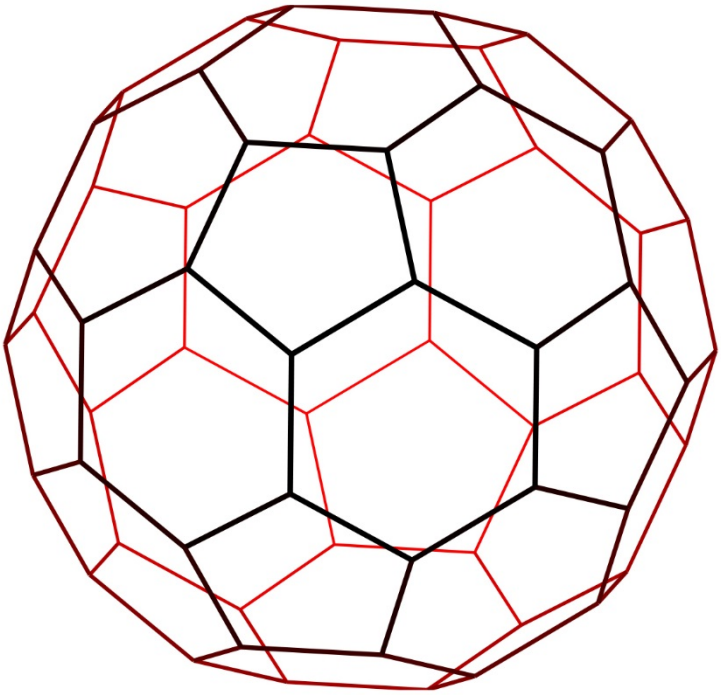
# Linear Algebra *in Situ*

CAAM 335
Fall 2018



Steven J. Cox

```
Poet, oracle and wit
Like unsuccessful anglers by
The ponds of apperception sit,
Baiting with the wrong request
The vectors of their interest;
At nightfall tell the angler's lie.

With time in tempest everywhere,
To rafts of frail assumption cling
The saintly and the insincere;
Enraged phenomena bear down
In overwhelming waves to drown
Both sufferer and suffering.

The waters long to hear our question put
Which would release their longed for answer, but.

                                    W.H. Auden
```

# Preface

This is a text where concrete physical problems are posed and the ensuing mathematical theory is developed, tested, applied and associated with existing theory. The problems I pose spring from questions of equilibria, dynamics, optimization and inference of large electrical, mechanical and chemical networks. Following Gil Strang, I demonstrate throughout that Linear Algebra is both a tool for expressing these questions and for achieving, computing and representing their solutions.

The theory needed to resolve the questions of network equilibria, optimization and inference is now well enshrined in the Fundamental Theorem of Linear Algebra, and it appears difficult to improve on this approach. Regarding dynamics however there are two distinct paths to the spectral theorem; one via zeros of the characteristic polynomial, $\det(zI - A)$, the other via poles of the resolvent, $(zI - A)^{-1}$. The first is common among introductory texts while the latter, to my knowledge, has yet to succeed at that level – although, since the treatise of Kato, it is well known to be considerably cleaner and more flexible. I feel strongly that students new to linear algebra can grasp the resolvent more readily than the determinant. For, with eigenvalues defined as those $z$ for which $(zI-A)$ does not have an inverse, the **direct approach** is to simply construct $(zI-A)^{-1}$ and observe the offending $z$. The construction of $(zI - A)^{-1}$, say via Gauss–Jordan, is straightforward though tedious. Once they understand the process however they may turn the tedium over to one of a number of "symbolic algebra" routines. I make systematic use of the symbolic toolbox in MATLAB. By contrast, the **indirect approach** ignores the inverse and relies on the determinant as a mere numerical test of invertibility. The approach via the resolvent comes however at the cost of presuming familiarity with the residue theorem of complex integration. I see this rather as a win–win situation, for the residue theorem is also key to making proper sense of the Inverse Laplace and Fourier Transforms. Hence, our two brief chapters on complex variables pay multiple dividends.

The reader will find here an introductory course, an advanced course, an array of intermediate courses, and a reference for self–study and/or use in advanced courses across Science, Engineering

and Mathematics. The general audience introductory course, assuming only one year of calculus, that I have taught to sophomores at Rice University for more than 20 years, is composed of the following sections from the first 13 chapters:

**Introductory Course**

1. Orientation, §§1–3
2. Electrical Networks, §§1–2
3. Mechanical Networks, §§1–3
4. The Column and Null Spaces, §§1–3
5. The Fundamental Theorem and Beyond, §§1–3
6. Least Squares, §§1–4
7. Metabolic Networks, §§1–3
8. Dynamical Systems, §§1–4
9. Complex Numbers, Vectors and Functions, §§1–3
10. Complex Integration, §§1–3
11. The Eigenvalue Problem, §§1–2
12. The Hermitian Eigenvalue Problem, §§1–2
13. The Singular Value Decomposition, §§1–2.

This course stresses applications, methods and computation over theory and algorithms. As the audience has been predominantly students of engineering and science I have used application chapters to motivate theory chapters and then used this theory to both revisit old applications and to embark on new ones. For example, the pseudo–inverse is invoked in Chapter 3 in order to ignore the rigid body motion of a mechanical network. This provokes discussion of null and column spaces but does not get resolved until the spectral representation and singular value decomposition in Chapters 11–13. Similary, the resolvent and eigenvalues arise naturally in our consideration, in Chapter 8, of dynamical systems but do not get resolved until the spectral representation is reached. As such the material, including the exercises, in the early sections of the first 13 Chapters (with the exception of Chapter 7 on Metabolic Networks) is highly integrated.

For audiences with either prior exposure to linear algebra or motivating applications one can skim Chapter 1 and the early sections of Chapters of 2, 3 and 7 and use the time saved to delve more deeply into the latter, more challenging, starred sections of Chapters 2–13 or perhaps into the more advanced material of Chapters 14–16. The starred sections offer short courses in Convex Analysis, §§5.4,7.6,12.6 and Fourier Analysis, §§9.5,9.6,10.4, 15.6.

The last three chapters, presuming a solid foundation in Linear Algebra, develop the Group, Representation and Graph Theory that underly the exact solution to three exciting problems concerning large networks. In particular: I provide a detailed derivation of the exact formulas, following Chung and Sternberg, for the 60 eigenvalues that govern the electronic structure of the Buckyball, and I provide detailed proofs that concrete constructions of Margulis achieve large girth in one case and establish a family of expander graphs in the other.

Steve Cox

# Contents

# 1. Orientation

You have likely encountered vectors, and perhaps matrices in your introductory calculus and/or physics courses. My goal in this chapter is to strengthen these encounters and so prepare you for the applications, computations and theory to come. I begin in §1.1 with a careful presentation of the basic objects – and the laws that govern their arithmetic combinations. I then introduce MATLAB in §1.2 as a means to visually explore the sense in which matrices transform vectors. I complete our orientation in §1.3 with an introduction to the principle methods of proof used in Linear Algebra. Throughout the chapter I introduce and reinforce concepts through examples and stress that you gain confidence and expertise by generating examples of your own. The exercises at the end of the chapter should help toward that end.

## 1.1. Objects

A **vector** is a column of real numbers, and is written, e.g.,

$$x = \begin{pmatrix} 2 \\ -4 \\ 1 \end{pmatrix}. \tag{1.1}$$

The vector has 3 elements and so lies in the class of all 3–element vectors, denoted, $\mathbb{R}^3$, where $\mathbb{R}$ stands for "real". We denote "is a member of" by the symbol $\in$. So, e.g., $x \in \mathbb{R}^3$. We denote the first element of $x$ by $x_1$, its second element by $x_2$ and so on. For example, $x_2 = -4$ in (1.1).

We will typically use the positive integer $n$ to denote the ambient dimension of our problem, and so will be working in $\mathbb{R}^n$. The sum of two vectors, $x$ and $y$, in $\mathbb{R}^n$ is defined elementwise by

$$z = x + y, \quad \text{where} \quad z_j = x_j + y_j, \quad j = 1, \ldots, n.$$

The multiplication of a vector, $x \in \mathbb{R}^n$, by a scalar $s \in \mathbb{R}$ is defined elementwise by

$$z = sx, \quad \text{where} \quad z_j = sx_j, \quad j = 1, \ldots, n.$$

For example,

$$\begin{pmatrix} 2 \\ 5 \end{pmatrix} + \begin{pmatrix} 1 \\ -3 \end{pmatrix} = \begin{pmatrix} 3 \\ 2 \end{pmatrix} \quad \text{and} \quad 6 \begin{pmatrix} 4 \\ 2 \end{pmatrix} = \begin{pmatrix} 24 \\ 12 \end{pmatrix}.$$

The most common product of two vectors, $x$ and $y$, in $\mathbb{R}^n$ is the **inner product**,

$$x^T y \equiv \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = \sum_{j=1}^{n} x_j y_j. \tag{1.2}$$

As $x_j y_j = y_j x_j$ for each $j$ it follows that $x^T y = y^T x$. For example,

$$\begin{pmatrix} 10 & 1 & 3 \end{pmatrix} \begin{pmatrix} 8 \\ 2 \\ -4 \end{pmatrix} = 10 \cdot 8 + 1 \cdot 2 + 3 \cdot (-4) = 70. \tag{1.3}$$

So, the inner product of two vectors is a number. The superscript $T$ on the $x$ on the far left of Eq. (1.2) stands for **transpose** and, when applied to a column yields a **row**. Columns are vertical

and rows are horizontal and so we see, in Eq. (1.2), that $x^T$ is $x$ laid on its side. We follow Euclid and measure the magnitude, or more commonly the **norm**, of a vector by the square root of the sum of the squares of its elements. In symbols,

$$\|x\| \equiv \sqrt{x^T x} = \sqrt{\sum_{j=1}^{n} x_j^2}. \tag{1.4}$$

For example, the norm of the vector in Eq. (1.1) is $\sqrt{21}$. As Eq. (1.4) is a direct generalization of the Euclidean distance of high school planar geometry we may expect that $\mathbb{R}^n$ has much the same "look." To be precise, let us consider the situation of Figure 1.1.

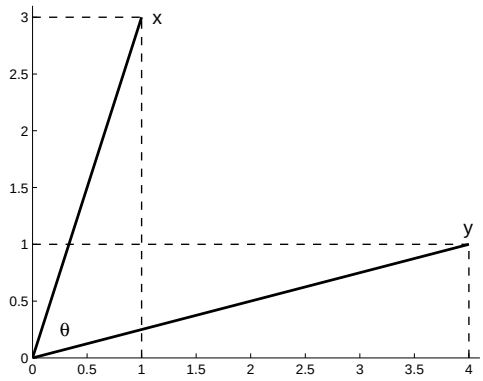



**Figure** 1.1. A guide to interpreting the inner product.

We have $x$ and $y$ in $\mathbb{R}^2$ and

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix} \quad \text{and} \quad y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 1 \end{pmatrix}$$

and we recognize that both $x$ and $y$ define right triangles with hypotenuses $\|x\|$ and $\|y\|$ respectively. We have denoted by $\theta$ the angle that $x$ makes with $y$. If $\theta_x$ and $\theta_y$ denotes the angles that $x$ and $y$ respectively make with the positive horizontal axis then $\theta = \theta_x - \theta_y$ and the Pythagorean Theorem permits us to note that

$$x_1 = \|x\| \cos(\theta_x), \quad x_2 = \|x\| \sin(\theta_x), \quad \text{and} \quad y_1 = \|y\| \cos(\theta_y), \quad y_2 = \|y\| \sin(\theta_y),$$

and these in turn permit us to express the inner product of $x$ and $y$ as

$$\begin{aligned} x^T y &= x_1 y_1 + x_2 y_2 \\ &= \|x\| \|y\| (\cos(\theta_x) \cos(\theta_y) + \sin(\theta_x) \sin(\theta_y)) \\ &= \|x\| \|y\| \cos(\theta_x - \theta_y) \\ &= \|x\| \|y\| \cos(\theta). \end{aligned} \tag{1.5}$$

We interpret this by saying that the inner product of two vectors is proportional to the cosine of the angle between them. Now given two vectors in say $\mathbb{R}^8$ we don't panic, rather we orient ourselves by observing that they together lie in a particular plane and that this plane, and the angle they make with one another is in no way different from the situation illustrated in Figure 1.1. And for this reason we say that $x$ and $y$ are perpendicular, or **orthogonal**, to one another whenever $x^T y = 0$.

In addition to the geometric interpretation of the inner product it is often important to be able to estimate it in terms of the products of the norms. Here is an argument that works for $x$ and $y$ in $\mathbb{R}^n$. Once we know where to start, we simply expand

$$\begin{aligned}\|(y^Ty)x-(x^Ty)y\|^2 &= ((y^Ty)x-(x^Ty)y)^T((y^Ty)x-(x^Ty)y)\\ &= \|y\|^4\|x\|^2-2\|y\|^2(x^Ty)^2+(x^Ty)^2\|y\|^2\\ &= \|y\|^2(\|x\|^2\|y\|^2-(x^Ty)^2)\end{aligned} \tag{1.6}$$

and then note that as the initial expression is nonnegative, the final expression requires (after taking square roots) that

$$\boxed{|x^Ty|\le\|x\|\|y\|.} \tag{1.7}$$

This is known as the **Cauchy–Schwarz inequality**.

As a vector is simply a column of numbers, a matrix is simply a row of columns, or a column of rows. This necessarily requires two numbers, the row and column indices, to specify each matrix element. For example

$$A=\begin{pmatrix}a_{11}&a_{12}&a_{13}\\a_{21}&a_{22}&a_{23}\end{pmatrix}=\begin{pmatrix}5&0&1\\2&3&4\end{pmatrix} \tag{1.8}$$

is a 2-by-3 matrix. The first dimension is the number of rows and the second is the number of columns and this ordering is also used to address individual elements. For example, the element in row 1 column 3 is $a_{13}=1$. We will consistently use upper–case letters to denote matrices.

The addition of two matrices (of the same size) and the multiplication of a matrix by a scalar proceed exactly as in the vector case. In particular,

$$(A+B)_{ij}=a_{ij}+b_{ij},\quad \text{e.g.,}\ \begin{pmatrix}5&0&1\\2&3&4\end{pmatrix}+\begin{pmatrix}2&4&6\\1&-3&4\end{pmatrix}=\begin{pmatrix}7&4&7\\3&0&8\end{pmatrix},$$

and

$$(cA)_{ij}=ca_{ij},\quad \text{e.g.,}\quad 3\begin{pmatrix}5&0&1\\2&3&4\end{pmatrix}=\begin{pmatrix}15&0&3\\6&9&12\end{pmatrix}.$$

The product of two commensurate matrices proceeds through a long sequence of inner products. In particular if $C=AB$ then the $ij$ element of $C$ is the product of the $i$th row of $A$ and the $j$th column of $B$. Hence, for two $A$ and $B$ to be commensurate it follows that each row of $A$ must have the same number of elements as each column of $B$. In other words, the number of columns of $A$ must match the number of rows of $B$. Hence, if $A$ is $m$-by-$n$ and $B$ is $n$-by-$p$ then the $ij$ element of their product $C=AB$ is

$$c_{ij}=\sum_{k=1}^{n}a_{ik}b_{kj}=A(i,:)B(:,k), \tag{1.9}$$

where $A(i,:)$ denotes row $i$ of $A$ and $B(:,k)$ denotes column $k$ of $B$. For example,

$$\begin{pmatrix}5&0&1\\2&3&4\end{pmatrix}\begin{pmatrix}2&4\\6&1\\-3&4\end{pmatrix}=\begin{pmatrix}5\cdot2+0\cdot6+1\cdot(-3)&5\cdot4+0\cdot1+1\cdot4\\2\cdot2+3\cdot6+4\cdot(-3)&2\cdot4+3\cdot1+4\cdot(-4)\end{pmatrix}=\begin{pmatrix}7&24\\10&-5\end{pmatrix}.$$

In this case, the product $BA$ is not even defined. If $A$ is $m$-by-$n$ and $B$ is $n$-by-$m$ then both $AB$ and $BA$ are defined, but unless $m=n$ they are of distinct dimensions and so not comparable. If

$m = n$ so $A$ and $B$ are square then we may ask if $AB = BA$? and learn that the answer is typically no. For example,

$$\begin{pmatrix} 5 & 0 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} 2 & 4 \\ 6 & 1 \end{pmatrix} = \begin{pmatrix} 10 & 20 \\ 22 & 11 \end{pmatrix} \neq \begin{pmatrix} 2 & 4 \\ 6 & 1 \end{pmatrix} \begin{pmatrix} 5 & 0 \\ 2 & 3 \end{pmatrix} = \begin{pmatrix} 18 & 12 \\ 32 & 3 \end{pmatrix}. \tag{1.10}$$

We will often abbreviate the awkward phrase "$A$ is $m$-by-$n$" with the declaration $A \in \mathbb{R}^{m\times n}$. The matrix algebra of multiplication, though tedious, is easy enough to follow. It stemmed from a row-centric point of view. It will help to consider the columns. If $A \in \mathbb{R}^{m\times n}$ and the $j$th column of $A$ is $A(:,j)$ and $x \in \mathbb{R}^n$ then we recognize the product

$$Ax = [A(:,1)\ A(:,2)\ \cdots\ A(:,n)] \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = x_1 A(:,1) + x_2 A(:,2) + \cdots + x_n A(:,n), \tag{1.11}$$

as a weighted sum of the columns of $A$. For example

$$\begin{pmatrix} 2 & 3 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \end{pmatrix} = 2 \begin{pmatrix} 2 \\ 1 \end{pmatrix} + 3 \begin{pmatrix} 3 \\ 4 \end{pmatrix} = \begin{pmatrix} 13 \\ 14 \end{pmatrix}. \tag{1.12}$$

We illustrate this in Figure 1.2(A) and then proceed to illustrate in the second panel the transformation by this $A$ of a representative collection of unit vectors.

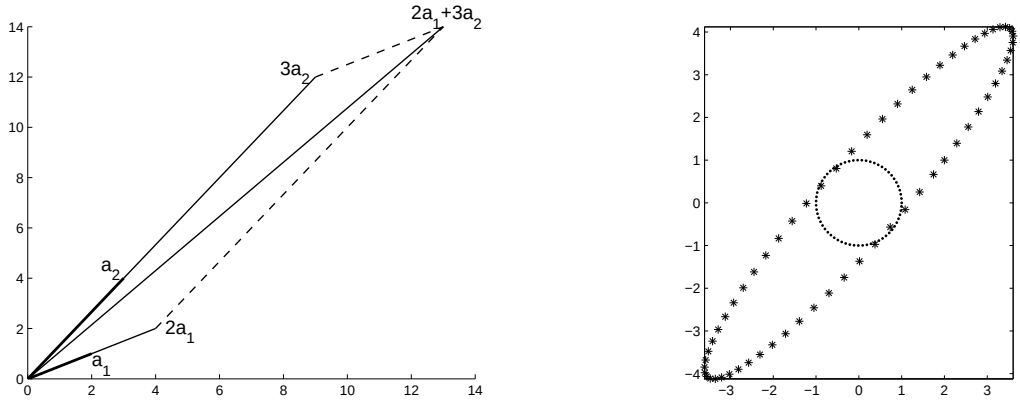



**Figure** 1.2. (A) An illustration of the matrix vector multiplication conducted in Eq. (1.12). Both $A(:,1)$ and $A(:,2)$ are plotted heavy for emphasis. We see that their multiples, by 2 and 3, simply extend them, while their weighted sum simply completes the natural parallelogram. (B) For a given $x$ on the unit circle (denoted by a dot) we plot its transformation by the $A$ matrix of Eq. (1.12) (denoted by an asterisk). `mymult1.m`

A common goal of matrix analysis is to describe $m$-by-$n$ matrices by many fewer than $mn$ numbers. The simplest such descriptor is the sum of the matrice's diagonal elements. We call this the **trace** and abbreviate it by

$$\mathrm{tr}(A) \equiv \sum_{i=1}^{n} a_{ii}. \tag{1.13}$$

Looking for matrices to trace you scan Eq. (1.10) and note that $10 + 11 = 18 + 3$ and you ask, knowing that $AB \neq BA$, whether

$$\boxed{\mathrm{tr}(AB) = \mathrm{tr}(BA)} \tag{1.14}$$

might possibly be true in general. For arbitrary $A$ and $B$ in $\mathbb{R}^{n\times n}$ we therefore construct $\mathrm{tr}(AB)$

$$(AB)_{ii} = \sum_{k=1}^{n} a_{ik}b_{ki} \quad \text{so} \quad \mathrm{tr}(AB) = \sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik}b_{ki},$$

and $\mathrm{tr}(BA)$

$$(BA)_{ii} = \sum_{k=1}^{n} b_{ik}a_{ki} \quad \text{so} \quad \mathrm{tr}(BA) = \sum_{i=1}^{n}\sum_{k=1}^{n} b_{ik}a_{ki}.$$

These sums indeed coincide, for both are simply the sum of the product of each element of $A$ and the reflected (interchange $i$ and $k$) element of $B$.

In general, if $A$ is $m$-by-$n$ then the matrix that results on exchanging its rows for its columns is called the **transpose** of $A$, denoted $A^T$. It follows that $A^T$ is $n$-by-$m$ and

$$(A^T)_{ij} = a_{ji}.$$

For example,

$$\begin{pmatrix} 5 & 0 & 1 \\ 2 & 3 & 4 \end{pmatrix}^T = \begin{pmatrix} 5 & 2 \\ 0 & 3 \\ 1 & 4 \end{pmatrix}.$$

We will have frequent need to transpose a product, so let us contrast

$$((AB)^T)_{ij} = \sum_{k=1}^{n} a_{jk}b_{ki} \quad \text{with} \quad (B^TA^T)_{ij} = \sum_{k=1}^{n} a_{jk}b_{ki} \tag{1.15}$$

and so conclude that

$$\boxed{(AB)^T = B^TA^T,} \tag{1.16}$$

i.e., that the transpose of a product is the product of the transposes in reverse order.

Regarding the norm of a matrix it seems natural, on recalling our definition of the norm of a vector, to simply define it as the square root of the sum of the squares of each element. This definition, where $A \in \mathbb{R}^{m\times n}$ is viewed as a collection of vectors, is associated with the name Frobenius and hence the subscript in the definition of the **Frobenius norm** of $A$,

$$\|A\|_F \equiv \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2\right)^{1/2}. \tag{1.17}$$

As scientific progress and mathematical insight most often come from seeing things from multiple angles we pause to note Eq. (1.17) may be seen as the trace of a product. In particular, with $B = A^T$ and $j = i$ in the general formula Eq. (1.15) we arrive immediately at

$$(AA^T)_{ii} = \sum_{k=1}^{n} a_{ik}^2.$$

As the sum over $i$ is precisely the trace of $AA^T$ we have established the equivalent definition

$$\|A\|_F = (\mathrm{tr}(AA^T))^{1/2}. \tag{1.18}$$

For example, the Frobenius norm of the $A$ in Eq. (1.8) is $\sqrt{55}$. Just as the vector norm can help us bound (recall Eq. (1.7)) the inner product of two vectors, this matrix norm can help us bound the product of a matrix and vector. More precisely, lets prove that

$$\|Ax\| \leq \|A\|_F \|x\|, \tag{1.19}$$

for arbitrary $A$ and $x$. To see this we complement Eq. (1.11) with a row representation

$$Ax = \begin{pmatrix} A(1,:)x \\ A(2,:)x \\ \vdots \\ A(m,:)x \end{pmatrix}$$

and so

$$\begin{aligned} \|Ax\| &= \sqrt{(A(1,:)x)^2 + (A(2,:)x)^2 + \cdots + (A(m,:)x)^2} \\ &\leq \sqrt{\|A(1,:)\|^2\|x\|^2 + \|A(2,:)\|^2\|x\|^2 + \cdots + \|A(:,n)\|^2\|x\|^2} \\ &= \|A\|_F \|x\|, \end{aligned}$$

where we have used Eq. (1.7) to conclude that each $|A(j,:)x| \leq \|A(j,:)\|\|x\|$. The simple rearrangement of Eq. (1.19),

$$\frac{\|Ax\|}{\|x\|} \leq \|A\|_F \qquad \forall\ x, \tag{1.20}$$

has the nice geometric interpretation: "The matrix $A$ can stretch no vector by more than $\|A\|_F$." We can reinforce this interpretation by returning to Figure 1.2 and noting that no vector in the ellipse is longer than $\|A\|_F = \sqrt{30}$.

### 1.2. Computations

The objects of the previous section turn stale and are easily forgotten unless handled. We are fortunate to work in a time in which both the tedium of their manipulation and the task of illustrating our "findings" have been automated – leaving one's imagination the only obstacle to discovery.

To prepare you to "handle" your own objects we now present a brief introduction to MATLAB via experiments on the innocent looking

$$A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}. \tag{1.21}$$

It is inert until it acts. Its action is spelled out in

$$Ax = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 + 2x_2 \\ x_2 \end{pmatrix} \tag{1.22}$$

but perhaps these symbols do not yet speak to you. To illustrate or animate this action we might turn to devices like Figure 1.2 where we plot its deformation of the unit circle. Though this gives a general sense of its influence it neglects to track the transformation of individual unit vectors. We correct for this and display our findings in Figure 1.3, by marking 12 unit vectors in black and their 12 deformations, under $A$, in red.

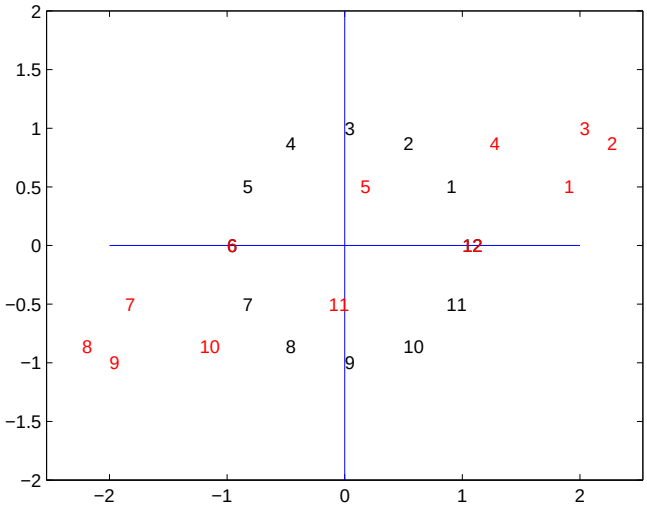



**Figure** 1.3. Illustration of the action, $Ax$ in red, specified in Eq. (1.22) for the twelve $x$ vectors (black). That is, $A$ takes the black 1 to the red 1, the black 2 to the red 2 and so on. Yes, both the black 6 and black 12 remain unmoved by $A$. (`mymult2.m`)

Now we are really on to something – for this figure suggests so many new questions! But before getting carried away lets take a careful look at the MATLAB script, `mymult2.m`, that generated Figure 1.3. For ease of reference we have numbered each line in our program.

```
1       A = [1 2; 0 1];                          % the matrix
2       plot([-2 2],[0 0])                       % plot the horizontal axis
3       hold on                                  % plot future info in same figure
4       plot([0 0],[-2 2])                       % plot the vertical axis
5       for j=1:12,                              % do what follows 12 times
6           ang = j*2*pi/12;                     % angle
7           x = [cos(ang); sin(ang)];            % a point on the unit circle
8           y = A*x;                             % transformed by A
9           text(x(1),x(2),num2str(j))           % place the counter value at x
10          s = text(y(1),y(2),num2str(j));      % place the counter value at y
11          set(s,'color','r')                   % paint that last value red
12      end
13      hold off                                 % let go of the picture
14      axis equal                               % fiddle with the axes
```

Our actor, `A`, gets line 1 billing. We specify matrices, and columns, between square brackets and terminate each row (except the last) with a semicolon. Note that line 1 is not an equation but rather an *assignment*. MATLAB assigns what it finds to the right of `=` to the symbol it finds at the left.

In line 2 we instruct MATLAB to plot a line in the plane from $(-2, 0)$ to $(2, 0)$ using the default color, blue. In line 4 we instruct MATLAB to plot a blue line from $(0, -2)$ to $(0, 2)$.

In line 5 we enter a loop that terminates at line 12 when the counter, `j`, reaches its terminal value. The colon is a powerful synonym for 'to,' in the sense that we read line 5 as "for j equal 1 to 12 execute lines 6 through 11." You see that `ang` will then take on multiples of $\pi/6$ and that `x` will

be the associated unit vector and `y` its transformation under `A`. In line 9 through 10 we take the important step of actually *marking* our tracks by turning the counter value to a text string that is then placed at `(x(1),x(2))` in the default (black) color and then again at `(y(1),y(2))`, but this time in red.

