathbb{R}^m = \mathcal{R}(A) \oplus \mathcal{N}(A^T)$. That is, the **left null space**, $\mathcal{N}(A^T)$, is the orthogonal complement of the column space, $\mathcal{R}(A)$. The word 'left' stems from the fact that $A^Ty = 0$ is equivalent to $y^TA = 0$, where y 'acts' on A from the left.

In order to compute a basis for $\mathcal{N}(A^T)$ we merely mimic the construction of the previous section. Namely, we compute $(A^T)_{\rm red}$ and then solve for the pivot variables in terms of the free ones.

With respect to the A matrix associated with the unstable ladder of §3.4, we find

$$A^T = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

and

$$(A^T)_{\rm red} = \texttt{rref}(A^T) = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We recognize the rank of A^T to be 6, with pivot and free indices

$$\{1,2,4,5,6,7\} \quad \text{and} \quad \{3,8\}$$

respectively. Solving $(A^T)_{\rm red}x = 0$ for the pivot variables in terms of the free we find

$$x_7 = x_8,\ x_6 = x_8,\ x_5 = 0,\ x_4 = 0,\ x_2 = x_3,\ x_1 = x_3,$$

or in vector form,

$$x = x_3 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + x_8 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

These two vectors constitute a basis for $\mathcal{N}(A^T)$ and indeed they are both orthogonal to every column of A. We have now exhibited means by which one may assemble bases for the four fundamental subspaces. In the process we have established

Fundamental Theorem of Linear Algebra. Suppose A is m-by-n and has rank r. One has the orthogonal direct sums

$$\mathbb{R}^n = \mathcal{R}(A^T) \oplus \mathcal{N}(A) \quad \text{and} \quad \mathbb{R}^m = \mathcal{R}(A) \oplus \mathcal{N}(A^T)$$

where the dimensions are

$$\begin{aligned}\dim\mathcal{R}(A) = \dim\mathcal{R}(A^T) &= r\\ \dim\mathcal{N}(A) &= n - r\\ \dim\mathcal{N}(A^T) &= m - r.\end{aligned}$$

We shall see many applications of this fundamental theorem. Perhaps one of the most common is the use of the orthogonality of $\mathcal{R}(A)$ and $\mathcal{N}(A^T)$ in the characterization of those b for which an x exists for which $Ax = b$. There are many instances for which $\mathcal{R}(A)$ is quite large and unwieldy while $\mathcal{N}(A^T)$ is small and therefore simpler to grasp. As an example, consider the $(n-1)$-by-n 'first order difference' matrix with -1 on the diagonal and 1 on the super diagonal,

$$A = \begin{pmatrix} -1 & 1 & 0 & 0 & \cdot\\ 0 & -1 & 1 & 0 & \cdot\\ \cdot & \cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & 0 & -1 & 1 \end{pmatrix}$$

It is not difficult to see that $\mathcal{N}(A^T) = \{0\}$ and so, $\mathcal{R}(A)$, being its orthogonal complement, is the entire space, $\mathbb{R}^{n-1}$. That is, for each $b \in \mathbb{R}^{n-1}$ there exists an $x \in \mathbb{R}^n$ such that $Ax = b$. The uniqueness of such an x is decided by $\mathcal{N}(A)$. We recognize this latter space as the span of the vector of ones. But you already knew that adding a constant to a function does not change its derivative.

4.3. Exercises

1. True or false: support your answer.
   (i) If A is square then $\mathcal{R}(A) = \mathcal{R}(A^T)$.
   (ii) If A and B have the same four fundamental subspaces then A=B.
2. Construct bases (by hand) for the four subspaces associated with
   $$A = \begin{pmatrix} 1 & 1 & -1\\ 1 & 0 & -1 \end{pmatrix}.$$
   Also provide a careful sketch of these subspaces.
3. Show that if $AB = 0$ then $\mathcal{R}(B) \subset \mathcal{N}(A)$.
4. Why is there no matrix whose row space and null space both contain the vector $[1\ 1\ 1]^T$?
5. Write down a matrix with the required property or explain why no such matrix exists.
   (a) Column space contains $[1\ 0\ 0]^T$ and $[0\ 0\ 1]^T$ while row space contains $[1\ 1]^T$ and $[1\ 2]^T$.
   (b) Column space has basis $[1\ 1\ 1]^T$ while null space has basis $[1\ 2\ 1]^T$.
   (c) Column space is $\mathbb{R}^4$ while row space is $\mathbb{R}^3$.
6. One often constructs matrices via outer products, e.g., given a n-by-1 vector v let us consider $A = vv^T$.
   (a) Show that v is a basis for $\mathcal{R}(A)$,
   (b) Show that $\mathcal{N}(A)$ coincides with all vectors perpendicular to v.
   (c) What is the rank of A?

5. Least Squares

We learned in the previous chapter that $Ax = b$ need not possess a solution when the number of rows of A exceeds its rank, i.e., $r < m$. As this situation arises quite often in practice, typically in the guise of 'more equations than unknowns,' we establish a rationale for the absurdity $Ax = b$.

5.1. The Normal Equations

The goal is to choose x such that Ax is as close as possible to b. Measuring closeness in terms of the sum of the squares of the components we arrive at the 'least squares' problem of minimizing

$$\|Ax - b\|^2 \equiv (Ax - b)^T(Ax - b) \tag{5.1}$$

over all $x \in \mathbb{R}^n$. The path to the solution is illuminated by the Fundamental Theorem. More precisely, we write

$$b = b_R + b_N \quad \text{where} \quad b_R \in \mathcal{R}(A) \quad \text{and} \quad b_N \in \mathcal{N}(A^T).$$

On noting that (i) $(Ax - b_R) \in \mathcal{R}(A)$ for every $x \in \mathbb{R}^n$ and (ii) $\mathcal{R}(A) \perp \mathcal{N}(A^T)$ we arrive at the Pythagorean Theorem

$$\|Ax - b\|^2 = \|Ax - b_R - b_N\|^2 = \|Ax - b_R\|^2 + \|b_N\|^2, \tag{5.2}$$

It is now clear from (5.2) that the best x is the one that satisfies

$$Ax = b_R. \tag{5.3}$$

As $b_R \in \mathcal{R}(A)$ this equation indeed possesses a solution. We have yet however to specify how one computes b_R given b. Although an explicit expression for b_R, the so called **orthogonal projection** of b onto $\mathcal{R}(A)$, in terms of A and b is within our grasp we shall, strictly speaking, not require it. To see this, let us note that if x satisfies (5.3) then

$$Ax - b = Ax - b_R - b_N = -b_N. \tag{5.4}$$

As b_N is no more easily computed than b_R you may claim that we are just going in circles. The 'practical' information in (5.4) however is that $(Ax - b) \in \mathcal{N}(A^T)$, i.e., $A^T(Ax - b) = 0$, i.e.,

$$A^TAx = A^Tb. \tag{5.5}$$

As $A^Tb \in \mathcal{R}(A^T)$ regardless of b this system, often referred to as the **normal equations**, indeed has a solution. This solution is unique so long as the columns of A^TA are linearly independent, i.e., so long as $\mathcal{N}(A^TA) = \{0\}$. Recalling Chapter 2, Exercise 2, we note that this is equivalent to $\mathcal{N}(A) = \{0\}$. We summarize our findings in

Proposition 5.1. The set of $x \in \mathbb{R}^n$ for which the misfit $\|Ax - b\|^2$ is smallest is composed of those x for which

$$A^TAx = A^Tb.$$

There is always at least one such x.
There is exactly one such x iff $\mathcal{N}(A) = \{0\}$.

As a concrete example, suppose with reference to the figure below that

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

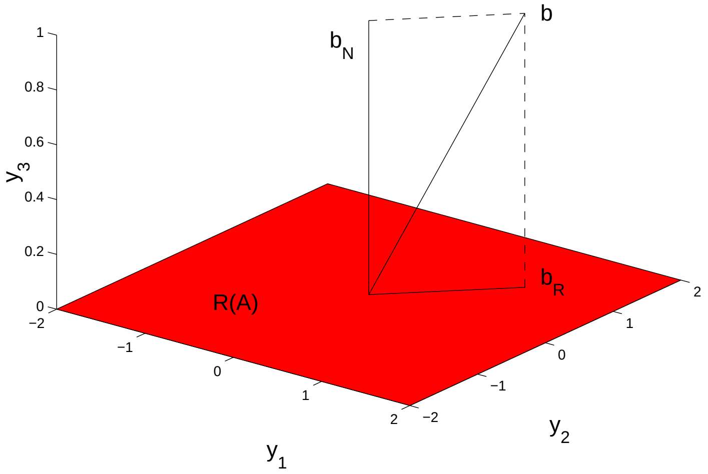


Figure 5.1. The decomposition of b.

As $b \neq \mathcal{R}(A)$ there is no x such that $Ax = b$. Indeed,

$$\|Ax - b\|^2 = (x_1 + x_2 - 1)^2 + (x_2 - 1)^2 + 1 \geq 1,$$

with the minimum uniquely attained at

$$x = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

in agreement with the unique solution of (5.5), for

$$A^T A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \quad \text{and} \quad A^T b = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

We now recognize, *a posteriori*, that

$$b_R = Ax = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

is the orthogonal projection of b onto the column space of A.

5.2. Applying Least Squares to the Biaxial Test Problem

We shall formulate the identification of the 20 fiber stiffnesses in Figure 2.3, as a least squares problem. We envision loading, f, the 9 nodes and measuring the associated 18 displacements, x. From knowledge of x and f we wish to infer the components of $K = \texttt{diag}(k)$ where k is the vector of unknown fiber stiffnesses. The first step is to recognize that

$$A^T K A x = f$$

may be written as

$$Bk = f \quad \text{where} \quad B = A^T \texttt{diag}(Ax). \tag{5.6}$$

Though conceptually simple this is not of great use in practice, for B is 18-by-20 and hence (5.6) possesses many solutions. The way out is to compute k as the result of more than one experiment. We shall see that, for our small sample, 2 experiments will suffice.

To be precise, we suppose that $x^{(1)}$ is the displacement produced by loading $f^{(1)}$ while $x^{(2)}$ is the displacement produced by loading $f^{(2)}$. We then piggyback the associated pieces in

$$B = \begin{pmatrix} A^T\mathtt{diag}(Ax^{(1)}) \\ A^T\mathtt{diag}(Ax^{(2)}) \end{pmatrix} \quad \text{and} \quad f = \begin{pmatrix} f^{(1)} \\ f^{(2)} \end{pmatrix}.$$

This B is 36-by-20 and so the system $Bk = f$ is overdetermined and hence ripe for least squares.

We proceed then to assemble B and f. We suppose $f^{(1)}$ and $f^{(2)}$ to correspond to horizontal and vertical stretching

$$\begin{aligned} f^{(1)} &= [-1\ 0\ 0\ 0\ 1\ 0\ -1\ 0\ 0\ 0\ 1\ 0\ -1\ 0\ 0\ 0\ 1\ 0]^T \\ f^{(2)} &= [0\ 1\ 0\ 1\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ -1\ 0\ -1\ 0\ -1]^T \end{aligned}$$

respectively. For the purpose of our example we suppose that each $k_j = 1$ except $k_8 = 5$. We assemble A^TKA as in Chapter 2 and solve

$$A^TKAx^{(j)} = f^{(j)}$$

with the help of the pseudoinverse. In order to impart some 'reality' to this problem we taint each $x^{(j)}$ with 10 percent noise prior to constructing B. Please see the attached M–file for details. Regarding

$$B^TBk = B^Tf$$

we note that Matlab solves this system when presented with `k=B\f` when B is rectangular. We have plotted the results of this procedure in the figure below

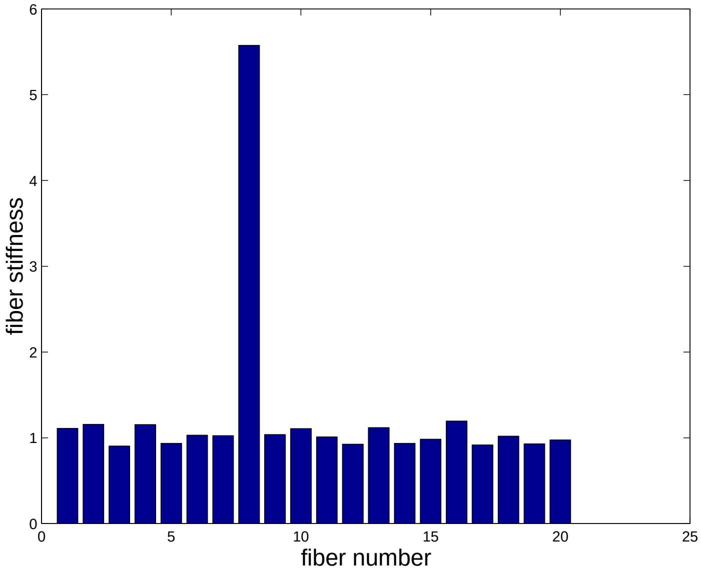


Figure 5.2. Results of a successful biaxial test.

The stiff fiber is readily identified.

5.3. Projections

From an algebraic point of view (5.5) is an elegant reformulation of the least squares problem. Though easy to remember it unfortunately obscures the geometric content, suggested by the word 'projection,' of (5.4). As projections arise frequently in many applications we pause here to develop them more carefully.

With respect to the normal equations we note that if $\mathcal{N}(A) = \{0\}$ then

$$x = (A^T A)^{-1} A^T b$$

and so the orthogonal projection of b onto $\mathcal{R}(A)$ is

$$b_R = Ax = A(A^T A)^{-1} A^T b. \tag{5.7}$$

Defining

$$P = A(A^T A)^{-1} A^T, \tag{5.8}$$

(5.7) takes the form $b_R = Pb$. Commensurate with our notion of what a 'projection' should be we expect that P map vectors not in $\mathcal{R}(A)$ onto $\mathcal{R}(A)$ while leaving vectors already in $\mathcal{R}(A)$ unscathed. More succinctly, we expect that $Pb_R = b_R$, i.e., $PPb = Pb$. As the latter should hold for all $b \in \mathbb{R}^m$ we expect that

$$P^2 = P. \tag{5.9}$$

With respect to (5.8) we find that indeed

$$P^2 = A(A^T A)^{-1} A^T A (A^T A)^{-1} A^T = A(A^T A)^{-1} A^T = P.$$

We also note that the P in (5.8) is symmetric. We dignify these properties through

Definition 5.1. A matrix P that satisfies $P^2 = P$ is called a **projection**. A symmetric projection is called an **orthogonal projection**.

We have taken some pains to motivate the use of the word 'projection.' You may be wondering however what symmetry has to do with orthogonality. We explain this in terms of the tautology

$$b = Pb + (I - P)b.$$

Now, if P is a projection then so too is $(I - P)$. Moreover, if P is symmetric then the dot product of b's two constituents is

$$(Pb)^T (I - P)b = b^T P^T (I - P)b = b^T (P - P^2) b = b^T 0 b = 0,$$

i.e., Pb is orthogonal to $(I - P)b$.

As examples of nonorthogonal projections we offer

$$\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 0 & 0 \\ -1/2 & 0 & 0 \\ -1/4 & -1/2 & 1 \end{pmatrix}$$

Finally, let us note that the central formula, $P = A(A^T A)^{-1} A^T$, is even a bit more general than advertised. It has been billed as the orthogonal projection onto the column space of A. The need often arises however for the orthogonal projection onto some arbitrary subspace M. The key to using the old P is simply to realize that **every** subspace is the column space of some matrix. More precisely, if

$$\{x_1,\ldots,x_m\}$$

is a basis for M then clearly if these x_j are placed into the columns of a matrix called A then $\mathcal{R}(A)=M$. For example, if M is the line through $[1\ 1]^T$ then

$$P=\begin{pmatrix}1\\1\end{pmatrix}\frac{1}{2}\begin{pmatrix}1 & 1\end{pmatrix}=\frac{1}{2}\begin{pmatrix}1 & 1\\1 & 1\end{pmatrix}$$

is orthogonal projection onto M.

5.4. Exercises

1. A steal beam was stretched to lengths $\ell=6,\ 7$, and 8 feet under applied forces of $f=1,\ 2$, and 4 tons. Assuming Hooke's law $\ell-L=cf$, find its compliance, c, and original length, L, by least squares.

2. With regard to the example of §5.3 note that, due to the the random generation of the noise that taints the displacements, one gets a different 'answer' every time the code is invoked.

   (i) Write a loop that invokes the code a statistically significant number of times and submit bar plots of the average fiber stiffness and its standard deviation for each fiber, along with the associated M–file.

   (ii) Experiment with various noise levels with the goal of determining the level above which it becomes difficult to discern the stiff fiber. Carefully explain your findings.

3. Find the matrix that projects $\mathbb{R}^3$ onto the line spanned by $[1\ 0\ 1]^T$.

4. Find the matrix that projects $\mathbb{R}^3$ onto the plane spanned by $[1\ 0\ 1]^T$ and $[1\ 1\ -1]^T$.

5. If P is the projection of $\mathbb{R}^m$ onto a k–dimensional subspace M, what is the rank of P and what is $\mathcal{R}(P)$?

6. Matrix Methods for Dynamical Systems

Up to this point we have largely been concerned with (i) deriving linear systems of algebraic equations (from considerations of static equilibrium) and (ii) the solution of such systems via Gaussian elimination.

In this section we hope to begin to persuade the reader that our tools extend in a natural fashion to the class of dynamic processes. More precisely, we shall argue that (i) Matrix Algebra plays a central role in the derivation of mathematical models of dynamical systems and that, with the aid of the Laplace transform in an analytical setting or the Backward Euler method in the numerical setting, (ii) Gaussian elimination indeed produces the solution.

6.1. Neurons and the Dynamic Strang Quartet

A nerve fiber's natural electrical stimulus is not direct current but rather a short burst of current, the so–called nervous impulse. In such a dynamic environment the cell's membrane behaves not only like a leaky conductor but also like a charge separator, or capacitor.

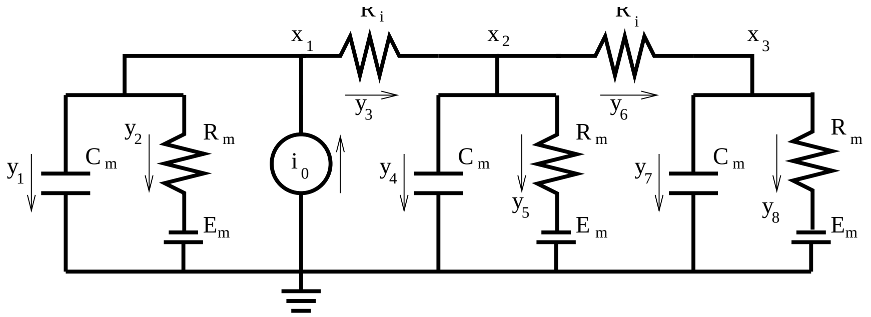


Figure 6.1. An RC model of a neuron.

The typical value of a cell's membrane capacitance is

$$c = 1\ (\mu F/cm^2)$$

where μF denotes micro–Farad. The capacitance of a single compartment is therefore

$$C_m = 2\pi a(\ell/N)c$$

and runs parallel to each R_m, see figure 6.1. We ask now how the static Strang Quartet of chapter one should be augmented. Regarding (S1') we proceed as before. The voltage drops are

$$e_1 = x_1, \quad e_2 = x_1 - E_m, \quad e_3 = x_1 - x_2, \quad e_4 = x_2,$$
$$e_5 = x_2 - E_m, \quad e_6 = x_2 - x_3, \quad e_7 = x_3, \quad e_8 = x_3 - E_m,$$

and so

$$e = b - Ax \quad \text{where} \quad b = -E_m \begin{pmatrix} 0\\1\\0\\0\\1\\0\\0\\1 \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix} -1 & 0 & 0\\ -1 & 0 & 0\\ -1 & 1 & 0\\ 0 & -1 & 0\\ 0 & -1 & 0\\ 0 & -1 & 1\\ 0 & 0 & -1\\ 0 & 0 & -1 \end{pmatrix}$$

In (S2) we must now augment Ohm's law with voltage–current law obeyed by a capacitor, namely – the current through a capacitor is proportional to the time rate of change of the potential across it. This yields, (denoting d/dt by $'$),

$$\begin{aligned} y_1 = C_m e_1', \quad y_2 = e_2/R_m, \quad y_3 = e_3/R_i, \quad y_4 = C_m e_4', \\ y_5 = e_5/R_m, \quad y_6 = e_6/R_i, \quad y_7 = C_m e_7', \quad y_8 = e_8/R_m, \end{aligned}$$

or, in matrix terms,

$$y = Ge + Ce'$$

where

$$\begin{aligned} G &= \text{diag}(0\ G_m\ G_i\ 0\ G_m\ G_i\ 0\ G_m) \\ C &= \text{diag}(C_m\ 0\ 0\ C_m\ 0\ 0\ C_m\ 0) \end{aligned}$$

are the conductance and capacitance matrices.