This script now belongs to your list of objects and as such invites experimentation. For example, What must change to up the action from 12 to 24 players? Once you've learned this script we can return to pondering Figure 1.3. Do you see that it `shears` the circle in the sense that it drags the top half to the right and bottom half to the left while the equator remains unmoved? Does this suggest that we could learn more be deforming shape other than circles? Though many shapes come to mind we might miss something if we stick to regular objects. One of the key advantages of computational experimentation is the ability to simultaneously observe the action upon many random players. One difficulty with `many` is that it becomes more difficult to mark our tracks. To get round this we will restrict our players to one half of the plane and paint each black while painting red their action by $A$. So how should we divide the plane. The simple guess of top, $x_2 > 0$, and bottom, $x_2 < 0$ does not seem to expose any new patterns and so one might instead tilt this guess to say align with diagonals and so divide the plane into the two bow-ties

$$E = \{x \in \mathbb{R}^2 : |x_2| > |x_1|\} \quad \text{and} \quad F = \{x \in \mathbb{R}^2 : |x_1| > |x_2|\}. \tag{1.23}$$

We illustrate our remarkable findings in Figure 1.4.

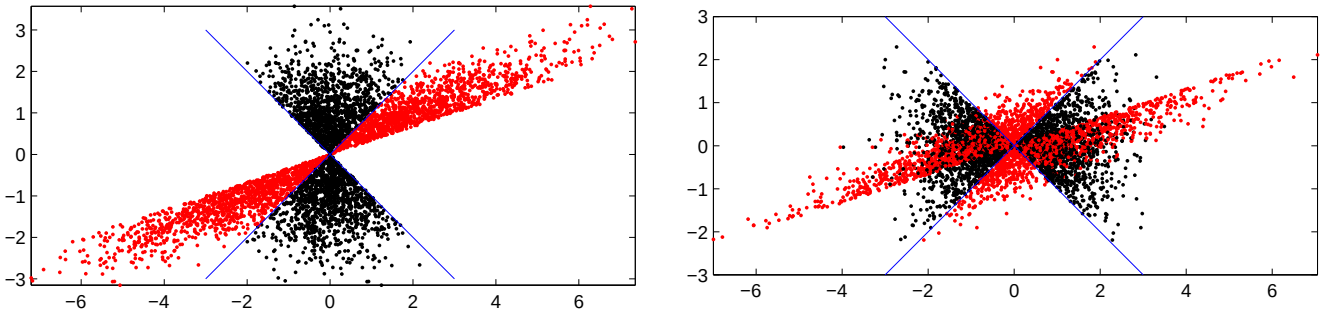


**Figure** 1.4. (A) The deformation (red) by $A$ of 2500 random vectors (black) from $E$. We surmise that $A$ takes $E$ to $F$. (B) The deformation (red) by $A$ of 2500 random vectors (black) from $F$. (`mymult3.m`)

The difference in clarity between between panels (A) and (B) is striking – for these are drawn from the *same* matrix. Panel (A) leads immediately, via Eq. (1.22), to the conjecture: if $|x_2| > |x_1|$ then $|x_1 + 2x_2| > |x_2|$. We leave its proof (and more) to Exer. 1.3 in order that we may explicate the script that generated Figure 1.4.

```
A = [1 2; 0 1];                              % the matrix
for n=1:2500,                                % do the following 2500 times
    x = randn(2,1);                          % generate a random point
    [sax,ord] = sort(abs(x),1,'ascend');     % sort their magnitudes
    x = x(ord);                              % reorder the elements
    y = A*x;                                 % transform via A
    plot(x(1),x(2),'k.')                     % mark the original point black
    hold on                                  % save this picture
    plot(y(1),y(2),'r.')                     % mark the transformed point red
```

```
end
plot([-3 3],[3 -3])                    % plot the NW-SE diagonal
plot([-3 3],[-3 3])                    % plot the SW-NE diagonal
axis equal                             % fiddle with axes
hold off                               % let go of the picture
```

There are two key differences with the previous script. Our $x$ vectors are now generated (and reordered) at random and we are plotting points rather than texting strings. The `x = randn(2,1)` places two random samples of the normal (or Gaussian, or bell–curve) distribution into the 2–by–1 vector `x`. In order to ensure that this `x` lies in $E$ we sort its absolute values via `sort` in an ascending fashion. The `sort` function returns two objects: `sax`, the sorted values and `ord`, the order in which they appeared. More precisely if `abs(x1)<abs(x2)` then `ord=[1 2]` and `x=x(ord)` changes nothing while if instead `abs(x1)>abs(x2)` then `ord=[2 1]` and `x=x(ord)` corrects their order. If instead we wish to restrict `x` to $F$, to generate panel (B), we switch `ascend` to `descend`.

Now that we understand how matrices like $A = [1\ 2; 0\ 1]$ act on objects like circles and bowties we may inspect their action on much more complicated objects. MATLAB has a large library of stock images that we may manipulate. We present such a before and after in Figure 1.5.

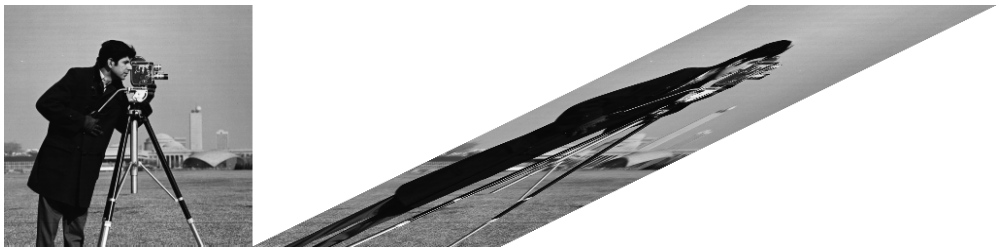


**Figure** 1.5. An image of a camerman, normal and sheared by the $A$ matrix in (1.22). (`mymult4.m`)

The code that achieves this transformation is

```
       P = imread('cameraman.tif');        % read the image
       [m,n] = size(P);                    % record its size
1      SP = 256*ones(m,2*m+n,'uint8');     % create a white canvas
       for i=1:m                           % inspect every pixel
       for j=1:n,                          %    of the original image
2          SP(i,2*m+j-2*i) = P(i,j);       % and shear it with the matrix A
       end
       end
       imshow([P SP])                      % display both images
```

We have numbered the "interesting lines." Regarding line 1, Why does 256 designate white? and Why have we added $2m$ columns? Regarding line 2, where exactly is $A$? You can discover the answers by observing the result of small changes to these lines.

### 1.3. Proofs

Regarding the **proofs** in the text, and more importantly in the exercises and exams, many will be of the type that brought us Eq. (1.14) and Eq. (1.16). These are what one might call **confirmations**. They require a clear head and may require a bit of rearrangement but as they follow directly from definitions they do not require magic, clairvoyance or even ingenuity. As further examples of confirmations let us prove (confirm) that

$$\mathrm{tr}(A) = \mathrm{tr}(A^T). \tag{1.24}$$

It would be acceptable to say that "As $A^T$ is the reflection of $A$ across its diagonal both $A$ and $A^T$ agree on the diagonal. As the trace of matrix is simply the sum of its diagonal terms we have confirmed Eq. (1.24)." It would also be acceptable to proceed in symbols and say "from $(A^T)_{ii} = a_{ii}$ for each $i$ it follows that

$$\mathrm{tr}(A^T) = \sum_{i=1}^{n} (A^T)_{ii} = \sum_{i=1} a_{ii} = \mathrm{tr}(A).\text{"}$$

It would not be acceptable to confirm Eq. (1.24) on a particular numerical matrix, nor even on a class of matrices of a particular size.

As a second example, lets confirm that

$$\text{if} \quad \|x\| = 0 \quad \text{then} \quad x = 0. \tag{1.25}$$

It would be acceptable to say that "As the sum of the squares of each element of $x$ is zero then in fact each element of $x$ must vanish." Or, in symbols, as

$$\sum_{i=1}^{n} x_i^2 = 0$$

we conclude that each $x_i = 0$.

Our third example is a slight variation on the second.

$$\text{if } x \in \mathbb{R}^n \quad \text{and} \quad x^T y = 0 \quad \text{for all } y \in \mathbb{R}^n \quad \text{then} \quad x = 0. \tag{1.26}$$

This says that the only vector that is orthogonal to every vector in the space is the zero vector. The most straightforward proof is probably the one that reduces this to the previous Proposition, Eq. (1.25). Namely, since $x^T y = 0$ for each $y$ we can simply use $y = x$ and discern that $x^T x = 0$ and conclude from Eq. (1.25) that indeed $x = 0$. As this section is meant to be an introduction to proving let us apply instead a different strategy, one that replaces a proposition with its equivalent contra–positive. More precisely, if your proposition reads "if $c$ then $d$" then its contrapositive reads "if not $d$ then not $c$." Do you see that a proposition is true if and only its contrapositive is true? Why bother? Sometimes the contrapositive is "easier" to prove, sometimes it throws new light on the original proposition, and it always expands our understanding of the landscape. So let us construct the contra–positive of Eq. (1.26). As clause $d$ is simply $x = 0$, not $d$ is simply $x \neq 0$. Clause $c$ is a bit more difficult, for it includes the clause "for all," that is often called the **universal quantifier** and abbreviated by $\forall$. So clause $c$ states $x^T y = 0 \ \forall\, y$. The negation of "some thing happens for every $y$" is that "there exists a $y$ for which that thing does not happen." This "there exists" is called the **existential quantifier** and is often abbreviated $\exists$. Hence, the contra–positive of Eq. (1.26) is

$$\text{if } x \in \mathbb{R}^n \quad \text{and} \quad x \neq 0 \quad \text{then} \quad \exists\, y \in \mathbb{R}^n \quad \text{such that} \quad x^T y \neq 0. \tag{1.27}$$

It is a matter of taste, guided by experience, that causes one to favor (or not) the contra–positive over the original. At first sight the student new to proofs and unsure of "where to start" may feel that the two are equally opaque. Mathematics however is that field that is, on first sight, opaque to everyone, but that on second (or third) thought begins to clarify, suggest pathways, and offer insight and rewards. The key for the beginner is not to despair but rather to generate as many starting paths as possible, in the hope that one of them will indeed lead to a fruitful second step, and on to a deeper understanding of what you are attempting to prove. So, investigating the contra–positive fits into our bigger strategy of generating multiple starting points and, even when a dead-end, is a great piece of guilt–free procrastination.

Back to the problem at hand I'd like to point out two avenues "suggested" by Eq. (1.27). The first is the old avenue – "take $y = x$" for then $x \neq 0$ surely implies that $x^T x \neq 0$. The second I feel is more concrete, more pedestrian, less clever and therefore hopefully contradicts the belief that one either "gets the proof or not." The concreteness I speak of is generated by the $\exists$ for it says we only have to find one – and I typically find that easier to do than finding many or all. To be precise, if $x \neq 0$ then a particular element $x_i \neq 0$. From here we can custom build a $y$, namely choose $y$ to be 0 at each element except for the $i$th in which you set $y_i = 1$. Now $x^T y = x_i$ which, by not $c$, is presumed nonzero.

As a final example lets prove that

$$\text{if } A \in \mathbb{R}^{n\times n} \quad \text{and} \quad Ax = 0 \ \forall\, x \in \mathbb{R}^n \quad \text{then} \quad A = 0. \tag{1.28}$$

In fact, lets offer three proofs.

The first is a "row proof." We denote row $j$ of $A$ by $A(j,:)$ and note that $Ax = 0$ implies that the inner product $A(j,:)x = 0$ for every $x$. By our proof of Eq. (1.26) it follows that the $j$th row vanishes, i.e., $A(j,:) = 0$. As this holds for each $j$ it follows that the entire matrix is 0.

Our second is a "column proof." We interpret $Ax = 0$, $\forall\, x$, in light of Eq. (1.11), to say that every weighted sum of the columns of $A$ must vanish. So lets get concrete and choose an $x$ that is zero in every element except the $j$th, for which we set $x_j = 1$. Now Eq. (1.11) and the if clause in Eq. (1.28) reveal that $A(:,j) = 0$, i.e., the $j$th column vanishes. As $j$ was arbitrary it follows that every column vanishes ans so the entire matrix is zero.

Our third proof will address the contrapositive,

$$\text{if } A \neq 0 \in \mathbb{R}^{n\times n} \quad \text{then} \quad \exists\ x \in \mathbb{R}^n \quad \text{such that} \quad Ax \neq 0. \tag{1.29}$$

We now move concretely and infer from $A \neq 0$ that for some particular $i$ and $j$ that $a_{ij} \neq 0$. We then construct (yet again) an $x$ of zeros except we set $x_j = 1$. It follows (from either the row or column interpretation of $Ax$) that the $i$th element of $Ax$ is $a_{ij}$. As this is not zero we have proven that $Ax \neq 0$.

We next move on to a class of propositions that involve infinity in a substantial way. If there are in fact an infinite number of claims we may use the Principle of Mathematical Induction, if it is a claim about equality of infinite sets then we may use the method of reciprocal inclusion, while if it is a claim about convergence of infinite sequences of vectors we may use the ordering of the reals.

The **Principle of Mathematical Induction** states that the truth of the infinite sequence of statements $\{P(n) : n = 1, 2, \ldots\}$ follows from establishing that
(PMI1) $P(1)$ is true.
(PMI2) if $P(n)$ is true then $P(n+1)$ is true, for arbitrary $n$.

For example, let us prove by induction that

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}^n = \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix} \qquad n = 1, 2, \ldots . \tag{1.30}$$

We first check the base case, here Eq. (1.30) holds by inspection when $n = 1$. We now suppose it holds for some $n$ then deduce its validity for $n + 1$. Namely

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}^{n+1} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}^n = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & n+1 \\ 0 & 1 \end{pmatrix}.$$

Regarding infinite sets, the **Principle of Mutual Inclusion** states that two sets coincide if each is a subset of the other. For example, given an $x \in \mathbb{R}^n$ lets consider the outer product matrix $xx^T \in \mathbb{R}^{n \times n}$ and let us prove that the two sets

$$N_1 \equiv \{y : x^T y = 0\} \quad \text{and} \quad N_2 \equiv \{z : xx^T z = 0\}$$

coincide. If $x = 0$ both sides are simply $\mathbb{R}^n$. So lets assume $x \neq 0$ and check the reciprocal inclusions, $N_1 \subset N_2$ and $N_2 \subset N_1$. The former here looks to be the "easy" direction. For if $x^T y = 0$ then surely $xx^T y = 0$. Next, if $xx^T z = 0$ then $x^T xx^T z = 0$, i.e., $\|x\|^2 x^T z = 0$ which, as $x \neq 0$ implies that $x^T z = 0$.

Regarding **Infinite Sequences** $\{x_n\}_{n=1}^{\infty} \subset \mathbb{R}$ we note that although the elements may change erratically with $n$ we may always extract a well ordered subsequence. For example, from the oscillatory $x_n = (-1)^n/n$ we may extract the decreasing $x_{n_k} \equiv x_{2k} = 1/(2k)$. More generally we call a sequence **monotone** if either $x_n \leq x_{n+1}$ for all $n$ or $x_n \geq x_{n+1}$ for all $n$. We state and prove the general case:

**Proposition** 1.1. Given $\{x_n\}_{n=1}^{\infty} \subset \mathbb{R}$ there exists a monotone subsequence $\{x_{n_k}\}_{k=1}^{\infty} \subset \mathbb{R}$.

Proof: Call $x_n$ a peak if $x_n > x_m$ for all $m < n$. If our sequence has no peaks then it is already monotone. If our sequence has an infinite number of peaks (as in our example above) at $n_1 < n_2 < \cdots$ then $x_{n_1} \geq x_{n_2} \geq \cdots$ is a monotone subsequence. It remains to study sequences with at least one but at most finitely many peaks. In this case, if $x_N$ is the peak with the biggest index then $x_{n_1}$ where $n_1 = N + 1$ is not a peak and so $\exists$ and $n_2 > n_1$ such that $x_{n_2} \geq x_{n_1}$. In the same fashion, as $n_2$ is not a peak $\exists$ and $n_3 > n_2$ such that $x_{n_3} \geq x_{n_2}$. On repetition this procedure generates an infinite monotone subsequence. End of Proof.

The great attraction of (bounded) monotone sequences is that they must converge to their smallest or largest value. To make this precise we call $u$ an upper bound for $\{x_n\}$ if $x_n \leq u$ for all $n$ and we denote by $x^u$ the **least upper bound**. For example, 1 is the least upper bound of $\{1 - 1/n\}_n$.

**Proposition** 1.2. If $\{x_n\}_n$ is monotonically nondecreasing and $x^u$ is its least upper bound then

$$\lim_{n \to \infty} x_n = x^u.$$

That is, given any $\varepsilon > 0$ $\exists$ $N > 0$ such that $|x_n - x^u| \leq \varepsilon$ $\forall$ $n > N$. We often abbreviate this as $x_n \to x^u$.

Proof: Given $\varepsilon > 0$ if there exists an $N > 0$ such that $x_n \leq x^u - \varepsilon$ for $n > N$ then $x^u - \varepsilon/2$ is an upper bound less than $x^u$, contrary to its definition. End of Proof.

In a similar fashion we call $\ell$ a lower bound for $\{x_n\}$ if $x_n \geq \ell$ for all $n$ and we denote by $x^\ell$ the **greatest lower bound**. For example, 0 is the greatest lower bound of $\{1/n\}_n$. If $\{x_n\}$ is nonincreasing then $x_n \to x^\ell$. Combining these last two propositions we find that every bounded sequence of real numbers has a convergent subsequence. Our argument in fact translates nicely to vectors.

**Proposition** 1.3. If $\{x_j\}_j \subset \mathbb{R}^n$ and there exists a finite $M$ for which $\|x_j\| \leq M$ for all $j$ then there exists a subsequence $\{x_{j_k}\} \subset \{x_j\}_j$ and an $x \in \mathbb{R}^n$ such that $x_{j_k} \to x$. That is given any $\varepsilon > 0$ $\exists$ $N > 0$ such that $\|x_{j_k} - x\| \leq \varepsilon$ $\forall$ $j_k > N$.

Proof: We note the elements of $x_j$ by $x_j(1)$ through $x_j(n)$. As $\{x_j(1)\}_j$ is a bounded sequence in $\mathbb{R}$ it has a subsequence, $\{x_{j_k}(1)\}_j$, that converges to a number that we label $x(1)$. As $\{x_{j_k}(2)\}_j$ is a bounded sequence in $\mathbb{R}$ it has a subsequence, $\{x_{j_{k_l}}(2)\}_l$, that converges to a number that we label $x(2)$. Moreover, this new subsequence does not affect the convergence of the first element. In particular, $x_{j_{k_l}}(1) \to x(1)$ as $l \to \infty$. We now continue to extract a subsequence from the previous sequence until we have exhausted all $n$ dimensions. End of Proof.

Our first application of this is to an alternate notion of matrix norm. We observed in Eq. (1.20) that the Frobenius norm is larger than the biggest stretch. The word "biggest" suggest that we are looking for the least upper bound. This three word phrase is a bit awkward and so is often rephrased as *supremum* which itself it abbreviated to sup. All this suggests that

$$\|A\| \equiv \sup_{\|x\|=1} \|Ax\| \tag{1.31}$$

is worthy of study. By definition there exists a sequence $\{x_j\}_j$ of unit vectors for which $\|Ax_j\| \to \|A\|$. By Prop. 1.3 there exists a convergent subsequence, $x_{j_k} \to \tilde{x}$. It follows that $\|x_{j_k}\| \to \|\tilde{x}\|$ and so $\|\tilde{x}\| = 1$. In addition,

$$\|Ax_{j_k} - A\tilde{x}\| = \|A(x_{j_k} - \tilde{x})\| \leq \|A\|_F \|x_{j_k} - \tilde{x}\|$$

permits us to conclude that $Ax_{j_k} \to A\tilde{x}$ and so $\|Ax_{j_k}\| \to \|A\tilde{x}\|$ and recalling $\|Ax_{j_k}\| \to \|A\|$ we conclude that $\|A\tilde{x}\| = \|A\|$. The upshot is that the *supremum* in Eq. (1.31) is actually attained. We distinguish this fact by writing

$$\boxed{\|A\| \equiv \max_{\|x\|=1} \|Ax\|.} \tag{1.32}$$

By definition we know that $\|A\| \leq \|A\|_F$ for every matrix. A simple example that shows up the disparity involves $I_n$, the identity matrix on $\mathbb{R}^n$. Please confirm that $\|I_n\| = 1$ while $\|I_n\|_F = \sqrt{n}$.

### 1.4. Notes and Exercises

For thousands more worked examples I recommend Lipschutz (1989). Higham and Higham (2005) is an excellent guide to MATLAB. For a more thorough guide to proofs please see Velleman (2006).

1. Consider the matrix
$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{1.33}$$
Evaluate the product $Ax$ for several choices of $x$. Sketch both $x$ and $Ax$ in the plane for several carefully marked $x$ and explain why $A$ is called a "rotation." Argue, on strictly geometric grounds, why $A^5 = A$.

2. Consider the matrix
$$A = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}. \tag{1.34}$$
Evaluate the product $Ax$ for several choices of $x$. Sketch both $x$ and $Ax$ in the plane for several carefully marked $x$ and explain why $A$ is called a "reflection." Argue, on strictly geometric grounds, why $A^3 = A$.

3. We will consider the action of
$$A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}, \tag{1.35}$$
on the bow-ties, $E$ and $F$, of Eq. (1.23).

   (a) Show that if $x \in E$ then $Ax \in F$,

   (b) Show that if $x \in F$ then $Bx \in E$.

   (c) Prove by induction that
$$A^n = \begin{pmatrix} 1 & 2n \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad B^n = \begin{pmatrix} 1 & 0 \\ 2n & 1 \end{pmatrix},$$
for positive integer $n$.

   (d) Use (c) to generalize (a) and (b). That is, show that if $x \in E$ then $A^n x \in F$ while if $x \in F$ then $B^n x \in E$ for all positive integer $n$.

4. We will make frequent use of the **identity matrix**, $I \in \mathbb{R}^{n\times n}$, comprised of zeros off the diagonal and ones on the diagonal. In symbols, $I_{ij} = 0$ if $i \neq j$, while $I_{ii} = 1$. Prove the two propositions, if $A \in \mathbb{R}^{n\times n}$ then $AI = IA = A$. The identity also gives us a means to define the **inverse** of a matrix. One (square) matrix is the inverse of another (square) matrix if their product is the identity matrix. Please show that
$$A^{-1} = \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad B^{-1} = \begin{pmatrix} 1 & 0 \\ -2 & 1 \end{pmatrix}, \tag{1.36}$$
are the inverses of the $A$ and $B$ matrices of Eq. (1.35).

5. Write a MATLAB program to investigate the shear of the integer diamond by the $A$ and $B$ matrices, Eq. (1.35), and their inverses, Eq. (1.36). More precisely, write a program that generates Figure 1.6.

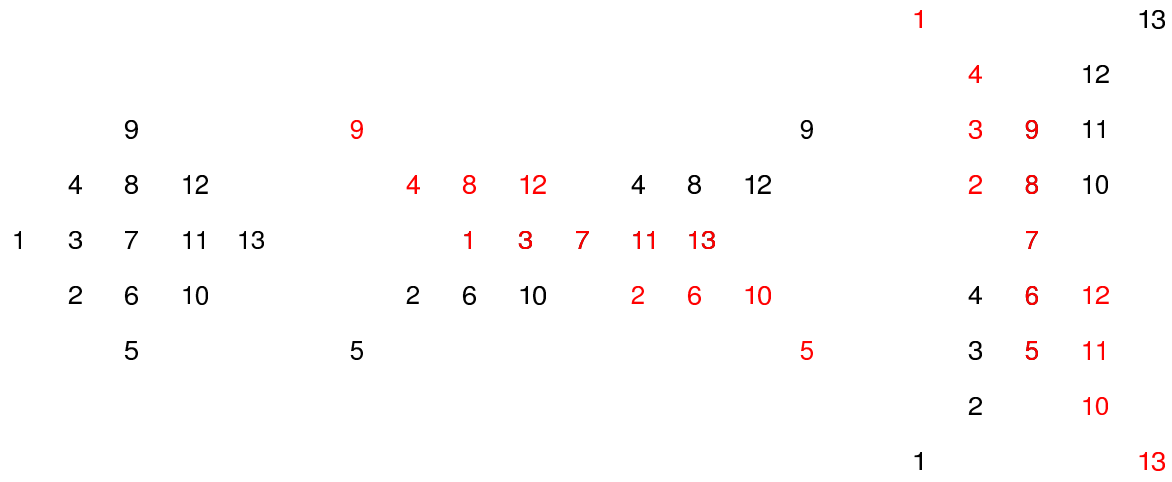



**Figure** 1.6. Shearing the integral diamond. (Left) The labels are at integral points, $1 = (-2, 0)$, $2 = (-1, -1)$, $3 = (-10)$, $4 = (-1, 1)$ and so on. (Center) Transformation by $A$ (black) and $A^{-1}$ (red) of the points in panel (Left). (Right) Transformation by $B$ (black) and $B^{-1}$ (red) of the points in panel (Left).

6. We can view, see Figure 1.7(A), vector sums as parallelogram generators. Please show that the area of this parallelogram is $ad - bc$. Show all of your work.

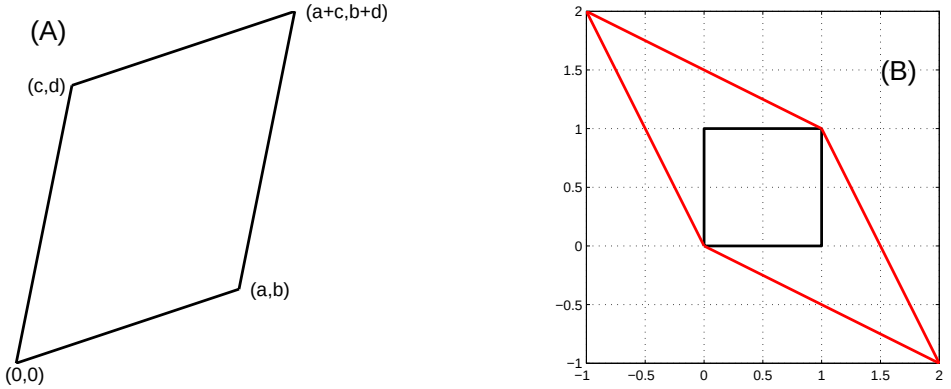



**Figure** 1.7. (A) The vectors $(a, b)$ and $(c, d)$ drawn from the origin, $(0, 0)$, sum to the fourth vertex of a parallelogram. (B) A black square and its deformation (red diamond) by the matrix in (1.37)

7. Show that

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \tag{1.37}$$

takes the black square to the red diamond in Figure 1.7(B). Use the previous exercise to compute the area of the red diamond.

8. Each of the following chapters will demonstrate the fundamental role that matrices play in modeling the world. Perhaps one of the simplest contexts is in the field of information retrieval. Here one has $m$ "terms" and $n$ "documents" and builds a so–called term-by-document matrix $A$ where $a_{ij}$ is the number of times that term $i$ appears in document $j$. In Figure 1.8(A) below we depict such a matrix where the documents are the 81 chapters of the *Tao Te Ching* and our

10 terms are *heaven, virtue, nature, life, knowledge, understand, fear, death, good*, and *right*. This matrix is then used to process new queries. For example, if the disciple is looking for the chapters most expressive of *virtue* and *good* then, as these are the second and ninth of our our terms we build the query vector

$$q = (0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0) \tag{1.38}$$

and search for means to compare this to the columns of $A$. The standard approach is to exploit the geometric interpretation (recall Eq. (1.5)) of the inner product and to so rank the chapters by the cosine of the angle they make with the query. More precisely, for the $j$th document we compute

$$\cos(\theta_j) = \frac{qa_j}{\|q\|\|a_j\|}. \tag{1.39}$$

and present these scores in Figure 1.8(B). As small angles correspond to values of cosine near 1 our analysis would direct the disciple to chapter 49 of the *Tao Te Ching*. Typically a threshold is chosen, e.g., 0.8, and a rank ordered list of all documents that exceed that threshold is returned.

Please change `tao.m` to find the chapter most expressive of *heaven, nature* and *knowledge*.

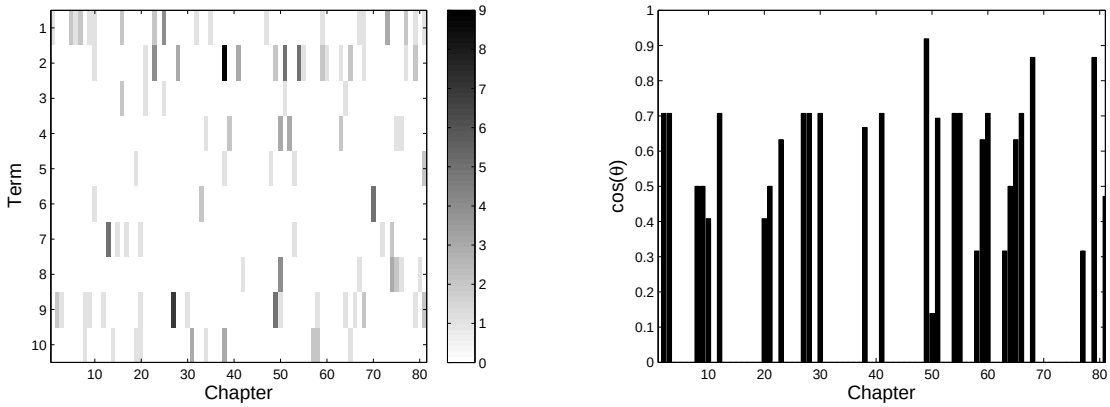



**Figure** 1.8. Query matching. (A) The $10 \times 81$ term-by-document matrix for the *Tao Te Ching*, illustrated with the help of the MATLAB command `imagesc`. (B) The cosine scores associated with the query in Eq. (1.38) as expressed in Eq. (1.39). `tao.m`

9. Prove that matrix multiplication is associative, i.e., that $(AB)C = A(BC)$.

10. Prove that if $x$ and $y$ lie in $\mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$ then

$$x^T Ay = y^T A^T x.$$

Hint: The left side is a number. Now argue as we did in achieving Eq. (1.16).

11. Suppose that $A \in \mathbb{R}^{n \times n}$ and $x^T Ax = 0\ \forall\ x \in \mathbb{R}^n$. Does this imply that $A = 0$? If so, prove it. If not, offer a counterexample.

12. Prove that $\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B)$.

13. Use Eq. (1.14) to prove that the fundamental commutator relation of Quantum Mechanics,

$$AB - BA = I,$$

can *not* hold for matrices.

14. Prove that $\mathrm{tr}(uv^T) = u^T v$ for $u$ and $v$ in $\mathbb{R}^n$.

15. Construct a nonzero $A \in \mathbb{R}^{2\times 2}$ for which $A^2 = 0$.

16. A matrix that equals its transpose is called **symmetric**. Suppose $S = A^T G A$ where $A \in \mathbb{R}^{m\times n}$ and $G \in \mathbb{R}^{m\times m}$. Prove that if $G = G^T$ then $S = S^T$.

17. Please confirm that the **polarization formula**

$$\|u+v\|^2 - \|u-v\|^2 = 4u^T v, \tag{1.40}$$

holds for all $u$ and $v$ in $\mathbb{R}^n$.

18. Establish the **triangle inequality**

$$\|x+y\| \le \|x\| + \|y\| \qquad \forall\ x, y \in \mathbb{R}^n. \tag{1.41}$$

First draw this for two concrete planar $x$ and $y$ and discuss the aptness of the name. Then, for the general case expand $\|x+y\|^2$, invoke the Cauchy–Schwarz inequality, Eq. (1.7), and finish with a square root.

19. The other natural vector product is the **outer product**. Note that if $x \in \mathbb{R}^n$ then the outer product of $x$ with itself, $xx^T$, lies in $\mathbb{R}^{n\times n}$. Please prove that $\|xx^T\|_F = \|x\|^2$.

20. The outer product is also a useful ingredient in the **Reflection Matrix**

$$H = I - 2xx^T, \tag{1.42}$$

associated with the unit vector $x$.

(a) How does $H$ transform vectors that are multiples of $x$?

(b) How does $H$ transform vectors that are orthogonal to $x$?

(c) How does $H$ transform vectors that are neither colinear with nor orthogonal to $x$? Illustrate your answers to (a-c) with a careful drawing.

(d) Confirm that $H^T = H$ and that $H^2 = I$.

21. There is a third way of computing the product of two vectors in $\mathbb{R}^3$, perhaps familiar from vector calculus. The **cross product** of $u$ and $v$ is written $u \times v$ and defined as the matrix vector product

$$u \times v \equiv \mathtt{X}(u)v = \begin{pmatrix} 0 & -u_3 & u_2 \\ u_3 & 0 & -u_1 \\ -u_2 & u_1 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} -u_3 v_2 + u_2 v_3 \\ u_3 v_1 - u_1 v_3 \\ -u_2 v_1 + u_1 v_2 \end{pmatrix}$$

(a) How does $\mathtt{X}(u)$ transform vectors that are multiples of $u$?

(b) How does $\mathtt{X}(u)$ transform vectors that are orthogonal to $u$?

(c) How does $\mathtt{X}(u)$ transform vectors that are neither colinear with nor orthogonal to $u$? Illustrate your answers to (a-c) with a careful drawing. You may wish to use the MATLAB function `cross`.

(d) Confirm that $\mathtt{X}(u)^T = -\mathtt{X}(u)$ and that $\mathtt{X}(u)^2 = uu^T - \|u\|^2 I$.

(e) Use (d) to derive

$$\|u \times v\|^2 = \|u\|^2\|v\|^2 - (u^T v)^2.$$

(f) If $\theta$ is the angle between $u$ and $v$ use (e) and (1.5) to show that

$$\|u \times v\| = \|u\|\|v\||\sin\theta|.$$

(g) Use (f) and Figure 1.9(A) to conclude that $\|u \times v\|$ is the area (base times height) of the parallelogram with sides $u$ and $v$.

(h) Use (g) and Figure 1.9(B) to conclude that $|w^T(u \times v)|$ is the volume (area of base times height) of the parallelepiped with sides $u$, $v$ and $w$. Hint: Let $u$ and $v$ define the base. Then $u \times v$ is parallel to the height vector obtained by drawing a perpendicular from $w$ to the base.

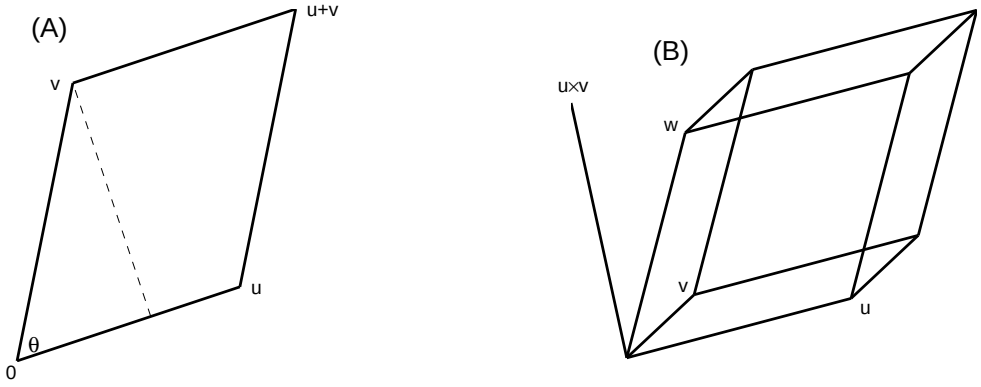



**Figure** 1.9. (A) Parallelogram. (B) Parallelepiped.

22. Show that if $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$ then $\|AB\|_F \le \|A\|_F\|B\|_F$. Hint: Adapt the proof of Eq. (1.19).

23. Via experimentation with small $n$ arrive (show your work) at a formula for $f_n$ in

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}^n = \begin{pmatrix} 1 & n & f_n \\ 0 & 1 & n \\ 0 & 0 & 1 \end{pmatrix}$$

and prove, via induction, that your formula holds true for all $n$.

24. Suppose that $\{a_j : j = 0, \pm1, \pm2, \ldots\}$ is a doubly infinite sequence. Prove, via induction, that

$$\sum_{j=0}^{n}\sum_{k=0}^{n} a_{j-k} = \sum_{m=-n}^{n} (n + 1 - |m|)a_m. \tag{1.43}$$

25. For the matrix of (1.37) compute, by hand and showing all work, that $\|A\| = 3$ and $\|A\|_F = \sqrt{10}$. Hint: For the former, choose $x = (\cos(\theta), \sin(\theta))^T$ and show that $\|Ax\|^2 = 5 - 8\cos(\theta)\sin(\theta)$. Now take a derivative in order to find the $\theta$ that gives the largest $\|Ax\|$.

# 2. Electrical Networks

The analysis and design of large electrical networks has always been conducted in the language of Linear Algebra. In fact, Gustav Kirchhoff, the law–giver of electrical circuit theory, had a significant impact on the creation of what we now call Linear Algebra.

We here develop his laws and algebra in the simple, but important, setting of nerve conduction. In §2.1 we construct a physical model of a neuron as a network of resistors and then build a mathematical model to predict its response to current injection. In §§2.2 and 2.3 we extend both the physical model and mathematical theory to encompass batteries and operational amplifiers. In each case we arrive at a linear system of equations for the internal (unknown) voltages and currents in terms of the known external voltages and currents. We develop this linear system in a general four part scheme known as a Strang Quartet.

### 2.1. Neurons and the Strang Quartet

The human brain is an electrical network of 100 billion neurons. A neuron is a spatially extended cell that, based upon inputs from its ten thousand upstream neighbors, signals its downstream neighbors. A neuron's spatial extent is typically idealized as simply a cylinder of radius $a$ and length $\ell$ 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. These are expressed via two quantities: $\rho_i$, the resistivity, in $\Omega\, cm$, of the cytoplasm that fills the cell, and $\rho_m$, the resistivity in $\Omega\, cm^2$ of the cell's lateral membrane.

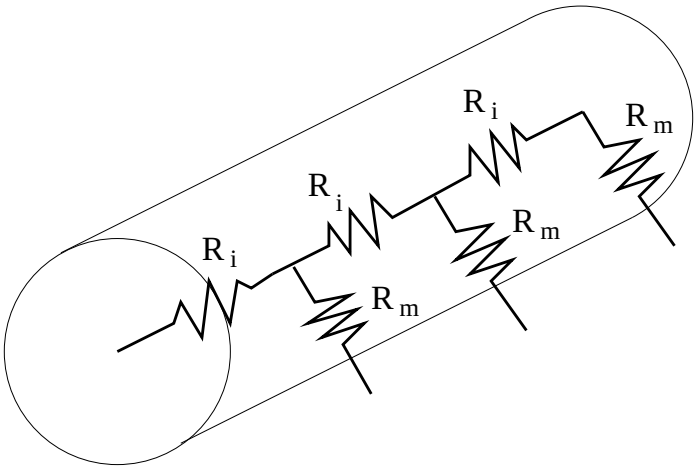



**Figure** 2.1. A 3 compartment model of a neuron.

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 $\ell/N$. Denoting a segment's axial and membrane resistance by

$$R_i = \frac{\rho_i \ell/N}{\pi a^2} \quad \text{and} \quad R_m = \frac{\rho_m}{2\pi a \ell/N}$$

respectively, we arrive at the lumped circuit model of Figure 2.1. For a neuron in a dish we may assume a constant extracellular potential, e.g., zero. We accomplish this by grounding the extracellular nodes, see Figure 2.2.

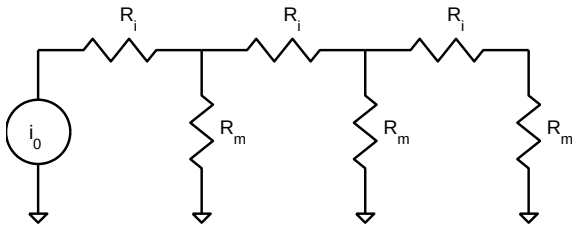



**Figure** 2.2. A rudimentary neuronal circuit model.

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 2.3 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.

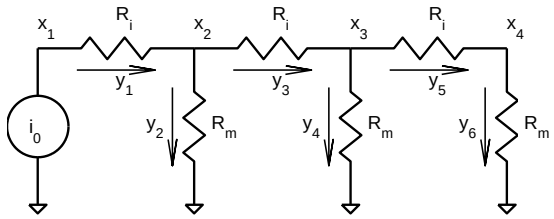



**Figure** 2.3 The fully dressed circuit model.

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

(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 incidence 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.

### 2.2. Resistor Nets with Current Sources and Batteries

With respect to the circuit of Figure 2.3, 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

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

where the reason for the leading minus sign will be revealed in the next section.

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,

$$y = Ge \quad \text{where} \quad G = \begin{pmatrix} 1/R_i & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/R_m & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/R_i & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/R_m & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/R_i & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/R_m \end{pmatrix}.$$

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

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

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

$$\boxed{A^T G A x = f.} \tag{2.1}$$

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 Eq. (2.1) may have either 1, 0, or infinitely many solutions, depending on $f$ and $A^T G A$. 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 2.4.

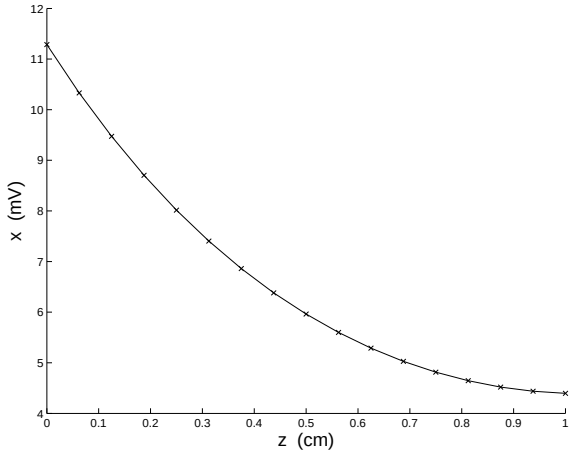



**Figure** 2.4. Results of a 16 compartment neuronal simulation. The voltage, as a function of distance from the left end, computed from (2.1) with $i_0 = 0.001\ mA$ (milliAmperes). `cab1.m`.

Once the structure of the constituents in the fundamental system Eq. (2.1) 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 2.4 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.

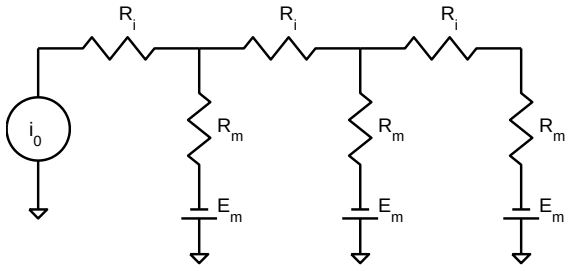



**Figure** 2.5 Circuit model with batteries associated with the rest potential.

We have seen 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 2.5.

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

$$\boxed{A^T G A x = A^T G b + f.} \tag{2.2}$$

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

Returning to Figure 2.5 we note that

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

To build and solve Eq. (2.2) requires only minor changes to our old code. The new program is called `cab2.m` and results of its use are indicated in Figure 2.6.

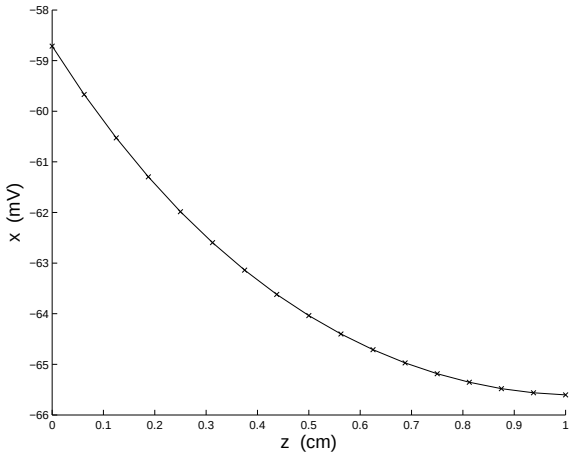



**Figure** 2.6. Results of a 16 compartment simulation with batteries, $E_m = -70\ mV$. `cab2.m`

### 2.3. Operational Amplifiers*

The true work horse of analog circuitry is the operational amplifier, or op-amp for short. It is an ingenious blend of nonlinear circuit elements (transistors) that yields straightforward linear combinations of its inputs. As resistor nets dissipate energy opamps actually increase energy. More precisely, they transmit energy to the circuit – for they are active devices that require (like neurons) their own energy source. The opamp symbol and a standard configuration are illustrated in Figure 2.7

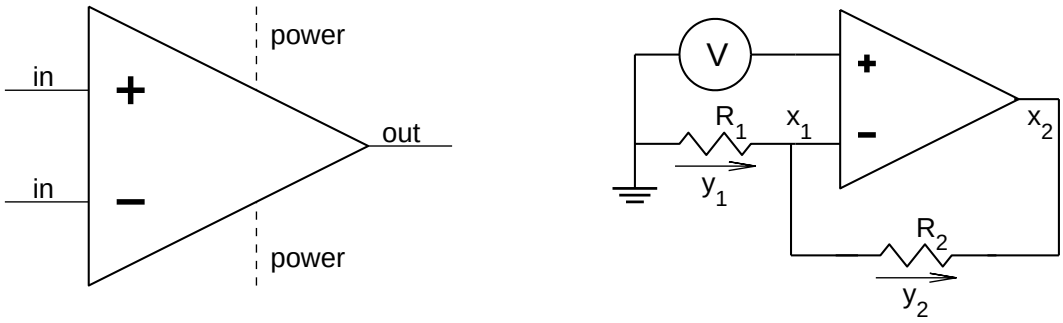



**Figure** 2.7. (A) An operational amplifier has two input terminals, labeled ±, and one output terminal. It also has two terminals for its power supply. In the future we will assume that all opamps are powered and ignore these connections. (B) This circuit, called the "noniverting amplifier," magnifies its input by a factor of $(1 + R_2/R_1)$. See Eq. (2.3).

The laws that govern the operation of opamps are

**(OA1)** The potentials at the two input terminals coincide, i.e., are equal to one another.
**(OA2)** The current entering the opamp at each input terminal is zero.

We now illustrate these laws in small and large networks. Starting with Figure 2.7 (OA1) dictates that $x_1 = V$ while (OA2) that $y_1 = y_2$. Together these two state that

$$\frac{0 - V}{R_1} = \frac{V - x_2}{R_2}$$

and so

$$\boxed{x_2 = (1 + R_2/R_1)V.} \tag{2.3}$$

This result causes us to speak of $(1 + R_2/R_1)$ as the **gain** of circuit Figure 2.7(B). As $R_1 \to 0$ the algebra suggests that we might achieve infinite gain. In reality, the opamp is typically powered, as in Figure 2.7(A) by $\pm V_{pow}$ volts and as such we can not get more than we put in. More precisely, (2.3) is actually

$$x_2 = \begin{cases} -V_{pow} & \text{if } (1 + R_2/R_1)V < -V_{pow} \\ (1 + R_2/R_1)V & \text{if } -V_{pow} \le (1 + R_2/R_1)V \le V_{pow} \\ V_{pow} & \text{if } (1 + R_2/R_1)V > V_{pow}. \end{cases} \tag{2.4}$$

In order to extend our analysis to larger circuits we need to reflect both on what we did and what we did not do en route to Eq. (2.3). What we did was to invoke the opamp rules (OA1) and (OA2). What we did not do was to balance current at the output terminal, for an opamp (being a powered device) does not conserve current between its input and output terminals. It is worth stating this:

**(OA3)** Do not apply KCL at the output terminal of any opamp.

Fortunately these rules do balance in the mathematical sense. That is, though they modify one step of the Strang Quartet they still lead to a consistent set of equations for all unknown potentials. To illustrate this we consider the differential amplifier of Figure 2.8

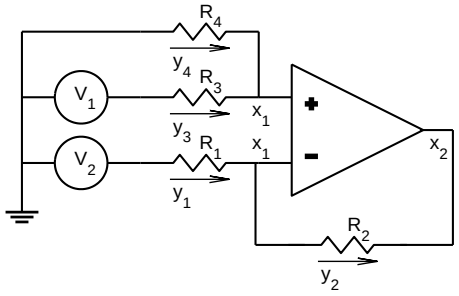



**Figure** 2.8. A differential amplifier. That is, a circuit that amplifies the difference $V_1 - V_2$.

Regarding the circuit of Figure 2.8, following (OA1) we have labeled both input terminals with the same unknown potential. The Strang Quartet begins, as usual, by expressing the voltage drops

$$\begin{aligned} e_1 &= V_2 - x_1 \\ e_2 &= x_1 - x_2 \\ e_3 &= V_1 - x_1 \\ e_4 &= 0 - x_1 \end{aligned} \quad \text{as} \quad e = b - Ax, \quad \text{where} \quad b = \begin{pmatrix} V_2 \\ 0 \\ V_1 \\ 0 \end{pmatrix} \quad \text{and} \quad A = \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}.$$

The next step, $y = Ge$, also proceeds, as before, with $G = \text{diag}(G_1, G_2, G_3, G_4)$ where $G_j \equiv 1/R_j$. Unlike the standard step of current balance we ignore the output terminal and yet balance currents at both input terminals even though their potentials are identical. In particular, current balance now takes the form

$$\begin{aligned} y_3 + y_4 &= 0 \\ y_1 - y_2 &= 0 \end{aligned} \quad \text{i.e.,} \quad By = 0 \quad \text{where} \quad B = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 1 & -1 & 0 & 0 \end{pmatrix}.$$

We note at once that $B \neq A^T$ but proceed to unpack $By = 0$ to $BGe = 0$ and then $BG(b - Ax)$, i.e.,

$$\boxed{BGAx = BGb.} \tag{2.5}$$

In the case of Figure 2.8 we find

$$BGA = \begin{pmatrix} G_3 + G_4 & 0 \\ G_1 + G_2 & -G_2 \end{pmatrix} \quad \text{and} \quad BGb = \begin{pmatrix} G_3 V_1 \\ G_1 V_2 \end{pmatrix}.$$

We can read off the solution to this triangular system

$$x_1 = \frac{G_3}{G_3 + G_4} V_1 \quad \text{and} \quad x_2 = (1 + G_1/G_2)x_1 - (G_1/G_2)V_2. \tag{2.6}$$

It follows that $x_2$ is a weighted difference of the two input voltages. We can simplify the algebra on choosing $R_1 = R_3$ and $R_2 = R_4 = \Gamma R_1$, for then Eq. (2.6) yields

$$\boxed{x_2 = \Gamma(V_1 - V_2),} \tag{2.7}$$

and we recognize Figure 2.8 as a **differential amplifier** with gain $\Gamma$.

### 2.4. Notes and Exercises

The modular, 4 part, approach to building mathematical models is due to Strang (2007). For a more thorough introduction to neuronal modeling please consult Gabbiani and Cox (2010).

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 Eq. (2.1). The second equation should read
$$(-x_1 + 2x_2 - x_3)/R_i + x_2/R_m = 0. \tag{2.8}$$

2. Let us work right to left in the circuit of Figure 2.3.
   (a) Deduce from KCL that $y_5 = y_6$ implies
   $$x_4 = \frac{R_m}{R_i + R_m} x_3 \quad \text{and} \quad y_5 = \frac{x_3}{R_i + R_m}.$$
   (b) Deduce from KCL that $y_3 - y_4 = y_5$ implies
   $$x_3 = \frac{R_m(R_i + R_m)}{R_i^2 + 3R_iR_m + R_m^2} x_2 \quad \text{and} \quad y_3 = x_2 \frac{R_i + 2R_m}{R_i^2 + 3R_iR_m + R_m^2}.$$
   (c) Deduce from KCL that $y_1 - y_2 = y_3$ implies
   $$x_2 = \frac{R_m^3 + 3R_iR_m^2 + R_i^2R_m}{5R_mR_i^2 + 6R_iR_m^2 + R_m^3 + R_i^3} x_1$$
   (d) Deduce from $i_0 = y_1$ that
   $$x_1 = i_0 \frac{R_i^3 + 5R_i^2R_m + 6R_iR_m^2 + R_m^3}{(R_i + R_m)(R_i + 3R_m)}$$
   and finally that the cell's input resistance is
   $$R_{in} \equiv \frac{x_1}{i_0} = \frac{R_i^3 + 5R_i^2R_m + 6R_iR_m^2 + R_m^3}{(R_i + R_m)(R_i + 3R_m)}$$

3. The pattern in the leading term of Eq. (2.8), twice self minus the contribution of the immediate neighbors, appears naturally when differentiating functions. In particular, please confirm that
   $$f(z+h) = f(z)+f'(z)h+f''(z)h^2/2+O(h^3) \quad \text{and} \quad f(z-h) = f(z)-f'(z)h+f''(z)h^2/2+O(h^3)$$
   where $O(h^3)$ indicates terms of order $h^3$. Now add these two expressions and arrive at
   $$f''(z) = \frac{f(z-h) - 2f(z) + f(z+h)}{h^2} + O(h^3) \tag{2.9}$$

4. 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, as in Figure 2.9, all of the potentials on the same graph, using different marker types for each. (You may wish to alter `cab1.m` so that it accepts $N$ and marker as arguments and then call it from a driver that uses `hold` and appends a `legend`).

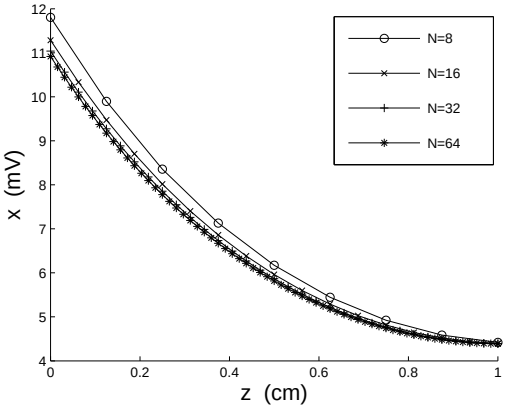



**Figure** 2.9. Apparent convergence of the cable response as the number of compartments grows.

Let us now interpret this convergence. The main observation is that the difference equation, Eq. (2.8), 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 Eq. (2.8) and recalling the definitions of $R_i$ and $R_m$ we see Eq. (2.8) 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 (thanks to Eq. (2.9)) at

$$\frac{d^2x(z)}{dz^2}-\frac{2\rho_i}{a\rho_m}x(z)=0. \tag{2.10}$$

(b) Recall that $2\cosh(t) = \exp(t)+\exp(-t)$ and $2\sinh(t) = \exp(t)-\exp(-t)$, set $\mu \equiv 2\rho_i/(a\rho_m)$ and show that

$$x(z) = \alpha\sinh(\sqrt{\mu}z)+\beta\cosh(\sqrt{\mu}z) \tag{2.11}$$

satisfies Eq. (2.10) regardless of $\alpha$ and $\beta$.

We shall determine $\alpha$ and $\beta$ 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

$$\frac{dx(0)}{dz}=-\frac{\rho_i i_0}{\pi a^2}. \tag{2.12}$$

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

$$\frac{dx(\ell)}{dz}=0. \tag{2.13}$$

(c) Substitute Eq. (2.11) into Eq. (2.12) and Eq. (2.13) and solve for $\alpha$ and $\beta$ and write out the final $x(z)$.

(d) Substitute into $x$ the $\ell, a, \rho_i$ and $\rho_m$ values used in `cab1.m`, plot the resulting function (using, e.g., `ezplot`) and compare this to the plot achieved in part (a).

(e) One distinct advantage of this exact solution is that it permits us to express the neuron's input resistance $R_{in} \equiv x(0)/i_0$ in terms of its fundamental material and geometric constants. Use part (c) to arrive at

$$R_{in}=\frac{\sqrt{\rho_i \rho_m/2}\cosh(\sqrt{\mu}\ell)}{\pi a^{3/2}\sinh(\sqrt{\mu}\ell)}.$$

Prove that this is an increasing function of $\ell$.

5. Suppose that we specify the potential, rather than inject current, at the neuron's left end...

6. Alter `cab2.m` to inject current at a specified node. Reproduce Figure 2.10.

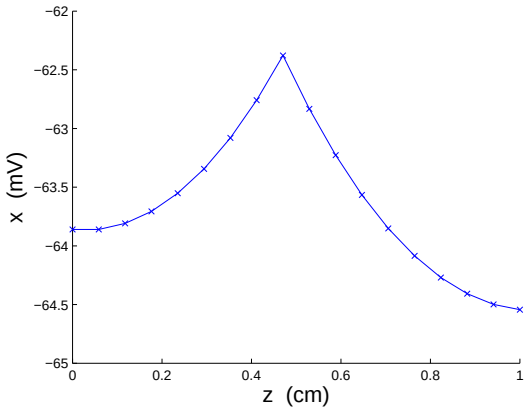



**Figure** 2.10. Response of a 16 node cable, with batteries, to stimulus at node 9.

7. Neurons are rarely straight. Instead, to maximize their contact with neighbors they branch, as in the fork of Figure 2.11. Derive the associated node–edge incidence matrix.

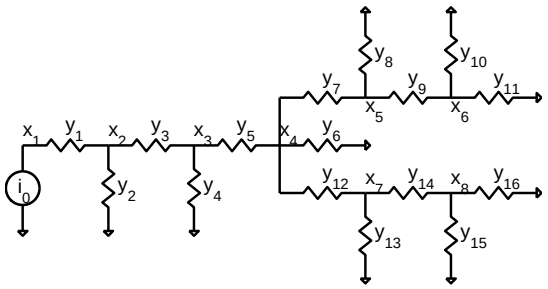



**Figure** 2.11. A compartmental model of a forked neuron. In order to reduce clutter we have neglected to orient the edges. To derive the incidence matrix please use rightward pointing arrows on the axial resistances and ground pointing arrows on the membrane conductances.