As Kirchhoff's Current law is insensitive to the type of device occupying an edge, step (S3) proceeds exactly as above.

$$i_0 - y_1 - y_2 - y_3 = 0 \quad y_3 - y_4 - y_5 - y_6 = 0 \quad y_6 - y_7 - y_8 = 0,$$

or, in matrix terms,

$$A^T y = -f \quad \text{where} \quad f = [i_0\ 0\ 0]^T.$$

Step (S4) remains one of assembling,

$$A^T y = -f \Rightarrow A^T(Ge + Ce') = -f \Rightarrow A^T(G(b - Ax) + C(b' - Ax')) = -f,$$

becomes

$$\boxed{A^T CAx' + A^T GAx = A^T Gb + f + A^T Cb'.} \tag{6.1}$$

This is the general form of the potential equations for an RC circuit. It presumes of the user knowledge of the initial value of each of the potentials,

$$x(0) = X. \tag{6.2}$$

Regarding the circuit of figure 6.1 we find

$$A^T CA = \begin{pmatrix} C_m & 0 & 0 \\ 0 & C_m & 0 \\ 0 & 0 & C_m \end{pmatrix} \quad A^T GA = \begin{pmatrix} G_i + G_m & -G_i & 0 \\ -G_i & 2G_i + G_m & -G_i \\ 0 & -G_i & G_i + G_m \end{pmatrix}$$

$$A^T Gb = E_m \begin{pmatrix} G_m \\ G_m \\ G_m \end{pmatrix} \quad \text{and} \quad A^T Cb' = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

and an initial (rest) potential of

$$x(0) = E_m[1\ 1\ 1]^T.$$

We shall now outline two modes of attack on such problems. The Laplace Transform is an analytical tool that produces exact, closed–form, solutions for small tractable systems and therefore offers insight into how larger systems 'should' behave. The Backward–Euler method is a technique for solving a discretized (and therefore approximate) version of (6.1). It is highly flexible, easy to code, and works on problems of great size. Both the Backward–Euler and Laplace Transform methods require, at their core, the algebraic solution of a linear system of equations. In deriving these methods we shall find it more convenient to proceed from the generic system

$$x' = Bx + g. \tag{6.3}$$

With respect to our fiber problem

$$\begin{aligned} B &= -(A^TCA)^{-1}A^TGA \\ &= \frac{1}{C_m}\begin{pmatrix} -(G_i+G_m) & G_i & 0 \\ G_i & -(2G_i+G_m) & G_i \\ 0 & G_i & -(G_i+G_m) \end{pmatrix} \end{aligned} \tag{6.4}$$

and

$$g = (A^TCA)^{-1}(A^TGb+f) = \frac{1}{C_m}\begin{pmatrix} E_mG_m + i_0 \\ E_mG_m \\ E_mG_m \end{pmatrix}.$$

6.2. The Laplace Transform

The Laplace Transform is typically credited with taking dynamical problems into static problems. Recall that the Laplace Transform of the function h is

$$(\mathcal{L}h)(s) \equiv \int_0^\infty \mathrm{e}^{-st}h(t)\,dt.$$

where s is a complex variable. We shall soon take a 2 chapter dive into the theory of complex variables and functions. But for now let us proceed calmly and confidently and follow the lead of Matlab . For example

```
>> syms t
>> laplace(exp(t))
ans = 1/(s-1)
>> laplace(t*exp(-t))
ans = 1/(s+1)^2
```

The Laplace Transform of a matrix of functions is simply the matrix of Laplace transforms of the individual elements. For example

$$\mathcal{L}\begin{pmatrix} \mathrm{e}^t \\ t\mathrm{e}^{-t} \end{pmatrix} = \begin{pmatrix} 1/(s-1) \\ 1/(s+1)^2 \end{pmatrix}.$$

Now, in preparing to apply the Laplace transform to (6.3) we write it as

$$\mathcal{L}x' = \mathcal{L}(Bx+g) \tag{6.5}$$

and so must determine how $\mathcal{L}$ acts on derivatives and sums. With respect to the latter it follows directly from the definition that

$$\mathcal{L}(Bx+g) = \mathcal{L}Bx + \mathcal{L}g = B\mathcal{L}x + \mathcal{L}g. \tag{6.6}$$

Regarding its effect on the derivative we find, on integrating by parts, that

$$\mathcal{L}x' = \int_0^\infty \mathrm{e}^{-st}x'(t)\,dt = x(t)\mathrm{e}^{-st}\Big|_0^\infty + s\int_0^\infty \mathrm{e}^{-st}x(t)\,dt.$$

Supposing that x and s are such that $x(t)\mathrm{e}^{-st} \to 0$ as $t \to \infty$ we arrive at

$$\mathcal{L}x' = s\mathcal{L}x - x(0). \tag{6.7}$$

Now, upon substituting (6.6) and (6.7) into (6.5) we find

$$s\mathcal{L}x - x(0) = B\mathcal{L}x + \mathcal{L}g,$$

which is easily recognized to be a linear system for $\mathcal{L}x$, namely

$$(sI - B)\mathcal{L}x = \mathcal{L}g + x(0). \tag{6.8}$$

The only thing that distinguishes this system from those encountered since chapter 1 is the presence of the complex variable s. This complicates the mechanical steps of Gaussian Elimination or the Gauss–Jordan Method but the methods indeed apply without change. Taking up the latter method, we write

$$\mathcal{L}x = (sI - B)^{-1}(\mathcal{L}g + x(0)).$$

The matrix $(sI - B)^{-1}$ is typically called the **resolvent** of B at s. We turn to Matlab for its **symbolic** calculation. For example,

```
>> syms s
>> B = [2 -1;-1 2]
>> R = inv(s*eye(2)-B)
R =
[ (s-2)/(s*s-4*s+3), -1/(s*s-4*s+3)]
[ -1/(s*s-4*s+3), (s-2)/(s*s-4*s+3)]
```

We note that $(sI - B)^{-1}$ is well defined except at the roots of the quadratic, $s^2 - 4s + 3$. This quadratic is the **determinant** of $(sI-B)$ and is often referred to as the **characteristic polynomial** of B. The roots of the characteristic polynomial are called the **eigenvalues** of B. We will develop each of these new mathematical objects over the coming chapters. We mention them here only to point out that they are all latent in the **resolvent**.

As a second example let us take the B matrix of (6.4) with the parameter choices specified in `fib3.m`, namely

$$B = \frac{1}{10}\begin{pmatrix} -5 & 3 & 0 \\ 3 & -8 & 3 \\ 0 & 3 & -5 \end{pmatrix}. \tag{6.9}$$

The associated resolvent is

$$(sI - B)^{-1} = \frac{1}{\chi_B(s)}\begin{pmatrix} s^2 + 1.3s + 0.31 & 0.3s + 0.15 & 0.09 \\ 0.3s + 0.15 & s^2 + s + 0.25 & 0.3s + 0.15 \\ 0.09 & 0.3s + 0.15 & s^2 + 1.3s + 0.31 \end{pmatrix}$$

where

$$\chi_B(s) = s^3 + 1.8s^2 + 0.87s + 0.11 \tag{6.10}$$

is the characteristic polynomial of B. Assuming a current stimulus of the form $i_0(t) = t\exp(-t/4)/1000$, and $E_m = 0$ brings

$$(\mathcal{L}g)(s) = \begin{pmatrix} 1.965/(s+1/4)^2 \\ 0 \\ 0 \end{pmatrix}$$

and so (6.10) persists in

$$\begin{aligned} \mathcal{L}x &= (sI - B)^{-1}\mathcal{L}g \\ &= \frac{1.965}{(s+1/4)^2(s^3+1.8s^2+0.87s+0.11)} \begin{pmatrix} s^2+1.3s+0.31 \\ 0.3s+0.15 \\ 0.09 \end{pmatrix} \end{aligned}$$

Now comes the rub. A simple linear solve (or inversion) has left us with the Laplace transform of x. The accursed No Free Lunch Theorem informs us that we shall have to do some work in order to recover x from $\mathcal{L}x$.

In coming sections we shall establish that the inverse Laplace transform of a function h is

$$(\mathcal{L}^{-1}h)(t) = \frac{1}{2\pi i}\int_{\mathcal{C}} h(s)\exp(st)\,ds, \tag{6.11}$$

where s runs along $\mathcal{C}$, a closed curve in the complex plane that encircles all of the **singularities** of h. We don't suppose the reader to have yet encountered integration in the complex plane and so please view (6.11) as a preview pf coming attractions.

With the inverse Laplace transform one may express the solution of (6.3) as

$$x(t) = \mathcal{L}^{-1}(sI - B)^{-1}(\mathcal{L}g + x(0)). \tag{6.12}$$

As an example, let us take the first component of $\mathcal{L}x$, namely

$$\mathcal{L}x_1(s) = \frac{1.965(s^2+1.3s+0.31)}{(s+1/4)^2(s^3+1.8s^2+0.87s+0.11))}. \tag{6.13}$$

The singularities, or **poles**, are the points s at which $\mathcal{L}x_1(s)$ blows up. These are clearly the roots of its denominator, namely

$$-11/10,\ -1/2,\ -1/5 \quad \text{and} \quad -1/4, \tag{6.14}$$

and hence the curve, $\mathcal{C}$, in (6.11), must encircle these. We turn to MATLAB however to actually evaluate (6.11). Referring to `fib3.m` for details we note that the `ilaplace` command produces

$$\begin{aligned} x_1(t) = 1.965\Big(&8e^{-t/2} - (1/17)e^{-t/4}(40912/17 + 76t) \\ &+ (400/3)e^{-t/5} + (200/867)e^{-11t/10}\Big) \end{aligned} \tag{6.15}$$

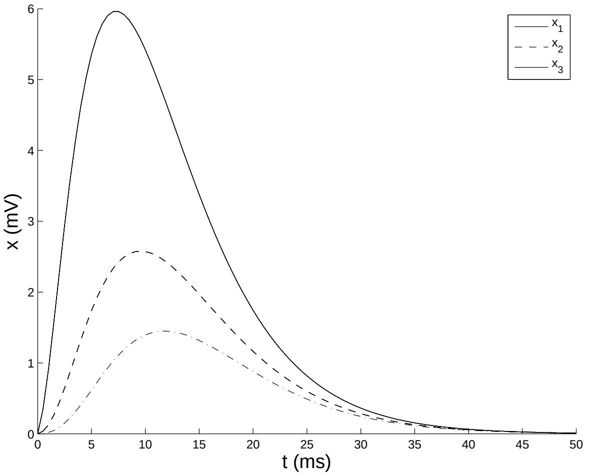


Figure 6.2. The 3 potentials associated with figure 6.1.

The other potentials, see the figure above, possess similar expressions. Please note that each of the poles of $\mathcal{L}x_1$ appear as exponents in x_1 and that the coefficients of the exponentials are polynomials whose degrees are determined by the **orders** of the respective poles. `fib3.m`

6.3. The Backward–Euler Method

Where in the previous section we tackled the derivative in (6.3) via an integral transform we pursue in this section a much simpler strategy, namely, replace the derivative with a finite difference quotient. That is, one chooses a small dt and 'replaces' (6.3) with

$$\frac{\tilde{x}(t)-\tilde{x}(t-dt)}{dt}=B\tilde{x}(t)+g(t). \tag{6.16}$$

The utility of (6.16) is that it gives a means of solving for $\tilde{x}$ at the present time, t, from knowledge of $\tilde{x}$ in the immediate past, $t-dt$. For example, as $\tilde{x}(0)=x(0)$ is supposed known we write (6.16) as

$$(I/dt-B)\tilde{x}(dt)=x(0)/dt+g(dt).$$

Solving this for $\tilde{x}(dt)$ we return to (6.16) and find

$$(I/dt-B)\tilde{x}(2dt)=\tilde{x}(dt)/dt+g(2dt)$$

and solve for $\tilde{x}(2dt)$. The general step from past to present,

$$\tilde{x}(jdt)=(I/dt-B)^{-1}(\tilde{x}((j-1)dt)/dt+g(jdt)), \tag{6.17}$$

is repeated until some desired final time, Tdt, is reached. This equation has been implemented in `fib3.m` with $dt=1$ and B and g as above. The resulting $\tilde{x}$ (run `fib3` yourself!) is indistinguishable from figure 6.2.

Comparing the two representations, (6.12) and (6.17), we see that they both produce the solution to the general linear system of ordinary equations, (6.3), by simply inverting a shifted copy of B. The former representation is hard but exact while the latter is easy but approximate. Of course we should expect the approximate solution, $\tilde{x}$, to approach the exact solution, x, as the time step, dt, approaches zero. To see this let us return to (6.17) and assume, for now, that $g\equiv 0$. In this case, one can reverse the above steps and arrive at the representation

$$\tilde{x}(jdt)=((I-dtB)^{-1})^{j}x(0). \tag{6.18}$$

Now, for a fixed time t we suppose that $dt = t/j$ and ask whether

$$x(t) = \lim_{j\to\infty} ((I - (t/j)B)^{-1})^j x(0).$$

This limit, at least when B is one-by-one, yields the exponential

$$x(t) = \exp(Bt)x(0),$$

clearly the correct solution to (6.3). A careful explication of the **matrix exponential** and its relationship to (6.12) will have to wait until we have mastered the inverse Laplace transform.

6.4. Dynamics of Mechanical Systems

Regarding the fiber nets of Chapter 2, we may move from the equilibrium equations, for the displacement x due to a constant force, f,

$$Sx = f, \quad \text{where} \quad S = A^T KA,$$

to the dynamical equations for the displacement, $x(t)$, due to a time varying force, $f(t)$, and or nonequilibrium initial conditions, by simply appending the Newtonian inertial terms, i.e.,

$$Mx''(t) + Sx(t) = f(t), \qquad x(0) = x_0, \quad x'(0) = v_0, \tag{6.19}$$

where M is the diagonal matrix of node masses, x_0 denotes their initial displacement and v_0 denotes their initial velocity.

We transform this system of second order differential equations to an equivalent first order system by introducing

$$u_1 \equiv x \quad \text{and} \quad u_2 \equiv u_1'$$

and then noting that (6.19) takes the form

$$u_2' = x'' = -M^{-1}Su_1 + M^{-1}f(t).$$

As such, we find that $u = (u_1\ u_2)^T$ obeys the familiar

$$u' = Bu + g, \quad u(0) = u_0 \tag{6.20}$$

where

$$B = \begin{pmatrix} 0 & I \\ -M^{-1}S & 0 \end{pmatrix}, \quad g = \begin{pmatrix} 0 \\ M^{-1}f \end{pmatrix}, \quad u_0 = \begin{pmatrix} x_0 \\ v_0 \end{pmatrix}. \tag{6.21}$$

Let us consider the concrete example of the chain of three masses in Fig. 2.1. If each node has mass m and each spring has stiffness k then

$$M^{-1}S = \frac{k}{m}\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}. \tag{6.22}$$

The associated characteristic polynomial of B is

$$\chi_B(s) = s^6 + 6cs^4 + 10c^2s^2 + 4c^3, \quad \text{where} \quad c \equiv k/m, \tag{6.23}$$

is a cubic in s^2 with simple roots at $-2c$ and $-2c \pm \sqrt{2}c$. And so the eigenvalues of B are the three purely imaginary numbers

$$\lambda_1 = i\sqrt{2c - \sqrt{2}c}, \quad \lambda_2 = i\sqrt{2c}, \quad \lambda_3 = i\sqrt{2c + \sqrt{2}c} \tag{6.24}$$

and their complex conjugates, $\lambda_4 = -\lambda_1$, $\lambda_5 = -\lambda_2$ and $\lambda_6 = -\lambda_3$. Next, if the exogenous force, f, is 0, and the initial disturbance is simply $x_1(0) = 1$ then

$$\mathcal{L}u_1(s) = \frac{1}{\chi_B(s)} \begin{pmatrix} 3c^2 s + 4cs^3 + s^5 \\ cs(s^2 + 2c) \\ c^2 s \end{pmatrix}. \tag{6.25}$$

On computing the inverse Laplace Transform we (will) find

$$x_1(t) = \sum_{j=1}^{6} \exp(\lambda_j t)(3c^2\lambda_j + 4c\lambda_j^3 + \lambda_j^5) \exp(\lambda_j t) \frac{(s - \lambda_j)}{\chi_B(s)}\bigg|_{s=\lambda_j}, \tag{6.26}$$

that is, x_1 is a weighted sum of exponentials. As each of the rates are purely imaginary it follows that our masses will simply oscillate according to weighted sums of three sinusoids. For example, note that

$$\exp(\lambda_2 t) = \cos(\sqrt{2}ct) + i\sin(\sqrt{2}ct) \quad \text{and} \quad \exp(\lambda_5 t) = \cos(\sqrt{2}ct) - i\sin(\sqrt{2}ct).$$

Of course such sums may reproduce quite complicated behavior. We have illustrated each of the displacements in Figure 6.3 in the case that $c = 1$. Rather than plotting the complex, yet explicit, expression (6.26) for x_1, we simply implement the Backward Euler scheme of the previous section.

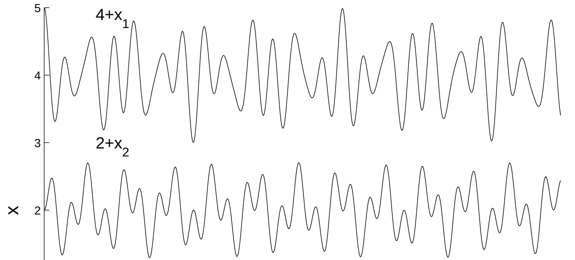


Figure 6.3. The displacements of the 3 masses in Fig. 2.1, with $k/m = 1$, following an initial displacement of the first mass. For viewing purposes we have offset the three displacements. `chain.m`

6.5. Exercises

1. Compute, *without* the aid of a machine, the Laplace transforms of e^t and te^{-t}. Show all of your work.

2. Extract from `fib3.m` analytical expressions for x_2 and x_3.

3. Use `eig` to compute the eigenvalues of B as given in (6.9). Use `poly` to compute the characteristic polynomial of B. Use `roots` to compute the roots of this characteristic polynomial. Compare these to the results of `eig`. How does MATLAB compute the roots of a polynomial? (type `type roots`) for the answer). Submit a MATLAB diary of your findings.

4. Adapt the Backward Euler portion of `fib3.m` so that one may specify an arbitrary number of compartments, as in `fib1.m`. As B, and so S, is now large and sparse please create the sparse B via `spdiags` and the sparse I via `speye`, and then prefactor S into LU and use $U\backslash L\backslash$ rather than $S\backslash$ in the time loop. Experiment to find the proper choice of dt. Submit your well documented M-file along with a plot of x_1 and x_{50} *versus* time (on the same well labeled graph) for a 100 compartment cable.

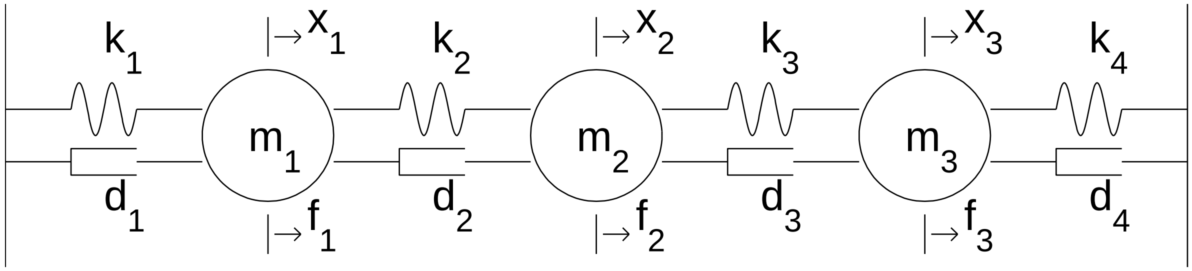
5. Derive (6.18) from (6.17) by working backwards toward $x(0)$. Along the way you should explain why $(I/dt - B)^{-1}/dt = (I - dtB)^{-1}$.

6. Show, for scalar B, that $((1-(t/j)B)^{-1})^j \to \exp(Bt)$ as $j \to \infty$. Hint: By definition

$$((1-(t/j)B)^{-1})^j = \exp(j\log(1/(1-(t/j)B)))$$

   now use L'Hôpital's rule to show that $j\log(1/(1-(t/j)B)) \to Bt$.

7. If we place a viscous damper in parallel with each spring in Figure 2.1 as below



   then the dynamical system (6.20) takes the form

$$Mx''(t) + Dx'(t) + Sx(t) = f(t) \tag{6.27}$$

   where $D = A^T\text{diag}(d)A$ where d is the vector of damping constants. Modify `chain.m` to solve this new system and use your code to reproduce the figure below.