8. The circuit depicted in Figure 2.12 is known as a **Wheatstone Bridge** .

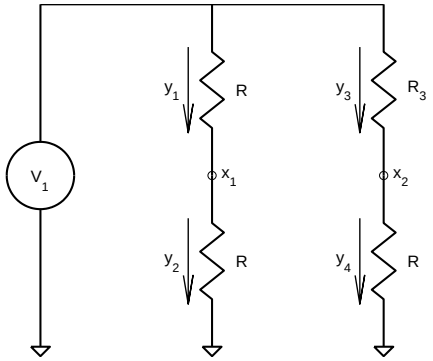



**Figure** 2.12. A wheatstone bridge. As we vary the odd resistance, $R_3$, we will arrive at a variable output voltage, $x_1 - x_2$.

(a) Carefully derive the equilibrium equations for the potentials $x_1$ and $x_2$.

(b) Solve the equations in (a) for

$$x_1 = V_1/2 \quad \text{and} \quad x_2 = V_1 \frac{R}{R + R_3}.$$

(c) If $V \equiv x_1 - x_2$ is now the voltage output of the bridge, show that it varies from $-V_1/2$ to $V_1/2$ as $R_3$ climbs from 0 to $\infty$.

9. The Instrumentation Amplifier of Figure 2.13(A) is an improvement over the differential amplifier

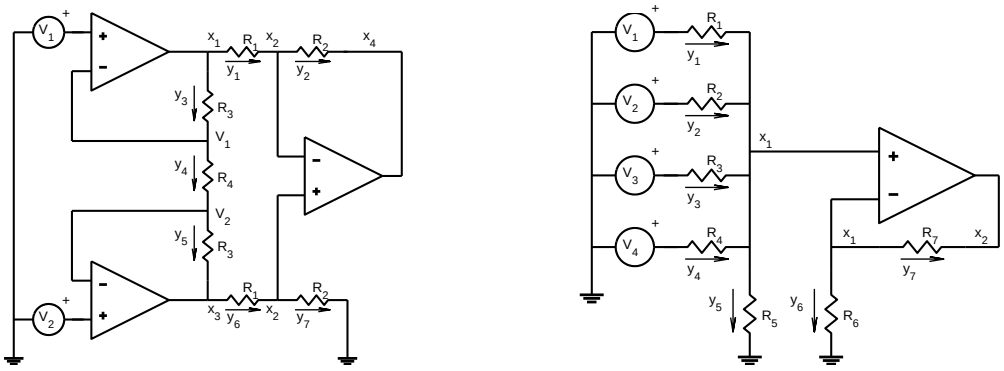



**Figure** 2.13. (A) The Instrumentation Amplifier. (B) The Summer.

Construct and solve Eq. (2.5) and conclude that

$$x_4 = (1 + 2R_3/R_4)(V_2 - V_1).$$

10. The circuit of Figure 2.13(B) is deemed a summer. Construct and solve Eq. (2.5) and conclude that

$$x_2 = \sum_{j=1}^{4} \gamma_j V_j \quad \text{where} \quad \gamma_j = \frac{1 + G_6/G_7}{G_1 + G_2 + G_3 + G_4 + G_5} G_j.$$

11. Note that our Summer and Differential Amplifier circuits are actually computing inner products of their inputs. As matrix vector multiplication is nothing more than a vector of inner products, we may now design circuits to implement matrix multiplication. Please combine two differential amplifiers to arrive at a circuit that performs $x = SV$ where

$$S = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \quad \text{and} \quad V = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix}.$$

Draw the circuit and specify all resistances.

# 3. Mechanical Networks

We derive the equations of mechanical equilibrium by developing and applying the matrix forms of Hooke's Law and Conservation of Force. We solve these equations, by hand, via Gaussian Elimination. This concrete elementary scheme reveals the importance of pivots and leads us to the Gauss–Jordan method of matrix inversion, the LU method of matrix factorization, and to the important notion of matrix determinant. In the final section we discover that the matrix governing mechanical equilibrium is positive definite. We then demonstrate that force balance is equivalent to the minimization of potential energy. Throughout the chapter we illustrate each of these ideas on mechanical networks of increasing complexity.

### 3.1. Elastic Fibers and the Strang Quartet

We connect 3 masses (nodes) with four springs (fibers) between two immobile walls, as in Figure 3.1, and apply forces at the masses and seek to determine the associated displacements.

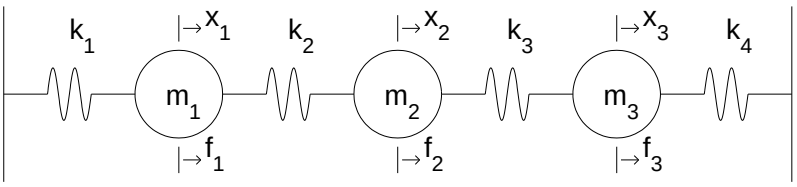



**Figure** 3.1. A fiber chain.

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,

$$k_j = \frac{Y_j a_j}{L_j} \tag{3.1}$$

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 $j$th 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,

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

where $A$ is the associated node–edge incidence matrix. 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. The constant of proportionality is the stiffness, $k_j$, in (3.1). If we denote the restoring force by $y_j$ Hooke's Law then reads $y_j = k_j e_j$, or,

in matrix terms

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

As (3.1) implies that each $k_j > 0$ we see that restoring forces echo the sign convention for elongations. Namely, $y_j$ is positive when spring $j$ is stretched, and negative when it is compressed. A positive restoring force will then "pull" on its ends and we arrive at the free body diagram in Figure 3.2 for each mass in Figure 3.1.

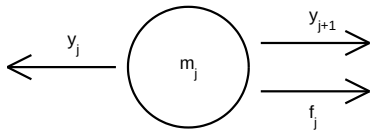



**Figure** 3.2. The free body diagram for the $j$th mass of Figure 3.1.

Balancing these forces at each mass we find

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

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

$$\begin{aligned} e &= Ax \\ y &= Ke \\ A^T y &= f \end{aligned} \tag{3.2}$$

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

$$A^T y = f \Rightarrow A^T K e = f \Rightarrow \boxed{A^T K A x = f.} \tag{3.3}$$

These four steps, (3.2)-(3.3), comprise the **Strang Quartet** for mechanical networks. Assembling $A^T K A$ we arrive at the final system

$$\begin{pmatrix} k_1 + k_2 & -k_2 & 0 \\ -k_2 & k_2 + k_3 & -k_3 \\ 0 & -k_3 & k_3 + k_4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix}. \tag{3.4}$$

For prescribed $k$ and $f$ values we view this as a linear system of three equations for the three unknown $x$ values. The key to solving such systems is to eliminate coefficients below the diagonal so that the remaining triangular system may be solved by back substitution. Its suite of variations on this idea of **Gaussian Elimination** is the workhorse within MATLAB. As we aim to develop a deeper understanding of Gaussian Elimination 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 physical systems that lack solutions or possess too many.

### 3.2. Gaussian Elimination and LU Decomposition

In Gaussian Elimination one first uses linear combinations of preceding rows to eliminate nonzeros below the main diagonal and then solves the resulting upper triangular system via back substitution. To firm up our understanding let us take up the case where each $k_j = 1$ and so (3.4) takes the form $Sx = f$, i.e.,

$$\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} \tag{3.5}$$

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

$$\text{new row } 2 = \text{old row } 2 + \frac{1}{2}\text{row } 1, \tag{3.6}$$

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

$$\text{new row } 3 = \text{old row } 3 + \frac{2}{3}\text{row } 2, \tag{3.7}$$

bringing the upper–triangular system

$$Ux = g, \tag{3.8}$$

or, more precisely,

$$\begin{pmatrix} 2 & -1 & 0 \\ 0 & 3/2 & -1 \\ 0 & 0 & 4/3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 + f_1/2 \\ f_3 + 2f_2/3 + f_1/3 \end{pmatrix} \tag{3.9}$$

One now simply reads off

$$x_3 = (f_1 + 2f_2 + 3f_3)/4. \tag{3.10}$$

This in turn permits, via so-called **back substitution**, the solution of the second equation

$$x_2 = 2(x_3 + f_2 + f_1/2)/3 = (f_1 + 2f_2 + f_3)/2, \tag{3.11}$$

and, in turn,

$$x_1 = (x_2 + f_1)/2 = (3f_1 + 2f_2 + f_3)/4. \tag{3.12}$$

One must say that Gaussian Elimination has succeeded here. For, regardless of the actual elements of $f$ we have produced an $x$ for which $A^T K A x = f$. On collecting (3.10)–(3.12) in matrix vector form we discover the beautifully symmetric form

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix}, \tag{3.13}$$

or $x = Zf$ for short. If we contrast this with our starting point, (3.5), or $Sx = f$ for short we find $f = Sx = SZf$. As this holds for every $f$ it follows that $SZ$ can only be the identity matrix, $I$. We can indeed confirm that

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

that is, $SZ = I$. Now whenever the product of two square matrices is the identity matrix then these two matrices are the inverses of one another. To make this explicit we use the notation $Z = S^{-1}$ and call it the **inverse** of $S$. Our *discovery* of $S^{-1}$ in (3.13) was aided by our solution of $Sx = f$ for general, variable, $f$. There is a systematic alternative to our process that works by simultaneously applying Gaussian Elimination to several "representative" $f$ vectors. More precisely, the **Gauss-Jordan method** computes the inverse of $S$ by augmenting it with the identity matrix, e.g.,

$$\left(\begin{array}{ccc|ccc} 2 & -1 & 0 & 1 & 0 & 0 \\ -1 & 2 & -1 & 0 & 1 & 0 \\ 0 & -1 & 2 & 0 & 0 & 1 \end{array}\right)$$

and then applying elementary row operations until $S$ has been transformed to $I$. In the process, the augmented $I$ will be transformed into to desired $S^{-1}$. This is easier done than said.

We first eliminate down, as in normal Gaussian Elimination, being careful to address each of the 3 $f$ vectors. This produces

$$\left(\begin{array}{ccc|ccc} 2 & -1 & 0 & 1 & 0 & 0 \\ 0 & 3/2 & -1 & 1/2 & 1 & 0 \\ 0 & 0 & 4/3 & 1/3 & 2/3 & 1 \end{array}\right).$$

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

$$\left(\begin{array}{ccc|ccc} 2 & -1 & 0 & 1 & 0 & 0 \\ 0 & 3/2 & 0 & 3/4 & 3/2 & 3/4 \\ 0 & 0 & 4/3 & 1/3 & 2/3 & 1 \end{array}\right)$$

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

$$\left(\begin{array}{ccc|ccc} 2 & 0 & 0 & 3/2 & 1 & 1/2 \\ 0 & 3/2 & 0 & 3/4 & 3/2 & 3/4 \\ 0 & 0 & 4/3 & 1/3 & 2/3 & 1 \end{array}\right)$$

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

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 3/4 & 1/2 & 1/4 \\ 0 & 1 & 0 & 1/2 & 1 & 1/2 \\ 0 & 0 & 1 & 1/4 & 1/2 & 3/4 \end{array}\right).$$

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.

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 (3.6),

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

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 analog of (3.7) 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 $E_2E_2^{-1} = I$. Now we may express the reduction of $S$ to $U$ (recall (3.8)) as

$$E_2E_1S = U \tag{3.14}$$

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 $S$. Do you agree that $S^{-1} = U^{-1}L^{-1}$?

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

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

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.

You may also be wondering if Gaussian Elimination always works so well. We first consider an example of trouble from which we can recover. If

$$B = \begin{pmatrix} 2 & 2 & 2 \\ 4 & 4 & 0 \\ 0 & 5 & 3 \end{pmatrix} \tag{3.15}$$

then elimination in column one brings

$$EB = \begin{pmatrix} 2 & 2 & 2 \\ 0 & 0 & -4 \\ 0 & 5 & 3 \end{pmatrix}.$$

The zero in the (2,2) position seems to defeat our simple implementation of Gaussian Elimination. A little thought brings two alternatives to eliminating the pesky 5. If we use row 1, instead of row 2, to eliminate the 5 then we will destroy the good work we did in column 1 in getting to $EB$. A better idea is to simply swap rows 2 and 3 in $EB$. This is a perfectly fine thing to do – for rows correspond to equations and the "order" in which the equations appear have no bearing on their solution. This row swap may itself be achieved by multiplication by the **elementary permutation** matrix

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}. \tag{3.16}$$

In particular

$$U = PEB = \begin{pmatrix} 2 & 2 & 2 \\ 0 & 5 & 3 \\ 0 & 0 & -4 \end{pmatrix}. \tag{3.17}$$

Regarding the associated lower triangular matrix we note that $P$ is its own inverse and so

$$PU = EB \quad \text{and} \quad E^{-1}PU = B.$$

If we define $L \equiv E^{-1}P$ then we agree with the MATLAB `lu` usage that $L$ is a "psychologically lower triangular matrix," i.e., a product of lower triangular and elementary permutation matrices. We could of course construct larger examples that require multiple row swaps. There are however many matrices in which even row swapping won't help.

A more careful look at back substitution reveals that the key to solving $Sx = f$ was the fact that no diagonal element of $U$ vanished. These quantities are so important that we pause to name them.

**Definition** 3.1. The diagonal elements of the upper triangular matrix, $U$, achieved via the application of Gaussian Elimination to $S$ are called the **pivots** of $S$.

If each pivot of $S$ is nonzero then $S$ is said to be **invertible**, or **nonsingular**. If one or more pivots of $S$ is zero then $S$ is said to be **noninvertible**, or **singular**.

Pivots also provide the most concrete setting by which to define and study the determinant. In what follows we define the determinant, by fiat, for two special classes of matrices and then use pivots to extend the definition to all square matrices. The special classes are **triangular** matrices, i.e, matrices whose elements are all zero either above or below the diagonal, and so called **elementary permutations**, i.e., matrices achieved by exchanging two rows in the identity matrix, as in (3.16).

**Definition** 3.2. If $A$ is square we denote the **determinant** of $A$ by $\det(A)$.
(i) If $A$ is triangular then $\det(A)$ is the product of its diagonal elements.
(ii) If $A$ is an elementary permutation of the identity then $\det(A) = -1$.
(iii) The **determinant** of an arbitrary square matrix $A$ is $(-1)^m$ times the product of the pivots of $A$, where $m$ is the number of requisite **row swaps**.

Looking back over our two examples we recognize that $\det(S) = 4$ for the $S$ in (3.5) and $\det(B) = 40$ for the $B$ in (3.15).

Finally, we ask what the lovely formulas, (3.14) and (3.17), tell us about the determinants of products. More precisely, as elimination of $S$ required no row swaps, from

$$\det(S) = \det(E_2E_1S) = \det(U) \quad \text{and} \quad \det(E_1) = \det(E_2) = 1$$

we infer that

$$\det(ES) = \det(E)\det(S) \tag{3.18}$$

for arbitrary $S$ so long as $E$ is an elementary elimination matrix. While, as elimination of $B$ required one row swap we infer from

$$\det(B) = -\det(U) = -\det(PEB) \quad \text{and} \quad \det(P) = -1 \tag{3.19}$$

that

$$\det(PB) = \det(P)\det(B) \tag{3.20}$$

for arbitrary $B$ so long as $P$ is an elementary permutation matrix. Hence, as the LU decomposition guarantees that every matrix is the product of elementary matrices it follows from (3.19) and (3.20) that

$$\boxed{\det(AB) = \det(A)\det(B)} \tag{3.21}$$

for every $A$ and $B$ in $\mathbb{R}^{n\times n}$.

### 3.3. Planar Network Examples

We move from uni-axial to biaxial elastic networks by first considering the frame in Figure 3.3.

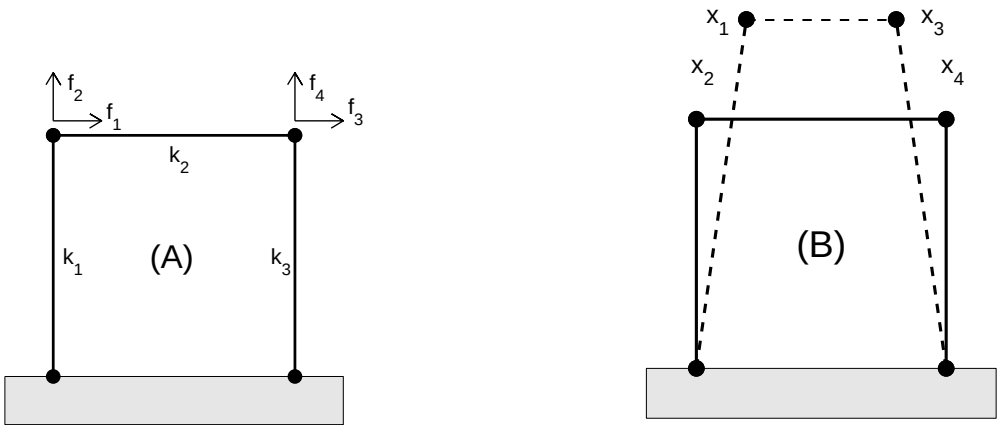



**Figure** 3.3. Deformation of a frame. In (A) we join 3 elastic members, with stiffnesses $k_1$, $k_2$ and $k_3$, at two joints, or nodes, and fix the other two ends to a foundation. Each (nonfixed) node is subject to a planar force, with components $(f_1, f_2)$ and $(f_3, f_4)$. On application of a particular force the frame is displaced as in (B). The respective components of displacement are $(x_1, x_2)$ and $(x_3, x_4)$.

Our first step, as in the previous section, is to express the elongation of each fiber in terms of the displacements of its ends. Begininng with fiber 1 in Figure 3.3(A), we suppose that it meets the foundation at position $(0, 0)$ and that, when at rest, its other end lies at $(0, L_1)$. When forced, as in Figure 3.3(B), the ends of the deformed fiber now lie at $(0, 0)$ and $(x_1, L_1 + x_2)$. As the elongation is simply the deformed length minus the undeformed length we find

$$e_1 = \sqrt{x_1^2 + (L_1 + x_2)^2} - L_1. \tag{3.22}$$

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 fiber lengths then we may effectively ignore the nonlinear contribution in (3.22). 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 that the remainder is of order $t^2$. 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

$$(x_1^2 + x_2^2)/(2L_1) \quad \text{is small compared to} \quad x_2 \tag{3.23}$$

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 fiber, 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 fiber 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. Hence, the restoring force in each fiber remains proportional to its elongation, i.e., $y_j = k_j e_j$ where $k_j$ is the stiffness of the $j$th spring. In matrix terms,

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

As in the uni-axial case, as positive $y_j$ pulls on its ends we find the free body diagrams of Figure 3.4.

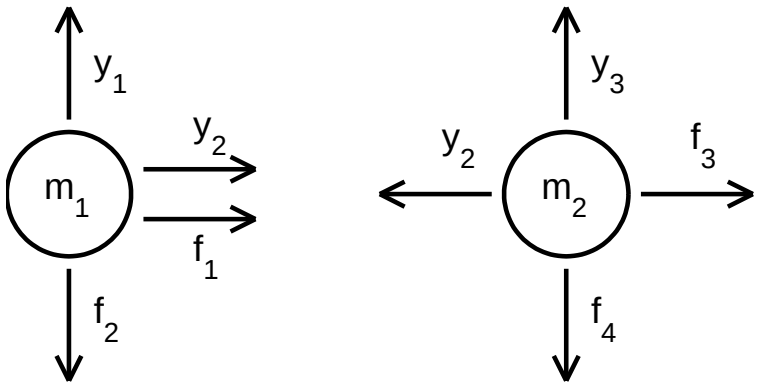



**Figure** 3.4. Free body diagrams for masses in Figure 3.3.

Balancing horizontal and vertical forces at $m_1$ then brings

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

while balancing horizontal and vertical forces at $m_2$ brings

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

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 frame. 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 structure. 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 4 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

$$z \equiv \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix} \tag{3.24}$$

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

$$Az = 0. \tag{3.25}$$

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 3.3. 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 4 that these modes are captured by the *null space* of $A$. For now we will deem our system **stable** if $z = 0$ is the only solution to (3.25).

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 (3.24). 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.

We close this section with the (scalable) example of the larger planar net in Figure 3.5. 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 up and right are positive we note that the elongation of fiber 1 is

$$e_1 = x_8 - x_2$$

while that of fiber 3 is

$$e_3 = x_3 - x_1.$$

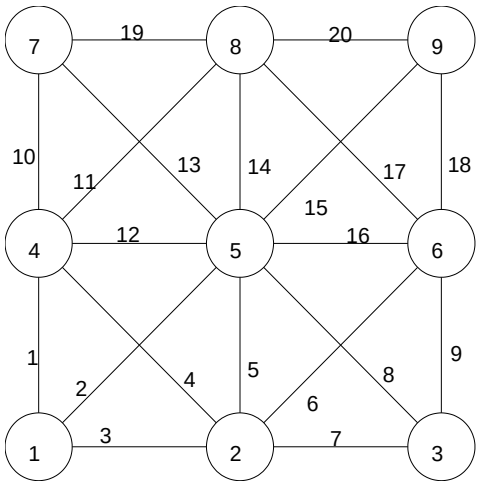



**Figure** 3.5. A crude tissue model.

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

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

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

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

These are both direct applications of the general formula

$$e_j = (x_{2n-1} - x_{2m-1})\cos(\theta_j) + (x_{2n} - x_{2m})\sin(\theta_j) \tag{3.26}$$

for fiber $j$, as depicted in Figure 3.6, connecting node $m$ to node $n$ and making the angle $\theta_j$ 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 one's 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 $x_{10} > x_2$. Does this jibe with (3.26)?

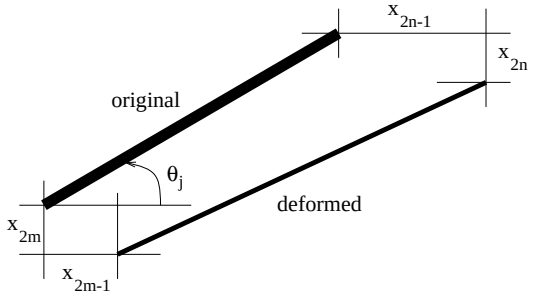



**Figure** 3.6. Elongation of a generic bar, see (3.26).

Applying (3.26) 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 `skin.m`. It begins with a matrix of raw data that anyone with a protractor could have keyed in directly from Figure 3.5. 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 (3.26) requires in order to know which columns of $A$ receive the proper cos or sin values. The nonzero structure of the final $A$ matrix is displayed in the Figure 3.7(A).

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 $A^T y = f$ and so one must solve $Sx = f$ where $S = A^T K A$. 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 5 on the Fundamental Theorem of Linear Algebra.

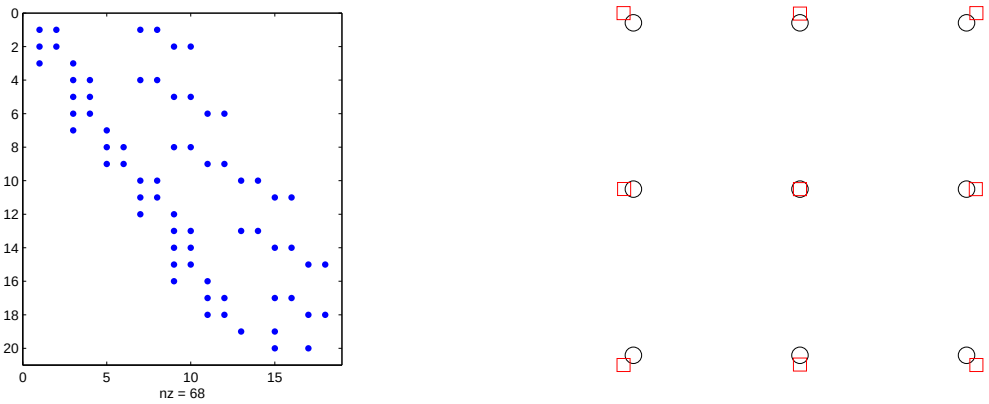



**Figure** 3.7. (A) The nonzero structure of the incidence matrix for the network of Figure 3.5. (B) The black circles lie at the centers of the nodes when the network is unloaded. The red squares mark the node centers after loading by the uniform traction force, $f$, in (3.27). (`skin.m`)

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

$$f = [-s \;\; -s \; 0 \;\; -1 \; s \;\; -s \;\; -1 \; 0 \; 0 \; 0 \; 1 \; 0 \;\; -s \; s \; 0 \; 1 \; s \; s]^T, \tag{3.27}$$

where $s = 1/\sqrt{2}$, we arrive at $x$ via `x=pinv(S)*f` and refer to Figure 3.7 for its graphical representation.

### 3.4. Equilibrium and Energy Minimization*

Given a stable mechanical system with stiffness matrix $S = A^T K A \in \mathbb{R}^{n\times n}$ and a load vector $f \in \mathbb{R}^n$ we rate **candidates** $u \in \mathbb{R}^n$ for its displacement based on their associated total potential energy. Where

$$\begin{aligned} \text{Total Potential Energy} &\equiv \text{Internal Strain Energy} - \text{Work Done by Load} \\ &= \tfrac{1}{2} u^T S u - u^T f. \end{aligned}$$

The resulting minimum principle hinges on two key properties of $S = A^T K A$, inherited from the physical fact that $K$ is a diagonal matrix with positive numbers on its diagonal. The first is that it is symmetric, for $S^T = (A^T K A)^T = A^T K^T A = S$, and the second is that it is **positive definite**, i.e.,

$$v^T S v = v^T A^T K A v = (Av)^T K (Av) = \sum_{j=1}^{n} k_j (Av)_j^2 > 0, \quad \forall\, v \in \mathbb{R}^n, \quad v \neq 0. \tag{3.28}$$

The inequality stems from the fact that each stiffness, $k_j > 0$, and as $A$ is stable that $Av \neq 0$. This also helps us see why $\frac{1}{2} v^T S v$ is identified as the Internal Strain Energy. For if $v$ is the candidate displacement then $e = Av$ is the associated elongation, or **strain**. The associated internal force is $y = Ke$ and hence $k_j (Av)_j^2/2 = e_j y_j/2$ is strain energy stored in the $j$th fiber.

**Proposition** 3.3. The candidate with the least total potential energy is precisely the equilibrium solution $x = S^{-1}f$. In other words

$$\boxed{-\tfrac{1}{2}f^T S^{-1} f = \min_{u\in\mathbb{R}^n} \tfrac{1}{2}u^T S u - u^T f.} \tag{3.29}$$

Proof: Suppose that $Sx = f$. Now for each $u \in \mathbb{R}^n$,

$$\begin{aligned}(\tfrac{1}{2}u^T S u - u^T f) - (\tfrac{1}{2}x^T S x - x^T f) &= \tfrac{1}{2}x^T S x - u^T S x + \tfrac{1}{2}u^T S u \\ &= \tfrac{1}{2}(x-u)^T S (x-u) \geq 0,\end{aligned} \tag{3.30}$$

where the last equality uses $S = S^T$ and the last inequality uses (3.28). It follows directly from (3.30) that $x = S^{-1}f$ indeed minimizes the total potential energy. On substitution of this $x$ into

$$\tfrac{1}{2}x^T S x - x^T f$$

we arrive at the left side of (3.29). End of Proof.

In addition to confirming our belief that equilibrium solutions should use less energy than other candidates, this principle can also be used to estimate important physical quantities *without ever having to solve* $Sx = f$. To see this, note from (3.30) that, $x^T f$, the actual work done by the load, obeys

$$x^T f = x^T S x \geq 2u^T f - u^T S u \quad \forall\, u \in \mathbb{R}^n. \tag{3.31}$$

The key point is that we are free to try *any* candidate on the right hand side. Each choice will provide a lower bound on the true work done. There are trivial choices, e.g., $u = 0$ informs is that $x^T f \geq 0$, and nontrivial choices, e.g., $u = f$ informs us that

$$x^T f \geq f^T(2I - A^T K A)f.$$

This inequality, in the context of our small example, (3.5), yields

$$x^T f \geq \begin{pmatrix} f_1 & f_2 & f_3 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} = \begin{pmatrix} f_1 & f_2 & f_3 \end{pmatrix} \begin{pmatrix} f_2 \\ f_1 + f_3 \\ f_2 \end{pmatrix} = 2f_2(f_1 + f_3).$$

In the constant load case, $f_1 = f_2 = f_3$, this reveals that the total displacement, $x_1 + x_2 + x_3$, exceeds $4f_1$.

Although developed (here) as a principle of mechanics this proposition has found use in many areas of physical equilibrium. We will also have occasion to invoke it as an analytical tool. Toward that end it seems best to formulate it in a general setting – and in a way too that removes the perhaps annoying $-1/2$ factor on the left side of (3.29).

**Proposition** 3.4 If $B \in \mathbb{R}^{n\times n}$ is symmetric and positive definite and $f \in \mathbb{R}^n$ then

$$f^T B^{-1} f = \max_{x\in\mathbb{R}^n} 2x^T f - x^T B x,$$

and the maximum is attained at that $x$ for which $Bx = f$.

Proof: This is a simple rearrangement of (3.29). In particular, note that

$$\max_{x\in\mathbb{R}^n}\{2x^T f - x^T B x\} = \max_{x\in\mathbb{R}^n}\{-2(\tfrac{1}{2}x^T B x - x^T f)\} = -2\min_{x\in\mathbb{R}^n}\{\tfrac{1}{2}x^T B x - x^T f\}.$$

End of Proof.

### 3.5. Notes and Exercises

As in the previous chapter, the Strang Quartet, is drawn from Strang (2007). That text is also an excellent source for a deeper investigation of the LU and Cholesky factorizations and Energy Minimization.

1. Deduce from $e = Ax$ and $A^T y = f$ that the work done by the load is precisely the work done by the springs. That is, show that
$$x^T f = y^T e.$$

2. With regard to Figure 3.1, (i) Derive the $A$ and $K$ matrices resulting from the removal of the fourth spring (but not the third mass) and assemble $S = A^T K A$.

   (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 $k_1 = k_2 = k_3 = k$,

   (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 $Ux = c$.

3. With regard to Figure 3.3

   (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 $S = A^T K A$.

   (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 $Ux = c$.

4. Prove that if $A$ and $B$ are invertible then $(AB)^{-1} = B^{-1} A^{-1}$.

5. Show that if $P$ is an elementary permutation of $I$ then $PP = I$ and use this to arrive at $P^{-1}$.

6. Both elimnation matrices and elementary permutation matrices can be inverted with ease. One more class of such matrices are those that are $A + uv^T$ where $A$ is invertible. Please show that
$$(A + uv^T)^{-1} = A^{-1} + \frac{1}{1 + v^T A^{-1} u} A^{-1} u v^T A^{-1}. \tag{3.32}$$

7. Note that $A$ is invertible then $AA^{-1} = I$. Use (3.21) to show that $\det(A^{-1}) = 1/\det(A)$.

8. (a) Compute the product of the pivots of
$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad a \neq 0,$$
   and compare your answer with Exer. 1.6 and discuss the relation of determinant to area.

(b) Compute the product of the pivots of

$$\begin{pmatrix} u_1 & v_1 & w_1 \\ u_2 & v_2 & w_2 \\ u_3 & v_3 & w_3 \end{pmatrix} \quad u_1 \neq 0, \quad u_1 v_2 \neq v_1 u_2$$

and compare your answer with $|w^T(u \times v)|$ (recall Exer. 1.21(h)) and discuss the relation of determinant to volume.

(c) Argue that if $X$ is a nice set in $\mathbb{R}^3$ (in the sense that for each $x \in X$ there is a cube centered at $x$ that also lies in $X$) and $A$ is a 3-by-3 matrix then

$$\frac{\text{volume}(AX)}{\text{volume}(X)} = |\det(A)|, \tag{3.33}$$

where $AX = \{Ax : x \in X\}$.

9. Show that

$$A = \frac{1}{2} \begin{pmatrix} -2 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 1 & -\sqrt{3} & -1 & \sqrt{3} \\ -1 & -\sqrt{3} & 0 & 0 & 1 & \sqrt{3} \end{pmatrix} \tag{3.34}$$

is the geometric incidence matrix for the equilateral triangle of Figure 3.8.

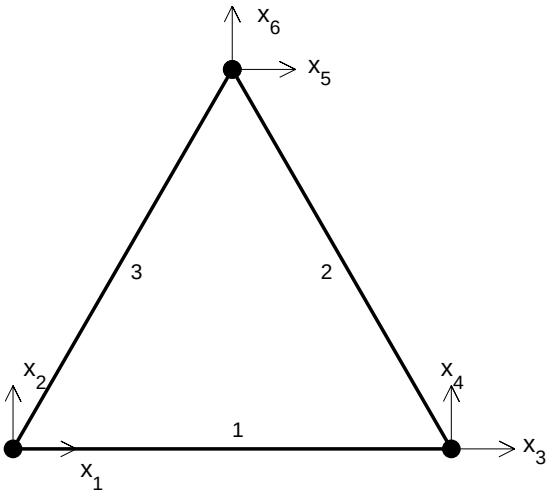



**Figure** 3.8. An equilateral triangle with labeled edges and degrees of freedom.

Show that the vectors

$$(1,0,1,0,1,0)^T, \quad (0,1,0,1,0,1)^T \quad \text{and} \quad (-1,\sqrt{3},-1,-\sqrt{3},2,0)^T$$

are displacements that do no work. Please draw them (i.e., draw the original triangle and its displacement) and explain (using phrases like "translation in direction ... by ..." and/or "rotation about ... by ...") what these displacements signify.

10. Generalize Figure 3.5 to the case of 16 nodes connected by 42 fibers by modifying `skin.m`. 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 3.7, from which you conclude the presence of a stiff fiber.

11. Generalize Figure 3.5 to permit ever finer meshes. In particular, with reference to Figure 3.9 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 $S = A^T K A$. 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 (3.1). 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 §3.1 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 as in Figure 3.10

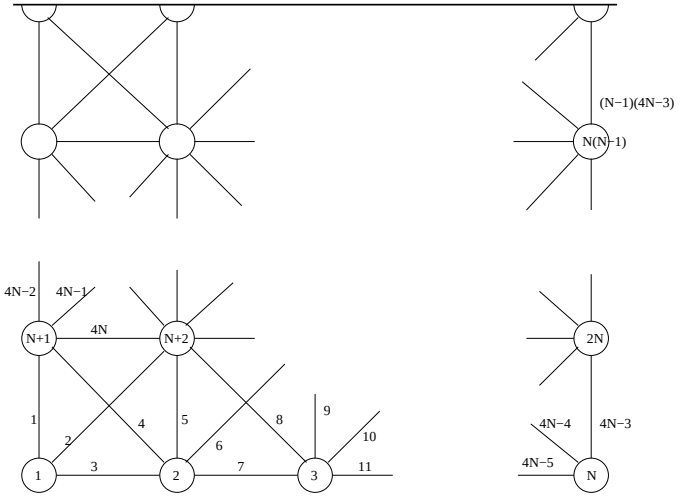



**Figure** 3.9. A fine anchored fiber network.

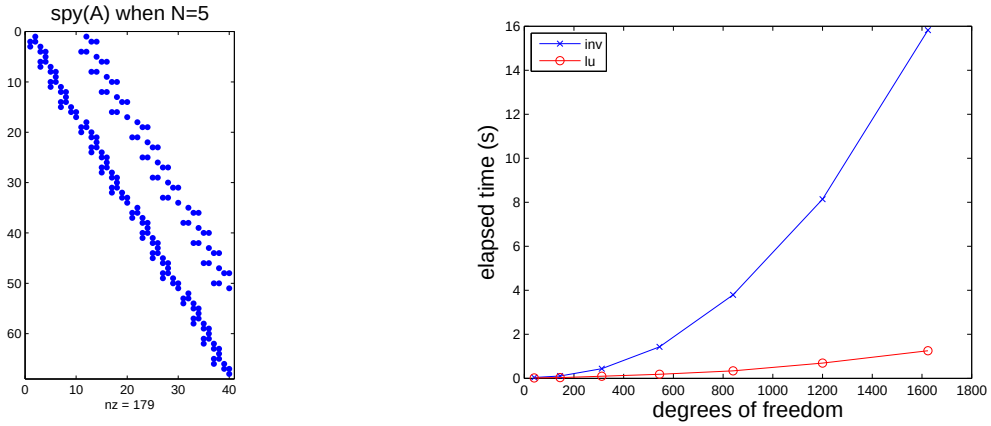



**Figure** 3.10. (A) The nonzeros of the incidence matrix of the fine net of Figure 3.9 with 5

levels. (B) Comparison of solution times for large fine nets.

**Submit** your well documented code, a spy plot of $S$ when $N = 9$, and a time comparison like (will vary with memory and cpu) Figure 3.10.

12. We consider the methane molecule of Figure 3.11. The carbon atom is at $(0, 0, 0)$ and hydrogen atoms are at

$$\begin{aligned} d &= (1,1,1) = (\sqrt{3}\cos\theta_d\cos\phi_d, \sqrt{3}\cos\theta_d\sin\phi_d, \sqrt{3}\sin\theta_d) \\ c &= (-1,-1,1) = (\sqrt{3}\cos\theta_c\cos\phi_c, \sqrt{3}\cos\theta_c\sin\phi_c, \sqrt{3}\sin\theta_c) \\ b &= (1,-1,-1) = (\sqrt{3}\cos\theta_b\cos\phi_b, \sqrt{3}\cos\theta_b\sin\phi_b, \sqrt{3}\sin\theta_b) \\ a &= (-1,1,-1) = (\sqrt{3}\cos\theta_a\cos\phi_a, \sqrt{3}\cos\theta_a\sin\phi_a, \sqrt{3}\sin\theta_a), \end{aligned}$$

where the angles are as illustrated in Figure 3.11. We can use (1.5) to compute their cosines. For example, with reference to Figure 3.11(B), as $d = (1,1,1)$ and $Pd = (1,1,0)$ and the dashed red vector is $e_1 = (1,0,0)$ we find

$$\cos(\theta_d) = \frac{d^T Pd}{\|d\|\|Pd\|} = \frac{2}{\sqrt{3}\sqrt{2}}. \quad \text{and} \quad \cos(\phi_d) = \frac{e_1^T Pd}{\|e_1\|\|Pd\|} = \frac{1}{\sqrt{2}}.$$

The respective sines are then completed by Pythagoras. Following this reasoning please derive

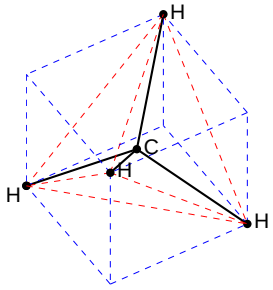
$$\begin{aligned} \cos\phi_d = \sin\phi_d = 1/\sqrt{2}, &\quad \sin\theta_d = 1/\sqrt{3}, \quad \cos\theta_d = \sqrt{2/3} \\ \cos\phi_c = \sin\phi_c = -1/\sqrt{2}, &\quad \sin\theta_c = 1/\sqrt{3}, \quad \cos\theta_c = \sqrt{2/3} \\ \cos\phi_b = -\sin\phi_b = 1/\sqrt{2}, &\quad \sin\theta_b = -1/\sqrt{3}, \quad \cos\theta_b = \sqrt{2/3} \\ -\cos\phi_a = \sin\phi_a = 1/\sqrt{2}, &\quad \sin\theta_a = -1/\sqrt{3}, \quad \cos\theta_a = \sqrt{2/3}. \end{aligned}$$

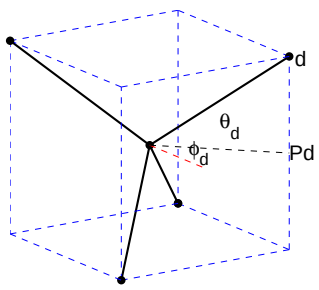
If the displacement of the carbon atom is $(x_1, x_2, x_3)$ and the displacement of the $d$ hydrogen atom is $(x_4, x_5, x_6)$ the elongation of the respective CH bond is

$$e_d = (x_4 - x_1)\cos\theta_d\cos\phi_d + (x_5 - x_2)\cos\theta_d\sin\phi_d + (x_6 - x_3)\sin\theta_d.$$

Apply this reasoning to find the elongation of the other bonds and so arrive at the following 4-by-15 incidence matrix

$$A = \frac{1}{\sqrt{3}}\begin{pmatrix} 1 & -1 & 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 1 & 0 & 0 & 0 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 1 & 0 & 0 & 0 \\ -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \tag{3.35}$$



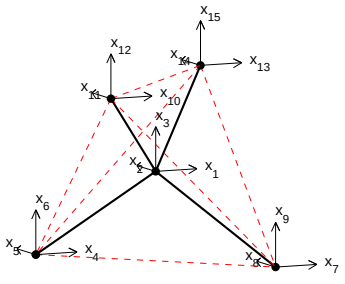



**Figure** 3.11 (A) A methane, $CH_4$, molecule. The 4 hydrogen atoms sit at the vertices of the red tetrahedron. In order to establish coordinates we have inscribed the tetrahedron in a blue box. (B) An illustration of the two angles needed to describe the hydrogen atom at point $d = (1, 1, 1)$. The red dashed line is the positive $x$-axis. The dashed black line is the projection of $d$ onto the $(x, y)$ plane. The angle $\phi_d$ is the angle between these dashed lines. The angle $\theta_d$ is the angle between the dashed black line and the $d$ vector. (C) Labeling the three degrees of freedom of each atom.

13. The minimum energy principal of §3.4 provides a means to define, and study, a scalar measure of network strength or, rather, weakness. Namely, given a load $f$ and associated displacement $x$ we define the **compliance** of the network to be simply $x^T f$, i.e., the work done by the load. In designing networks we naturally choose fiber stiffnesses to lessen the compliance, i.e., to strengthen the network. Toward that end, for fixed incidence matrix $A$ and load $f$ we denote by $C(k)$ the compliance of the network with stiffnesses $k = (k_1, \ldots, k_m)$.

    (a) Use Prop. 3.4 to show that

    $$C(k) = \max_{v \in \mathbb{R}^n} \; 2v^T f - v^T A^T K A v.$$

    (b) Use (a) to show that if $\kappa = (\kappa_1, \ldots, \kappa_m)$ is a stiffness vector for which $\kappa_j \geq k_j$ for each $j$ then $C(\kappa) \leq C(k)$. That is, stiffening each of the fibers lessens the work done by the load. Hint: (3.28).

    (c) Show that the compliance is a **convex function** of $k$. That is, show that

    $$C(tk + (1-t)\kappa) \leq tC(k) + (1-t)C(\kappa) \tag{3.36}$$

    for all $0 \leq t \leq 1$ and all $k$ and $\kappa$. Hint: First show that the max of a sum can not exceed the sum of the max's. That is,

    $$\max_v \left(\alpha(v) + \beta(v)\right) \leq \max_v \alpha(v) + \max_u \beta(u).$$

    (d) Show that if $\|f\| = 1$ and $A^T K A f = \lambda f$ for some $\lambda \in \mathbb{R}$ then $C(k) \geq 1/\lambda$. Hint: Choose $u = \alpha f$ and maximize over $\alpha \in \mathbb{R}$.

# 4. The Column and Null Spaces

The previous chapter revealed pivots to be the crucial determinants of the solvability of linear systems of equilibrium equations. In this and the following chapter we consider the general problem $Sx = f$ for arbitrary $S \in \mathbb{R}^{m\times n}$ and $f \in \mathbb{R}^m$. Pivots will remain pivotal here, bringing us to the important notions of linear independence, basis, rank and dimension. We apply these notions both to determination of stability of mechanical networks and to the detailed structure of nilpotent matrices.

## 4.1. The Column Space

We begin with the direct geometric interpretation of matrix–vector multiplication. Recalling (1.11), the multiplication of the vector $x \in \mathbb{R}^n$ by the matrix $S \in \mathbb{R}^{m\times n}$ produces a linear combination of the columns of $S$,

$$Sx = [S(:,1)\ S(:,2)\ \cdots\ S(:,n)] \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = x_1 S(:,1) + x_2 S(:,2) + \cdots + x_n S(:,n). \tag{4.1}$$

The collection of all such linear combinations is know as the **span of the columns** of $S$, or, equivalently, the **range** of $S$ or, equivalently the **column space** of $S$. Its formal definition is

**Definition** 4.1. The **column space** of the matrix $S \in \mathbb{R}^{m\times n}$ is the span of its columns, i.e.,

$$\mathcal{R}(S) \equiv \{Sx : x \in \mathbb{R}^n\}. \tag{4.2}$$

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

Hopefully our opening chapter has prepared you to parse the set notation used in (4.2). The braces, $\{\}$ denote **set** and the colon denotes **such that, for which, or where**. Hence, an English translation of $\{Sx : x \in \mathbb{R}^n\}$ would be "the set of all products of the form $Sx$ where $x$ lies in $\mathbb{R}^n$." But lets not over analyze, we learn by doing.

The column space of the single column

$$S = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

is the line in the plane through the point $(1, 1)$, while the column space of

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

is the entire plane, i.e., all of $\mathbb{R}^2$. Can you "see" how each vector in the plane can be written as a linear combination (weighted sum) of these two columns? We are early in the chapter and so wish to build intuition and confidence so that when we venture into higher dimensions your vertigo is balanced by your sense of wonder.

For example, the column space of the $S$ matrix associated with the frame in Figure 3.3 is, by definition,

$$\mathcal{R}(S) = \left\{ x_1 \begin{pmatrix} 1 \\ 0 \\ -1 \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} -1 \\ 0 \\ 1 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} : x \in \mathbb{R}^4 \right\}.$$

And now, although you can not fully visualize this set you can see that the first and third columns are colinear, i.e., lie on the same line. As a result we can get by with the more compact description

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

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 real 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** 4.2. A subset $M$ of $\mathbb{R}^n$ is a **subspace** of $\mathbb{R}^n$ when
$(\mathcal{S}_1)$ $p + q \in M$ whenever $p \in M$ and $q \in M$, and
$(\mathcal{S}_2)$ $tp \in M$ whenever $p \in M$ and $t \in \mathbb{R}$.

Let us confirm now that the column space, $\mathcal{R}(S)$, is indeed a subspace. Regarding
$(\mathcal{S}_1)$ 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., $(p+q) \in \mathcal{R}(S)$. Regarding
$(\mathcal{S}_2)$, $tp = tSx = S(tx)$ so $tp \in \mathcal{R}(S)$.

Note that we used only the definition of $\mathcal{R}(S)$ and did not make mention of any particular $S$.

To show that something is not a subspace it suffices to produce one instance that violates one of the two conditions. For example, to prove that the circle

$$C = \{x \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1\}$$

is not a subspace we note that $(1,0) \in C$ and $(0,1) \in C$ while their sum $(1,1) \notin C$ and so $(\mathcal{S}_1)$ is violated. We could, for good measure, violate condition $(\mathcal{S}_2)$ by noting that $2(1,0) \notin C$.

### 4.2. The Null Space

If the product of two real numbers is zero then we know that one of them must be zero. This inference is false in higher dimensions. For example

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

Given a matrix $S$, we will see that it pays to keep track of those vectors that $S$ annihilates.

**Definition** 4.3. The **null space** of $S \in \mathbb{R}^{m\times n}$ 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 $S(x + y) = 0$, i.e., $x + y \in \mathcal{N}(S)$. 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 3.3 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 addresses 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 (4.1) 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** 4.4. 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** 4.5. 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
($\mathcal{B}_1$) $M = \mathcal{R}(S)$, and
($\mathcal{B}_2$) $\mathcal{N}(S) = \{0\}$.

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

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

comprise a basis for $\mathbb{R}^2$, while the columns of neither

$$\begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix} \quad \text{nor} \quad \begin{pmatrix} 1 & 2 & 3 \\ 3 & 6 & 9 \end{pmatrix}$$

comprise bases for $\mathbb{R}^2$.

### 4.3. Pivots, Rank and Dimension

To appreciate the importance of the notions of the past two sections it probably best to move beyond its "textbook" examples. To begin, we compute bases for the null and column spaces of the incidence matrix associated with the ladder in Figure 4.1

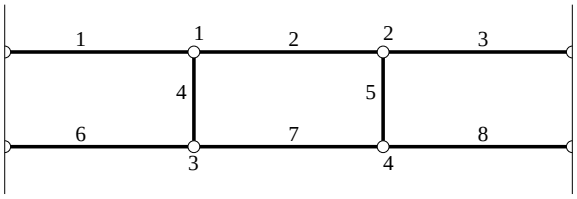



**Figure** 4.1. An unstable ladder?

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 incidence 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.

Each nonzero row of $A_{\text{red}}$ is called a **pivot row**. The first nonzero in each row of $A_{\text{red}}$ is called a **pivot**. Each column that contains a pivot is called a **pivot column**. On account of the staircase nature of $A_{\text{red}}$ 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 $A_{\text{red}}$. One now asks how this might help us distinguish the independent columns of $A$. For, although the rows of $A_{\text{red}}$ 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** 4.6. Suppose $A \in \mathbb{R}^{m\times n}$ If the pivot columns of $A_{\rm red}$ are columns $c_1, c_2, \ldots, c_r$ then columns $c_1, c_2, \ldots, c_r$ of $A$ constitute a basis for $\mathcal{R}(A)$.

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

$$x_k = 0, \quad k \notin \{c_j : j = 1, \ldots, r\}. \tag{4.3}$$

Now $Ax = 0$ necessarily implies that $A_{\rm red}x = 0$, contrary to the fact that the $\{A_{\rm red}(:, c_j) : j = 1, \ldots, r\}$ are the pivot columns of $A_{\rm red}$. (The implication $Ax = 0 \Rightarrow A_{\rm red}x = 0$ 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 $TA = A_{\rm red}$ and you see why $Ax = 0 \Rightarrow A_{\rm red}x = 0$. 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 $\{A(:, c_j) : j = 1, \ldots, r\}$ is indeed $\mathcal{R}(A)$. This is obvious if $r = n$, i.e., if *all* of the columns are linearly independent. If $r < n$ there exists a $q \notin \{c_j : j = 1, \ldots, r\}$. Looking back at $A_{\rm red}$ we note that its $q$th column is a linear combination of the pivot columns with indices not exceeding $q$. Hence, there exists an $x$ satisfying (4.3) and $A_{\rm red}x = 0$ and $x_q = 1$. This $x$ then necessarily satisfies $Ax = 0$. This states that the $q$th column of $A$ is a linear combination of those in $\{A(:, c_j) : j = 1, \ldots, r\}$. End of Proof.

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

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

and **free variables**

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

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

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

One solves $A_{\rm red}x = 0$ 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,

$$x = x_6 \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_8 \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \tag{4.4}$$

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 (4.4) 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** 4.7. Suppose that $A \in \mathbb{R}^{m\times n}$ has 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 we 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 (4.4) 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 4.1? I hope not, for vertical motion of bar 4 must "stretch" bars 1,2,6 and 7. This apparent contradiction is resolved by reflecting on the "linearization" performed in (3.23).

To secure your understanding we close this section with a few more (modest size) 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. Prop. 4.6 then informs us that the first two columns of $A$, namely

$$\left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\} \tag{4.5}$$

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

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

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

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 (4.5) 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

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

constitutes a basis for $\mathcal{N}(B)$. From these examples we may abstract one useful generality. If $m < n$ then there will always be at least one free variable. As a result,

**Proposition** 4.8. If $A \in \mathbb{R}^{m \times n}$ and $m < n$ then there exists a nonzero $x \in \mathbb{R}^n$ for which $Ax = 0$. In other words, any collection of more than $m$ vectors in $\mathbb{R}^m$ will be linearly dependent.

The number of pivots, $r$, of $A \in \mathbb{R}^{m \times n}$ 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. This of course presumes that every basis is the same size. Let us confirm that.

**Proposition** 4.9. If $M$ is a subspace of $\mathbb{R}^n$ and both $X = \{x_1, \ldots, x_p\}$ and $Y = \{y_1, \ldots, y_q\}$ are bases for $M$ then $p = q$.

Proof: We establish the contrapositive by supposing $p \neq q$, or without loss, $p < q$ and then arguing that the $y_j$ are not linearly independent. As each $y_j$ is in the span of $X$ it follows that $y_j = Xa_j$ for some nonzero $a_j \in \mathbb{R}^p$. We collect these in $A \in \mathbb{R}^{p \times q}$. As $q > p$ we may infer from the previous

proposition the existence of a nonzero $z \in \mathbb{R}^q$ for which $Az = 0$. Now $Yz = XAz = 0$ informs us that the columns of $Y$ are linearly dependent. End of Proof.

The results of this section now permit us concise answers to the two questions with which we began the chapter; when can we solve $Ax = b$ and when is the solution unique? The most concise answers relate the rank, $r$, of $A$ to its number of rows, $m$, and its number of columns, $n$. Namely,

**Existence:** If $r = m$ then $\mathcal{R}(A) = \mathbb{R}^m$ and so $Ax = b$ has a solution $x$ for every $b \in \mathbb{R}^m$.
**Uniqueness:** If $r = n$ then $\mathcal{N}(A) = \{0\}$ and so $Ax = b$ has at most one solution.

We note that if $m = n = r$ the matrix is invertible and $Ax = b$ has the solution $x = A^{-1}b$.

### 4.4. The Structure of Nilpotent Matrices*

A square matrix $A \in \mathbb{R}^n$ is said to be **nilpotent** when $A^m = 0$ while $A^{m-1} \neq 0$ for some integer $m > 1$. For example,

$$A = \begin{pmatrix} 2 & 2 & -2 \\ 5 & 1 & -3 \\ 1 & 5 & -3 \end{pmatrix} \tag{4.6}$$

obeys $A^3 = 0$ while $A^2 \neq 0$. Nilpotent matrices arise naturally when studying the eigenvalue problem. In fact, as we will see in Chapter 11, they are the only obstruction to a clean eigen–decomposition. We study them here because they may be fully decomposed by their growing sequence of null spaces

$$\mathcal{N}(A) \subset \mathcal{N}(A^2) \subset \cdots \subset \mathcal{N}(A^{m-1}) \subset \mathcal{N}(A^m) = \mathbb{R}^n. \tag{4.7}$$

In fact, we will show that the structure of a nilpotent matrix is completely encoded in the sequence of dimensions,

$$d_j \equiv \dim \mathcal{N}(A^j), \quad \text{for} \quad j = 1, \ldots, m.$$

Our first interesting observation is that these dimensions can not grow too quickly. More precisely,

$$d_j - d_{j-1} \leq d_1, \quad \text{for} \quad j = 2, \ldots, m. \tag{4.8}$$

We build up to the full decomposition by first putting a nilpotent matrix in strictly (only zeros on the diagonal) upper triangular form. We begin with a basis, $X_1 = \{x_j : j = 1, \ldots, d_1\}$ for $\mathcal{N}(A)$ and then expand this to basis of $\mathcal{N}(A^2)$ by appending $X_2 = \{x_j : j = d_1 + 1, \ldots, d_2\}$. We continue this process until we arrive at a basis $\{x_j : j = 1, \ldots, n\}$ for all of $\mathbb{R}^n$. Regarding the example in (4.6) we find

$$x_1 = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}$$

to be a basis for $\mathcal{N}(A)$ and

$$v_1 = \begin{pmatrix} 2 \\ 3 \\ 3 \end{pmatrix} \quad \text{and} \quad v_2 = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$$

to be a basis for $\mathcal{N}(A^2)$. As $v_1 = 2x_1 + v_2$ we choose $x_2 = v_2$. To find $x_3$ we simply look for anything not in the span of $x_1$ and $x_2$. Note that

$$x_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

will do. Let us now examine the action of $A$ onto our basis vectors. The first is easy, $Ax_1 = 0$. For the second, as $x_2 \in \mathcal{N}(A^2)$ we find that $Ax_2 \in \mathcal{N}(A)$ and hence $Ax_2$ must be a multiple of $x_1$. In fact

$$Ax_2 = 4x_1.$$

Finally, by the same token, $Ax_3 \in \mathcal{N}(A^2)$ and so $Ax_3$ is a linear combination of $x_1$ and $x_2$. In fact

$$Ax_3 = -2x_1 - x_2.$$

If we gather our findings we find

$$A[x_1\ x_2\ x_3] = [0\ 4x_1\ \ -2x_1 - x_2] = [x_1\ x_2\ x_3]U \tag{4.9}$$

where $U$ is the strictly upper triangular matrix

$$U = \begin{pmatrix} 0 & 4 & -2 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix}. \tag{4.10}$$

We write (4.9) as $AX = XU$ and note that by construction the columns of $X$ are linearly independent. As such $X$ is invertible and we may multiply across by $X^{-1}$ and arrive at

$$\boxed{X^{-1}AX = U.} \tag{4.11}$$

It is no more difficult to establish the general case.

**Proposition** 4.10. If $A$ is nilpotent then there exists an invertible matrix $X$ and a strictly upper triangular matrix, $U$ for which (4.11) holds.

Proof: We begin with a basis, $X_1 = \{x_j : j = 1, \ldots, d_1\}$ for $\mathcal{N}(A)$ and then expand this to basis of $\mathcal{N}(A^2)$ by appending $X_2 = \{x_j : j = d_1 + 1, \ldots, d_2\}$. The key observation is that, by construction, for each $x_j \in X_2$ we know that

$$Ax_j \neq 0 \quad \text{while} \quad A(Ax_j) = 0.$$

In other words, $Ax_j$ is a nontrivial linear combination of elements of $\mathcal{N}(A)$. As as result, $Ax_j = 0$ for the $j \leq d_1$ while $Ax_j$ for $d_1 + 1 \leq j \leq d_2$ is a linear combination of the strictly prior $x_j$, i.e., for $j \leq d_1$. Proceeding on, if necessary, the $x_j$ added to complete a basis for $\mathcal{N}(A^3)$ can be expressed as a linear combination of strictly prior $x_j$. End of Proof.