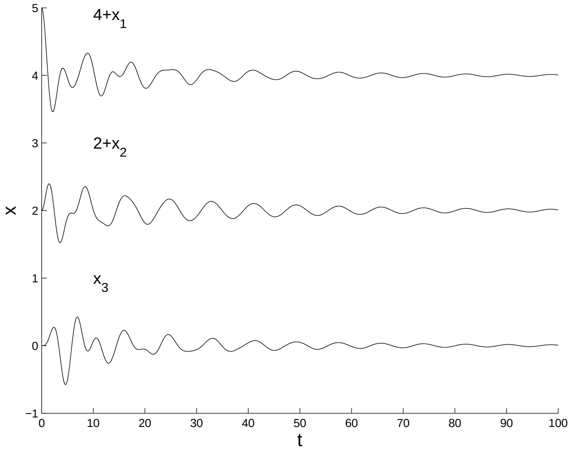


Figure 6.1. The displacement of the three masses in the weakly damped chain, where $k/m = 1$ and $d/m = 1/10$.

7. Complex Numbers and Functions

7.1. Complex Algebra

A complex number is simply a pair of real numbers. In order to stress however that the two algebras differ we separate the two real pieces by the symbol $+i$. More precisely, each complex number, z, may be uniquely expressed by the combination $x + iy$ where x and y are real and i denotes $\sqrt{-1}$. We call x the **real** part and y the **imaginary** part of z. We now summarize the main rules of complex arithmetic.

If

$$z_1 = x_1 + iy_1 \quad \text{and} \quad z_2 = x_2 + iy_2$$

then

$$\begin{aligned}
z_1 + z_2 &\equiv (x_1 + x_2) + i(y_1 + y_2) \\
z_1 z_2 &\equiv (x_1 + iy_1)(x_2 + iy_2) = (x_1x_2 - y_1y_2) + i(x_1y_2 + x_2y_1) \\
\overline{z}_1 &\equiv x_1 - iy_1, \\
\frac{z_1}{z_2} &\equiv \frac{z_1}{z_2}\frac{\overline{z}_2}{\overline{z}_2} = \frac{(x_1x_2 + y_1y_2) + i(x_2y_1 - x_1y_2)}{x_2^2 + y_2^2} \\
|z_1| &\equiv \sqrt{z_1\overline{z}_1} = \sqrt{x_1^2 + y_1^2},
\end{aligned}$$

In addition to the Cartesian representation $z = x + iy$ one also has the polar form

$$z = |z|(\cos\theta + i\sin\theta), \quad \text{where } \theta \in (-\pi, \pi] \text{ and}$$

$$\theta = \text{atan2}(y, x) \equiv \begin{cases} \pi/2, & \text{if } x = 0,\ y > 0, \\ -\pi/2, & \text{if } x = 0,\ y < 0, \\ \arctan(y/x) & \text{if } x > 0, \\ \arctan(y/x) + \pi & \text{if } x < 0,\ y \geq 0, \\ \arctan(y/x) - \pi & \text{if } x < 0,\ y < 0. \end{cases}$$

This form is especially convenient with regards to multiplication. More precisely,

$$\begin{aligned}
z_1z_2 &= |z_1||z_2|\{(\cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2) + i(\cos\theta_1\sin\theta_2 + \sin\theta_1\cos\theta_2)\} \\
&= |z_1||z_2|\{\cos(\theta_1 + \theta_2) + i\sin(\theta_1 + \theta_2)\}.
\end{aligned}$$

As a result,

$$z^n = |z|^n(\cos(n\theta) + i\sin(n\theta)).$$

A complex vector (matrix) is simply a vector (matrix) of complex numbers. Vector and matrix addition proceed, as in the real case, from elementwise addition. The dot or inner product of two complex vectors requires, however, a little modification. This is evident when we try to use the old notion to define the length of complex vector. To wit, note that if

$$z = \begin{pmatrix} 1 + i \\ 1 - i \end{pmatrix}$$

then

$$z^Tz = (1 + i)^2 + (1 - i)^2 = 1 + 2i - 1 + 1 - 2i - 1 = 0.$$

Now length **should** measure the distance from a point to the origin and should only be zero for the zero vector. The fix, as you have probably guessed, is to sum the squares of the **magnitudes** of the components of z. This is accomplished by simply conjugating one of the vectors. Namely, we define the length of a complex vector via

$$\|z\| \equiv \sqrt{\overline{z}^T z}. \tag{7.1}$$

In the example above this produces

$$\sqrt{|1+i|^2 + |1-i|^2} = \sqrt{4} = 2.$$

As each real number is the conjugate of itself, this new definition subsumes its real counterpart.

The notion of magnitude also gives us a way to define limits and hence will permit us to introduce complex calculus. We say that the sequence of complex numbers, $\{z_n : n = 1, 2, \ldots\}$, converges to the complex number z_0 and write

$$z_n \longrightarrow z_0 \quad \text{or} \quad z_0 = \lim_{n\to\infty} z_n,$$

when, presented with any $\varepsilon > 0$ one can produce an integer N for which $|z_n - z_0| < \varepsilon$ when $n \geq N$. As an example, we note that $(i/2)^n \to 0$.

7.2. Complex Functions

A complex function is merely a rule for assigning certain complex numbers to other complex numbers. The simplest (nonconstant) assignment is the identity function $f(z) \equiv z$. Perhaps the next simplest function assigns to each number its square, i.e., $f(z) \equiv z^2$. As we decomposed the **argument** of f, namely z, into its real and imaginary parts, we shall also find it convenient to partition the **value** of f, z^2 in this case, into its real and imaginary parts. In general, we write

$$f(x+iy) = u(x,y) + iv(x,y)$$

where u and v are both real–valued functions of two real variables. In the case that $f(z) \equiv z^2$ we find

$$u(x,y) = x^2 - y^2 \quad \text{and} \quad v(x,y) = 2xy.$$

With the tools of the previous section we may produce complex polynomials

$$f(z) = z^m + c_{m-1}z^{m-1} + \cdots + c_1 z + c_0.$$

We say that such an f is of **degree** m. We shall often find it convenient to represent polynomials as the product of their factors, namely

$$f(z) = (z-\lambda_1)^{m_1}(z-\lambda_2)^{m_2}\cdots(z-\lambda_h)^{m_h}. \tag{7.2}$$

Each λ_j is a **root** of f of **degree** m_j. Here h is the number of **distinct** roots of f. We call λ_j a **simple** root when $m_j = 1$. In chapter 6 we observed the appearance of ratios of polynomials or so called **rational** functions. Suppose

$$r(z) = \frac{f(z)}{g(z)}$$

is rational, that f is of order at most $m-1$ while g is of order m with the simple roots $\{\lambda_1,\ldots,\lambda_m\}$. It should come as no surprise that such a r should admit a **Partial Fraction Expansion**

$$r(z) = \sum_{j=1}^{m} \frac{r_j}{z-\lambda_j}.$$

One uncovers the r_j by first multiplying each side by $(z-\lambda_j)$ and then letting z tend to λ_j. For example, if

$$\frac{1}{z^2+1} = \frac{r_1}{z+i} + \frac{r_2}{z-i} \tag{7.3}$$

then multiplying each side by $(z+i)$ produces

$$\frac{1}{z-i} = r_1 + \frac{r_2(z+i)}{z-i}.$$

Now, in order to isolate r_1 it is clear that we should set $z=-i$. So doing we find $r_1 = i/2$. In order to find r_2 we multiply (7.3) by $(z-i)$ and then set $z=i$. So doing we find $r_2 = -i/2$, and so

$$\frac{1}{z^2+1} = \frac{i/2}{z+i} + \frac{-i/2}{z-i}. \tag{7.4}$$

Returning to the general case, we encode the above in the simple formula

$$r_j = \lim_{z\to\lambda_j} (z-\lambda_j) r(z). \tag{7.5}$$

You should be able to use this to confirm that

$$\frac{z}{z^2+1} = \frac{1/2}{z+i} + \frac{1/2}{z-i}. \tag{7.6}$$

Recall that the resolvent we met in Chapter 6 was in fact a matrix of rational functions. Now, the partial fraction expansion of a matrix of rational functions is simply the matrix of partial fraction expansions of each of its elements. This is easier done than said. For example, the resolvent of

$$B = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

is

$$\begin{aligned}(zI-B)^{-1} &= \frac{1}{z^2+1}\begin{pmatrix} z & 1 \\ -1 & z \end{pmatrix} \\ &= \frac{1}{z+i}\begin{pmatrix} 1/2 & i/2 \\ -i/2 & 1/2 \end{pmatrix} + \frac{1}{z-i}\begin{pmatrix} 1/2 & -i/2 \\ i/2 & 1/2 \end{pmatrix}.\end{aligned} \tag{7.7}$$

The first line comes from either Gauss-Jordan by hand or via the symbolic toolbox in Matlab . More importantly, the second line is simply an amalgamation of (7.3) and (7.4). Complex matrices have finally entered the picture. We shall devote all of Chapter 9 to uncovering the remarkable properties enjoyed by the matrices that appear in the partial fraction expansion of $(zI-B)^{-1}$. Have you noticed that, in our example, the two matrices are each projections, that they sum to I, and that their product is 0? Could this be an accident? To answer this we will also need to develop $(zI-B)^{-1}$ in a geometric expansion. At its simplest, the n-term geometric series, for $z \neq 1$, is

$$\sum_{k=0}^{n-1} z^k = \frac{1-z^n}{1-z}. \tag{7.8}$$

We will prove this in the exercises and use it to appreciate the beautiful orthogonality of the columns of the Fourier Matrix.

In Chapter 6 we were confronted with the complex exponential when considering the Laplace Transform. By analogy to the real exponential we define

$$\mathrm{e}^z \equiv \sum_{n=0}^{\infty} \frac{z^n}{n!}$$

and find that

$$\begin{aligned}\mathrm{e}^{i\theta} &= 1 + i\theta + (i\theta)^2/2 + (i\theta)^3/3! + (i\theta)^4/4! + \cdots \\ &= (1 - \theta^2/2 + \theta^4/4! - \cdots) + i(\theta - \theta^3/3! + \theta^5/5! - \cdots) \\ &= \cos\theta + i\sin\theta.\end{aligned}$$

With this observation, the polar form is now simply $z = |z|\mathrm{e}^{i\theta}$. One may just as easily verify that

$$\cos\theta = \frac{\mathrm{e}^{i\theta} + \mathrm{e}^{-i\theta}}{2} \quad \text{and} \quad \sin\theta = \frac{\mathrm{e}^{i\theta} - \mathrm{e}^{-i\theta}}{2i}.$$

These suggest the definitions, for complex z, of

$$\cos z \equiv \frac{\mathrm{e}^{iz} + \mathrm{e}^{-iz}}{2} \quad \text{and} \quad \sin z \equiv \frac{\mathrm{e}^{iz} - \mathrm{e}^{-iz}}{2i}.$$

As in the real case the exponential enjoys the property that

$$\mathrm{e}^{z_1+z_2} = \mathrm{e}^{z_1}\mathrm{e}^{z_2}$$

and in particular

$$\mathrm{e}^{x+iy} = \mathrm{e}^x\mathrm{e}^{iy} = \mathrm{e}^x\cos y + i\mathrm{e}^x\sin y.$$

Finally, the inverse of the complex exponential is the complex logarithm,

$$\ln z \equiv \ln(|z|) + i\theta, \quad \text{for } z = |z|\mathrm{e}^{i\theta}.$$

One finds that $\ln(-1+i) = \ln\sqrt{2} + i3\pi/4$.

7.3. Complex Differentiation

The complex function f is said to be differentiable at z_0 if

$$\lim_{z\to z_0} \frac{f(z) - f(z_0)}{z - z_0}$$

exists, by which we mean that

$$\frac{f(z_n) - f(z_0)}{z_n - z_0}$$

converges to the same value **for every** sequence $\{z_n\}$ that converges to z_0. In this case we naturally call the limit $f'(z_0)$.

Example: The derivative of z^2 is $2z$.

$$\lim_{z\to z_0} \frac{z^2 - z_0^2}{z - z_0} = \lim_{z\to z_0} \frac{(z - z_0)(z + z_0)}{z - z_0} = 2z_0.$$

Example: The exponential is its own derivative.

$$\lim_{z\to z_0} \frac{\mathrm{e}^z - \mathrm{e}^{z_0}}{z - z_0} = \mathrm{e}^{z_0} \lim_{z\to z_0} \frac{\mathrm{e}^{z-z_0} - 1}{z - z_0} = \mathrm{e}^{z_0} \lim_{z\to z_0} \sum_{n=0}^{\infty} \frac{(z-z_0)^n}{(n+1)!} = \mathrm{e}^{z_0}.$$

Example: The real part of z is **not** a differentiable function of z.

We show that the limit depends on the angle of approach. First, when $z_n \to z_0$ on a line parallel to the real axis, e.g., $z_n = x_0 + 1/n + iy_0$, we find

$$\lim_{n\to\infty} \frac{x_0 + 1/n - x_0}{x_0 + 1/n + iy_0 - (x_0 + iy_0)} = 1,$$

while if $z_n \to z_0$ in the imaginary direction, e.g., $z_n = x_0 + i(y_0 + 1/n)$, then

$$\lim_{n\to\infty} \frac{x_0 - x_0}{x_0 + i(y_0 + 1/n) - (x_0 + iy_0)} = 0.$$

This last example suggests that when f is differentiable a simple relationship must bind its partial derivatives in x and y.

Proposition 7.1. If f is differentiable at z_0 then

$$f'(z_0) = \frac{\partial f}{\partial x}(z_0) = -i\frac{\partial f}{\partial y}(z_0).$$

Proof: With $z = x + iy_0$,

$$f'(z_0) = \lim_{z\to z_0} \frac{f(z) - f(z_0)}{z - z_0} = \lim_{x\to x_0} \frac{f(x + iy_0) - f(x_0 + iy_0)}{x - x_0} = \frac{\partial f}{\partial x}(z_0).$$

Alternatively, when $z = x_0 + iy$ then

$$f'(z_0) = \lim_{z\to z_0} \frac{f(z) - f(z_0)}{z - z_0} = \lim_{y\to y_0} \frac{f(x_0 + iy) - f(x_0 + iy_0)}{i(y - y_0)} = -i\frac{\partial f}{\partial y}(z_0).$$

End of Proof.

In terms of the real and imaginary parts of f this result brings the **Cauchy–Riemann equations**

$$\boxed{\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \quad \text{and} \quad \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}.} \tag{7.9}$$

Regarding the converse proposition we note that when f has continuous partial derivatives in region obeying the Cauchy–Riemann equations then f is in fact differentiable in the region.

We remark that with no more energy than that expended on their real cousins one may uncover the rules for differentiating complex sums, products, quotients, and compositions.

As one important application of the derivative let us attempt to expand in partial fractions a rational function whose denominator has a root with degree larger than one. As a warm-up let us try to find $r_{1,1}$ and $r_{1,2}$ in the expansion

$$\frac{z+2}{(z+1)^2} = \frac{r_{1,1}}{z+1} + \frac{r_{1,2}}{(z+1)^2}.$$

Arguing as above it seems wise to multiply through by $(z+1)^2$ and so arrive at

$$z+2 = r_{1,1}(z+1) + r_{1,2}. \tag{7.10}$$

On setting $z = -1$ this gives $r_{1,2} = 1$. With $r_{1,2}$ computed (7.10) takes the simple form $z+1 = r_{1,1}(z+1)$ and so $r_{1,1} = 1$ as well. Hence

$$\frac{z+2}{(z+1)^2} = \frac{1}{z+1} + \frac{1}{(z+1)^2}.$$

This latter step grows more cumbersome for roots of higher degree. Let us consider

$$\frac{(z+2)^2}{(z+1)^3} = \frac{r_{1,1}}{z+1} + \frac{r_{1,2}}{(z+1)^2} + \frac{r_{1,3}}{(z+1)^3}.$$

The first step is still correct: multiply through by the factor at its highest degree, here 3. This leaves us with

$$(z+2)^2 = r_{1,1}(z+1)^2 + r_{1,2}(z+1) + r_{1,3}. \tag{7.11}$$

Setting $z = -1$ again produces the last coefficient, here $r_{1,3} = 1$. We are left however with one equation in two unknowns. Well, not really one equation, for (7.11) is to hold for **all** z. We exploit this by taking two derivatives, with respect to z, of (7.11). This produces

$$2(z+2) = 2r_{1,1}(z+1) + r_{1,2} \quad \text{and} \quad 2 = 2r_{1,1}.$$

The latter of course needs no comment. We derive $r_{1,2}$ from the former by setting $z = -1$. We generalize from this example and arrive at

Proposition 7.2. The First Residue Theorem. The ratio, $r = f/g$, of two polynomials where the order of f is less than that of g and g has h distinct roots $\{\lambda_1, \ldots, \lambda_h\}$ of respective degrees $\{m_1, \ldots, m_h\}$, may be expanded in partial fractions

$$r(z) = \sum_{j=1}^{h} \sum_{k=1}^{m_j} \frac{r_{j,k}}{(z-\lambda_j)^k} \tag{7.12}$$

where, as above, the **residue** $r_{j,k}$ is computed by first clearing the fraction and then taking the proper number of derivatives and finally clearing their powers. That is,

$$r_{j,k} = \lim_{z \to \lambda_j} \frac{1}{(m_j - k)!} \frac{d^{m_j - k}}{dz^{m_j - k}} \{(z-\lambda_j)^{m_j} r(z)\}. \tag{7.13}$$

As an application, we consider

$$B = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 3 & 0 \\ 0 & 1 & 1 \end{pmatrix} \tag{7.14}$$

and compute the $R_{j,k}$ matrices in the expansion of its resolvent

$$\begin{aligned} R(z) = (zI - B)^{-1} &= \begin{pmatrix} \frac{1}{z-1} & 0 & 0 \\ \frac{1}{(z-1)(z-3)} & \frac{1}{z-3} & 0 \\ \frac{1}{(z-1)^2(z-3)} & \frac{1}{(z-1)(z-3)} & \frac{1}{z-1} \end{pmatrix} \\ &= \frac{1}{z-1} R_{1,1} + \frac{1}{(z-1)^2} R_{1,2} + \frac{1}{z-3} R_{2,1}. \end{aligned}$$

The only challenging term is the (3,1) element. We write

$$\frac{1}{(z-1)^2(z-3)} = \frac{r_{1,1}}{z-1} + \frac{r_{1,2}}{(z-1)^2} + \frac{r_{2,1}}{z-3}.$$

It follows from (7.13) that

$$r_{1,1} = \left(\frac{1}{z-3}\right)'(1) = -\frac{1}{4} \quad \text{and} \quad r_{1,2} = \left(\frac{1}{z-3}\right)(1) = -\frac{1}{2} \tag{7.15}$$

and

$$r_{2,1} = \left(\frac{1}{(z-1)^2}\right)(3) = \frac{1}{4}. \tag{7.16}$$

It now follows that

$$\begin{aligned}(zI-B)^{-1} = \frac{1}{z-1}\begin{pmatrix}1 & 0 & 0\\ -1/2 & 0 & 0\\ -1/4 & -1/2 & 1\end{pmatrix} + \frac{1}{(z-1)^2}\begin{pmatrix}0 & 0 & 0\\ 0 & 0 & 0\\ -1/2 & 0 & 0\end{pmatrix}\\ + \frac{1}{z-3}\begin{pmatrix}0 & 0 & 0\\ 1/2 & 1 & 0\\ 1/4 & 1/2 & 0\end{pmatrix}.\end{aligned} \tag{7.17}$$

In closing, we that the method of partial fraction expansions has been implemented in Matlab . In fact, (7.15) and (7.16) follow from the single command

```
[r,p]=residue([0 0 0 1],[1 -5 7 -3]).
```

The first input argument is Matlab -speak for the polynomial $f(z) = 1$ while the second argument corresponds to the denominator

$$g(z) = (z-1)^2(z-3) = z^3 - 5z^2 + 7z - 3.$$

7.4. Exercises

1. Express $|e^{x+iy}|$ in terms of x and/or y.

2. Suppose $z \neq 1$ and define the n-term geometric series

$$\sigma \equiv \sum_{k=0}^{n-1} z^k,$$

and show, by brute force, that $\sigma - z\sigma = 1 - z^n$. Derive (7.8) from this result.

3. Confirm that $e^{\ln z} = z$ and $\ln e^z = z$.

4. Find the real and imaginary parts of $\cos z$ and $\sin z$. Express your answers in terms of regular and hyperbolic trigonometric functions.

5. Show that $\cos^2 z + \sin^2 z = 1$.

6. As in the real calculus, the exponential and natural log permit one to define arbitrary powers. Please compute $\sqrt{i}$ via the definition $z^p \equiv \mathrm{e}^{p \ln z}$.