This transformation of $A$ into $U$ by $X$ in (4.11) is called a **similarity transformation** and in this case $A$ and $U$ are said to be **similar**. It follows from (1.14) and (3.21) that if two matrices are similar then they have the same determinant and trace. This then proves,

**Corollary** 4.11. If $A$ is nilpotent then $\det(A) = \mathrm{tr}(A) = 0$.

To reveal the finer structure of nilpotent $A$ we recognize that we have a lot a freedom in how we choose the many bases. The key point is to start at the right end. It also helps to have the right definition.

**Definition** 4.12. Let $U$ and $V$ be subspaces of $\mathbb{R}^n$ and $U \subset V$. We say that the set of vectors $\{v_1, \cdots, v_p\} \subset V$ is **linearly independent mod** $U$ when

$$\sum_{i=1}^{p} a_i v_i \in U \text{ implies that each } a_i = 0.$$

We say that $\{v_1, \cdots, v_p\}$ is a **basis of** $V$ **mod** $U$ when it is linearly independent mod $U$ and $V = U \oplus \text{span}\{v_1, \cdots, v_p\}$

It is easy to find a basis of $V$ mod $U$: simply choose a basis $\{u_1, \ldots, u_k\}$ for $U$ and extend it to a basis $\{u_1, \ldots, u_k, v_1, \ldots, v_p\}$ for $V$. The vectors $\{v_1, \ldots, v_p\}$ are then a basis of $V$ mod $U$.

A **Jordan block** of size $s$ is an $s$-by-$s$ matrix of zeros save the superdiagonal, that is populated by ones. For example, the Jordan blocks of sizes 1, 2 and 3 are

$$J_1 = 0, \quad J_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad J_3 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}. \tag{4.12}$$

We may now establish the fundamental result.

**Proposition** 4.13. If $A^{m-1} \neq 0$ and $A^m = 0$ then there exists an invertible matrix $X$ and a Jordan matrix $J$ for which

$$\boxed{X^{-1}AX = J.} \tag{4.13}$$

$J$ is the block diagonal matrix beginning with $c_1$ Jordan blocks of size 1, $c_2$ Jordan blocks of size 2 up through $c_m$ Jordan blocks of size $m$. The chain numbers, $c_j$, are determined by the null space dimensions, $d_j$, via

$$c = S_m d \quad \text{where} \quad S_m = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix} \tag{4.14}$$

is the stiffness matrix (recall Exer. 3.2) for the hanging elastic chain with $m$ unit stiffnesses.

Proof: We follow the recipe,

**Step 1:** Choose a basis $\{v_m^1, \ldots, v_m^{c_m}\}$ of $\mathcal{N}(A^m)$ mod $\mathcal{N}(A^{m-1})$. There are $c_m = d_m - d_{m-1}$ such basis vectors. Each of these vectors will generate a chain of length $m$ and a Jordan block of size $m$.

**Step 2:** If $m = 1$, stop. Else, apply $A$ to the vectors constructed above, obtaining $Av_m^i \in \mathcal{N}(A^{m-1})$. Then the key point is that

$$\{Av_m^1, \ldots, Av_m^{c_m}\} \quad \text{is linearly independent mod} \quad \mathcal{N}(A^{m-2}).$$

Now extend these (if necessary) to a basis of $\mathcal{N}(A^{m-1})$ mod $\mathcal{N}(A^{m-2})$, by appending $\{v_{m-1}^1, \ldots, v_{m-1}^{c_{m-1}}\}$. Each of these new appended vectors will generate a chain of length $m-1$ and a Jordan block of size $m-1$. Since we have extended a list of size $c_m$ to reach a length $d_{m-1} - d_{m-2}$, there are $c_{m-1} = (d_{m-1} - d_{m-2}) - (d_m - d_{m-1})$ chains of length $m-1$.

**Step 3:** Repeat Step 2, with $m$ replaced by $m-1$.

Once the algorithm terminates, we may arrange the chosen basis as follows:

$$X = \{v_1^1, \ldots, v_1^{c_1}, Av_2^1, v_2^1, \ldots, Av_2^{c_2}, v_2^{c_2}, , \ldots, A^{i-1}v_i^1, \ldots, v_i^1, \ldots, A^{i-1}v_i^{c_i}, , \ldots, v_i^{c_i}, \ldots\}$$

and so $AX = XJ$ where $J$ is the block diagonal matrix beginning with $c_1$ zeros, then $c_2$ blocks of size 2, then $c_3$ blocks of size 3 up through $c_m$ blocks of size $m$. The formulas for $c_m$ and $c_{m-1}$ in Steps 1 and 2 are precisely those implemented by (4.14). End of Proof.

For example,

$$A = \begin{pmatrix} 1 & -1 & 1 \\ 2 & -2 & 2 \\ 1 & -1 & 1 \end{pmatrix}$$

obeys $A^2 = 0$, and has null space dimensions, $d_1 = 2$ and $d_2 = 3$. The proposition then dictates that we will encounter $c_1 = 2d_1 - d_2 = 1$ chain of length 1, and $c_2 = d_2 - d_1 = 1$ chain of length 2. We compute the associated basis vectors according to the recipe above. Following Step 1 we note (any vector not in $\mathcal{N}(A)$ will suffice) that

$$v_2^1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

is a basis of $\mathcal{N}(A^2)$ mod $\mathcal{N}(A)$. Step 2 leads us to

$$Av_2^1 = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$$

which, we are happy to confirm, indeed lies in $\mathcal{N}(A)$. We continue within Step 2 to complete a basis for $\mathcal{N}(A)$ by appending

$$v_1^1 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}.$$

This completes all of the steps and so we assemble

$$X = [v_1^1 \; Av_2^1 \; v_2^1]$$

and develop

$$AX = [0 \; 0 \; x_2] = XJ \quad \text{where} \quad J = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

We have been a little cavalier regarding the "extending" of bases. Our examples have been so small that we have been able to proceed by visual inspection. What we need is something like a recipe.

We begin with with the decreasing shrinking sequence of column spaces

$$\mathcal{R}(A) \supset \mathcal{R}(A^2) \supset \cdots \supset \mathcal{R}(A^{m-1}) \supset \mathcal{R}(A^m) = 0$$

and note that

$$\dim(\mathcal{N}(A^m) \bmod \mathcal{N}(A^{m-1})) = n - d_{m-1}$$

is precisely the dimension of $\mathcal{R}(A^{m-1})$. Now for each pivot column $y_k$ of $A^{m-1}$ we set $x_k$ to the associated coordinate vector. It follows that $A^{m-1}x_k = y_k$ which ensures that $x_k \neq \mathcal{N}(A^{m-1})$. As the $x_k$ are linearly independent by construction, they comprise a basis for $\mathcal{N}(A^m)$ mod $\mathcal{N}(A^{m-1})$. As a result, Step 1 may be accomplished concretely via

**Step 1':** Let $p_j$ denote the index of the $j$th pivot column of $A^{m-1}$, where $j = 1, \ldots, \dim \mathcal{R}(A^{m-1})$. Define $v_m^j$ to be zero in every element except $p_j$, where it take the value 1.

Each of the extensions in the subsequent steps may be done in a similar fashion.

### 4.5. Notes and Exercises

Our focus on equilibria of network equations like $Sx = f$ in the previous two chapters lead us to ask two questions, in this chapter, about general $S$. The first; *for which $f$ does there exist an $x$ such that $Sx = f$?* received the answer; *for those $f$ in the column space of $S$.* The second; *when such an $x$ exists can there be many such $x$?* received the answer; *no, not if the null space of $S$ is simply the zero vector.*

These answers required careful definition of the basic concepts of linear algebra; subspace, basis, dimension and rank. We saw that subspaces are natural generalizations of lines and planes and that the rank of a matrix is revealed by Gaussian Elimination.

Our investigation of nilpotent matrices makes use of notes of Marco Gualtieri for the theory and notes of Idris Mercer for the large clean examples.

1. Which of the following subsets of $\mathbb{R}^3$ are actually subspaces? Check both conditions 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 $x_1 x_2 = 0$.
   (d) The vectors whose first two components obey $x_1 + x_2 = 0$..
   (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. Suppose $M$ and $Q$ are subspaces of $\mathbb{R}^n$. Which of the following subsets of $\mathbb{R}^n$ are actually subspaces? Check both conditions and show your work.
   (a) The union of $M$ and $Q$.
   (b) The intersection of $M$ and $Q$.

3. True or false, with justification.
   (a) $(1, 1, 1)^T$ lies in the span of $(1, 2, 1)^T$ and $(2, 1, 2)^T$.
   (b) $(1, 0, 1)^T$ lies in the span of $(1, 2, 1)^T$ and $(2, 1, 2)^T$.
   (c) $(1, 0, 0)^T$ lies in the span of $(1, 2, 1)^T$ and $(2, 1, 2)^T$.
   (d) If $u$ lies in the span of $v$ and $w$ then $v$ lies in the span of $u$ and $w$.

4. True or false, with justification.
   (a) $(1, 1, 1)^T$, $(1, 2, 1)^T$ and $(2, 1, 2)^T$ are linearly independent.
   (b) $(1, 0, 1)^T$, $(1, 2, 1)^T$ and $(2, 1, 2)^T$ are linearly independent.

(c) $(1,0,0)^T$, $(1,2,1)^T$ and $(2,1,2)^T$ are linearly independent.

(d) If $u$ and $v$ are linearly independent and $v$ and $w$ are linearly independent then $u$ and $w$ are linearly independent.

5. Prove that if $A \in \mathbb{R}^{n\times n}$ and has $n$ pivots then $A$ is invertible.

6. Prove that if the rank of $A \in \mathbb{R}^{n\times n}$ is one then $A = uv^T$ for some $u$ and $v$ in $\mathbb{R}^n$.

7. I encourage you to use `rref` and `null` for the following. (i) Add a diagonal crossbar between nodes 3 and 2 in Figure 4.1 and compute bases for the column and null spaces of the new incidence 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 incidence matrix.

8. Compute bases for the null and column spaces of the triangle incidence matrix, (3.34). Offer mechanical interpretations of the null space basis vectors.

9. Compute bases for the null and column spaces of the methane incidence matrix, (3.35). Offer mechanical interpretations of the null space basis vectors.

10. 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 4.1.

    (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 this says something about $\|Ax\|$.)

11. Suppose that $A \in \mathbb{R}^{m\times n}$ and $\mathcal{N}(A) = \mathbb{R}^n$. Argue that $A$ must be the zero matrix. Hint: Prove the contrapositive.

12. Show that if $AB = 0$ then $\mathcal{R}(B) \subset \mathcal{N}(A)$.

13. Show that if $A^{m-1}v \neq 0$ and $A^mv = 0$ then the chain $\{v, Av, \ldots, A^{m-1}v\}$ is linearly independent.

14. Suppose that $A \in \mathbb{R}^{n\times n}$ obeys $A^m = 0$ while $A^{m-1} \neq 0$. (i) Prove that (4.7) holds. (ii) Prove that (4.8) holds.

15. Suppose that $A$ and $B$ are both $n$-by-$n$. Show that if one of them is invertible then $AB$ is similar to $BA$.

16. Consider the nilpotent matrix

$$A = \begin{pmatrix} 2 & 2 & 2 & 2 & -4 \\ 7 & 1 & 1 & 1 & -5 \\ 1 & 7 & 1 & 1 & -5 \\ 1 & 1 & 7 & 1 & -5 \\ 1 & 1 & 1 & 7 & -5 \end{pmatrix}.$$

    Use MATLAB to confirm that $d_j = \dim \mathcal{N}(A^j) = j$ for $j = 1, \ldots, 5$. Conclude that $A$ is similar to $J_5$. Find a basis for $\mathcal{N}(A^5)$ mod $\mathcal{N}(A^4)$ via Steps 1' of §4.4. Use this basis vector to complete *the* Jordan chain and write out the the full transformation matrix, $X$.

# 5. The Fundamental Theorem and Beyond

The previous chapter, in a sense, only told half of the story. In particular, $A \in \mathbb{R}^{m\times n}$ 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 complementary to $\mathcal{N}(A)$. These questions are answered by the column space and null space of $A^T$. This then completes the Fundamental Theorem of Linear Algebra. We review its relevance to the problems of electrical and mechanical equilibrium and then embark on two very useful generalizations. First to Vector Spaces and Linear Operators and then to convex sets.

### 5.1. 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** 5.1. The **row space** of $A \in \mathbb{R}^{m\times n}$ is 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 $A_{\text{red}}$ are linearly independent and that all remaining rows of $A_{\text{red}}$ are zero leads us to

**Proposition** 5.2. The pivot rows of $A_{\text{red}}$ comprise a basis for $\mathcal{R}(A^T)$.

As there are $r$ pivot rows of $A_{\text{red}}$ we find that the dimension of $\mathcal{R}(A^T)$ is $r$. Recalling Prop. 4.7 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 inner 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

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

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

$$x_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \; x_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \; x_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \; x_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ -1 \end{pmatrix}, \; x_5 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \; x_6 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}.$$

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

**Proposition** 5.3. Fundamental Theorem of Linear Algebra (Preliminary). Suppose $A \in \mathbb{R}^{m\times n}$ 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

$$x = x_R + x_N, \quad \text{where} \quad x_R \in \mathcal{R}(A^T) \quad \text{and} \quad x_N \in \mathcal{N}(A). \tag{5.1}$$

We often express (5.1) more succinctly as

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

where $U \oplus V \equiv \{u + v : u \in U, \; v \in V\}$ denotes the **direct sum** of two subspaces that intersect only at 0. 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 $\mathcal{N}(A)$.

We have worked with a large example that flows from a "real" problem. It is however difficult to fully visualize the resulting spaces. So we consider

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \tag{5.2}$$

note that its row space is spanned by

$$x_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad \text{and} \quad x_2 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$$

and that its null space is spanned by

$$x_3 = \begin{pmatrix} -1 \\ 0 \\ -1 \end{pmatrix}.$$

We illustrate in Figure 5.1 the complementary piercing of $\mathcal{R}(A^T)$ by $\mathcal{N}(A)$.

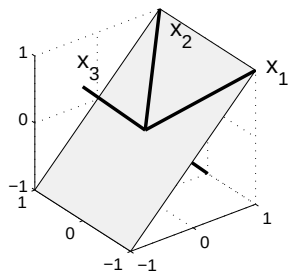



**Figure** 5.1. A depiction of $\mathbb{R}^3 = \mathcal{R}(A^T) \oplus \mathcal{N}(A)$ for the $A$ of (5.2). Here a finite piece (gray) of the $\mathcal{R}(A^T)$ plane is generated by $\pm x_1$ and $\pm x_2$ while the a finite piece of the $\mathcal{N}(A)$ line is generated by $\pm x_3$.

### 5.2. The Fundamental Theorem

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 $B \in \mathbb{R}^{p\times q}$ 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 $(A^T)_{\text{red}}$ 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

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

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 $(A^T)_{\text{red}}x = 0$ 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

**Proposition** 5.4. Fundamental Theorem of Linear Algebra. Suppose $A \in \mathbb{R}^{m\times n}$ 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 $\dim \mathcal{R}(A) = \dim \mathcal{R}(A^T) = r$, $\dim \mathcal{N}(A) = n - r$ and $\dim \mathcal{N}(A^T) = m - r$.

We illustrate this result in Figure 5.2. We see $A$ mapping $\mathbb{R}^n$ to $\mathcal{R}(A)$, a subspace of $\mathbb{R}^m$ and $A^T$ mapping $\mathbb{R}^m$ to $\mathcal{R}(A^T)$, a subspace of $\mathbb{R}^n$. These subspaces have the same dimensions and are orthogonal to the respective null spaces.

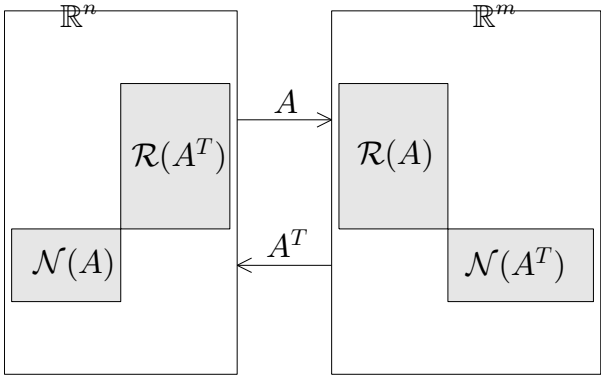



**Figure** 5.2. An illustration of the Fundamental Theorem of Linear Algebra.

We will make frequient use of this Fundamental Theorem throughout the remainder of the text. Notably, we will see at the start of the next chapter that the theory and algorithms behind the solution of least squares problems springs entirely from the fact that $\mathbb{R}^m = \mathcal{R}(A) \oplus \mathcal{N}(A^T)$.

### 5.3. Vector Spaces and Linear Transformations*

Our study has been motivated by the equations of equilibrium of electrical and mechanical systems. We then abstracted these square systems to rectangular systems and captured the action of an $m \times n$ matrix via its four fundamental subspaces of two Euclidean spaces, $\mathbb{R}^n$ and $\mathbb{R}^m$. These spaces were introduced as the collection of real vectors with $n$ and $m$ components respectively. It is natural, and we shall see necessary, to abstract further to vectors of complex numbers, complex matrices, complex functions and beyond. All of these are captured by the notion of vector space.

**Definition** 5.5 A **vector space** over the complex numbers, $\mathbb{C}$, is a set $V$ together with two operations, one of vector addition and one of scalar multiplication, that satisfy the eight axioms listed below. In the list below, let $u$, $v$ and $w$ be arbitrary vectors in $V$, and $a$ and $b$ arbitrary complex numbers.

1. Associativity of vector addition, $u + (v + w) = (u + v) + w$.
2. Commutativity of vector addition, $u + v = v + u$.
3. Identity element of vector addition. There exists an element $0 \in V$, called the zero vector, such that $v + 0 = v$ for each $v \in V$.
4. Inverse elements of vector addition. For every $v \in V$, there exists an element $-v \in V$, called the additive inverse of $v$, such that $v + (-v) = 0$.
5. Distributivity of scalar multiplication with respect to vector addition, $c(u + v) = cu + cv$.
6. Distributivity of scalar multiplication with respect to complex addition, $(a + b)v = av + bv$.
7. Compatibility of scalar multiplication with complex multiplication, $a(bv) = (ab)v$.
8. Identity element of scalar multiplication, $1v = v$.

If you are not familiar with complex arithmetic you may, for now, assume that $V$ is built over the reals, $\mathbb{R}$. We will cover complex arithmetic, from the beginning, in Chapter 9. Do you see that $\mathbb{R}^n$ is a vector space? Also note that the collection of $m$-by-$n$ matrices is a vector space. As is the collection of all polynomials with complex coefficients. All of the key "spatial" concepts so far discussed in $\mathbb{R}^n$, notably, subspace, basis and dimension extend naturally to all vector spaces. For example the set of 2–by–2 upper triangular matrices is a 3 dimensional subspace of $\mathbb{R}^{2\times 2}$ with basis

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

If $A$ is a rule for transforming elements of the vector space $V$ into elements of the vector space $W$ we call it a **linear transformation** if

$$A(av + bw) = aAv + bAw, \qquad \forall \quad v \in V,\ w \in V,\ a \in \mathbb{C},\ b \in \mathbb{C}. \tag{5.3}$$

If $V$ and $W$ are finite dimensional, with bases $\{v_1, v_2, \ldots, v_n\}$ and $\{w_1, w_2, \ldots, w_m\}$ respectively then (5.3) implies that

$$A(c_1v_1 + c_2v_2 + \cdots + c_nv_n) = c_1A(v_1) + c_2A(v_2) + \cdots + c_nA(v_n),$$

for arbitrary $c \in \mathbb{C}^n$. Now, as each $A(v_j) \in W$ it may be expressed in terms of the $w_i$. In particular, we find

$$A(v_j) = a_{1j}w_1 + a_{2j}w_2 + \cdots + a_{mj}w_m.$$

And so

$$\begin{aligned} A(c_1v_1 + c_2v_2 + \cdots + c_nv_n) = (a_{11}c_1 + a_{12}c_2 + \cdots + a_{1n}c_n)w_1 \\ + (a_{21}c_1 + a_{22}c_2 + \cdots + a_{2n}c_n)w_2 \\ + \cdots + (a_{m1}c_1 + a_{m2}c_2 + \cdots + a_{mn}c_n)w_m. \end{aligned}$$

We now view $A$ as transforming the coefficients of the $v_j$ into the coefficients of the $w_i$ via matrix–vector multiplication.

For example, the transformation of the plane that reflects vectors over the horizontal axis,

$$A(x_1, x_2) \equiv (x_1, -x_2)$$

is linear. In the "standard" basis,

$$v_1 = (1, 0), \qquad v_2 = (0, 1)$$

we find $A(v_1) = v_1$ and so $a_{11} = 1$ and $a_{12} = 0$. Similarly $A(v_2) = -v_2$ and so $a_{21} = 0$ and $a_{22} = -1$ and we arrive at the standard matrix representation

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

If instead however we choose the "tilted" basis

$$w_1 = (1, 1), \qquad w_2 = (1, -1) \tag{5.4}$$

then $A(w_1) = w_2$ and $A(w_2) = w_1$ and we arrive at the tilted matrix representation

$$A_t = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

As $A_s$ and $A_t$ represent the same transformation it should not be surprising that the **change of basis** matrix serves as a similarity transformation between the representations. In particular, note that

$$B \equiv \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \tag{5.5}$$

is the basis changer, i.e., $w_1 = Bv_1$, $w_2 = Bv_2$, and as a result the similarity transformation

$$A_t = B^{-1}A_sB.$$

This notion of basis change via similarity transformation is central to the second theme of this text, i.e., the spectral representation of a matrix. We prepared much of the ground in Prop. 4.13 when we proved that each nilpotent matrix is similar to a block diagonal matrix composed of null Jordan blocks. The remaining theory will construct similarity transformations from bases of invariant subspaces.

To be precise, we call the subset $U$ of the vector space $V$ an **invariant subspace** of the linear transformation $A$ if $Au \in U$ for each $u \in U$. For example, the invariant subspaces of

$$A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \tag{5.6}$$

are span $(w_1)$ and span $(w_2)$ where $w_1$ and $w_2$ are the tilted basis in (5.4). In this case the $B$ in (5.5) yields

$$B^{-1}AB = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$$

and we say that $B$ diagonalizes $A$.

### 5.4. Linear Inequalities and Convex Sets*

Convex sets are natural generalizations of subspaces and so worthy of study in their own right. Beyond mere generalization we will use a simple question about mixing of elements in the inner product, $x^T y$, to motivate the study of convex sets and to illustrate some of their amazing properties.

Given $A$ and $b$, to know whether there exists an $x$ such that $Ax = b$ we need only know whether $b$ lies in the span of the columns of $A$. In a number of important applications we desire additional properties of $x$. The simplest of these being that $x \geq 0$, by which we mean that each element of $x$ is nonzero. The question of the existence of nonnegative $x$ for which $Ax = b$ is then answered by whether $b$ lies in the **cone** of $A$, i.e., in

$$\text{cone}(A) \equiv \{Ax : x \geq 0\}.$$

Of course this is more like naming the answer rather than providing an answer. The fundamental theorem of linear algebra helped us out of the related word game by identifying $\mathcal{R}(A)$ with the orthogonal complement of $\mathcal{N}(A^T)$. This observation is often expressed as a **Fredholm Alternative**

**Proposition** 5.6. Given $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ precisely one of these alternatives pertains.
(1) There exists an $x \in \mathbb{R}^n$ such that $Ax = b$.
(2) There exists a $y \in \mathbb{R}^m$ such that $A^T y = 0$ and $y^T b \neq 0$.

This says that either $b$ lies in the span of the columns of $A$ or $b$ is not orthogonal to element of the left null space of $A$. In asking then for conditions on $b$ such that there exists an $x \geq 0$ such that $Ax = b$, beyond the tautological $b \in \text{cone}(A)$, we find the **Farkas Alternative**.

**Proposition** 5.7. Given $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ precisely one of these alternatives pertains.
(1) There exists an $x \in \mathbb{R}^n$ such that $x \geq 0$ and $Ax = b$.
(2) There exists a $y \in \mathbb{R}^m$ such that $y^T A \geq 0$ and $y^T b < 0$.

This says that either $b$ lies in the cone of the columns of $A$ or there exists a subspace that separates $b$ from the columns of $A$. Before proving this alternative we illustrate its new terms in the case that

$$A = \begin{pmatrix} -1 & 3 \\ 1 & 1 \end{pmatrix}.$$

To find $\text{cone}(A)$ we now solve $Ax = b$ and derive conditions on $b$ that guarantee that $x \geq 0$. Gaussian elimination in column 1 brings $-x_1 + 3x_2 = b_1$ and $4x_2 = b_1 + b_2$. Back substitution then yields $x_2 = (b_1 + b_2)/4$ and $x_1 = (3b_2 - b_1)/4$. Hence

$$\text{cone}(A) = \{b \in \mathbb{R}^2 : b_1 + b_2 \geq 0 \ \&\ 3b_2 \geq b_1\}.$$

We see in Figure !!!! that $\text{cone}(A)$ is indeed the shaded cone "between" the columns of $A$ and that it coincides with the intersection of half spaces.

For $b \notin \text{cone}(A)$, e.g., $b = (3,\ 0)^T$ we note that the line through $y^\perp = (3,\ 1/2)^T$ separates $b$ from the columns of $A$ in the sense that, with $y = (-1/2,\ 3)^T$,

$$y^T b = -3/2, \quad y^T A = (7/2,\ 3/2).$$

In general, if $b \notin \text{cone}(A)$ then $b \notin H_i$

We now take up the proof of the Farkas Alternative.

Proof: If (1) and (2) both hold then $x^T A^T y = y^T b$ and yet the right side is negative while the left is nonnegative. This shows that (1) and (2) can not both be true.

Now (1) is true iff $b \in \text{cone}(A)$ and so it suffices to show that if $b \notin \text{cone}(A)$ then (2) holds.

The mixing question asks, given a nonnegative, nonincreasing vector $x \in \mathbb{R}^n$, i.e.,

$$x_1 \geq x_2 \geq \cdots x_n \geq 0, \tag{5.7}$$

and a second nonnegative vector, $y \in \mathbb{R}^n$, how should I mix the elements of $y$ in order to maximize its interaction with $x$? To mix the elements of $y$ we mean to multiply it by a **doubly stochastic matrix**, i.e., a matrix with nonnegative elements such that each row and column sum to 1. In symbols, $D$ belongs to $\text{DS}_n$, the class of doubly stochastic matrices of size $n$, when

$$\text{each } d_{i,j} \geq 0, \qquad \sum_{i=1}^{n} d_{i,j} = 1 \quad \text{and} \quad \sum_{j=1}^{n} d_{i,j} = 1. \tag{5.8}$$

With this preparation we can now properly pose and answer the mixing question.

**Proposition** 5.8. **Mixing Theorem.** If $x$ and $y$ are two nonnegative, nonincreasing vectors in $\mathbb{R}^n$ then

$$x^T D y \leq x^T y \quad \forall\ D \in \text{DS}_n. \tag{5.9}$$

This states that the interaction between two nonnegative vectors is greatest when they are similarly ordered. This type of result is central to the field of Quantum Information Theory where tools for quantifying disorder, or entropy, are required. We will derive the Mixing Theorem as a consequence of

**Proposition** 5.9. **Birkhoff's Theorem.** Every doubly stochastic matrix can be written as a convex combination of permutation matrices.

A matrix is a **permutation matrix** if it can be achieved by interchanging columns of the identity matrix. Or equivalently, if it is a product of the elementary permutation matrices defined in §3.2. We will denote the class of $n$-by-$n$ permutation matrices by $\text{Per}_n$.

Finally a **convex combination** of objects is simply a linear combination in which each of the coefficients is nonnegative and for which their sum is 1. For example the proof of Birkhoff's Theorem, in the 2–by–2 case, is the one–liner

$$\begin{pmatrix} d & 1-d \\ 1-d & d \end{pmatrix} = d \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + (1-d) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad 0 \leq d \leq 1, \tag{5.10}$$

for the matrix on the left is the most general doubly stochastic matrix in $\text{DS}_2$ while the two matrices on the right are the only permutations in $\text{Per}_2$. To prove Birkhoff's Theorem in general will require a deeper understanding of convex sets. With that motivation we now begin with

**Definition** 5.10. The subset $C$ of a vector space is said to be **convex** if whenever $c_1$ and $c_2$ lie in $C$ so too does the chord between them, i.e, $tc_1 + (1-t)c_2 \in C$ for every $0 < t < 1$.

Please confirm that $\mathrm{DS}_n$ is convex. Birkhoff's Theorem says, in a sense, that the permutation matrices are the corners, or vertices, of the class of doubly stochastic matrices. This notion is made precise by

**Definition** 5.11. A point $c$ in a convex set $C$ is called an **extreme point** of $C$ if it does not lie on any chord in $C$. That is, if there do not exist distinct $c_1$ and $c_2$ in $C$ and a $0 < t < 1$ for which $c = tc_1 + (1-t)c_2$.