7. Carefully sketch, by hand, the complex numbers $\omega_n = \exp(2\pi i/n)$ for $n = 2$, 4 and 8. ω_n is called an nth root of unity. Please compute ω_n^n by hand. Construct the n-by-n Fourier Matrix, F_n, via
$$F_n(j,k) = \frac{1}{\sqrt{n}}\omega_n^{(j-1)(k-1)}, \tag{7.18}$$
where the row index, j, and the column index, k, each run from 1 to n. We will now prove that the conjugate of F_n is in fact the inverse of F_n, i.e., that $F_n F_n^* = I$. To do this first show that the (j,m) element of the product $F_n F_n^*$ is
$$\sum_{k=1}^{n} F(j,k)F^*(k,m) = \frac{1}{n}\sum_{k=1}^{n} \exp(2\pi i(k-1)(j-m)/n). \tag{7.19}$$
Conclude that this sum is 1 when $j = m$. To show that the sum is zero when $j \neq m$ set $z = \exp(2\pi i(j-m)/n)$ and recognize in (7.19) the n-term geometric series of (7.8).

8. Verify that $\sin z$ and $\cos z$ satisfy the Cauchy-Riemann equations (7.9) and use Proposition 7.1 to evaluate their derivatives.

9. Compute, by hand the partial fraction expansion of the rational function that we arrived at in (6.13). That is, find $r_{1,1}$, $r_{1,2}$, r_2, r_3 and r_4 in
$$\begin{aligned} r(s) &= \frac{s^2 + 1.3s + 0.31}{(s+1/4)^2(s+1/2)(s+1/5)(s+11/10)} \\ &= \frac{r_{1,1}}{s+1/4} + \frac{r_{1,2}}{(s+1/4)^2} + \frac{r_2}{s+1/2} + \frac{r_3}{s+1/5} + \frac{r_4}{s+11/10}. \end{aligned}$$
Contrast your answer with the explicit expression in (6.15).

10. Submit a Matlab diary documenting your use of `residue` in the partial fraction expansion of resolvent of
$$B = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$
You should achieve
$$\begin{aligned} (sI-B)^{-1} &= \frac{1}{s-(2+\sqrt{2})}\frac{1}{4}\begin{pmatrix} 1 & -\sqrt{2} & 1 \\ -\sqrt{2} & 2 & -\sqrt{2} \\ 1 & -\sqrt{2} & 1 \end{pmatrix} \\ &+ \frac{1}{s-2}\frac{1}{2}\begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix} + \frac{1}{s-(2-\sqrt{2})}\frac{1}{4}\begin{pmatrix} 1 & \sqrt{2} & 1 \\ \sqrt{2} & 2 & \sqrt{2} \\ 1 & \sqrt{2} & 1 \end{pmatrix} \end{aligned}$$

8. Complex Integration

Our main goal is a better understanding of the partial fraction expansion of a given resolvent. With respect to the example that closed the last chapter, see (7.14)–(7.17), we found

$$(zI-B)^{-1} = \frac{1}{z-\lambda_1}P_1 + \frac{1}{(z-\lambda_1)^2}D_1 + \frac{1}{z-\lambda_2}P_2$$

where the P_j and D_j enjoy the amazing properties

$$\begin{gathered} BP_1 = P_1B = \lambda_1 P_1 + D_1 \quad \text{and} \quad BP_2 = P_2B = \lambda_2 P_2 \\ P_1 + P_2 = I, \quad P_1^2 = P_1, \quad P_2^2 = P_2, \quad \text{and} \quad D_1^2 = 0, \\ P_1D_1 = D_1P_1 = D_1 \quad \text{and} \quad P_2D_1 = D_1P_2 = 0. \end{gathered}$$

In order to show that this **always** happens, i.e., that it is not a quirk produced by the particular B in (7.14), we require a few additional tools from the theory of complex variables. In particular, we need the fact that partial fraction expansions may be carried out through complex integration.

8.1. Cauchy's Theorem

We shall be integrating complex functions over complex curves. Such a curve is parametrized by one complex valued or, equivalently, two real valued, function(s) of a real parameter (typically denoted by t). More precisely,

$$C \equiv \{z(t) = x(t) + iy(t) : t_1 \le t \le t_2\}.$$

For example, if $x(t) = y(t) = t$ from $t_1 = 0$ to $t_2 = 1$, then C is the line segment joining $0 + i0$ to $1 + i$.

We now define

$$\int_C f(z)\,dz \equiv \int_{t_1}^{t_2} f(z(t))z'(t)\,dt.$$

For example, if $C = \{t + it : 0 \le t \le 1\}$ as above and $f(z) = z$ then

$$\int_C z\,dz = \int_0^1 (t+it)(1+i)\,dt = \int_0^1 \{(t-t) + i2t\}\,dt = i,$$

while if C is the unit circle $\{\mathrm{e}^{it} : 0 \le t \le 2\pi\}$ then

$$\int_C z\,dz = \int_0^{2\pi} \mathrm{e}^{it} i \mathrm{e}^{it}\,dt = i\int_0^{2\pi} \mathrm{e}^{i2t}\,dt = i\int_0^{2\pi} \{\cos(2t) + i\sin(2t)\}\,dt = 0.$$

Remaining with the unit circle but now integrating $f(z) = 1/z$ we find

$$\int_C z^{-1}\,dz = \int_0^{2\pi} \mathrm{e}^{-it} i \mathrm{e}^{it}\,dt = 2\pi i.$$

We generalize this calculation to arbitrary (integer) powers over arbitrary circles. More precisely, for integer m and fixed complex a we integrate $(z-a)^m$ over

$$C(a,\rho) \equiv \{a + \rho\mathrm{e}^{it} : 0 \le t \le 2\pi\},$$

the circle of radius ρ centered at a. We find

$$\begin{aligned}\int_{C(a,\rho)} (z-a)^m\,dz &= \int_0^{2\pi} (a+\rho e^{it}-a)^m \rho i e^{it}\,dt \\ &= i\rho^{m+1}\int_0^{2\pi} e^{i(m+1)t}\,dt \\ &= i\rho^{m+1}\int_0^{2\pi} \{\cos((m+1)t)+i\sin((m+1)t)\}\,dt \\ &= \begin{cases} 2\pi i & \text{if } m=-1, \\ 0 & \text{otherwise,}\end{cases}\end{aligned} \tag{8.1}$$

regardless of the size of ρ!

When integrating more general functions it is often convenient to express the integral in terms of its real and imaginary parts. More precisely

$$\begin{aligned}\int_C f(z)\,dz &= \int_C \{u(x,y)+iv(x,y)\}\{dx+idy\} \\ &= \int_C \{u(x,y)\,dx - v(x,y)\,dy\} + i\int_C \{u(x,y)\,dy + v(x,y)\,dx\} \\ &= \int_a^b \{u(x(t),y(t))x'(t) - v(x(t),y(t))y'(t)\}\,dt \\ &\quad + i\int_a^b \{u(x(t),y(t))y'(t) + v(x(t),y(t))x'(t)\}\,dt.\end{aligned}$$

The second line should invoke memories of

Proposition 8.1. Green's Theorem. If C is a closed curve and M and N are continuously differentiable real–valued functions on C_{in}, the region enclosed by C, then

$$\int_C \{M\,dx + N\,dy\} = \iint_{C_{in}} \left(\frac{\partial N}{\partial x} - \frac{\partial M}{\partial y}\right)\,dxdy$$

Applying this proposition to the situation above, we find, so long as C is closed, that

$$\int_C f(z)\,dz = -\iint_{C_{in}} \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)\,dxdy + i\iint_{C_{in}} \left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y}\right)\,dxdy.$$

At first glance it appears that Green's Theorem only serves to muddy the waters. Recalling the Cauchy–Riemann equations however we find that each of these double integrals is in fact identically zero! In brief, we have proven

Proposition 8.2. Cauchy's Theorem. If f is differentiable on and in the closed curve C then

$$\int_C f(z)\,dz = 0.$$

Strictly speaking, in order to invoke Green's Theorem we require not only that f be differentiable but that its derivative in fact be continuous. This however is simply a limitation of our simple mode of proof, Cauchy's Theorem is true as stated.

This theorem, together with (8.1), permits us to integrate every proper rational function. More precisely, if $r = f/g$ where f is a polynomial of degree at most $m-1$ and g is an mth degree polynomial with h distinct zeros at $\{\lambda_j\}_{j=1}^h$ with respective multiplicities of $\{m_j\}_{j=1}^h$ we found that

$$r(z) = \sum_{j=1}^{h} \sum_{k=1}^{m_j} \frac{r_{j,k}}{(z-\lambda_j)^k}. \tag{8.2}$$

Observe now that if we choose the radius ρ_j so small that λ_j is the only zero of g encircled by $C_j \equiv C(\lambda_j, \rho_j)$ then by Cauchy's Theorem

$$\int_{C_j} r(z)\,dz = \sum_{k=1}^{m_j} r_{j,k} \int_{C_j} \frac{1}{(z-\lambda_j)^k}\,dz.$$

In (8.1) we found that each, save the first, of the integrals under the sum is in fact zero. Hence

$$\int_{C_j} r(z)\,dz = 2\pi i r_{j,1}. \tag{8.3}$$

With $r_{j,1}$ in hand, say from (7.13) or `residue`, one may view (8.3) as a means for computing the indicated integral. The opposite reading, i.e., that the integral is a convenient means of expressing $r_{j,1}$, will prove just as useful. With that in mind, we note that the remaining residues may be computed as integrals of the product of r and the appropriate factor. More precisely,

$$\int_{C_j} r(z)(z-\lambda_j)^{k-1}\,dz = 2\pi i r_{j,k}. \tag{8.4}$$

One may be led to believe that the precision of this result is due to the very special choice of curve and function. We shall see ...

8.2. The Second Residue Theorem

After (8.3) and (8.4) perhaps the most useful consequence of Cauchy's Theorem is the freedom it grants one to choose the most advantageous curve over which to integrate. More precisely,

Proposition 8.3. Suppose that C_2 is a closed curve that lies inside the region encircled by the closed curve C_1. If f is differentiable in the annular region outside C_2 and inside C_1 then

$$\int_{C_1} f(z)\,dz = \int_{C_2} f(z)\,dz.$$

Proof: With reference to the figure below we introduce two vertical segments and define the closed curves $C_3 = abcda$ (where the bc arc is clockwise and the da arc is counter-clockwise) and $C_4 = adcba$ (where the ad arc is counter-clockwise and the cb arc is clockwise). By merely following the arrows we learn that

$$\int_{C_1} f(z)\,dz = \int_{C_2} f(z)\,dz + \int_{C_3} f(z)\,dz + \int_{C_4} f(z)\,dz.$$

As Cauchy's Theorem implies that the integrals over C_3 and C_4 each vanish, we have our result. End of Proof.

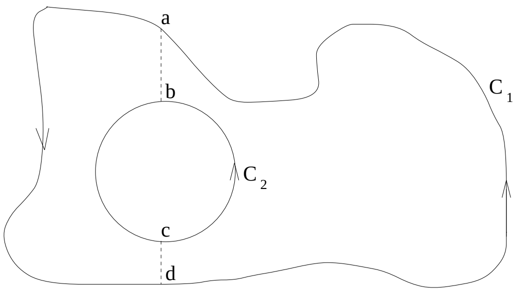


Figure 8.1. The Curve Replacement Lemma.

This proposition states that in order to integrate a function it suffices to integrate it over regions where it is singular, i.e., nondifferentiable.

Let us apply this reasoning to the integral

$$\int_C \frac{z}{(z-\lambda_1)(z-\lambda_2)}\,dz$$

where C encircles both λ_1 and λ_2 as depicted in the cartoon on the next page. We find that

$$\int_C \frac{z}{(z-\lambda_1)(z-\lambda_2)}\,dz = \int_{C_1} \frac{z\,dz}{(z-\lambda_1)(z-\lambda_2)} + \int_{C_2} \frac{z\,dz}{(z-\lambda_1)(z-\lambda_2)}.$$

Developing the integrand in partial fractions we find

$$\begin{aligned}\int_{C_1} \frac{z\,dz}{(z-\lambda_1)(z-\lambda_2)} &= \frac{\lambda_1}{\lambda_1-\lambda_2}\int_{C_1}\frac{dz}{z-\lambda_1} + \frac{\lambda_2}{\lambda_2-\lambda_1}\int_{C_1}\frac{dz}{z-\lambda_2} \\ &= \frac{2\pi i\lambda_1}{\lambda_1-\lambda_2}.\end{aligned}$$

Similarly,

$$\int_{C_2} \frac{z\,dz}{(z-\lambda_1)(z-\lambda_2)} = \frac{2\pi i\lambda_2}{\lambda_2-\lambda_1}.$$

Putting things back together we find

$$\int_C \frac{z}{(z-\lambda_1)(z-\lambda_2)}\,dz = 2\pi i\left(\frac{\lambda_1}{\lambda_1-\lambda_2}+\frac{\lambda_2}{\lambda_2-\lambda_1}\right) = 2\pi i. \tag{8.5}$$

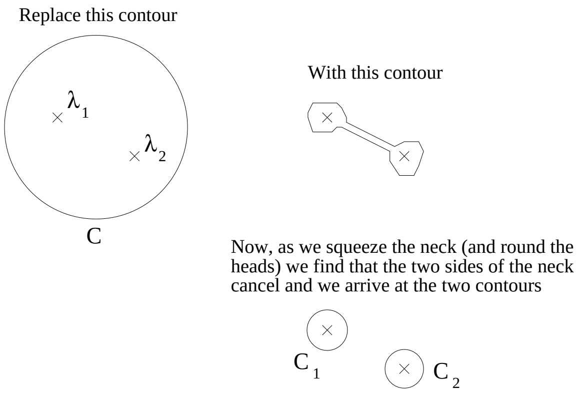

Replace this contour
λ_1
λ_2
C
With this contour
Now, as we squeeze the neck (and round the heads) we find that the two sides of the neck cancel and we arrive at the two contours
C $_1$
C $_2$

Figure 8.2. Concentrating on the poles.

We may view (8.5) as a special instance of integrating a rational function around a curve that encircles all of the zeros of its denominator. In particular, recalling (8.2) and (8.3), we find

$$\int_C r(z)\,dz = \sum_{j=1}^{h}\sum_{k=1}^{m_j}\int_{C_j}\frac{r_{j,k}}{(z-\lambda_j)^k}\,dz = 2\pi i\sum_{j=1}^{h} r_{j,1}. \tag{8.6}$$

To take a slightly more complicated example let us integrate $f(z)/(z-a)$ over some closed curve C inside of which f is differentiable and a resides. Our Curve Replacement Lemma now permits us to claim that

$$\int_C \frac{f(z)}{z-a}\,dz = \int_{C(a,\rho)}\frac{f(z)}{z-a}\,dz.$$

It appears that one can go no further without specifying f. The alert reader however recognizes that the integral over $C(a,\rho)$ is independent of r and so proceeds to let $r \to 0$, in which case $z \to a$ and $f(z) \to f(a)$. Computing the integral of $1/(z-a)$ along the way we are lead to the hope that

$$\int_C \frac{f(z)}{z-a}\,dz = f(a)2\pi i$$

In support of this conclusion we note that

$$\begin{aligned}\int_{C(a,\rho)}\frac{f(z)}{z-a}\,dz &= \int_{C(a,\rho)}\left\{\frac{f(z)}{z-a}+\frac{f(a)}{z-a}-\frac{f(a)}{z-a}\right\}dz\\ &= f(a)\int_{C(a,\rho)}\frac{1}{z-a}\,dz + \int_{C(a,\rho)}\frac{f(z)-f(a)}{z-a}\,dz.\end{aligned}$$

Now the first term is $f(a)2\pi i$ regardless of ρ while, as $\rho \to 0$, the integrand of the second term approaches $f'(a)$ and the region of integration approaches the point a. Regarding this second term, as the integrand remains bounded as the perimeter of $C(a,\rho)$ approaches zero the value of the integral must itself be zero. End of Proof.

This result is typically known as

Proposition 8.4. Cauchy's Integral Formula. If f is differentiable on and in the closed curve C then

$$f(a) = \frac{1}{2\pi i}\int_C \frac{f(z)}{z-a}\,dz \tag{8.7}$$

for each a lying inside C.

The consequences of such a formula run far and deep. We shall delve into only one or two. First, we note that, as a does not lie on C, the right hand side is a perfectly smooth function of a. Hence, differentiating each side, we find

$$f'(a) = \frac{df(a)}{da} = \frac{1}{2\pi i}\int_C \frac{d}{da}\frac{f(z)}{z-a}\,dz = \frac{1}{2\pi i}\int_C \frac{f(z)}{(z-a)^2}\,dz \tag{8.8}$$

for each a lying inside C. Applying this reasoning n times we arrive at a formula for the nth derivative of f at a,

$$\frac{d^n f}{da^n}(a) = \frac{n!}{2\pi i}\int_C \frac{f(z)}{(z-a)^{1+n}}\,dz \tag{8.9}$$

for each a lying inside C. The upshot is that once f is shown to be differentiable it must in fact be infinitely differentiable. As a simple extension let us consider

$$\frac{1}{2\pi i}\int_C \frac{f(z)}{(z-\lambda_1)(z-\lambda_2)^2}\,dz$$

where f is still assumed differentiable on and in C and that C encircles both λ_1 and λ_2. By the curve replacement lemma this integral is the sum

$$\frac{1}{2\pi i}\int_{C_1} \frac{f(z)}{(z-\lambda_1)(z-\lambda_2)^2}\,dz + \frac{1}{2\pi i}\int_{C_2} \frac{f(z)}{(z-\lambda_1)(z-\lambda_2)^2}\,dz$$

where λ_j now lies in only C_j. As $f(z)/(z-\lambda_2)$ is well behaved in C_1 we may use (8.7) to conclude that

$$\frac{1}{2\pi i}\int_{C_1} \frac{f(z)}{(z-\lambda_1)(z-\lambda_2)^2}\,dz = \frac{f(\lambda_1)}{(\lambda_1-\lambda_2)^2}.$$

Similarly, As $f(z)/(z-\lambda_1)$ is well behaved in C_2 we may use (8.8) to conclude that

$$\frac{1}{2\pi i}\int_{C_2} \frac{f(z)}{(z-\lambda_1)(z-\lambda_2)^2}\,dz = \frac{d}{da}\frac{f(a)}{(a-\lambda_1)}\Big|_{a=\lambda_2}.$$

These calculations can be read as a concrete instance of

Proposition 8.5. The Second Residue Theorem. If g is a polynomial with roots $\{\lambda_j\}_{j=1}^h$ of degree $\{m_j\}_{j=1}^h$ and C is a closed curve encircling each of the λ_j and f is differentiable on and in C then

$$\int_C \frac{f(z)}{g(z)}\,dz = 2\pi i\sum_{j=1}^h \mathrm{res}(f/g,\lambda_j)$$

where

$$\mathrm{res}(f/g,\lambda_j) = \lim_{z\to\lambda_j}\frac{1}{(m_j-1)!}\frac{d^{m_j-1}}{dz^{m_j-1}}\left((z-\lambda_j)^{m_j}\frac{f(z)}{g(z)}\right)$$

is called the **residue** of f/g at λ_j by extension of (7.13).

One of the most important ‘instances’ of this theorem is the formula for

8.3. The Inverse Laplace Transform

If r is a rational function with poles $\{\lambda_j\}_{j=1}^h$ then the inverse Laplace transform of r is

$$(\mathcal{L}^{-1}r)(t) \equiv \frac{1}{2\pi i}\int_C r(z)\mathrm{e}^{zt}\,dz \tag{8.10}$$

where C is a curve that encloses each of the poles of r. As a result

$$(\mathcal{L}^{-1}r)(t) = \sum_{j=1}^h \mathrm{res}(r(z)\mathrm{e}^{zt},\lambda_j). \tag{8.11}$$

Let us put this lovely formula to the test. We take our examples from chapter 6. Let us first compute the inverse Laplace Transform of

$$r(z) = \frac{1}{(z+1)^2}.$$

According to (8.11) it is simply the residue of $r(z)\mathrm{e}^{zt}$ at $z=-1$, i.e.,

$$\mathrm{res}(-1) = \lim_{z\to -1} \frac{d}{dz}\mathrm{e}^{zt} = t\mathrm{e}^{-t}.$$

This closes the circle on the example begun in §6.3 and continued in exercise 6.1. For our next example we recall from (6.13) (ignoring the leading 1.965),

$$\begin{aligned}\mathcal{L}x_1(s) &= \frac{(s^2+1.3s+0.31)}{(s+1/4)^2(s^3+1.8s^2+0.87s+0.11))} \\ &= \frac{(s^2+1.3s+0.31)}{(s+1/4)^2(s+1/2)(s+1/5)(s+11/10)},\end{aligned}$$

and so (8.11) dictates that

$$\begin{aligned}x_1(t) = \exp(-t/4)&\frac{d}{ds}\frac{(s^2+1.3s+0.31)}{(s+1/2)(s+1/5)(s+11/10)}\bigg|_{s=-1/4} \\ + \exp(-t/2)&\frac{(s^2+1.3s+0.31)}{(s+1/4)^2(s+1/5)(s+11/10)}\bigg|_{s=-1/2} \\ + \exp(-t/5)&\frac{(s^2+1.3s+0.31)}{(s+1/4)^2(s+1/2)(s+11/10)}\bigg|_{s=-1/5} \\ + \exp(-11t/10)&\frac{(s^2+1.3s+0.31)}{(s+1/4)^2(s+1/2)(s+1/5)}\bigg|_{s=-10/11}.\end{aligned} \tag{8.12}$$

Evaluation of these terms indeed confirms our declaration, (6.15), and the work in exercise 7.9.

The curve replacement lemma of course gives us considerable freedom in our choice of the curve C used to define the inverse Laplace transform, (8.10). As in applications the poles of r are typically in the left half of the complex plane (why?) it is common to choose C to be the half circle

$$C = C_L(\rho) \cup C_A(\rho),$$

comprised of the line segment, C_L, and arc, C_A,

$$C_L(\rho) \equiv \{i\omega : -\rho \le \omega \le \rho\} \quad \text{and} \quad C_A(\rho) \equiv \{\rho\mathrm{e}^{i\theta} : \pi/2 \le \theta \le 3\pi/2\},$$

where ρ is chosen large enough to encircle the poles of r. With this concrete choice, (8.10) takes the form

$$\begin{aligned}(\mathcal{L}^{-1}r)(t) &= \frac{1}{2\pi i}\int_{C_L} r(z)\mathrm{e}^{zt}\,dz + \frac{1}{2\pi i}\int_{C_A} r(z)\mathrm{e}^{zt}\,dz \\ &= \frac{1}{2\pi}\int_{-\rho}^{\rho} r(i\omega)\mathrm{e}^{i\omega t}\,d\omega + \frac{\rho}{2\pi}\int_{\pi/2}^{3\pi/2} r(\rho\mathrm{e}^{i\theta})\mathrm{e}^{\rho\mathrm{e}^{i\theta}t}\mathrm{e}^{i\theta}\,d\theta.\end{aligned} \tag{8.13}$$

Although this second term appears unwieldy it can be shown to vanish as $\rho\to\infty$, in which case we arrive at

$$(\mathcal{L}^{-1}r)(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} r(i\omega)\mathrm{e}^{i\omega t}\,d\omega, \tag{8.14}$$

the conventional definition of the inverse Laplace transform.

8.4. Exercises

1. Compute the integral of z^2 along the parabolic segment $z(t) = t + it^2$ as t ranges from 0 to 1.

2. Evaluate each of the integrals below and state which result you are using, e.g., The bare–handed calculation (8.1), Cauchy's Theorem, The Cauchy Integral Formula, The Second Residue Theorem, and show all of your work.

$$\int_{C(2,1)} \frac{\cos(z)}{z-2}\,dz, \quad \int_{C(2,1)} \frac{\cos(z)}{z(z-2)}\,dz, \quad \int_{C(2,1)} \frac{\cos(z)}{z(z+2)}\,dz,$$
$$\int_{C(0,2)} \frac{\cos(z)}{z^3+z}\,dz, \quad \int_{C(0,2)} \frac{\cos(z)}{z^3}\,dz, \quad \int_{C(0,2)} \frac{z\cos(z)}{z-1}\,dz.$$

3. Let us confirm the representation (8.4) in the matrix case. More precisely, if $R(z) \equiv (zI - B)^{-1}$ is the resolvent associated with B then (8.4) states that

$$R(z) = \sum_{j=1}^{h} \sum_{k=1}^{m_j} \frac{R_{j,k}}{(z-\lambda_j)^k}$$

where

$$R_{j,k} = \frac{1}{2\pi i} \int_{C_j} R(z)(z-\lambda_j)^{k-1}\,dz. \tag{8.15}$$

Compute the $R_{j,k}$ per (8.15) for the B in (7.14). Confirm that they agree with those appearing in (7.17).

4. Use (8.11) to compute the inverse Laplace transform of $1/(s^2 + 2s + 2)$.

5. Use the result of the previous exercise to solve, via the Laplace transform, the differential equation

$$x'(t) + x(t) = \mathrm{e}^{-t}\sin t, \qquad x(0) = 0.$$

Hint: Take the Laplace transform of each side.

6. Evaluate all expressions in (8.12) in Matlab 's symbolic toolbox via `syms`, `diff` and `subs` and confirm that the final result jibes with (6.15).

7. Return to §6.6 and argue how one deduces (6.26) from (6.25). Evaluate (6.26) and graph it with `ezplot` and contrast it with Fig. 6.3.

8. Let us check the limit we declared in going from (8.13) to (8.14). First show that

$$|\mathrm{e}^{\rho \mathrm{e}^{i\theta} t}| = \mathrm{e}^{\rho t \cos\theta}.$$

Next show (perhaps graphically) that

$$\cos\theta \le 1 - 2\theta/\pi \quad \text{when} \quad \pi/2 \le \theta \le \pi.$$

Now confirm each step in

$$\begin{aligned}
\rho\left|\int_{\pi/2}^{3\pi/2} r(\rho\mathrm{e}^{i\theta})\mathrm{e}^{\rho\mathrm{e}^{i\theta}t}\mathrm{e}^{i\theta}\,d\theta\right| &\le \rho\max_{\theta}|r(\rho\mathrm{e}^{i\theta})|\int_{\pi/2}^{3\pi/2}|\mathrm{e}^{\rho\mathrm{e}^{i\theta}t}|\,d\theta \\
&= \rho\max_{\theta}|r(\rho\mathrm{e}^{i\theta})|2\int_{\pi/2}^{\pi}\mathrm{e}^{\rho t\cos\theta}\,d\theta \\
&\le \rho\max_{\theta}|r(\rho\mathrm{e}^{i\theta})|2\int_{\pi/2}^{\pi}\mathrm{e}^{\rho t(1-2\theta/\pi)}\,d\theta \\
&= \max_{\theta}|r(\rho\mathrm{e}^{i\theta})|(\pi/t)(1-\mathrm{e}^{-\rho t}),
\end{aligned}$$

and finally argue why

$$\max_{\theta}|r(\rho\mathrm{e}^{i\theta})|\to 0$$

as $\rho\to\infty$.

9. The Eigenvalue Problem

Harking back to chapter 6 we labeled the complex number λ an **eigenvalue** of B if $\lambda I - B$ was not invertible. In order to find such λ one has only to find those s for which $(sI - B)^{-1}$ is not defined. To take a concrete example we note that if

$$B = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} \tag{9.1}$$

then

$$(sI - B)^{-1} = \frac{1}{(s-1)^2(s-2)} \begin{pmatrix} (s-1)(s-2) & s-2 & 0 \\ 0 & (s-1)(s-2) & 0 \\ 0 & 0 & (s-1)^2 \end{pmatrix} \tag{9.2}$$

and so $\lambda_1 = 1$ and $\lambda_2 = 2$ are the two eigenvalues of B. Now, to say that $\lambda_j I - B$ is not invertible is to say that its columns are linearly dependent, or, equivalently, that the null space $\mathcal{N}(\lambda_j I - B)$ contains more than just the zero vector. We call $\mathcal{N}(\lambda_j I - B)$ the jth **eigenspace** and call each of its nonzero members a jth **eigenvector**. The dimension of $\mathcal{N}(\lambda_j I - B)$ is referred to as the **geometric multiplicity** of λ_j. With respect to B above we compute $\mathcal{N}(\lambda_1 I - B)$ by solving $(I - B)x = 0$, i.e.,

$$\begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Clearly

$$\mathcal{N}(\lambda_1 I - B) = \{c[1\ 0\ 0]^T : c \in \mathbf{R}\}.$$

Arguing along the same lines we also find

$$\mathcal{N}(\lambda_2 I - B) = \{c[0\ 0\ 1]^T : c \in \mathbf{R}\}.$$

That B is 3-by-3 but possesses only 2 linearly independent eigenvectors, much like our patient on the front cover, suggests that matrices can not necessarily be judged by the number of their eigenvectors. After a little probing one might surmise that B's condition is related to the fact that λ_1 is a double pole of $(sI - B)^{-1}$. In order to flesh out that remark and find a proper replacement for the missing eigenvector we must take a much closer look at the resolvent.

9.1. The Resolvent

One means by which to come to grips with $(sI - B)^{-1}$ is to treat it as the matrix analog of the scalar function

$$\frac{1}{s - b}. \tag{9.3}$$

This function is a scaled version of the even simpler function $1/(1 - z)$. This latter function satisfies (recall the n-term geometric series (7.8))

$$\frac{1}{1-z} = 1 + z + z^2 + \cdots + z^{n-1} + \frac{z^n}{1-z} \tag{9.4}$$

for each positive integer n. Furthermore, if $|z| < 1$ then $z^n \to 0$ as $n \to \infty$ and so (9.4) becomes, in the limit,

$$\frac{1}{1-z} = \sum_{n=0}^{\infty} z^n,$$

the full geometric series. Returning to (9.3) we write

$$\frac{1}{s-b} = \frac{1/s}{1-b/s} = \frac{1}{s} + \frac{b}{s^2} + \cdots + \frac{b^{n-1}}{s^n} + \frac{b^n}{s^n}\frac{1}{s-b},$$

and hence, so long as $|s| > |b|$ we find,

$$\frac{1}{s-b} = \frac{1}{s}\sum_{n=0}^{\infty}\left(\frac{b}{s}\right)^n.$$

This same line of reasoning may be applied in the matrix case. That is,

$$(sI-B)^{-1} = s^{-1}(I-B/s)^{-1} = \frac{1}{s} + \frac{B}{s^2} + \cdots + \frac{B^{n-1}}{s^n} + \frac{B^n}{s^n}(sI-B)^{-1}, \tag{9.5}$$

and hence, so long as $|s| > \|B\|$ where $\|B\|$ is the magnitude of the largest eigenvalue of B, we find

$$(sI-B)^{-1} = s^{-1}\sum_{n=0}^{\infty}(B/s)^n. \tag{9.6}$$

Although (9.6) is indeed a formula for the resolvent you may, regarding computation, not find it any more attractive than the Gauss-Jordan method. We view (9.6) however as an analytical rather than computational tool. More precisely, it facilitates the computation of integrals of $R(s)$. For example, if C_ρ is the circle of radius ρ centered at the origin and $\rho > \|B\|$ then

$$\int_{C_\rho}(sI-B)^{-1}\,ds = \sum_{n=0}^{\infty} B^n \int_{C_\rho} s^{-1-n}\,ds = 2\pi i I. \tag{9.7}$$

This result is essential to our study of the eigenvalue problem. As are the two resolvent identities. Regarding the first we deduce from the simple observation

$$(s_2I-B)^{-1} - (s_1I-B)^{-1} = (s_2I-B)^{-1}(s_1I - B - s_2I + B)(s_1I-B)^{-1}$$

that

$$R(s_2) - R(s_1) = (s_1 - s_2)R(s_2)R(s_1). \tag{9.8}$$

The second identity is simply a rewriting of

$$(sI-B)(sI-B)^{-1} = (sI-B)^{-1}(sI-B) = I,$$

namely,

$$BR(s) = R(s)B = sR(s) - I. \tag{9.9}$$

9.2. The Partial Fraction Expansion of the Resolvent

The Gauss–Jordan method informs us that $R(s)$ will be a matrix of rational functions of s, with a common denominator. In keeping with the notation of the previous chapters we assume the denominator to have the h distinct roots, $\{\lambda_j\}_{j=1}^h$ with associated orders $\{m_j\}_{j=1}^h$.

Now, assembling the partial fraction expansions of each element of R we arrive at

$$R(s) = \sum_{j=1}^{h} \sum_{k=1}^{m_j} \frac{R_{j,k}}{(s-\lambda_j)^k} \tag{9.10}$$

where, recalling equation (8.4), $R_{j,k}$ is the matrix

$$R_{j,k} = \frac{1}{2\pi i} \int_{C_j} R(z)(z-\lambda_j)^{k-1}\, dz. \tag{9.11}$$

To take a concrete example, with respect to (9.1) and (9.2) we find

$$R_{1,1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad R_{1,2} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad R_{2,1} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

One notes immediately that these matrices enjoy some amazing properties. For example

$$R_{1,1}^2 = R_{1,1}, \quad R_{2,1}^2 = R_{2,1}, \quad R_{1,1}R_{2,1} = 0, \quad \text{and} \quad R_{1,2}^2 = 0. \tag{9.12}$$

We now show that this is no accident. We shall find it true in general as a consequence of (9.11) and the first resolvent identity.

Proposition 9.1. $R_{j,1}^2 = R_{j,1}$.

Proof: Recall that the C_j appearing in (9.11) is any circle about λ_j that neither touches nor encircles any other root. Suppose that C_j and C_j' are two such circles and C_j' encloses C_j. Now

$$R_{j,1} = \frac{1}{2\pi i} \int_{C_j} R(z)\, dz = \frac{1}{2\pi i} \int_{C_j'} R(z)\, dz$$

and so

$$\begin{aligned} R_{j,1}^2 &= \frac{1}{(2\pi i)^2} \int_{C_j} R(z)\, dz \int_{C_j'} R(w)\, dw \\ &= \frac{1}{(2\pi i)^2} \int_{C_j} \int_{C_j'} R(z)R(w)\, dw\, dz \\ &= \frac{1}{(2\pi i)^2} \int_{C_j} \int_{C_j'} \frac{R(z)-R(w)}{w-z}\, dw\, dz \\ &= \frac{1}{(2\pi i)^2} \left\{ \int_{C_j} R(z) \int_{C_j'} \frac{1}{w-z}\, dw\, dz - \int_{C_j'} R(w) \int_{C_j} \frac{1}{w-z}\, dz\, dw \right\} \\ &= \frac{1}{2\pi i} \int_{C_j} R(z)\, dz = R_{j,1}. \end{aligned}$$

We used the first resolvent identity, (9.8), in moving from the second to the third line. In moving from the fourth to the fifth we used only

$$\int_{C_j'} \frac{1}{w-z}\,dw = 2\pi i \quad \text{and} \quad \int_{C_j} \frac{1}{w-z}\,dz = 0. \tag{9.13}$$

The latter integrates to zero because C_j *does not* encircle w. End of Proof.

Recalling definition 5.1 that matrices that equal their squares are projections we adopt the abbreviation

$$P_j \equiv R_{j,1}.$$

With respect to the product P_jP_k, for $j \neq k$, the calculation runs along the same lines. The difference comes in (9.13) where, as C_j lies completely outside of C_k, both integrals are zero. Hence,

Proposition 9.2. If $j \neq k$ then $P_jP_k = 0$.

Along the same lines we define

$$D_j \equiv R_{j,2}$$

and prove

Proposition 9.3. If $1 \leq k \leq m_j - 1$ then $D_j^k = R_{j,k+1}$. $D_j^{m_j} = 0$.

Proof: For k and ℓ greater than or equal to one,

$$\begin{aligned}
R_{j,k+1}R_{j,\ell+1} &= \frac{1}{(2\pi i)^2}\int_{C_j} R(z)(z-\lambda_j)^k\,dz \int_{C_j'} R(w)(w-\lambda_j)^\ell\,dw \\
&= \frac{1}{(2\pi i)^2}\int_{C_j}\int_{C_j'} R(z)R(w)(z-\lambda_j)^k(w-\lambda_j)^\ell\,dw\,dz \\
&= \frac{1}{(2\pi i)^2}\int_{C_j'}\int_{C_j} \frac{R(z)-R(w)}{w-z}(z-\lambda_j)^k(w-\lambda_j)^\ell\,dw\,dz \\
&= \frac{1}{(2\pi i)^2}\int_{C_j} R(z)(z-\lambda_j)^k \int_{C_j'} \frac{(w-\lambda_j)^\ell}{w-z}\,dw\,dz \\
&\quad - \frac{1}{(2\pi i)^2}\int_{C_j'} R(w)(w-\lambda_j)^\ell \int_{C_j} \frac{(z-\lambda_j)^k}{w-z}\,dz\,dw \\
&= \frac{1}{2\pi i}\int_{C_j} R(z)(z-\lambda_j)^{k+\ell}\,dz = R_{j,k+\ell+1}.
\end{aligned}$$

because

$$\int_{C_j'} \frac{(w-\lambda_j)^\ell}{w-z}\,dw = 2\pi i(z-\lambda_j)^\ell \quad \text{and} \quad \int_{C_j} \frac{(z-\lambda_j)^k}{w-z}\,dz = 0. \tag{9.14}$$

With $k = \ell = 1$ we have shown $R_{j,2}^2 = R_{j,3}$, i.e., $D_j^2 = R_{j,3}$. Similarly, with $k = 1$ and $\ell = 2$ we find $R_{j,2}R_{j,3} = R_{j,4}$, i.e., $D_j^3 = R_{j,4}$, and so on. Finally, at $k = m_j$ we find

$$D_j^{m_j} = R_{j,m_j+1} = \frac{1}{2\pi i}\int_{C_j} R(z)(z-\lambda_j)^{m_j}\,dz = 0$$

by Cauchy's Theorem. End of Proof.

Of course this last result would be trivial if in fact $D_j = 0$. Note that if $m_j > 1$ then

$$D_j^{m_j-1} = R_{j,m_j} = \int_{C_j} R(z)(z-\lambda_j)^{m_j-1}\,dz \neq 0$$

for the integrand then has a term proportional to $1/(z-\lambda_j)$, which we know, by (8.1), leaves a nonzero residue.

With this we now have the sought after expansion

$$R(z) = \sum_{j=1}^{h} \left\{ \frac{1}{z-\lambda_j} P_j + \sum_{k=1}^{m_j-1} \frac{1}{(z-\lambda_j)^{k+1}} D_j^k \right\}, \tag{9.15}$$

along with verification of a number of the properties enjoyed by the P_j and D_j.

In the case that each eigenvalue is simple pole of the resolvent, i.e., $m_j = 1$, we arrive at the compact representation

$$(zI - B)^{-1} = \sum_{j=1}^{n} \frac{P_j}{z-\lambda_j}. \tag{9.16}$$

As a quick application of this we return to a curious limit that arose in our discussion of Backward Euler. Namely, we would like to evaluate

$$\lim_{k\to\infty} (I - (t/k)B)^{-k}.$$

To find this we note that $(zI - B)^{-1} = (I - B/z)^{-1}/z$ and so (9.16) may be written

$$(I - B/z)^{-1} = \sum_{j=1}^{h} \frac{zP_j}{z-\lambda_j} = \sum_{j=1}^{h} \frac{P_j}{1-\lambda_j/z} \tag{9.17}$$

If we now set $z = k/t$, where k is a positive integer, and use the fact that the P_j are projections that annihilate one another then we arrive first at

$$(I - (t/k)B)^{-k} = \sum_{j=1}^{h} \frac{P_j}{(1-(t/k)\lambda_j)^k} \tag{9.18}$$

and so, in the limit find

$$\lim_{k\to\infty} (I - (t/k)B)^{-k} = \sum_{j=1}^{h} \exp(\lambda_j t) P_j \tag{9.19}$$

and naturally refer to this limit as $\exp(Bt)$. We will confirm that this solve our dynamics problems by checking, in the Exercises, that $(\exp(Bt))' = B\exp(Bt)$.

9.3. The Spectral Representation

With just a little bit more work we shall arrive at a similar expansion for B itself. We begin by applying the second resolvent identity, (9.9), to P_j. More precisely, we note that (9.9) implies that

$$\begin{aligned} BP_j = P_jB &= \frac{1}{2\pi i}\int_{C_j}(zR(z)-I)\,dz \\ &= \frac{1}{2\pi i}\int_{C_j} zR(z)\,dz \\ &= \frac{1}{2\pi i}\int_{C_j} R(z)(z-\lambda_j)\,dz + \frac{\lambda_j}{2\pi i}\int_{C_j} R(z)\,dz \\ &= D_j + \lambda_j P_j, \end{aligned} \tag{9.20}$$

where the second equality is due to Cauchy's Theorem and the third arises from adding and subtracting $\lambda_j R(z)$. Summing (9.20) over j we find

$$B\sum_{j=1}^h P_j = \sum_{j=1}^h \lambda_j P_j + \sum_{j=1}^h D_j. \tag{9.21}$$

We can go one step further, namely the evaluation of the first sum. This stems from (9.7) where we integrated $R(s)$ over a circle C_ρ where $\rho > \|B\|$. The connection to the P_j is made by the residue theorem. More precisely,

$$\int_{C_\rho} R(z)\,dz = 2\pi i\sum_{j=1}^h P_j.$$

Comparing this to (9.7) we find

$$\sum_{j=1}^h P_j = I, \tag{9.22}$$

and so (9.21) takes the form

$$B = \sum_{j=1}^h \lambda_j P_j + \sum_{j=1}^h D_j. \tag{9.23}$$

It is this formula that we refer to as the **Spectral Representation** of B. To the numerous connections between the P_j and D_j we wish to add one more. We first write (9.20) as

$$(B-\lambda_j I)P_j = D_j \tag{9.24}$$

and then raise each side to the m_j power. As $P_j^{m_j} = P_j$ and $D_j^{m_j} = 0$ we find

$$(B-\lambda_j I)^{m_j}P_j = 0. \tag{9.25}$$

For this reason we call the range of P_j the jth **generalized eigenspace**, call each of its nonzero members a jth **generalized eigenvector** and refer to the dimension of $\mathcal{R}(P_j)$ as the **algebraic multiplicity** of λ_j. Let us conform that the eigenspace is indeed a subspace of the generalized eigenspace.