Our proof that $\mathrm{Per}_n$ is the set of extreme points of $\mathrm{DS}_n$ will follow from the polyhedral nature of $\mathrm{DS}_n$. A **polyhedron** is the intersection of a finite number of half spaces. A **half space** is the set of points lying on one side of a hyperplane and a **hyperplane** in $\mathbb{R}^n$ is a fixed translate of all vectors orthogonal to some nonzero fixed $a \in \mathbb{R}^n$. In symbols, a hyperplane in $\mathbb{R}^n$ is

$$\{x \in \mathbb{R}^n : a^T x = b\}.$$

Here $b \in \mathbb{R}$ specifies the amount that the subspace $a^\perp$ is to be translated. It follows that the associated half spaces are

$$\{x \in \mathbb{R}^n : a^T x \leq b\} \quad \text{and} \quad \{x \in \mathbb{R}^n : a^T x \geq b\}.$$

The intersection of $m$ half spaces, individually parametrized by $a_i$ and $b_i$, describes the general polyhedron

$$\mathcal{P} = \bigcap_{i=1}^{m} \{x \in \mathbb{R}^n : a_i^T x \leq b_i\}.$$

It is customary to lay these $a_i^T$ into the rows of a matrix, $A \in \mathbb{R}^{m \times n}$, and to collect the $b_i$ into a vector $b \in \mathbb{R}^n$. In that way we arrive at the concise definition of a polyhedron as

$$\mathcal{P} \equiv \{x \in \mathbb{R}^n : Ax \leq b\}, \quad \text{where } A \in \mathbb{R}^{m \times n} \quad \text{and} \quad b \in \mathbb{R}^m. \tag{5.11}$$

For example, the triangle in Figure 5.3 is specified by

$$A = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 1 & 1 & 0 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \tag{5.12}$$

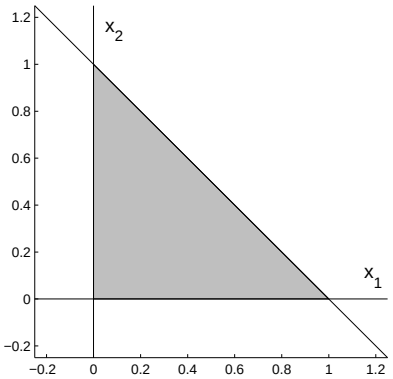



**Figure** 5.3. The intersection of the three half spaces parametrized by (5.12).

We intuitively expect the extreme points of the triangle in Figure 5.3 to lie at the three vertices. Geometrically these occur at the intersection of any two of the defining hyperplanes. Algebraically this means that two (distinct) rows in $Ax \leq b$ actually hold with equality. The general form of this statement looks like

**Proposition** 5.12. For a polyhedron, $\mathcal{P}$, described by (5.11), a vector $x$ is an extreme point of $\mathcal{P}$ iff $n$ linearly independent inequalities of the system $Ax \leq b$ are equalities at $x$.

Proof: . Let $a_i^T,\ i = 1, \ldots, m$, be the rows of $A$. Suppose that $x$ is an extreme point of $\mathcal{P}$, and let $\mathcal{I}$ be the set of those indices $i$ for which $a_i^T x = b_i$ and set $\mathcal{A} = \text{span}\{a_i : i \in \mathcal{I}\}$. If $\mathcal{A} \neq \mathbb{R}^n$ then there exists a nonzero vector $h \in \mathcal{A}^\perp$. We will use this to find a $t > 0$ for which the segment $S_t = [x - th, x + th] \subset \mathcal{P}$. This will contradict the assumption that $x$ is an extreme point of $\mathcal{P}$ and so prove that $\mathcal{A} = \mathbb{R}^n$. For $i \in \mathcal{I}$ as $h \perp \mathcal{A}$ it follows that $a_i^T(x \pm th) = b_i$ for all $t > 0$. To handle the other rows we define

$$t_0 \equiv \min_{i \notin \mathcal{I}} \frac{b_i - a_i^T x}{|a_i^T h|} > 0$$

and note for $t < t_0$,

$$a_i^T(x \pm th) \leq a_i^T x + t_0 |a_i^T h| \leq a_i^T x + \frac{b_i - a_i^T x}{|a_i^T h|} |a_i^T h| = b_i \quad \forall\ i \notin \mathcal{I}.$$

This yields the desired contradiction, $S_{t_0} \subset \mathcal{P}$.

Conversely let us prove that if $\mathcal{A} = \mathbb{R}^n$ then $x$ is an extreme point. We show that if $x = tp + (1-t)q$ for $p$ and $q$ in $\mathcal{P}$ and $0 < t < 1$ then $p = q$. Note that for each $i \in \mathcal{I}$

$$b_i = a_i^T x = t a_i^T p + (1-t) a_i^T q \leq b_i$$

and so equality holds and so $a_i^T p = a_i^T q = b_i$ and so $(p - q) \perp \mathcal{A}$ and so $p = q$. End of Proof.

With this criterion we can now identify the extreme points of the set of doubly stochastic matrices.

**Proposition** 5.13. $\text{Per}_n$ is the set of extreme points of $\text{DS}_n$.

Proof: We have already observed that each $P \in \text{Per}_n$ is an extreme point of $\text{DS}_n$. To prove the converse, we note that only $2n - 1$ of the $2n$ sums in (5.8) are independent. Now let us prove the claim by induction in $n$. The base case $n = 2$ has been verified by (5.10). Let us justify the inductive step from $n - 1$ to $n$. Thus, let $X$ be an extreme point of $\text{DS}_n$. By the preceding Proposition, among the constraints defining $\text{DS}_n$ (i.e., $2n - 1$ equalities and $n^2$ inequalities $d_{ij} \geq 0$) there should be $n^2$ linearly independent which are satisfied at $X$ as equations. Thus, at least $n^2 - (2n - 1) = (n - 1)^2$ entries in $X$ should be zeros. It follows that at least one of the columns of $X$ contains at most one nonzero entry (since otherwise the number of zero entries in $X$ would be at most $n(n-2) < (n-1)^2$). Thus, there exists at least one column with at most 1 nonzero entry. As the sum of entries in this column is 1, this nonzero entry, say $X_{i'j'}$, is equal to 1. Since the entries in row $i'$ are nonnegative and sum to 1 it follows that $X_{i'j'}$ is the only nonzero in row $i'$. Removing row $i'$ and column $j'$ from $X$ produces an $X' \in \text{DS}_{n-1}$. By the inductive hypothesis, this $X'$ is a convex combination of elements of $\text{Per}_{n-1}$. Augmenting $X'$ and each of these permutation matrices by the column and the row removed from $X$ we recover $X$ as a convex combination of elements of

$\mathrm{Per}_n$. But as $X$ is an extreme point this combination must degenerate to a single summand, i.e., $X$ is a single permutation matrix. This completes the inductive step. End of Proof.

Our definition presumes that extreme points of a set are members of the set. So, for example, the open interval $(0, 1)$ has no extreme points. To curtail this in general we will assume that our convex set is **closed**. That is, the limit of every convergent sequence drawn from $C$ has a limit in $C$. Please confirm that $\mathrm{DS}_n$ is closed and that each element of $\mathrm{Per}_n$ is an extreme point of $\mathrm{DS}_n$. To complete the proof of Birkhoff's we need to know that the elements of a convex set can always be expressed as convex combinations of its extreme points. This is answered by

**Proposition** 5.14. **Krein Millman.** If $C \subset \mathbb{R}^n$ is bounded, closed, nonempty and convex then each point in $C$ is a convex combination of the extreme points of $C$.

We will prove this by induction on the dimension of $C$. We note that the "ambient" dimension of $\mathrm{DS}_2$ is 4, as it lies in the class of 2–by–2 matrices. As a set however it is the simple chord (5.10) – and hence it appears that the true dimension of $\mathrm{DS}_2$ is 1. To generalize our notion of dimension we first generalize our notion of span.

The **Affine span** is the usual span subject to the additional constraint that the coefficients sum to one. For example the affine span of $\mathrm{DS}_2$ is the line

$$\mathrm{Aspan}(\mathrm{DS}_2) = \left\{ \begin{pmatrix} d & 1-d \\ 1-d & d \end{pmatrix} : d \in \mathbb{R} \right\}.$$

It is a line in the sense that it simply extends the chord defined in (5.10). To see its dimension we note that every affine span is merely a shifted, or translated, subspace. As such, for an arbitrary element $s \in \mathrm{Aspan}(C)$, we define

$$\mathrm{Shift}(\mathrm{Aspan}(C)) \equiv \{c - s : c \in \mathrm{Aspan}(C)\}. \tag{5.13}$$

Returning to $\mathrm{DS}_2$ we from $\mathrm{Aspan}(\mathrm{DS}_2)$ pick

$$s = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and find} \quad \mathrm{Shift}(\mathrm{Aspan}(\mathrm{DS}_2)) = \left\{ d \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} : d \in \mathbb{R} \right\},$$

and recognize this to be a classical one dimensional subspace of $\mathbb{R}^{2\times 2}$. We will prove in Exer. 5.16 that $\mathrm{Shift}(\mathrm{Aspan}(C))$ is always a subspace – and that this subspace does not depend on the choice of $s$. This then justifies our definition of **Affine dimension**

$$\mathrm{Adim}(C) \equiv \dim(\mathrm{Shift}(\mathrm{Aspan}(C))). \tag{5.14}$$

From this definition it follows, for example, that $\mathrm{Adim}(\mathrm{DS}_k) = (k-1)^2$.

The Affine span is also the crucial notion in defining the boundary of a closed convex set. We first define the **Affine Interior** of a convex set $C \subset \mathbb{R}^n$ to be all those points $c \in C$ for which $C$ also contains a restricted ball around $c$ – where the restriction is to $\mathrm{Aspan}(C)$. More precisely,

$$\mathrm{Aint}(C) \equiv \{c \in C : \exists\, r > 0 \text{ such that if } x \in \mathrm{Aspan}(C) \text{ and } \|x - c\| < r \text{ then } x \in C\}.$$

With this we define the **Affine boundary** of the closed convex $C$ to be those points left in $C$ when we exclude its Affine interior. That is,

$$\mathrm{Abdry}(C) \equiv C \setminus \mathrm{Aint}(C) = \{x \in C : x \notin \mathrm{Aint}(C)\}.$$

For example, please confirm that

$$\begin{aligned} \text{Aint}(\text{DS}_2) &= \left\{ d \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + (1-d) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} : 0 < d < 1 \right\} \quad \text{and} \\ \text{Abdry}(\text{DS}_2) &= \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\}. \end{aligned} \tag{5.15}$$

The inductive step requires a beautiful geometric construction, see Figure 5.4, that makes definite the sense in which linear algebra supports convex geometry. It hinges on the following

**Lemma** 5.15. If $C$ is closed and convex and $d \notin C$ then there exists a unique closest point in $C$ to $d$.

Proof: The distance from $C$ to this point is the least distance between $d$ and any point of $C$. That is

$$\text{dist}(C, d) = \inf_{c \in C} \|c - d\|,$$

where inf denotes the gteatest lower bound. Build a sequence by choosing $c_k \in C$ such that $\|c_k - d\| < \text{dist}(C, d) + 1/k$. Now

$$\|c_k\| = \|(c_k - d) + d\| \leq \|c_k - d\| + \|d\| \leq \text{dist}(C, d) + 1 + 1/k.$$

This bound then establishes a subsequence, $\{c_{n_k}\} \subset \{c_k\}$ and a limit $\overline{c}$ such that $c_{n_k} \to \overline{c}$. As $C$ is closed it follows that $\overline{c} \in C$. By construction it follows that $\|c_{n_k} - d\| \to \|\overline{c} - d\| = \text{dist}(C, d)$. If there are two closest points, $\overline{c}_1$ and $\overline{c}_2$ then the midpoint $m = (\overline{c}_1 + \overline{c}_2)/2 \in C$. As $\|d - m\|$ is the height of the isocoles triangle with vertices $d$, $\overline{c}_1$ and $\overline{c}_2$ it follows that $\|d - m\| < \|d - \overline{c}_1\|$, in contradiction of $\overline{c}_1$. hence, the closest point is unique. End of Proof.

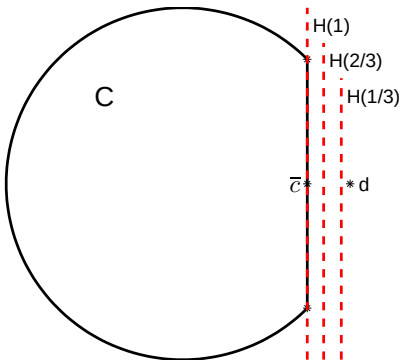



**Figure** 5.4. Illustration of the hyperplane $H(1)$ that supports the truncated disk, $C$, at $\overline{c}$. Given the point $d$, its closest point in $C$ is $\overline{c}$. The separating hyperplanes, $H(1/3)$ and $H(2/3)$, are described in (5.17).

**Proposition** 5.16. If $C \subset \mathbb{R}^n$ is closed, bounded and convex and $\overline{c} \in \text{Abdry}(C)$ then there exists a nonzero vector $a \in \mathbb{R}^n$ such that

$$a^T c \leq a^T \overline{c} \quad \forall\ c \in C. \tag{5.16}$$

The associated hyperplane

$$H = \{y \in \mathbb{R}^n : a^T y = a^T \overline{c}\}$$

is said to **support** $C$ at $\overline{c}$. (See Figure 5.4.)

Proof: There exists a sequence of points, $d_n \in \text{Aspan}(C) \setminus C$ such that $d_n \to \overline{c}$. Let $c_n$ be the unique closest point in $C$ to $d_n$.

Now define the vector, scalar (parametrized by $0 < t < 1$) and hyperplane

$$\begin{aligned} a_n &= d_n - c_n, \\ b_n(t) &= (1-t)\|d_n\|^2 - t\|c_n\|^2 + (2t-1)c_n^T d_n, \\ H_n(t) &= \{x : a_n^T x = b_n(t)\}. \end{aligned} \tag{5.17}$$

We now prove that $H_n(t)$ separates $C$ and $d_n$ in the sense that

$$\begin{aligned} f_n(x,t) \equiv a_n^T x - b_n(t) &= (d_n - c_n)^T (x - d_n) + t\|c_n - d_n\|^2 \\ &= (d_n - c_n)^T (x - c_n) - (1-t)\|c_n - d_n\|^2 \end{aligned} \tag{5.18}$$

takes different signs in $C$ and $d_n$, for fixed $t$. Using the top form of (5.18) we find that $f_n(d,t) > 0$. To see that $f_n(x,t) < 0$ for $x \in C$ suppose otherwise, i.e., that there exists a $u \in C$ such that $f(u,t) > 0$. In this case, by the bottom form of (5.18), it follows that $(d - \overline{c})^T (u - \overline{c}) > 0$. So we consider $g(\varepsilon) \equiv \|\varepsilon u + (1-\varepsilon)\overline{c} - d\|^2$ and find that

$$g'(0) = 2(\overline{c} - d)^T (u - \overline{c}) < 0$$

and so $g$ is decreasing at zero and there is a small $0 < \varepsilon_0 < 1$ such that $g(\varepsilon_0) < g(0)$ but this means $\|\varepsilon_0 u + (1-\varepsilon_0)\overline{c} - d\| < \|d - \overline{c}\|$ which contradicts the definition of $\overline{c}$.

Now slide $t$ as illustrated in Figure 5.4. In particular as $t \to 1$ we find that $H(1)$ supports $C$ at $\overline{c}$ in the sense that $b(1) = \overline{c}^T d - \|\overline{c}\|^2 = a^T \overline{c}$ and so

$$H(1) = \{x : a^T x = a^T \overline{c}\}$$

and $a^T c \leq a^T \overline{c}$ for all $c \in C$. Now finally slide $d$ along this line to $\overline{c}$ and observe that every point in $\text{Abdry}(C)$ has a supporting hyperplane. End of Proof.

The inductive utility of this construction is indicated in Figure 5.4. In particular, while $\text{Adim}(C) = 2$ we find $\text{Adim}(H(1) \cap C) = 1$ and the two extreme points of $H(1) \cap C$, indicated by the unlabeled asterisks, are also extreme points of $C$. Of course this is just a picture. We must prove

**Proposition** 5.17. Suppose that $C$ is nonempty closed convex set and that $H$ is a hyperplane that supports $C$ at $\overline{c} \in \text{Abdry}(C)$. If $\overline{y}$ is an extreme point of $H \cap C$ then $\overline{y}$ is an extreme point of $C$.

Proof: Let $a$ be the direction associated with $H = \{y \in \mathbb{R}^n : a^T y = a^T \overline{c}\}$, so that

$$a^T c \leq a^T \overline{c} \quad \forall c \in C. \tag{5.19}$$

Assume that $\overline{y}$ is an extreme point of $H \cap C$, but can be written as a proper convex combination in $C$, i.e.,

$$\overline{y} = \lambda c_1 + (1-\lambda)c_2, \quad 0 < \lambda < 1 \tag{5.20}$$

for some $c_1$, $c_2 \in C$. We will now argue that $c_1$ and $c_2$ lie in $H$ and the use the fact that $\overline{y}$ is an extreme point of $H \cap C$ to conclude that $c_1 = c_2 = \overline{y}$.

As $\overline{y} \in H$ it follows from (5.19) that

$$a^T\overline{y} = a^T\overline{c} \geq \max\{a^Tc_1, a^Tc_2\}$$

while (5.20) requires

$$a^T\overline{y} = \lambda a^Tc_1 + (1-\lambda)a^Tc_2.$$

The latter is however strictly less than $\max\{a^Tc_1, a^Tc_2\}$ unless $a^Tc_1 = a^Tc_2 = a^T\overline{y}$. These equalities state that each $c_j \in H$, and, as we had already presumed $c_j \in C$ it follows that each $c_j \in H \cap C$. But now, since $\overline{y}$ is an extreme point of $H \cap C$ it follows that (5.20) can only imply that $c_1 = c_2 = \overline{y}$ and hence we have shown that $\overline{y}$ is an extreme point of $C$. End of Proof.

We now have all of the ingredients necessary to prove the Krein–Millman Theorem.

Proof of Prop. 5.14. As the extreme points of $C$ lie in $C$ and $C$ is convex it follows that the convex combinations of extreme points also lie in $C$. So it remains to show that every $c \in C$ is a convex combination of extreme points of $C$.

We use induction on the affine dimension of $C$. The case $\mathrm{Adim}(C) = 1$ has already been established by our explicit study of $\mathrm{DS}_2$. We assume then that the statement in question is valid for all convex, closed, and bounded sets with affine dimension $k$, and let $C$ be a convex, closed, and bounded set with $\mathrm{Adim}(C) = k+1$.

Given $c \in C$ we choose a nonzero direction $h \in \mathbb{R}^n$ and form the line

$$\ell = \{c + \lambda h : \lambda \in \mathbb{R}\} \quad h \neq 0.$$

Moving along this line from $c$ in each of the two possible directions, we eventually leave $C$ (since $C$ is bounded). Hence there exist $\lambda_+ > 0$ and $\lambda_- < 0$ such that

$$\overline{c}_\pm = c + \lambda_\pm h \in \mathrm{Abdry}(C).$$

As $c$ lies on the chord from $\overline{c}_-$ to $\overline{c}_+$ it follows that $c$ is a convex combination of $\overline{c}_+$ and $\overline{c}_-$. It remains then to confirm that $\overline{c}_\pm$ are themselves convex combinations of the extreme points of $C$.

We denote by $H_+$ a hyperplane that supports $C$ at $\overline{c}_+$. The set $H_+ \cap C$ (which clearly is convex, closed and bounded) is of affine dimension $k$; by the inductive hypothesis, the point $\overline{c}_+$ of this set is a convex combination of extreme points of the set, and by Prop. . 5.17 all these extreme points are extreme points of $C$ as well. Thus, $\overline{c}_+$ is a convex combination of extreme points of $C$. Similar reasoning is valid for $\overline{c}_-$. End of Proof.

Birkhoff's Theorem now follows directly from Props. 5.12 and 5.14. Regarding the Mixing Theorem we note that if $x$ and $y$ are nonnegative and nonincreasing then, by direct inspection (write it out)

$$x^TPy \leq x^Ty \quad \forall\ P \in \mathrm{Per}_n. \tag{5.21}$$

Now, by Birkhoff's Theorem, if $D \in \mathrm{DS}_n$ then there exist $m$ permutation matrices, $P_i$, and $m$ nonnegative $\lambda_i$ that sum to one, such that

$$D = \sum_{i=1}^{m} \lambda_i P_i.$$

To this we apply $x^T$ from the left and $y$ from the right and deduce from (5.21) that

$$x^T D y = \sum_{i=1}^{m} \lambda_i x^T P_i y \leq \sum_{i=1}^{m} \lambda_i x^T y = x^T y,$$

as claimed by the Mixing Theorem.

### 5.5. Tensegrities

We consider the concrete prismatic tensegrity $T_3$ (of Zhang Guest Ohsaki) with 6 nodes, 3 lateral struts, 3 lateral cables, and 3 horizontal bottom cables and 3 horizontal top cables. When the top is shifted by $a$ from the bottom our nodes are at

$$\begin{aligned}
p_0 &= (r\cos(\pi/3 - a), r\sin(\pi/3 - a), 0) \\
p_1 &= (r\cos(\pi - a), r\sin(\pi - a), 0) \\
p_2 &= (r\cos(5\pi/3 - a), r\sin(5\pi/3 - a), 0) \\
p_3 &= (r\cos(\pi/3), r\sin(\pi/3), 1) \\
p_4 &= (r\cos(\pi), r\sin(\pi), 1) \\
p_5 &= (r\cos(5\pi/3), r\sin(5\pi/3), 1)
\end{aligned}$$

and so the edge matrix (following Williams 2.30) reads

$$\Pi(a) = \begin{pmatrix}
p_0 - p_1 & 0 & p_0 - p_2 & 0 & 0 & 0 & p_0 - p_4 & 0 & 0 & p_0 - p_3 & 0 & 0 \\
p_1 - p_0 & p_1 - p_2 & 0 & 0 & 0 & 0 & 0 & p_1 - p_5 & 0 & 0 & p_1 - p_4 & 0 \\
0 & p_2 - p_1 & p_2 - p_0 & 0 & 0 & 0 & 0 & 0 & p_2 - p_3 & 0 & 0 & p_2 - p_5 \\
0 & 0 & 0 & p_3 - p_4 & 0 & p_3 - p_5 & 0 & 0 & p_3 - p_2 & p_3 - p_0 & 0 & 0 \\
0 & 0 & 0 & p_4 - p_3 & p_4 - p_5 & 0 & p_4 - p_0 & 0 & 0 & 0 & p_4 - p_1 & 0 \\
0 & 0 & 0 & 0 & p_5 - p_4 & p_5 - p_3 & 0 & p_5 - p_1 & 0 & 0 & 0 & p_5 - p_2
\end{pmatrix}$$

A smooth function $t \mapsto q(t)$ where $q(0) = p$ is called a motion from $p$. It is called **admissible** if it does not change the length of bars and does not stretch any cable. The motion is called **rigid** if

$$q(t) = Q(t)p + r(t), \quad \text{where} \quad Q(t)^T Q(t) = I \quad \text{and} \quad \det(Q(t)) = 1. \tag{5.22}$$

**Definition** 5.18. The placement $p$ is **stable** when the only motion from $p$ is rigid.

**Proposition** 5.19. If $p$ is a stable placement then there is a nonzero proper prestress for $p$.

The prestress vector $\omega$ lies in its null space. In `tensegrity.m` we find the product of the pivots of $\Pi(a)$ to be

$$\frac{729}{32}(3\sin(a) - \sqrt{3}\cos(a))$$

hence the shift $a = \pi/6$ gives rise to the prestress

$$\omega = \begin{pmatrix} \text{ones}(6,1)/\sqrt{3} \\ -\text{ones}(3,1) \\ \text{ones}(3,1) \end{pmatrix}$$

This prestress generates the stress matrix

$$\Omega = \sum_e \omega_e B_e$$

where, e.g.,

$$B_1 = \begin{pmatrix} I & -I & 0 & 0 & 0 & 0 \\ -I & I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad B_2 = \begin{pmatrix} I & 0 & -I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -I & 0 & I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and so with $s = 1/\sqrt{3}$

$$\Omega = \begin{pmatrix} 2sI & -sI & -sI & -I & I & 0 \\ -sI & 2sI & -sI & 0 & -I & I \\ -sI & -sI & 2sI & I & 0 & -I \\ -I & 0 & I & 2sI & -sI & -sI \\ I & -I & 0 & -sI & 2sI & -sI \\ 0 & I & -I & -sI & -sI & 2sI \end{pmatrix}$$

**Proposition** 5.20. If $p$ has a strict prestress then $\mathcal{N}(\Pi^T)$ is the space of admissible velocities.

A velocity is called rigid if it is the derivative of a rigid motion (5.22), at $t = 0$. So suppose $Q(t) = I + tW + O(t^2)$ and $r(t) = tv + O(t^2)$ then

$$q'(0) = Wp + v. \tag{5.23}$$

The constraint that $Q^T(t)Q(t) = I$ of course constrains $W$ that appear in (5.23). In particular,

$$(I + tW + O(t^2))^T(I + tW + O(t^2)) = I \Rightarrow W^T = -W. \tag{5.24}$$

A **flexure** is an element of $\mathcal{N}(\Pi^T)$ that is orthogonal to every rigid velocity.

**Proposition** 5.21. The placement $p$ is stable if there exists a flexure, $v$ for which $v^T\Omega v > 0$.

So we compute the flexures of $T_3$. Let $\{v_1, \ldots, v_7\}$ form a basis for $\mathcal{N}(\Pi^T)$ and $\{w_1, \ldots, w_6\}$ for a basis for $V_{rig}$. The flexure coefficients lie in the null space of $G \in \mathbb{R}^{6\times 7}$ where

$$G_{ij} = w_i^T v_j.$$

We solve $Gx = 0$ and arrive at the flexure

$$v = \sum_{j=1}^{7} x_i v_i = (-1, s, -1, -1, -s, -1, 2, 0, -1, s, -1, 1, 0, 2, 1, -s, -1, 1)^T \tag{5.25}$$

where $s = \sqrt{3}$ and we then compute $v^T\Omega v = 48\sqrt{3}$ and conclude stability.

In order to prove the two propositions it will be helpful to normalize our class of tensegrities.

**Proposition** 5.22. Let the tensegrity have at least one bar, $b = 1 - 2$, and at least two edges. Given a placement $\mathbf{p}$ there is exactly one $\mathbf{p}^* \in \text{Euc}(\mathbf{p})$ such that $\mathbf{p}_1 = (0,0,0)$, $\mathbf{p}_2 = (x_2,0,0)$ and $\mathbf{p}_3 = (x_3,y_3,0)$ where $y_3 > 0$.

Proof: We introduce the right–handed ordered triple of orthonormal vectors, $\mathbf{e}$, $\mathbf{f}$, $\mathbf{g}$ in $\mathbb{R}^3$ such that

$$\mathbf{p}_2 = \mathbf{p}_1 + \mu\mathbf{e} \quad \text{and} \quad \mathbf{p}_3 = \mathbf{p}_1 + \nu\mathbf{e} + \phi\mathbf{f}.$$

We then construct the proper orthogonal

$$\mathbf{Q} = \mathbf{e}_1\mathbf{e}^T + \mathbf{e}_2\mathbf{f}^T + \mathbf{e}_3\mathbf{g}^T \quad \text{and} \quad \mathbf{r} = -\mathbf{Q}\mathbf{p}_1,$$

where $\mathbf{e}_j$ is the $j$th column of the the 3-by-3 $\mathbf{I}$. And set $\mathbf{p}^* = \mathbf{Q}(\mathbf{p} - \mathbf{p}_1)$ and check that indeed $\mathbf{p}_1^* = \mathbf{Q}(\mathbf{p}_1 - \mathbf{p}_1) = 0$ and $\mathbf{p}_2^* = \mathbf{Q}(\mathbf{p}_2 - \mathbf{p}_1) = \mu\mathbf{Q}\mathbf{e} = \mu\mathbf{e}_1$ and $\mathbf{p}_3^* = \mathbf{Q}(\mathbf{p}_3 - \mathbf{p}_1) = \mathbf{Q}(\nu\mathbf{e} + \phi\mathbf{f}) = \nu\mathbf{e}_1 + \phi\mathbf{e}_2$. This point is unique in $\text{Euc}(\mathbf{p})$, as the rigid motion that carries the first placement to the second must leave nodes 1 and 2 on the x-axis; a rotation which carries node 3 into a new position must leave the axis fixed and hence move node 3 from the plane. End of Proof.

**Definition** 5.23. For a given set of nodes 1,2,3, the class of all placements have the properties in Prop. 5.22 is denoted **Rep**.