Proposition 9.4. $\mathcal{N}(\lambda_j I - B) \subset \mathcal{R}(P_j)$ with equality if and only if $m_j = 1$.

Proof: For each $e \in \mathcal{N}(\lambda_j I - B)$ we show that $P_j e = e$. To see this note that $Be = \lambda_j e$ and so $(sI-B)e = (s-\lambda_j)e$ so $(sI-B)^{-1}e = (s-\lambda_j)^{-1}e$ and so

$$P_j e = \frac{1}{2\pi i}\int_{C_j}(sI-B)^{-1}e\,ds = \frac{1}{2\pi i}\int_{C_j}\frac{1}{s-\lambda_j}e\,ds = e.$$

By (9.24) we see that $\mathcal{R}(P_j) \subset \mathcal{N}(\lambda_j I - B)$ if and only if $m_j = 1$. End of Proof.

With regard to the example with which we began the chapter we note that although it has only two linearly independent eigenvectors the span of the associated generalized eigenspaces indeed fills out $\mathbf{R}^3$, i.e., $\mathcal{R}(P_1) \oplus \mathcal{R}(P_2) = \mathbf{R}^3$. One may view this as a consequence of $P_1 + P_2 = I$, or, perhaps more concretely, as appending the generalized first eigenvector $[0\ 1\ 0]^T$ to the original two eigenvectors $[1\ 0\ 0]^T$ and $[0\ 0\ 1]^T$. In still other words, the algebraic multiplicities sum to the ambient dimension (here 3), while the sum of geometric multiplicities falls short (here 2).

9.4. Diagonalization of a Semisimple Matrix

If $m_j = 1$ then we call λ_j semisimple. If *each* $m_j = 1$ we call B semisimple. In this case we see that

$$\dim \mathcal{N}(\lambda_j I - B) = \dim \mathcal{R}(P_j) \equiv n_j$$

and that

$$\sum_{j=1}^{h} n_j = n.$$

We then denote by $E_j = [e_{j,1}\ e_{j,2}\ \cdots\ e_{j,n_j}]$ a matrix composed of basis vectors of $\mathcal{R}(P_j)$. We note that

$$Be_{j,k} = \lambda_j e_{j,k},$$

and so

$$BE = E\Lambda \quad \text{where} \quad E = [E_1\ E_2\ \cdots E_h] \tag{9.26}$$

and Λ is the diagonal matrix of eigenvalues,

$$\Lambda = \mathrm{diag}(\lambda_1 \mathtt{ones}(n_1,1)\ \lambda_2 \mathtt{ones}(n_2,1)\ \cdots \lambda_h \mathtt{ones}(n_h,1)).$$

As the eigenmatrix, E, is square and composed of linearly independent columns it is invertible and so (9.26) may be written

$$B = E\Lambda E^{-1} \quad \text{and} \quad \Lambda = E^{-1}BE. \tag{9.27}$$

In this sense we say that E diagonalizes B. Let us work out a few examples. With respect to the rotation matrix

$$B = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

we recall, see (8.3), that

$$\begin{aligned} R(s) &= \frac{1}{s^2+1}\begin{pmatrix} s & 1 \\ -1 & s \end{pmatrix} \\ &= \frac{1}{s-i}\begin{pmatrix} 1/2 & -i/2 \\ i/2 & 1/2 \end{pmatrix} + \frac{1}{s+i}\begin{pmatrix} 1/2 & i/2 \\ -i/2 & 1/2 \end{pmatrix} \\ &= \frac{1}{s-\lambda_1}P_1 + \frac{1}{s-\lambda_2}P_2, \end{aligned} \tag{9.28}$$

and so

$$B = \lambda_1 P_1 + \lambda_2 P_2 = i\begin{pmatrix} 1/2 & -i/2 \\ i/2 & 1/2 \end{pmatrix} - i\begin{pmatrix} 1/2 & i/2 \\ -i/2 & 1/2 \end{pmatrix}.$$

As $m_1 = m_2 = 1$ we see that B is semisimple. By inspection we see that $\mathcal{R}(P_1)$ and $\mathcal{R}(P_2)$ are spanned by

$$e_1 = \begin{pmatrix} 1 \\ i \end{pmatrix} \quad \text{and} \quad e_2 = \begin{pmatrix} 1 \\ -i \end{pmatrix},$$

respectively. It follows that

$$E = \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix}, \quad E^{-1} = \frac{1}{2}\begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix}, \quad \Lambda = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}$$

and so

$$B = E\Lambda E^{-1} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix}\begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}\begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix}.$$

Consider next,

$$B = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{pmatrix} \tag{9.29}$$

and note that `[E,L] = eig(B)` returns

$$E = \begin{pmatrix} 1 & 0 & 1/\sqrt{3} \\ 0 & 1 & 1/\sqrt{3} \\ 0 & 0 & 1/\sqrt{3} \end{pmatrix} \quad L = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

and so $B = ELE^{-1}$. We note that although we did not explicitly compute the resolvent we do not have explicit information about the orders of its poles. However, as Matlab has returned 2 eigenvalues, $\lambda_1 = 1$ and $\lambda_2 = 2$, with geometric multiplicities $n_1 = 2$ (the first two columns of E are linearly independent) and $n_2 = 1$ it follows that $m_1 = m_2 = 1$.

9.5. The Characteristic Polynomial

Our understanding of eigenvalues as poles of the resolvent has permitted us to exploit the beautiful field of complex integration and arrive at a (hopefully) clear understanding of the Spectral Representation. It is time however to take note of alternate paths to this fundamental result. In fact the most common understanding of eigenvalues is not via poles of the resolvent but rather as zeros of the characteristic polynomial. Of course an eigenvalue is an eigenvalue and mathematics will not permit contradictory definitions. The starting point is as above - an eigenvalue of the n-by-n matrix B is a value of s for which $(sI - B)$ is not invertible. Of the many tests for invertibility let us here recall that a matrix is not invertible if and only if it has a zero pivot. If we denote the jth pivot of B by $p_j(B)$ we may define the **determinant** of B as

$$\det(B) \equiv \prod_{j=1}^{n} p_j(B)$$

For example, the pivots of

$$B = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

are a and $d - bc/a$ and so

$$\det(B) \equiv ad - bc.$$

Similarly, the pivots of

$$sI - B = \begin{pmatrix} s-a & -b \\ -c & s-d \end{pmatrix}$$

are $s-a$ and $(s-d) - bc/(s-a)$ and so

$$\chi_B(s) = (s-a)(s-d) - bc.$$

This is a degree-2 polynomial for the 2-by-2 matrix B. The formula that defines the determinant of a general n-by-n matrix, though complicated, always involves an alternating sum of products of n elements of B, and hence

$$\chi_B(s) \equiv \det(sI - B)$$

is a polynomial of degree n. It is called the **characteristic polynomial** of B. Since the zeros of $\chi_B(s)$ are the points where $\det(sI - B) = 0$, those zeros are precisely the points where $sI - B$ is not invertible, i.e., the eigenvalues of B. The Fundamental Theorem of Algebra tells us that a degree-n polynomial has no more than n roots, and thus an n-by-n matrix can never have more than n distinct eigenvalues. In the language of §9.3, we must have $h \le n$. Let us look at a few examples. First, consider the matrix from §9.5,

$$B = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

The simple formula for the 2-by-2 determinant leads to

$$\det(sI - B) = s^2 + 1 = (s - i)(s + i),$$

confirming that the eigenvalues of B are $\pm i$.

For larger matrices we turn to MATLAB . Its `poly` function directly delivers the coefficients of the characteristic polynomial. Returning to the 3-by-3 example of (9.1), we find

```
>> B = [1 1 0; 0 1 0; 0 0 2]
>> chi_B = poly(B)
chi_B = 1 -4 5 -2
>> roots(p)
2.0000
1.0000 + 0.0000i
1.0000 - 0.0000i
```

which means $\chi_B(s) = s^3 - 4s^2 + 5s - 2 = (s-1)^2(s-2)$ which is none other than the common denominator in our resolvent (9.2). It is interesting to contrast this with the example of (9.29). They have the same characteristic polynomials and therefore the same eigenvalues, although the latter is semisimple while the former is not.

9.6. Exercises

1. Argue as in Proposition 9.1 that if $j \ne k$ then $D_j P_k = P_j D_k = 0$.

2. Argue from (9.24) and $P_j^2 = P_j$ that $D_j P_j = P_j D_j = D_j$.

3. We suggested in Chapter 6 that $x(t) = \exp(Bt)x(0)$ is the solution of the dynamical system $x'(t) = Bx(t)$. Let us confirm that our representation (for semisimple B)

$$\exp(Bt) = \sum_{j=1}^{h} \exp(\lambda_j t) P_j \tag{9.30}$$

indeed satisfies

$$(\exp(Bt))' = B\exp(Bt).$$

To do this merely compare the t derivative of (9.30) with the product of

$$B = \sum_{j=1}^{h} \lambda_j P_j$$

and (9.30).

4. Let us compute the matrix exponential, $\exp(Bt)$, for the nonsemisimple matrix B in (9.1) by following the argument that took us from (9.16) to (9.19). To begin, deduce from

$$(zI - B)^{-1} = \frac{1}{z-\lambda_1}P_1 + \frac{1}{(z-\lambda_1)^2}D_1 + \frac{1}{z-\lambda_2}P_2$$

that

$$(I - B/z)^{-1} = \frac{1}{1-\lambda_1/z}P_1 + \frac{1}{z(1-\lambda_1/z)^2}D_1 + \frac{1}{1-\lambda_2/z}P_2.$$

Next take explicit products of this matrix with itself and deduce (using exercise 2) that

$$(I - B/z)^{-k} = \frac{1}{(1-\lambda_1/z)^k}P_1 + \frac{k}{z(1-\lambda_1/z)^{k+1}}D_1 + \frac{1}{(1-\lambda_2/z)^k}P_2.$$

Next set $z = k/t$ and find that

$$(I - (t/k)B)^{-k} = \frac{1}{(1-(t/k)\lambda_1)^k}P_1 + \frac{t}{(1-(t/k)\lambda_1)^{k+1}}D_1 + \frac{1}{(1-(t/k)\lambda_2)^k}P_2,$$

and deduce that

$$\exp(Bt) = \lim_{k\to\infty}(I - (t/k)B)^{-k} = \exp(\lambda_1 t)P_1 + t\exp(\lambda_1 t)D_1 + \exp(\lambda_2 t)P_2.$$

Finally, argue as in the previous exercise that $(\exp(Bt))' = B\exp(Bt)$.

5. We would now like to argue, by example, that the Fourier Matrix of Exer. 7.7 diagonalizes every circulant matrix. We call this matrix

$$B = \begin{pmatrix} 2 & 8 & 6 & 4 \\ 4 & 2 & 8 & 6 \\ 6 & 4 & 2 & 8 \\ 8 & 6 & 4 & 2 \end{pmatrix}$$

circulant because each column is a shifted version of its predecessor. First compare the results of `eig(B)` and $F_4^* B(:,1)$ and then confirm that

$$B = F_4 \text{diag}(F_4^* B(:,1)) F_4^*/4.$$

Why must we divide by 4? Now check the analogous formula on a circulant 5-by-5 circulant matrix of your choice. Submit a marked-up diary of your computations.

6. Let us return exercise 6.7 and study the eigenvalues of B as functions of the damping d when each mass and stiffness is 1. In this case

$$B = \begin{pmatrix} 0 & I \\ -S & -dS \end{pmatrix} \quad \text{where} \quad S = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

(i) Write and execute a MATLAB program that plots, as below, the 6 eigenvalues of B as d ranges from 0 to 1.1 in increments of 0.005.

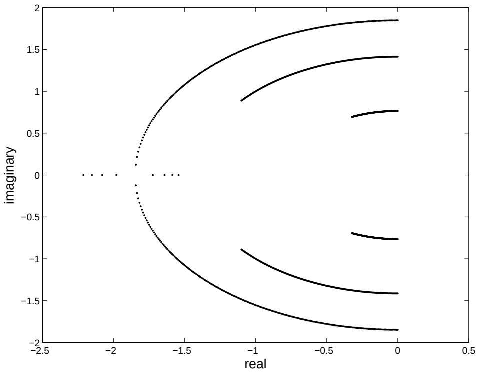


Figure 9.1. Trajectories of eigenvalues of the damped chain as the damping increased.

(ii) Argue that if $[u; v]^T$ is an eigenvalue of B with eigenvalue λ then $v = \lambda u$ and $-Su - dSv = \lambda v$. Substitute the former into the latter and deduce that

$$Su = \frac{-\lambda^2}{1 + d\lambda} u.$$

(iii) Confirm, from Exercise 7.10, that the eigenvalues of S are $\mu_1 = 2 + \sqrt{2}$, $\mu_2 = 2$ and $\mu_3 = 2 - \sqrt{2}$ and hence that the six eigenvalues of B are the roots of the 3 quadratics

$$\lambda^2 + d\mu_j\lambda + \mu_j = 0, \quad \text{i.e.,}\ \lambda_{\pm j} = \frac{-d\mu_j \pm \sqrt{(d\mu_j)^2 - 4\mu_j}}{2}.$$

Deduce from the projections in Exer. 7.10 the 6 associated eigenvectors of B.

(iv) Now argue that when d obeys $(d\mu_j)^2 = 4\mu_j$ that a complex pair of eigenvalues of B collide on the real line and give rise to a nonsemisimple eigenvalue. Describe Figure 9.1 in light of your analysis.

10. The Spectral Representation of a Symmetric Matrix

Our goal in this chapter is to show that if B is symmetric then

- each eigenvalue, λ_j, is real,
- each eigenprojection, P_j, is symmetric and
- each eigennilpotent, D_j, vanishes.

Let us begin with an example. The resolvent of

$$B = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

is

$$\begin{aligned} R(s) &= \frac{1}{s(s-3)} \begin{pmatrix} s-2 & 1 & 1 \\ 1 & s-2 & 1 \\ 1 & 1 & s-2 \end{pmatrix} \\ &= \frac{1}{s} \begin{pmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{pmatrix} + \frac{1}{s-3} \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}. \\ &= \frac{1}{s-\lambda_1} P_1 + \frac{1}{s-\lambda_2} P_2 \end{aligned}$$

and so indeed each of the bullets holds true. With each of the D_j falling by the wayside you may also expect that the respective geometric and algebraic multiplicities coincide.

10.1. The Spectral Representation

We begin with

Proposition 10.1. If B is real and $B = B^T$ then the eigenvalues of B are real.

Proof: We suppose that λ and x comprise an eigenpair of B, i.e., $Bx = \lambda x$ and show that $\lambda = \overline{\lambda}$. On conjugating each side of this eigen equation we find

$$Bx = \lambda x \quad \text{and} \quad B\overline{x} = \overline{\lambda}\overline{x} \tag{10.1}$$

where we have used the fact that B is real. We now take the inner product of the first with $\overline{x}$ and the second with x and find

$$\overline{x}^T Bx = \lambda \|x\|^2 \quad \text{and} \quad x^T B\overline{x} = \overline{\lambda}\|x\|^2. \tag{10.2}$$

The left hand side of the first is a scalar and so equal to its transpose

$$\overline{x}^T Bx = (\overline{x}^T Bx)^T = x^T B^T \overline{x} = x^T B\overline{x} \tag{10.3}$$

where in the last step we used $B = B^T$. Combining (10.2) and (10.3) we see

$$\lambda \|x\|^2 = \overline{\lambda}\|x\|^2.$$

Finally, as $\|x\| > 0$ we may conclude that $\lambda = \overline{\lambda}$. End of Proof.

We next establish

Proposition 10.2. If B is real and $B = B^T$ then each eigenprojection, P_j, and each eigennilpotent, D_j, is real and symmetric.

Proof: From the fact (please prove it) that "the transpose of the inverse is the inverse of the transpose" we learn that

$$\{(sI-B)^{-1}\}^T = \{(sI-B)^T\}^{-1} = (sI-B)^{-1}$$

i.e., the resolvent of a symmetric matrix is symmetric. Hence

$$\begin{aligned}P_j^T &= \left(\frac{1}{2\pi i}\int_{C_j}(sI-B)^{-1}\,ds\right)^T = \frac{1}{2\pi i}\int_{C_j}\{(sI-B)^{-1}\}^T\,ds \\ &= \frac{1}{2\pi i}\int_{C_j}(sI-B)^{-1}\,ds = P_j.\end{aligned}$$

By the same token, we find that each $D_j^T = D_j$. The elements of P_j and D_j are real because they are residues of real functions evaluated at real poles. End of Proof.

The next result will banish the nilpotent component.

Proposition 10.3. The zero matrix is the only real symmetric nilpotent matrix.

Proof: Suppose that D is n-by-n and $D = D^T$ and $D^m = 0$ for some positive integer m. We show that $D^{m-1} = 0$ by showing that every vector lies in its null space. To wit, if $x \in \mathbf{R}^n$ then

$$\begin{aligned}\|D^{m-1}x\|^2 &= x^T(D^{m-1})^T D^{m-1}x \\ &= x^T D^{m-1}D^{m-1}x \\ &= x^T D^{m-2}D^m x \\ &= 0.\end{aligned}$$

As $D^{m-1}x = 0$ for every x it follows (recall Exercise 3.3) that $D^{m-1} = 0$. Continuing in this fashion we find $D^{m-2} = 0$ and so, eventually, $D = 0$. End of Proof.

We have now established the key result of this chapter.

Proposition 10.4. If B is real and symmetric then

$$B = \sum_{j=1}^{h} \lambda_j P_j \tag{10.4}$$

where the λ_j are real and the P_j are real orthogonal projections that sum to the identity and whose pairwise products vanish.

One indication that things are simpler when using the spectral representation is

$$B^{100} = \sum_{j=1}^{h} \lambda_j^{100} P_j. \tag{10.5}$$

As this holds for all powers it even holds for power series. As a result,

$$\exp(B) = \sum_{j=1}^{h} \exp(\lambda_j) P_j.$$

It is also extremely useful in attempting to solve

$$Bx = b$$

for x. Replacing B by its spectral representation and b by Ib or, more to the point by $\sum_j P_j b$ we find

$$\sum_{j=1}^{h} \lambda_j P_j x = \sum_{j=1}^{h} P_j b.$$

Multiplying through by P_1 gives $\lambda_1 P_1 x = P_1 b$ or $P_1 x = P_1 b/\lambda_1$. Multiplying through by the subsequent P_js gives $P_j x = P_j b/\lambda_j$. Hence,

$$x = \sum_{j=1}^{h} P_j x = \sum_{j=1}^{h} \frac{1}{\lambda_j} P_j b. \tag{10.6}$$

We clearly run in to trouble when one of the eigenvalues vanishes. This, of course, is to be expected. For a zero eigenvalue indicates a nontrivial null space which signifies dependencies in the columns of B and hence the lack of a unique solution to $Bx = b$.

Another way in which (10.6) may be viewed is to note that, when B is symmetric, (9.15) takes the form

$$(zI - B)^{-1} = \sum_{j=1}^{h} \frac{1}{z - \lambda_j} P_j.$$

Now if 0 is not an eigenvalue we may set $z = 0$ in the above and arrive at

$$B^{-1} = \sum_{j=1}^{h} \frac{1}{\lambda_j} P_j. \tag{10.7}$$

Hence, the solution to $Bx = b$ is

$$x = B^{-1} b = \sum_{j=1}^{h} \frac{1}{\lambda_j} P_j b,$$

as in (10.6). With (10.7) we have finally reached a point where we can begin to define an inverse even for matrices with dependent columns, i.e., with a zero eigenvalue. We simply exclude the offending term in (10.7). Supposing that $\lambda_h = 0$ we define the **pseudo–inverse** of B to be

$$B^+ \equiv \sum_{j=1}^{h-1} \frac{1}{\lambda_j} P_j.$$

Let us now see whether it is deserving of its name. More precisely, when $b \in \mathcal{R}(B)$ we would expect that $x = B^+ b$ indeed satisfies $Bx = b$. Well

$$BB^+ b = B \sum_{j=1}^{h-1} \frac{1}{\lambda_j} P_j b = \sum_{j=1}^{h-1} \frac{1}{\lambda_j} B P_j b = \sum_{j=1}^{h-1} \frac{1}{\lambda_j} \lambda_j P_j b = \sum_{j=1}^{h-1} P_j b.$$

It remains to argue that the latter sum really is b. We know that

$$b = \sum_{j=1}^{h} P_j b \quad \text{and} \quad b \in \mathcal{R}(B).$$

The latter informs us that $b \perp \mathcal{N}(B^T)$. As $B = B^T$ we have in fact that $b \perp \mathcal{N}(B)$. As P_h is nothing but orthogonal projection onto $\mathcal{N}(B)$ it follows that $P_h b = 0$ and so $B(B^+ b) = b$, that is, $x = B^+ b$ is a solution to $Bx = b$.

The representation (10.5) is unarguably terse and in fact is often written out in terms of individual eigenvectors. Let us see how this is done. Note that if $x \in \mathcal{R}(P_1)$ then $x = P_1 x$ and so

$$Bx = BP_1 x = \sum_{j=1}^{h} \lambda_j P_j P_1 x = \lambda_1 P_1 x = \lambda_1 x,$$

i.e., x is an eigenvector of B associated with λ_1. Similarly, every (nonzero) vector in $\mathcal{R}(P_j)$ is an eigenvector of B associated with λ_j.

Next let us demonstrate that each element of $\mathcal{R}(P_j)$ is orthogonal to each element of $\mathcal{R}(P_k)$ when $j \neq k$. If $x \in \mathcal{R}(P_j)$ and $y \in \mathcal{R}(P_k)$ then

$$x^T y = (P_j x)^T P_k y = x^T P_j P_k y = 0.$$

With this we note that if $\{x_{j,1}, x_{j,2}, \ldots, x_{j,n_j}\}$ constitutes a basis for $\mathcal{R}(P_j)$ then in fact the union of such bases,

$$\{x_{j,p} : 1 \leq j \leq h,\ 1 \leq p \leq n_j\}, \tag{10.8}$$

forms a linearly independent set. Notice now that this set has

$$\sum_{j=1}^{h} n_j$$

elements. That these dimensions indeed sum to the ambient dimension, n, follows directly from the fact that the underlying P_j sum to the n-by-n identity matrix. We have just proven

Proposition 10.5. If B is real and symmetric and n-by-n then B has a set of n linearly independent eigenvectors.

Getting back to a more concrete version of (10.5) we now assemble matrices from the individual bases

$$E_j \equiv [x_{j,1}\ x_{j,2}\ \ldots\ x_{j,n_j}]$$

and note, once again, that $P_j = E_j(E_j^T E_j)^{-1} E_j^T$, and so

$$B = \sum_{j=1}^{h} \lambda_j E_j (E_j^T E_j)^{-1} E_j^T. \tag{10.9}$$

I understand that you may feel a little underwhelmed with this formula. If we work a bit harder we can remove the presence of the annoying inverse. What I mean is that it is possible to choose a basis for each $\mathcal{R}(P_j)$ for which each of the corresponding E_j satisfy $E_j^T E_j = I$. As this construction is fairly general let us devote a separate section to it.

10.2. Gram–Schmidt Orthogonalization

Suppose that M is an m-dimensional subspace with basis

$$\{x_1, \ldots, x_m\}.$$

We transform this into an orthonormal basis

$$\{q_1, \ldots, q_m\}$$

for M via
(GS1) Set $y_1 = x_1$ and $q_1 = y_1/\|y_1\|$.
(GS2) $y_2 = x_2$ minus the projection of x_2 onto the line spanned by q_1. That is

$$y_2 = x_2 - q_1(q_1^T q_1)^{-1} q_1^T x_2 = x_2 - q_1 q_1^T x_2.$$

Set $q_2 = y_2/\|y_2\|$ and $Q_2 = [q_1\ q_2]$.
(GS3) $y_3 = x_3$ minus the projection of x_3 onto the plane spanned by q_1 and q_2. That is

$$\begin{aligned} y_3 &= x_3 - Q_2(Q_2^T Q_2)^{-1} Q_2^T x_3 \\ &= x_3 - q_1 q_1^T x_3 - q_2 q_2^T x_3. \end{aligned}$$

Set $q_3 = y_3/\|y_3\|$ and $Q_3 = [q_1\ q_2\ q_3]$.
Continue in this fashion through step
(GSm) $y_m = x_m$ minus its projection onto the subspace spanned by the columns of Q_{m-1}. That is

$$\begin{aligned} y_m &= x_m - Q_{m-1}(Q_{m-1}^T Q_{m-1})^{-1} Q_{m-1}^T x_m \\ &= x_m - \sum_{j=1}^{m-1} q_j q_j^T x_m. \end{aligned}$$

Set $q_m = y_m/\|y_m\|$.

To take a simple example, let us orthogonalize the following basis for $\mathbf{R}^3$,

$$x_1 = [1\ 0\ 0]^T, \quad x_2 = [1\ 1\ 0]^T, \quad x_3 = [1\ 1\ 1]^T.$$

(GS1) $q_1 = y_1 = x_1$.
(GS2) $y_2 = x_2 - q_1 q_1^T x_2 = [0\ 1\ 0]^T$, and so $q_2 = y_2$.
(GS3) $y_3 = x_3 - q_1 q_1^T x_3 - q_2 q_2^T x_3 = [0\ 0\ 1]^T$, and so $q_3 = y_3$.
We have arrived at

$$q_1 = [1\ 0\ 0]^T, \quad q_2 = [0\ 1\ 0]^T, \quad q_3 = [0\ 0\ 1]^T. \tag{10.10}$$

Once the idea is grasped the actual calculations are best left to a machine. Matlab accomplishes this via the `orth` command. Its implementation is a bit more sophisticated than a blind run through our steps (GS1) through (GSm). As a result, there is no guarantee that it will return the same basis. For example

```
>> X=[1 1 1;0 1 1;0 0 1];
>> Q=orth(X)
Q =
  0.7370 -0.5910  0.3280
  0.5910  0.3280 -0.7370
  0.3280  0.7370  0.5910
```

This ambiguity does not bother us, for one orthogonal basis is as good as another. We next put this into practice, via (10.9).

10.3. The Diagonalization of a Symmetric Matrix

By choosing an orthogonal basis $\{q_{j,k} : 1 \le k \le n_j\}$ for each $\mathcal{R}(P_j)$ and collecting the basis vectors in

$$Q_j = [q_{j,1}\ q_{j,2}\ \cdots\ q_{j,n_j}]$$

we find that

$$P_j = Q_j Q_j^T = \sum_{k=1}^{n_j} q_{j,k} q_{j,k}^T.$$

As a result, the spectral representation (10.9) takes the form

$$B = \sum_{j=1}^{h} \lambda_j Q_j Q_j^T = \sum_{j=1}^{h} \lambda_j \sum_{k=1}^{n_j} q_{j,k} q_{j,k}^T. \tag{10.11}$$

This is the spectral representation in perhaps its most detailed dress. There exists, however, still another form! It is a form that you are likely to see in future engineering courses and is achieved by assembling the Q_j into a single n-by-n orthonormal matrix

$$Q = [Q_1\ \cdots\ Q_h].$$

Having orthonormal columns it follows that $Q^TQ = I$. Q being square, it follows in addition that $Q^T = Q^{-1}$. Now

$$Bq_{j,k} = \lambda_j q_{j,k}$$

may be encoded in matrix terms via

$$BQ = Q\Lambda \tag{10.12}$$

where Λ is the n-by-n diagonal matrix whose first n_1 diagonal terms are λ_1, whose next n_2 diagonal terms are λ_2, and so on. That is, each λ_j is repeated according to its multiplicity. Multiplying each side of (10.12), from the right, by Q^T we arrive at

$$\boxed{B = Q\Lambda Q^T.} \tag{10.13}$$

Because one may just as easily write

$$\boxed{Q^T BQ = \Lambda} \tag{10.14}$$

one says that Q **diagonalizes** B.

Let us return to our example

$$B = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

of the last chapter. Recall that the eigenspace associated with $\lambda_1 = 0$ had

$$e_{1,1} = [-1\ 1\ 0]^T \quad \text{and} \quad e_{1,2} = [-1\ 0\ 1]^T$$

for a basis. Via Gram–Schmidt we may replace this with

$$q_{1,1} = \frac{1}{\sqrt{2}}[-1\ 1\ 0]^T \quad \text{and} \quad q_{1,2} = \frac{1}{\sqrt{6}}[-1\ -1\ 2]^T.$$

Normalizing the vector associated with $\lambda_2 = 3$ we arrive at

$$q_2 = \frac{1}{\sqrt{3}}[1\ 1\ 1]^T,$$

and hence

$$Q = [q_{1,1}\ q_{1,2}\ q_2] = \frac{1}{\sqrt{6}}\begin{pmatrix} -\sqrt{3} & -1 & \sqrt{2} \\ \sqrt{3} & -1 & \sqrt{2} \\ 0 & 2 & \sqrt{2} \end{pmatrix} \quad \text{and} \quad \Lambda = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

10.4. Rayleigh's Principle and the Power Method

As the eigenvalues of an n-by-n symmetric matrix, B, are real we may order them from high to low,

$$\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n.$$

In this section we will derive two extremely useful characterizations of the largest eigenvalue. The first was discovered by Lord Rayleigh in his research on sound and vibration. To begin, we denote the associated orthonormal eigenvectors of B by

$$q_1,\ q_2,\ \ldots,\ q_n$$

and note that each $x \in \mathbf{R}^n$ enjoys the expansion

$$x = (x^T q_1)q_1 + (x^T q_2)q_2 + \cdots + (x^T q_n)q_n. \tag{10.15}$$

Applying B to each side we find

$$Bx = (x^T q_1)\lambda_1 q_1 + (x^T q_2)\lambda_2 q_2 + \cdots + (x^T q_n)\lambda_n q_n. \tag{10.16}$$

Now taking the inner product of (10.15) and (10.16) we find

$$\begin{aligned} x^T Bx &= (x^T q_1)^2\lambda_1 + (x^T q_2)^2\lambda_2 + \cdots + (x^T q_n)^2\lambda_n \\ &\le \lambda_1\{(x^T q_1)^2 + (x^T q_2)^2 + \cdots + (x^T q_n)^2\} \\ &= \lambda_1 x^T x. \end{aligned}$$

That is, $x^T Bx \le \lambda_1 x^T x$ for every $x \in \mathbf{R}^n$. This, together with the fact that $q_1^T Bq_1 = \lambda_1 q_1^T q_1$ establishes

Proposition 10.6. Rayleigh's Principle. If B is symmetric then its largest eigenvalue is

$$\lambda_1 = \max_{x \ne 0} \frac{x^T Bx}{x^T x}$$

and the maximum is attained on the line through q_1.

For our next characterization we return to (10.16) and record higher powers of B onto x

$$\begin{aligned} B^k x &= (x^T q_1)\lambda_1^k q_1 + (x^T q_2)\lambda_2^k q_2 + \cdots + (x^T q_n)\lambda_n^k q_n \\ &= (x^T q_1)\lambda_1^k \left(q_1 + \frac{x^T q_2}{x^T q_1}\frac{\lambda_2^k}{\lambda_1^k} q_2 + \cdots + \frac{x^T q_n}{x^T q_1}\frac{\lambda_n^k}{\lambda_1^k} q_n \right). \end{aligned} \tag{10.17}$$

And so, using $\text{sign}(t) \equiv t/|t|$,

$$\frac{B^k x}{\|B^k x\|} = \text{sign}(\lambda_1^k x^T q_1)q_1 + O((\lambda_2/\lambda_1)^k)$$

and note that the latter term goes to zero with increasing k so long as λ_1 is strictly greater than λ_2. If, in addition, we assume that $\lambda_1 > 0$, then the first term does not depend on k and we arrive at

Proposition 10.7. The Power Method. If B is symmetric and its greatest eigenvalue is simple and positive and the initial guess, x, is not orthogonal to q_1 then

$$\lim_{k\to\infty} \frac{B^k x}{\|B^k x\|} = \text{sign}(x^T q_1)q_1$$

10.5. Exercises

1. The stiffness matrix associated with the unstable swing of figure 2.2 is
$$S = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$
(i) Find the three distinct eigenvalues, $\lambda_1 = 1, \lambda_2 = 2, \lambda_3 = 0$, along with their associated eigenvectors $e_{1,1}, e_{1,2}, e_{2,1}, e_{3,1}$, and projection matrices, P_1, P_2, P_3. What are the respective geometric multiplicities?

   (ii) Use the folk theorem that states "in order to transform a matrix it suffices to transform its eigenvalues" to arrive at a guess for $S^{1/2}$. Show that your guess indeed satisfies $S^{1/2}S^{1/2} = S$. Does your $S^{1/2}$ have anything in common with the element-wise square root of S?

   (iii) Show that $\mathcal{R}(P_3) = \mathcal{N}(S)$.

   (iv) Assemble
$$S^+ = \frac{1}{\lambda_1}P_1 + \frac{1}{\lambda_2}P_2$$
and check your result against `pinv(S)` in MATLAB .

   (v) Use S^+ to solve $Sx = f$ where $f = [0\ 1\ 0\ 2]^T$ and carefully draw before and after pictures of the unloaded and loaded swing.

   (vi) It can be very useful to sketch each of the eigenvectors in this fashion. In fact, a movie is the way to go. Please run the MATLAB truss demo by typing `truss` and view all 12 of the movies. Please sketch the 4 eigenvectors of (i) by showing how they deform the swing.

2. The Cayley-Hamilton Theorem. Use Proposition 10.4 to show that if B is real and symmetric and $p(z)$ is a polynomial then
$$p(B) = \sum_{j=1}^{h} p(\lambda_j)P_j. \tag{10.18}$$
Confirm this in the case that B is the matrix in Exercise 1 and $p(z) = z^2 + 1$. Deduce from (10.18) that if p is the characteristic polynomial of B, i.e., $p(z) = \det(zI - B)$, then $p(B) = 0$. Confirm this, via the MATLAB command, `poly`, on the S matrix in Exercise 1.

3. Argue that the least eigenvalue of a symmetric matrix obeys the minimum principle

$$\lambda_n = \min_{x \neq 0} \frac{x^T B x}{x^T x} \tag{10.19}$$

and that the minimum is attained at q_n.

4. The stiffness of a fiber net is typically equated with the least eigenvalue of its stiffness matrix. Let B be the stiffness matrix constructed in Exercise 2 of §2.5, and let C denote the stiffness matrix of the same net, except that the stiffness of fiber 4 is twice that used in constructing B. Show that $x^T B x \leq x^T C x$ for all $x \in \mathbf{R}^4$ and then deduce from (10.19) that the stiffness of the C network exceeds the stiffness of the B network.

5. The Power Method. Submit a MATLAB diary of your application of the Power Method to the S matrix in Exercise 1.

11. The Singular Value Decomposition

11.1. Introduction

The singular value decomposition is, in a sense, the spectral representation of a rectangular matrix. Of course if A is m-by-n and $m \neq n$ then it does not make sense to speak of the eigenvalues of A. We may, however, rely on the previous section to give us relevant spectral representations of the two symmetric matrices

$$A^T A \quad \text{and} \quad AA^T.$$

That these two matrices together may indeed tell us 'everything' about A can be gleaned from

$$\begin{aligned} \mathcal{N}(A^T A) = \mathcal{N}(A), \quad \mathcal{N}(AA^T) = \mathcal{N}(A^T), \\ \mathcal{R}(A^T A) = \mathcal{R}(A^T), \quad \text{and} \quad \mathcal{R}(AA^T) = \mathcal{R}(A). \end{aligned}$$

You have proven the first of these in a previous exercise. The proof of the second is identical. The row and column space results follow from the first two via orthogonality.

On the spectral side, we shall now see that the eigenvalues of AA^T and $A^T A$ are nonnegative and that their nonzero eigenvalues coincide. Let us first confirm this on the A matrix associated with the unstable swing (see figure 2.2)

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{11.1}$$

The respective products are

$$AA^T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad A^T A = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Analysis of the first is particularly simple. Its null space is clearly just the zero vector while $\lambda_1 = 2$ and $\lambda_2 = 1$ are its eigenvalues. Their geometric multiplicities are $n_1 = 1$ and $n_2 = 2$. In $A^T A$ we recognize the S matrix from exercise 10.1 and recall that its eigenvalues are $\lambda_1 = 2$, $\lambda_2 = 1$, and $\lambda_3 = 0$ with multiplicities $n_1 = 1$, $n_2 = 2$, and $n_3 = 1$. Hence, at least for this A, the eigenvalues of AA^T and $A^T A$ are nonnegative and their nonzero eigenvalues coincide. In addition, the geometric multiplicities of the nonzero eigenvalues sum to 3, the rank of A.

Proposition 11.1. The eigenvalues of AA^T and $A^T A$ are nonnegative. Their nonzero eigenvalues, including geometric multiplicities, coincide. The geometric multiplicities of the nonzero eigenvalues sum to the rank of A.

Proof: If $A^T Ax = \lambda x$ then $x^T A^T Ax = \lambda x^T x$, i.e., $\|Ax\|^2 = \lambda \|x\|^2$ and so $\lambda \geq 0$. A similar argument works for AA^T.

Now suppose that $\lambda_j > 0$ and that $\{x_{j,k}\}_{k=1}^{n_j}$ constitutes an orthogonal basis for the eigenspace $\mathcal{R}(P_j)$. Starting from

$$A^T Ax_{j,k} = \lambda_j x_{j,k} \tag{11.2}$$

we find, on multiplying through (from the left) by A that

$$AA^T Ax_{j,k} = \lambda_j Ax_{j,k},$$

i.e., λ_j is an eigenvalue of AA^T with eigenvector $Ax_{j,k}$, so long as $Ax_{j,k} \neq 0$. It follows from the first paragraph of this proof that $\|Ax_{j,k}\| = \sqrt{\lambda_j}$, which, by hypothesis, is nonzero. Hence,

$$y_{j,k} \equiv \frac{Ax_{j,k}}{\sqrt{\lambda_j}}, \quad 1 \le k \le n_j \tag{11.3}$$

is a collection of unit eigenvectors of AA^T associated with λ_j. Let us now show that these vectors are orthonormal for fixed j.

$$y_{j,i}^T y_{j,k} = \frac{1}{\lambda_j} x_{j,i}^T A^T A x_{j,k} = x_{j,i}^T x_{j,k} = 0.$$

We have now demonstrated that if $\lambda_j > 0$ is an eigenvalue of A^TA of geometric multiplicity n_j then it is an eigenvalue of AA^T of geometric multiplicity at least n_j. Reversing the argument, i.e., generating eigenvectors of A^TA from those of AA^T we find that the geometric multiplicities must indeed coincide.

Regarding the rank statement, we discern from (11.2) that if $\lambda_j > 0$ then $x_{j,k} \in \mathcal{R}(A^TA)$. The union of these vectors indeed constitutes a basis for $\mathcal{R}(A^TA)$, for anything orthogonal to each of these $x_{j,k}$ necessarily lies in the eigenspace corresponding to a zero eigenvalue, i.e., in $\mathcal{N}(A^TA)$. As $\mathcal{R}(A^TA) = \mathcal{R}(A^T)$ it follows that $\dim \mathcal{R}(A^TA) = r$ and hence the n_j, for $\lambda_j > 0$, sum to r. End of Proof.

Let us now gather together some of the separate pieces of the proof. For starters, we order the eigenvalues of A^TA from high to low,

$$\lambda_1 > \lambda_2 > \cdots > \lambda_h$$

and write

$$A^TA = X\Lambda_n X^T \tag{11.4}$$

where

$$X = [X_1 \cdots X_h], \quad \text{where} \quad X_j = [x_{j,1} \cdots x_{j,n_j}]$$

and Λ_n is the n-by-n diagonal matrix with λ_1 in the first n_1 slots, λ_2 in the next n_2 slots, *etc.* Similarly

$$AA^T = Y\Lambda_m Y^T \tag{11.5}$$

where

$$Y = [Y_1 \cdots Y_h], \quad \text{where} \quad Y_j = [y_{j,1} \cdots y_{j,n_j}].$$

and Λ_m is the m-by-m diagonal matrix with λ_1 in the first n_1 slots, λ_2 in the next n_2 slots, *etc.* The $y_{j,k}$ were defined in (11.3) under the assumption that $\lambda_j > 0$. If $\lambda_j = 0$ let Y_j denote an orthonormal basis for $\mathcal{N}(AA^T)$. Finally, call

$$\sigma_j = \sqrt{\lambda_j}$$

and let Σ denote the m-by-n matrix diagonal matrix with σ_1 in the first n_1 slots and σ_2 in the next n_2 slots, *etc.* Notice that

$$\Sigma^T\Sigma = \Lambda_n \quad \text{and} \quad \Sigma\Sigma^T = \Lambda_m. \tag{11.6}$$

Now recognize that (11.3) may be written

$$Ax_{j,k} = \sigma_j y_{j,k}$$

and that this is simply the column by column rendition of

$$AX = Y\Sigma.$$

As $XX^T = I$ we may multiply through (from the right) by X^T and arrive at the **singular value decomposition** of A,

$$A = Y\Sigma X^T. \tag{11.7}$$

Let us confirm this on the A matrix in (11.1). We have

$$\Lambda_4 = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad X = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 & 0 & 0 & 1 \\ 0 & \sqrt{2} & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & \sqrt{2} & 0 \end{pmatrix}$$

and

$$\Lambda_3 = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad Y = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Hence,

$$\Sigma = \begin{pmatrix} \sqrt{2} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{11.8}$$

and so $A = Y\Sigma X^T$ says that A should coincide with

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{2} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} -1/\sqrt{2} & 0 & 1/\sqrt{2} & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} & 0 \end{pmatrix}.$$

This indeed agrees with A. It also agrees (up to sign changes in the columns of X) with what one receives upon typing `[Y,SIG,X]=svd(A)` in Matlab .

You now ask what we get for our troubles. I express the first dividend as a proposition that looks to me like a quantitative version of the fundamental theorem of linear algebra.

Proposition 11.2. If $Y\Sigma X^T$ is the singular value decomposition of A then
(i) The rank of A, call it r, is the number of nonzero elements in Σ.
(ii) The first r columns of X constitute an orthonormal basis for $\mathcal{R}(A^T)$. The $n - r$ last columns of X constitute an orthonormal basis for $\mathcal{N}(A)$.
(iii) The first r columns of Y constitute an orthonormal basis for $\mathcal{R}(A)$. The $m - r$ last columns of Y constitute an orthonormal basis for $\mathcal{N}(A^T)$.

Let us now 'solve' $Ax = b$ with the help of the pseudo–inverse of A. You know the 'right' thing to do, namely reciprocate all of the nonzero singular values. Because m is not necessarily n we must also be careful with dimensions. To be precise, let Σ^+ denote the n-by-m matrix whose first n_1 diagonal elements are $1/\sigma_1$, whose next n_2 diagonal elements are $1/\sigma_2$ and so on. In the case that $\sigma_h = 0$, set the final n_h diagonal elements of Σ^+ to zero. Now, one defines the **pseudo-inverse** of A to be

$$A^+ \equiv X\Sigma^+ Y^T.$$

Taking the A of (11.1) we find

$$\Sigma^+ = \begin{pmatrix} 1/\sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

and so

$$\begin{aligned} A^+ &= \begin{pmatrix} -1/\sqrt{2} & 0 & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 & 0 \\ 1/\sqrt{2} & 0 & 0 & 1/\sqrt{2} \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 0 & -1/2 & 0 \\ 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{aligned}$$

in agreement with what appears from `pinv(A)`. Let us now investigate the sense in which A^+ is the inverse of A. Suppose that $b \in \mathbf{R}^m$ and that we wish to solve $Ax = b$. We suspect that A^+b should be a good candidate. Observe now that

$$\begin{aligned} (A^TA)A^+b &= X\Lambda_n X^T X\Sigma^+ Y^T b \qquad \text{by (11.4)} \\ &= X\Lambda_n\Sigma^+Y^Tb \qquad \text{because } X^TX = I \\ &= X\Sigma^T\Sigma\Sigma^+Y^Tb \qquad \text{by (11.6)} \\ &= X\Sigma^TY^Tb \qquad \text{because } \Sigma^T\Sigma\Sigma^+ = \Sigma^T \\ &= A^Tb \qquad \text{by (11.7),} \end{aligned}$$

that is, A^+b satisfies the **least-squares problem** $A^TAx = A^Tb$.

11.2. The SVD in Image Compression

Most applications of the SVD are manifestations of the folk theorem, *The singular vectors associated with the largest singular values capture the essence of the matrix.* This is most easily seen when applied to gray scale images. For example, the `jpeg` associated with the image in the top left of figure 11.2 is 262-by-165 matrix. Such a matrix-image is read, displayed and 'decomposed' by

```
M = imread('jb.jpg'); imagesc(M); colormap(gray);
[Y,S,X] = svd(double(M));
```

The singular values lie on the diagonal of S and are arranged in decreasing order. We see their rapid decline in Figure 11.1.

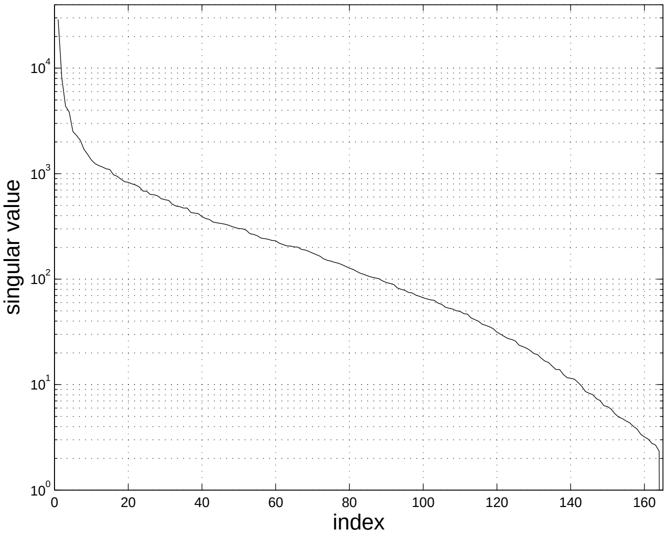


Figure 11.1. The singular values of John Brown.

We now experiment with quantitative versions of the folk theorem. In particular, we examine the result of keeping but the first k singular vectors and values. That is we construct

```
  Ak = Y(:,1:k)*S(1:k,1:k)*X(:,1:k)';
```

for decreasing values of k.

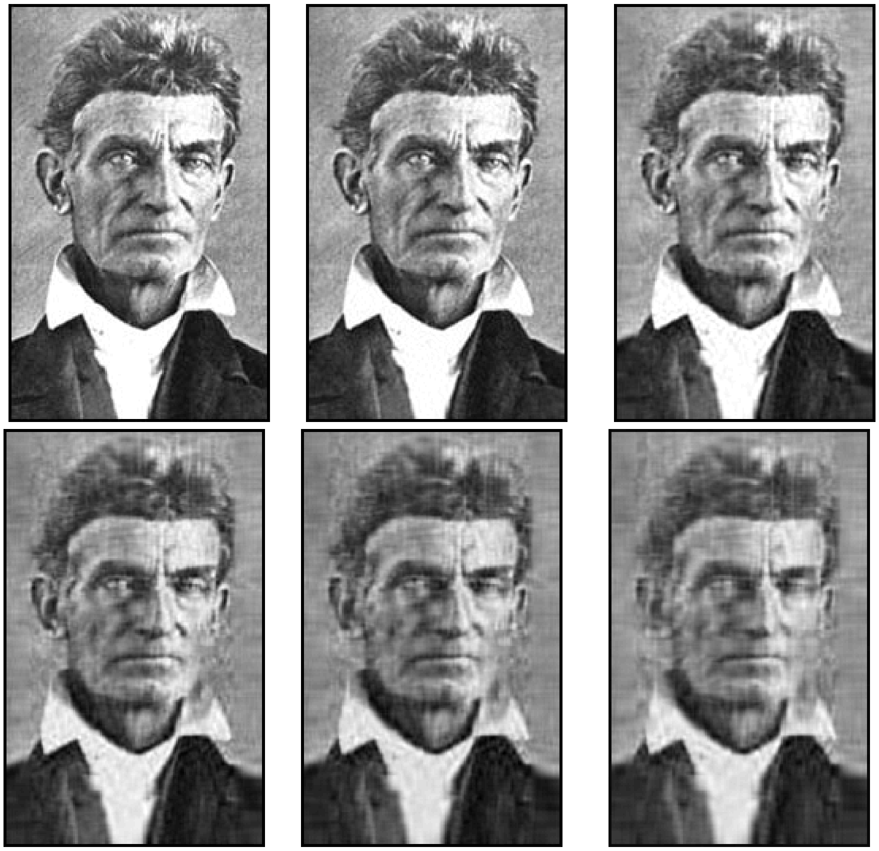


Figure 11.2. The results of `imagesc(Ak)` for, starting at the top left, `k=165, 64, 32` and moving right, and then starting at the bottom left and moving right, `k=24,20,16`.

11.3. Trace, Norm and Low Rank Approximation

In this section we establish a precise version of last section's folk theorem. To prepare the way we will need the notion of matrix trace and norm. The trace of an n-by-n matrix B is simply the sum of its diagonal elements

$$\mathrm{tr}(B) \equiv \sum_{j=1}^{n} B_{jj}. \tag{11.9}$$

It enjoys a very useful property, namely, if A is also n-by-n then

$$\mathrm{tr}(AB) = \mathrm{tr}(BA). \tag{11.10}$$

From here we arrive at a lovely identity for symmetric matrices. Namely, if $B = B^T$ and we list its eigenvalues, including multiplicities, $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ then, from (10.13) it follows that

$$\mathrm{tr}(B) = \mathrm{tr}(Q\Lambda Q^T) = \mathrm{tr}(\Lambda QQ^T) = \mathrm{tr}(\Lambda) = \sum_{j=1}^{n} \lambda_j. \tag{11.11}$$

You might wish to confirm that this holds for the many examples we've accumulated. From here we may now study the properties of the natural, or Frobenius, norm of an m-by-n matrix A,

$$\|A\|_F \equiv \left(\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2 \right)^{1/2}. \tag{11.12}$$

It is not difficult to establish that in fact

$$\|A\|_F = (\mathrm{tr}(AA^T))^{1/2} \tag{11.13}$$

from which we may deduce from (11.11) that

$$\|A\|_F^2 = \sum_{i=1}^{m} \lambda_i(AA^T) = \sum_{i=1}^{m} \sigma_i^2,$$

i.e., the Frobenius norm of a matrix is the square root of the sum of the squares of its singular values. We will also need the fact that if Q is square and $Q^TQ = I$ then

$$\|QA\|_F^2 = \mathrm{tr}(QAA^TQ^T) = \mathrm{tr}(AA^TQ^TQ) = \mathrm{tr}(AA^T) = \|A\|_F^2. \tag{11.14}$$

The same argument reveals that $\|AQ\|_F = \|A\|_F$. We may now establish

Proposition 11.3. Given an m-by-n matrix A (with SVD $A = Y\Sigma X^T$) and a whole number $k \le \min\{m, n\}$ then the best (in terms of Frobenius distance) rank k approximation of A is

$$A_k = Y(:, 1:k)\Sigma(1:k, 1:k)X(:, 1:k)^T.$$

The square of the associated approximation error is

$$\|A - A_k\|_F^2 = \min_{\mathrm{rank}(B)=k} \|A - B\|_F^2 = \sum_{j>k} \sigma_j^2.$$

Proof: If B is m-by-n then

$$\|A - B\|_F^2 = \|Y\Sigma X^T - B\|_F^2 = \|Y(\Sigma - Y^TBX)X^T\|_F^2 = \|\Sigma - Y^TBX\|_F^2.$$

If B is to be chosen, among matrices of rank k, to minimize this distance then, as Σ is diagonal, so too must Y^TBX. If we denote this diagonal matrix by S then

$$Y^TBX = S \quad \text{implies} \quad B = YSX^T,$$

and so

$$\|A - B\|_F^2 = \sum_j (\sigma_j - s_j)^2.$$

As the rank of B is k it follows that the best choice of the s_j is $s_j = \sigma_j$ for $j = 1 : k$ and $s_j = 0$ there after. End of Proof.

11.4. Exercises

1. Suppose that A is m-by-n and b is m-by-1. Set $x^+ = A^+b$ and suppose x satisfies $A^TAx = A^Tb$. Prove that $\|x^+\| \le \|x\|$. (Hint: decompose $x = x_R + x_N$ into its row space and null space components. Likewise $x^+ = x_R^+ + x_N^+$. Now argue that $x_R = x_R^+$ and $x_N^+ = 0$ and recognize that you are almost home.)

2. Experiment with compressing the bike image below (also under Resources on our Owlspace page). Submit labeled figures corresponding to several low rank approximations. Note: `bike.jpg` is really a color file, so after saving it to your directory and entering Matlab you might say `M = imread('bike.jpg')` and then `M = M(:,:,1)` prior to `imagesc(M)` and `colormap('gray')`.

3. Please prove the identity (11.10).

4. Please prove the identity (11.13).

12. Reduced Dynamical Systems

We saw in the last chapter how to build low rank approximations of matrices using the Singular Value Decomposition, with applications to image compression. We devote this chapter to building reduced models of dynamical systems from variations of the power method, with applications to neuronal modeling.

12.1. The Full Model

We consider distributed current injection into passive RC cable of Chapter 6. We suppose it has length ℓ and radius a and we choose a compartment length dx and arrive at $n = \ell/dx$ compartments. The cell's axial resistivity is R. Its membrane capacitance and conductance are C_m and G_L.

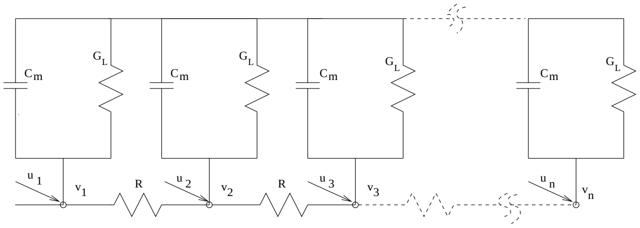


Figure 12.1. The full RC cable.

If u_j is the synaptic current injected at node j and v_j is the transmembrane potential there then (arguing as in Chapter 6) v obeys the differential equation

$$Cv'(t) + Gv(t) = u(t), \tag{12.1}$$

where

$$C = (2\pi a dx C_m)I_n, \quad G = (2\pi a dx G_L)I_n - \frac{\pi a^2}{Rdx}S_n,$$

I_n is the n-by-n identity matrix and S_n is the n-by-n second–difference matrix

$$S_n = \begin{pmatrix} -1 & 1 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdots & 0 & 1 & -2 & 1 \\ 0 & \cdots & 0 & 0 & 1 & -1 \end{pmatrix} \tag{12.2}$$

We divide through by $2\pi a dx C_m$ and arrive at

$$v'(t) = Bv(t) + g(t)$$

where

$$B = \frac{a}{2RC_m dx^2}S_n - (G_L/C_m)I_n \quad \text{and} \quad g = u/(2\pi a dx C_m). \tag{12.3}$$

We will construct a reduced model that faithfully reproduces the response at a specified compartment (the putative *spike initiation zone*), assumed without loss to be the first,

$$y(t) = q^T v(t), \quad \text{where} \quad q^T = (1\ 0\ 0\ \cdots\ 0),$$

to an arbitrary stimulus, u. On applying the Laplace Transform to each side of (12.1) we find

$$s\mathcal{L}v = B\mathcal{L}v + \mathcal{L}g$$

and so the output

$$\mathcal{L}y = q^T\mathcal{L}v = q^T(sI - B)^{-1}\mathcal{L}g \equiv H(s)\mathcal{L}g$$

may be expressed in terms of the transfer function

$$H(s) \equiv q^T(sI - B)^{-1}.$$

12.2. The Reduced Model

We choose a reduced dimension k and an $n \times k$ orthonormal *reducer* X, suppose $v = X\hat{v}$ and note that $\hat{v}$ is governed by the reduced system

$$\begin{aligned}\hat{v}'(t) &= X^TBX\hat{v}(t) + X^Tg(t),\\ \hat{y}(t) &= q^TX\hat{v}(t),\end{aligned}$$

We choose X so that the associated transfer function,

$$\hat{H}(s) = q^TX(sI_k - X^TBX)^{-1}X^T,$$

is close to H. We will see that it suffices to set

$$\begin{aligned}X(:,1) &= B^{-1}q/\|B^{-1}q\|,\\ X(:,j+1) &= B^{-1}X(:,j)/\|B^{-1}X(:,j)\|, \quad j = 1,\ldots,k-1,\end{aligned} \tag{12.4}$$

followed by orthonormalization,

$$X = \mathrm{orth}(X).$$

We note that (12.4) is nothing more than the Power Method applied to B^{-1}, and so as j increases the associated column of X approaches the (constant) eigenvector associated with the least eigenvalue of B. We apply this algorithm and then return to prove that it indeed makes H close to $\hat{H}$.

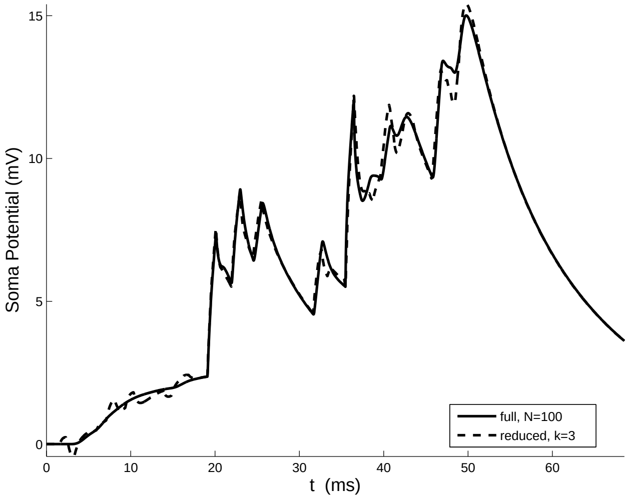


Figure 12.2. Response of the first compartment in a 100 compartment cell and the corresponding 3 compartment reduced cell to identical current injections distributed randomly in space and time. `cabred.m`

Now, in order to understand why this method works so well we examine the associated transfer functions. In particular, we examine their respective Taylor series about $s = 0$. To begin, we note that the first resolvent identity, (9.8), permits us to evaluate

$$R'(s) = \lim_{h\to 0}\frac{R(s+h)-R(s)}{h} = \lim_{h\to 0}\frac{(-h)R(s+h)R(s)}{h} = -R^2(s).$$

and so

$$R''(s) = -2R(s)R'(s) = 2R^3(s), \quad R'''(s) = 6R^2(s)R'(s) = -6R^4(s),$$

and in general

$$\frac{d^jR(s)}{ds^j} = (-1)^j j! R^{j+1}(s) \quad \text{and so} \quad \frac{d^jR(0)}{ds^j} = -j!(B^{-1})^{j+1}$$

and so

$$H(s) = q^TR(s) = -\sum_{j=0}^{\infty} s^jM_j, \quad \text{where} \quad M_j = q^T(B^{-1})^{j+1}$$

are the associated **moments**. The reduced transfer function is analogously found to be

$$\hat{H}(s) = -\sum_{j=0}^{\infty} s^j\hat{M}_j, \quad \text{where} \quad \hat{M}_j = q^TX((X^TBX)^{-1})^{j+1}X^T. \tag{12.5}$$

If $\hat{M}_j = M_j$ for $0 \le j < k$ then $\hat{H}(s) = H(s) + O(s^k)$ and so we now show that the reducer built according to (12.4) indeed matches the first k moments.

Proposition 12.1. If $(B^{-1})^{j+1}q \in \mathcal{R}(X)$ for $0 \le j < k$ then $\hat{M}_j = M_j$ for $0 \le j < k$.

Proof: We will show that $M_j^T = \hat{M}_j^T$, beginning with $j = 0$ where

$$M_0^T = B^{-1}q \quad \text{and} \quad \hat{M}_0^T = X(X^TBX)^{-1}X^Tq.$$

We note that $P = X(X^TBX)^{-1}X^TB$ obeys $P^2 = P$ and $PX = X$ and proceed to develop

$$\begin{aligned}\hat{M}_0^T &= X(X^TBX)^{-1}X^Tq \\ &= X(X^TBX)^{-1}X^TBB^{-1}q \\ &= PB^{-1}q = B^{-1}q = M_0^T,\end{aligned}$$

where the penultimate equality follows from the assumption that $B^{-1}q \in \mathcal{R}(X)$ and the fact that P acts like the identity on X. When $j = 1$ we find

$$\begin{aligned}\hat{M}_1^T &= X(X^TBX)^{-1}(X^TBX)^{-1}X^Tq \\ &= X(X^TBX)^{-1}X^TBB^{-1}X(X^TBX)^{-1}X^TBB^{-1}q \\ &= PB^{-1}PB^{-1}q = (B^{-1})^2q = M_1^T\end{aligned}$$

where the penultimate equality follows from the assumption that $(B^{-1})^2q \in \mathcal{R}(X)$. The remaining moment equalities follow in identical fashion. End of Proof.

12.3. Exercises

1. Show that the n-by-n second difference S_n in (12.2) obeys

   $$x^T S_n x = -\sum_{j=1}^{n-1}(x_j - x_{j+1})^2$$

   for every $x \in \mathbf{R}^n$. Why does it follow that the eigenvalues of S_n are nonpositive? Why does it then follow that eigenvalues of the B matrix in (12.3) are negative? If we then label the eigenvalues of B as

   $$\lambda_n(B) \le \lambda_{n-1}(B) \le \cdots \le \lambda_2(B) \le \lambda_1(B) < 0,$$

   which eigenvector of B does (12.4) approach as j increases?

2. Please derive the Taylor expansion displayed in (12.5).

3. Please confirm that $X(X^TBX)^{-1}X^TB$ is indeed projection onto $\mathcal{R}(X)$.