**Proposition** 5.24. **Rep** is an affine subspace of $\mathbb{R}^{3N}$ with tangent space

$$\mathcal{U} = \{\mathbf{v} \in \mathbb{R}^{3N} : \mathbf{v}_1 = 0,\ \mathbf{v}_2^T\mathbf{e}_2 = \mathbf{v}_2^T\mathbf{e}_3 = 0,\ \mathbf{v}_3^T\mathbf{e}_3 = 0\}. \tag{5.26}$$

**Proposition** 5.25. For any $\mathbf{p}^* \in \mathbf{Rep}$, $\mathcal{U} \oplus \check{V}_R = \mathbb{R}^{3N}$, that is

$$\mathcal{U} + \check{V}_R = \mathbb{R}^{3N} \quad \text{and} \quad \mathcal{U} \cap \check{V}_R = \emptyset.$$

Proof: Given $\mathbf{v} \in \mathbb{R}^{3N}$ we construct $\mathbf{u} \in \mathcal{U}$ and $\mathbf{r} \in \check{V}_R$ such that $\mathbf{u} + \mathbf{r} = \mathbf{v}$. From the three conditions on $\mathbf{u}$ it follows that

$$\mathbf{r}_1 = \mathbf{v}_1,\ \mathbf{r}_2(2) = \mathbf{v}_2(2),\ \mathbf{r}_2(3) = \mathbf{v}_2(3) \quad \text{and} \quad \mathbf{r}_3(3) = \mathbf{v}_3(3). \tag{5.27}$$

Now as $\mathbf{r} \in \check{V}_R(\mathbf{p}^*)$ we need $\mathbf{r} = \mathbf{W}\mathbf{p}^* + \mathbf{b}$ where $\mathbf{W}$ is skew

$$\mathbf{W} = w_3\mathbf{e}_1 \times \mathbf{e}_2 + w_2\mathbf{e}_1 \times \mathbf{e}_3 + w_1\mathbf{e}_2 \times \mathbf{e}_3.$$

Now, as $\mathbf{p}_1^* = 0$ we find $\mathbf{b} = \mathbf{r}_1 = \mathbf{v}_1$. Next $\mathbf{p}_2^* = \lambda\mathbf{e}_x$ with $\lambda > 0$ brings

$$\mathbf{r}_2 = \mathbf{W}\mathbf{p}_2^* + \mathbf{b} = \lambda\mathbf{W}\mathbf{e}_1 + \mathbf{v}_1 = -\lambda w_z\mathbf{e}_3 - \lambda w_y\mathbf{e}_3 + \mathbf{v}_1.$$

To reconcile this with (5.27) brings

$$w_z = (v_1(2) - v_2(2))/\lambda \quad \text{and} \quad w_y = (v_1(3) - v_2(3))/\lambda.$$

Similarly, $\mathbf{p}_3^* = \mu\mathbf{e}_1 + \nu\mathbf{e}_2$ with $\nu > 0$ so

$$\mathbf{r}_3 = \nu w_z\mathbf{e}_1 - \mu w_z\mathbf{e}_2 - (\mu w_y + \nu w_x)\mathbf{e}_3 + \mathbf{v}_1$$

which we reconcile with (5.27) and solve for

$$w_x = \frac{v_1(3) - v_3(3) - \mu w_y}{\nu}.$$

It follows that $\mathbf{u} \equiv \mathbf{v} - \mathbf{r} \in \mathcal{U}$. End of Proof.

**Proposition** 5.26. A placement $\mathbf{p}^* \in$ Rep is stable iff there are no admissible motions from $\mathbf{p}^*$ that remain in Rep.

Proof: If $\mathbf{p}^*$ is stable then the only admissible motions starting from $\mathbf{p}^*$ are rigid motions. But no rigid motions stay in Rep.

If $p^*$ is not stable then there exists a non-rigid admissible $q(t)$ from $\mathbf{p}^*$. But we can use

$$\mathbf{q}^*(t) = \mathbf{Q}(t)(\mathbf{q}(t) - \mathbf{q}_1(t)) \tag{5.28}$$

to construct an equivalent motion from $\mathbf{p}^*$. It is admissible, since the rigid mappings used in the construction all conserve lengths. Existence of this motion will show $\mathbf{p}^*$ unstable, once we verify that the construction does not create a constant-valued motion. But were it constant,

$$\mathbf{q}^*(t) = \mathbf{Q}(t)(\mathbf{q}(t) - \mathbf{q}_1(t)) = \mathbf{p}^*$$

and hence $q(t) = \mathbf{Q}(t)^T\mathbf{p}^* + \mathbf{q}_1(t)$ would be a rigid motion. End of Proof.

**Proposition** 5.27. A placement $\mathbf{p}$ is stable iff its equivalent $\mathbf{p}^* \in$ Rep is stable.

Proof: Suppose the motion $\mathbf{q}(t)$ starts from $\mathbf{p}$ and is admissible but not rigid. We convert it into a mtion in Rep from $\mathbf{p}^*$.

End of Proof.

**Proposition** 5.28. If $\mathbf{p}^* \in$ Rep is rigidly equivalent to $\mathbf{p}$ with rotation $\mathbf{Q}$, and if velocities and accelerations are related by (*** and (*** then

$$\mathbf{B}_e\mathbf{p}^* \cdot \mathbf{v}^* = \mathbf{B}_e\mathbf{p} \cdot \mathbf{v}$$

and

$$\mathbf{B}_e\mathbf{v}^* \cdot \mathbf{v}^* + \mathbf{B}_e\mathbf{p}^* \cdot \mathbf{a}^* = \mathbf{B}_e\mathbf{v} \cdot \mathbf{v} + \mathbf{B}_e\mathbf{p} \cdot \mathbf{a}.$$

Hence the placement $\mathbf{p}$ is second order stable iff its equivalent $\mathbf{p}^* \in$ Rep is second order stable.

### 5.6. Notes and Exercises

Our discussion of convex sets follows Ben-Tal and Nemirovski (2001). For a simpler proof of the Mixing Theorem and an introduction to the broader field of rearrangements see **?**.

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 $v \in \mathbb{R}^n$ 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 $A$?
   (d) Show that if $\|v\| = 1$ then $A^2 = A$ and $\mathrm{tr}(A) = 1$.

7. Show that invariant subspaces are indeed subspaces. That is, given a vector space $V$ and a linear transformation $A$ from $V$ to $V$ and a set $U \subset V$ such that for each $u \in U$ it follows that $Au \in U$, show that $U$ is a subspace.

8. Show that the $w_1$ and $w_2$ in (5.4) are indeed invariant vectors of the $A$ in (5.6). That is, show that each $Aw_j$ is proportional to $w_j$.

9. The matrix

$$R_{e_1,\theta} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix}$$

   rotates each points in $\mathbb{R}^3$ by $\theta$ radians about the $e_1 = (1, 0, 0)$ axis. If instead we wish to rotate points about the unit vector $v_1$ then we choose a unit vector, $v_2$, orthogonal to $v_1$ and then use the cross product, $v_3 = v_1 \times v_2$ to define the third. Consider the matrix $V = [v_1\ v_2\ v_3]$ and show that
   (a) $V^T V = I$, and so, as $v_j = Ve_j$ it follows that $e_j = V^T v_j$.
   (b) With $V$ as the change of basis matrix show that

$$R_{v_1,\theta} = V R_{e_1,\theta} V^T, \tag{5.29}$$

   is indeed rotation about $v_1$ by $\theta$.

10. Let $\wp_n$ denote the set of complex polynomials of degree $n$,

$$\wp_n \equiv \{c_1 z^n + c_2 z^{n-1} + \cdots + c_n z + c_{n+1} : c_j \in \mathbb{C}\}.$$

    Let $D$ denote differentiation with respect to $z$.

(a) Show that $D$ is a linear transformation from $\wp_n$ to $\wp_{n-1}$.

(b) Choose bases for $\wp_n$ and $\wp_{n-1}$ from the columns of the identity matrices of order $n+1$ and $n$ respectively and express the matrix representation of $D$.

11. Show that $\mathrm{DS}_n$, the set of doubly stochastic $n$-by-$n$ matrices, defined in (5.8), is convex and closed.

12. Show that the ball $B \equiv \{v \in \mathbb{R}^n : \|v\| \leq 1\}$ is convex and closed. Show that its set of extreme points is $S \equiv \{v \in \mathbb{R}^n : \|v\| = 1\}$. For each $v \in S$ show that $\{y \in \mathbb{R}^n : v^T y = 1\}$ is the supporting hyperplane to $B$ at $v$.

13. Generalize the previous exercise by supposing that $A \in \mathbb{R}^{n\times n}$ is symmetric and positive definite and $c > 0$. Show that the $B \equiv \{v \in \mathbb{R}^n : v^T A v \leq c\}$ is convex and closed and that its set of extreme points is $S \equiv \{v \in \mathbb{R}^n : v^T A v = c\}$.

14. Show that $\{A \in \mathbb{R}^{n\times n} : |\mathrm{tr}\, A| \leq 1\}$ is convex. Draw this set when $n = 2$ and identify its four extreme points.

15. In applications to metabolic networks we will see that its natural to suppose that each of our unknowns is nonnegative. Show that if $S \in \mathbb{R}^{m\times n}$ and $f \in \mathbb{R}^m$ then $\{v \in \mathbb{R}^n : Sv = f,\ v \geq 0\}$ is a polyhedron.

16. Prove that $\mathrm{Shift}(\mathrm{Aspan}(C))$, as defined in (5.13), is a subspace – and that this subspace does not depend on the choice of $s$.

17. Show that $\mathrm{Adim}(\mathrm{DS}_n) = (n-1)^2$.

18. The Mixing Theorem states that rearrangements with similar order maximize the inner product. Show that rearrangements with opposite order minimize the inner product.

# 6. Least Squares

We learned in the previous chapter that for a given $A \in \mathbb{R}^{m\times n}$, the equation $Ax = b$ need not possess a solution when the number of rows of $A$ exceeds its rank, i.e., when $r < m$. We detail two natural engineering contexts, where material parameters are to be inferred from experiment, in which the governing $A$ matrix has $r < m$, and offer probabilistic and statistical interpretations. To resolve the inconsistent system $Ax = b$ we project the faulty $b$ into the column space of $A$ and proceed to solve the consistent, so-called normal equations, $A^TAx = A^Tb$. We then develop the associated theory of projection matrices and show how this permits us to transform linearly independent collections of vectors into orthonormal collections.

This theory has many, far reaching applications. Among these we have chosen to focus on Orthogonal Polynomials, Detecting Integer Relations, and constructing autoregressive models of stationary time series.

### 6.1. The Normal Equations

When faced with a matrix $A$ and vector $b \notin \mathcal{R}(A)$, 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

$$\|Ax - b\|^2 \equiv (Ax - b)^T(Ax - b) \tag{6.1}$$

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

$$\|Ax - b\|^2 = \|Ax - b_R - b_N\|^2 = \|Ax - b_R\|^2 + \|b_N\|^2, \tag{6.2}$$

As $b_N$ is what it is, (6.2) states that the best $x$ is the one that satisfies

$$Ax = b_R. \tag{6.3}$$

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 (6.3) then

$$Ax - b = Ax - b_R - b_N = -b_N. \tag{6.4}$$

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 (6.4) however is that $(Ax - b) \in \mathcal{N}(A^T)$, i.e., $A^T(Ax - b) = 0$, i.e.,

$$\boxed{A^TAx = A^Tb.} \tag{6.5}$$

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 Exer. 4.10, we note that this is equivalent to $\mathcal{N}(A) = \{0\}$. We summarize our findings in

**Proposition** 6.1. Suppose that $A \in \mathbb{R}^{m\times n}$. The set of $x \in \mathbb{R}^n$ for which the misfit $\|Ax - b\|^2$ is smallest is composed of those $x$ for which

$$\boxed{A^T Ax = A^T b.}$$

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

As a concrete example, we take

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}. \tag{6.6}$$

and plot $\mathcal{R}(A)$ and the associated decomposition of $b$ in Figure 6.1.

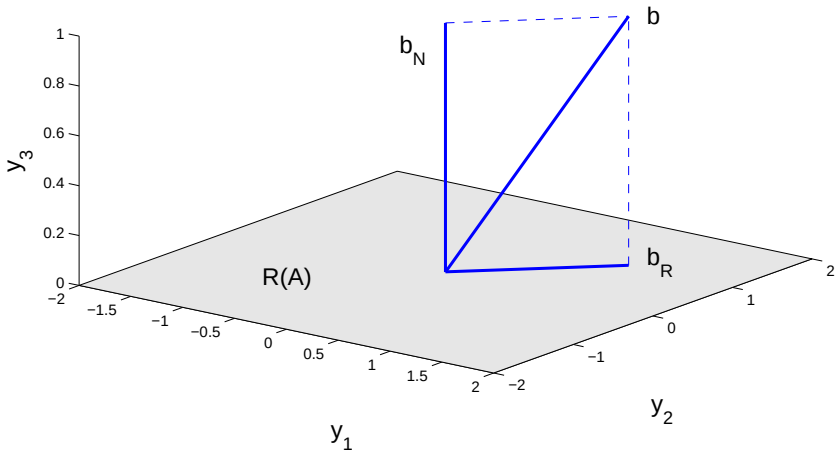



**Figure** 6.1. Decomposition of $b = b_R + b_N$ into its column space and left null space components.

As this $b \neq \mathcal{R}(A)$ 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 (6.5), 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 $A$.

We consider a more typical example, attempting to determine the neuronal membrane conductance, $G$, and reversal potential, $E$, by fitting Ohm's law for the putative neuronal membrane current,

$$b = G(v - E), \tag{6.7}$$

where we have $m$ current measurements, $b_j$, at $m$ prescribed voltage levels, $v_j$. The key step is to recognize (6.7) as an $m$-by-2 system of equations for the unknown biophysical parameters $x_1 = G$ and $x_2 = GE$, via

$$v_j x_1 - x_2 = b_j, \quad j = 1, \ldots, m$$

and to translate this into the least squares problem $Ax = b$ where

$$A = \begin{pmatrix} v_1 & -1 \\ v_2 & -1 \\ \vdots & \vdots \\ v_m & -1 \end{pmatrix}.$$

We illustrate this in Figure 6.2 on noisy synthetic data generated by the "true values", $G = 1.2\,mS$ and $E = -70\,mV$. We suppose we have 21 accurate voltage measurements, between $-75\,mV$ and $-65\,mV$ in steps of $0.5\,mV$. We then generate the 21 true values of $b$ via (6.7) and then soil these with `randn` noise.

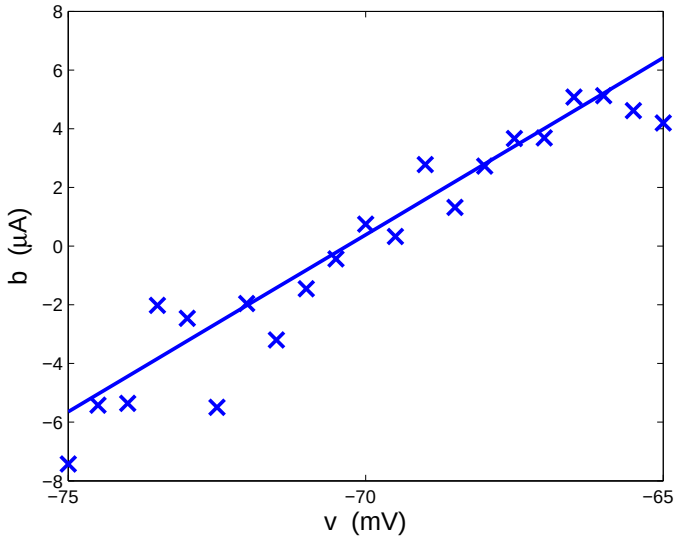



**Figure** 6.2. The 'X' points are the synthetic noisy measurements. On solving the associated least squares problem, $A^T Ax = A^T b$ we "recover" $G = 1.18$ and $E = -68.75$, which when used in (6.7), yields the solid straight line.

### 6.2. Application to a Biaxial Test Problem

We progress from identifying 2 electrical parameters from noisy voltage–current measurements to identifying 20 fiber stiffness in Figure 3.5 from noisy force–displacement measurements.

We envision loading the 9 nodes with a known force vectors, $f \in \mathbb{R}^{18}$, and measuring the associated 18 displacements, $x$. From knowledge of $x$ and $f$ we wish to infer the twenty components

of $K = \texttt{diag}(k)$ where $k$ is the vector of unknown fiber stiffnesses. The first, and most important, step is to recognize that

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

may be written as

$$Bk = f \quad \text{where} \quad B = A^T \texttt{diag}(Ax). \tag{6.8}$$

Though conceptually simple this is not of great use in practice, for $B$ is 18-by-20 and hence (6.8) possesses many solutions. The way out (as in our previous example) is to conduct more experiments. We will 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

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

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

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

respectively. For the purpose of our example we suppose that each $k_j = 1$ except $k_8 = 5$. We assemble $A^T K A$ as in Chapter 3 and solve

$$A^T K A x^{(j)} = f^{(j)}$$

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

$$B^T B k = B^T f$$

we note that MATLAB solves this system when presented with `k=B\f` when $B$ is rectangular. We have illustrated the results of this procedure in Figure 6.3.

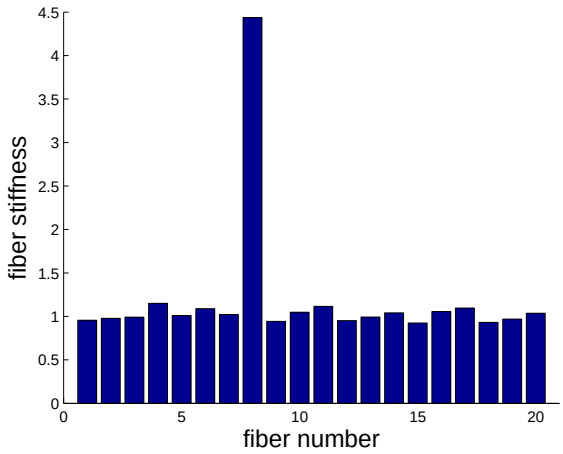



**Figure** 6.3. Results of a successful biaxial test.

We see that even in the presense of noise that the stiff fiber is readily identified.

### 6.3. Projections

From an algebraic point of view (6.5) 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 (6.4). 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

$$b_R = Ax = A(A^T A)^{-1} A^T b. \tag{6.9}$$

Defining

$$P = A(A^T A)^{-1} A^T, \tag{6.10}$$

(6.9) 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

$$P^2 = P. \tag{6.11}$$

With respect to (6.10) 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 (6.10) is symmetric. We dignify these properties through

**Definition** 6.2. 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 inner 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 an example, for the $A$ of (6.6) we find

$$P = A(A^T A)^{-1} A^T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

as suggested by Figure 6.1. It is very useful to even develop expressions for the projection onto a line. In this case $A$ is the single column, let us denote it by $a$, and the associated projection matrix is the scaled outer product

$$P = aa^T/(a^Ta). \tag{6.12}$$

For example,

$$P = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \tag{6.13}$$

is orthogonal projection onto the line through $a = (1, 1)$. We illustrate this in Figure 6.4.

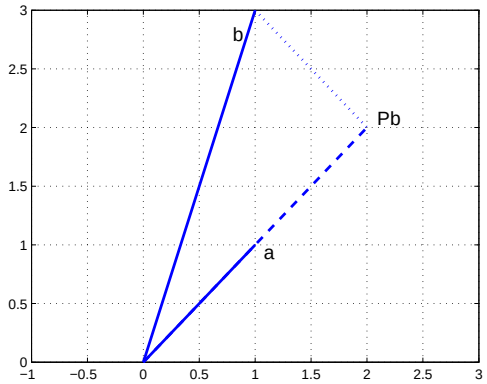



**Figure** 6.4. The projection of $b = (1, 3)$ onto the line through $a = (1, 1)$ via the projection matrix, (6.13).

### 6.4. The QR Decomposition

While linear independence remains a central concept we shall see that representations are often more concise and computations are always more robust when the relevant basis vectors are chosen to be "maximally independent," i.e., orthogonal. Projections are a natural means for transforming a basis for a space $M$ to an orthonormal basis for $M$.

This process, known as the **Gram–Schmidt Procedure**, takes $n$ basis vectors, $x_j$, for a subspace $M$ and returns $n$ orthonormal vectors, $q_j$, in $M$.

**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

$$y_2 = x_2 - q_1(q_1^Tq_1)^{-1}q_1^Tx_2 = x_2 - q_1q_1^Tx_2.$$

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

$$\begin{aligned} y_3 &= x_3 - Q_2(Q_2^TQ_2)^{-1}Q_2^Tx_3 \\ &= x_3 - q_1q_1^Tx_3 - q_2q_2^Tx_3. \end{aligned}$$

Set $q_3 = y_3/\|y_3\|$ and $Q_3 = [q_1\ q_2\ q_3]$.
Continue in this fashion through step

**GSn:** $y_n = x_n$ minus its projection onto the subspace spanned by the columns of $Q_{n-1}$. That is

$$\begin{aligned} y_n &= x_n - Q_{n-1}(Q_{n-1}^T Q_{n-1})^{-1} Q_{n-1}^T x_n \\ &= x_n - \sum_{j=1}^{n-1} q_j q_j^T x_n. \end{aligned}$$

Set $q_n = y_n/\|y_n\|$ and $Q = [q_1, q_2, \ldots, q_n]$.

As the resulting $Q \in \mathbb{R}^{m\times n}$ has orthonormal columns it follows that $Q^TQ = I$. We call such a $Q$ an **orthogonal matrix**. It follows that if $m = n$ then $Q^T = Q^{-1}$.

To take a simple example, let us orthogonalize the following basis for $\mathbb{R}^3$,

$$x_1 = [1\ 0\ 0]^T, \quad x_2 = [1\ 1\ 0]^T, \quad x_3 = [1\ 1\ 1]^T. \tag{6.14}$$

**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 the canonical basis

$$q_1 = [1\ 0\ 0]^T, \quad q_2 = [0\ 1\ 0]^T, \quad q_3 = [0\ 0\ 1]^T. \tag{6.15}$$

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 of steps **GS1**–**n**. As a result, there is no guarantee that it will return the same basis. For example, here is a MATLAB diary of `orth` applied to (6.15),

```
>> 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. In fact, we can often get by just knowing that an orthonormal basis exists. For example, lets show that orthogonal bases permit us to easily "see" the rank of a projection.

**Proposition** 6.3. If $P = P^2$ then $\mathrm{tr}(P) = \mathrm{rank}(P)$.

Proof: We suppose that $P \in \mathbb{R}^{n\times n}$ and denote by $r$ the rank of $P$. We suppose that $\{q_1, \ldots, q_r\}$ is an orthonormal basis for $\mathcal{R}(P)$ and that $\{q_{r+1}, \ldots, q_n\}$ is an orthonormal basis for $\mathcal{N}(P^T)$.

We set $Q = [q_1\ q_2\ \cdots q_n]$ and proceed to compute the trace of $Q^TPQ$. We note that the diagonal element $(Q^TPQ)_{j,j} = q_j^T P q_j$. If $j \le r$ we find $Pq_j = q_j$ and so $(Q^TPQ)_{j,j} = 1$ while if $j > r$ then $q_j^T P = 0$ and hence $(Q^TPQ)_{j,j} = 0$. It follows that $\mathrm{tr}(Q^TPQ) = r$. To connect this to $\mathrm{tr}(P)$ we invoke the product formula, Eq. (1.14), and find

$$\mathrm{tr}(Q^TPQ) = \mathrm{tr}(Q^TQP) = \mathrm{tr}(IP) = \mathrm{tr}(P), \tag{6.16}$$

where we've used the fundamental theorem of linear algebra to ensure that $Q^TQ = I$. End of Proof.

A more concrete use for the Gram–Schmidt Procedure will follow from viewing it as a factorization of $X$. More precisely, we wish to interpret the procedure as expressing each $x_j$ as a linear combination of the $q_i$ for $i \leq j$. This is simple for $j = 1$, namely **GS1** states

$$x_1 = (q_1^T x_1) q_1. \tag{6.17}$$

Unpacking **GS2** we next find that

$$x_2 = (q_1^T x_2) q_1 + \|x_2 - (q_1^T x_2) q_1\| q_2, \tag{6.18}$$

is indeed a linear combination of $q_1$ of $q_2$. The awkward norm term can be reduced by simply taking the inner product of each side of (6.18) with $q_2$. As $q_i^T q_j = \delta_{ij}$ this yields

$$q_2^T x_2 = \|x_2 - (q_1^T x_2) q_1\|$$

and so, in fact (6.18) takes the form

$$x_2 = (q_1^T x_2) q_1 + (q_2^T x_2) q_2. \tag{6.19}$$

We next continue this line of argument and find that

$$x_j = (q_1^T x_j) q_1 + (q_2^T x_j) q_2 + \cdots + (q_j^T x_j) q_j. \tag{6.20}$$

for each $j$ up to $n$. That each $x_j$ is expressed in terms of $q_i$ for $i \leq j$ manifests itself in a triangular decomposition. Namely, we recognize that (6.17), (6.19) and (6.20), when collected, take the form

$$[x_1, x_2, \ldots, x_n] = [q_1, q_2, \ldots, q_n] \begin{pmatrix} q_1^T x_1 & q_1^T x_2 & \cdots & \cdots & q_1^T x_n \\ 0 & q_2^T x_2 & q_2^T x_3 & \cdots & q_2^T x_n \\ 0 & 0 & q_3^T x_3 & \cdots & q_3^T x_n \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & q_n^T x_n \end{pmatrix}.$$

This matrix of $q_j$ is simply the $Q$ produced by Gram–Schmidt. The remaining upper triangular matrix is typically denoted $R$. As none of its diagonal elements may vanish, it is invertible. We have now established

**Proposition** 6.4. If $X \in \mathbb{R}^{m \times n}$ has linearly independent columns then there exists an orthonormal $Q \in \mathbb{R}^{m \times n}$ and a nonsingular upper triangular $R \in \mathbb{R}^{n \times n}$ such that

$$\boxed{X = QR.} \tag{6.21}$$

This result offers a potentially dramatic rewrite of the least squares problem, $Ax = b$. For recall that if $x \neq \mathcal{R}(A)$ we instead must solve the normal equations, $A^T Ax = A^T b$. Of course we may solve this via the LU factorization of Chapter 3. If we instead factor $A = QR$ then the normal equations become $R^T Rx = R^T Q^T b$. However, $R^T$ inherits its nonsingularity from $R$ and so we may multiply each side by $(R^T)^{-1}$ and arrive at the reduced normal equations

$$\boxed{Rx = Q^T b.} \tag{6.22}$$

As $R$ is upper triangular, this may be solved by a single sweep of back substitution, without ever even having to construct $A^T A$. We note that MATLAB generates $Q$ and $R$ via its `qr` function and that it indeed solves (6.22) when confronted with the least squares problem `x=A\b`.

To take a concrete example, with the $A$ and $b$ of (6.6) we find

$$Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad R = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad Q^T b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

and so (6.22) takes the form

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

and $x_2 = 1$ and $x_1 = 0$ as expected.

### 6.5. Orthogonal Polynomials*

We consider the polynomials

$$e_0(x) = 1, \quad e_1(x) = x, \ldots, \ e_n(x) = x^n \tag{6.23}$$

and note they constitute a basis for their span, the space of polynomials of degree $n$,

$$\mathcal{P}_n \equiv \{a_0 e_0(x) + a_1 e_1(x) + \cdots + a_n e_n(x) : a \in \mathbb{R}^{n+1}\}.$$

Given a function, say $f$, over some interval, say $[-1, 1]$, it is very common to attempt to approximate $f$ by a member of $\mathcal{P}_n$ by solving the least squares problem

$$\min_{p \in \mathcal{P}_n} \int_{-1}^{1} (f(x) - p(x))^2 \, dx. \tag{6.24}$$

As suggested by the last section, it might be considerably simpler if the basis vectors of $\mathcal{P}_n$ were orthogonal in a sense consistent with the criterion in Eq. (6.24). More precisely, we consider the **inner product**

$$\langle p, q \rangle \equiv \int_{-1}^{1} p(x) q(x) \, dx, \tag{6.25}$$

and use it to orthogonalize the $e_k$ of Eq. (6.23).

$$\begin{aligned} q_0(x) &= e_0(x) = 1 \\ q_1(x) &= e_1(x) - \langle e_1, q_0 \rangle q_0 / \langle q_0, q_0 \rangle = x \\ q_2(x) &= e_2(x) - \langle e_2, q_1 \rangle q_1 / \langle q_1, q_1 \rangle - \langle e_2, q_0 \rangle q_0 / \langle q_0, q_0 \rangle = x^2 - 1/3 \\ q_3(x) &= e_3(x) - \langle e_3, q_2 \rangle q_2 / \langle q_2, q_2 \rangle - \langle e_3, q_1 \rangle q_1 / \langle q_1, q_1 \rangle - \langle e_3, q_0 \rangle q_0 / \langle q_0, q_0 \rangle = x^3 - 3x/5. \end{aligned} \tag{6.26}$$

These are called the **Legendre polynomials** . We illustrate the first few nonconstants in Figure 6.5(A).

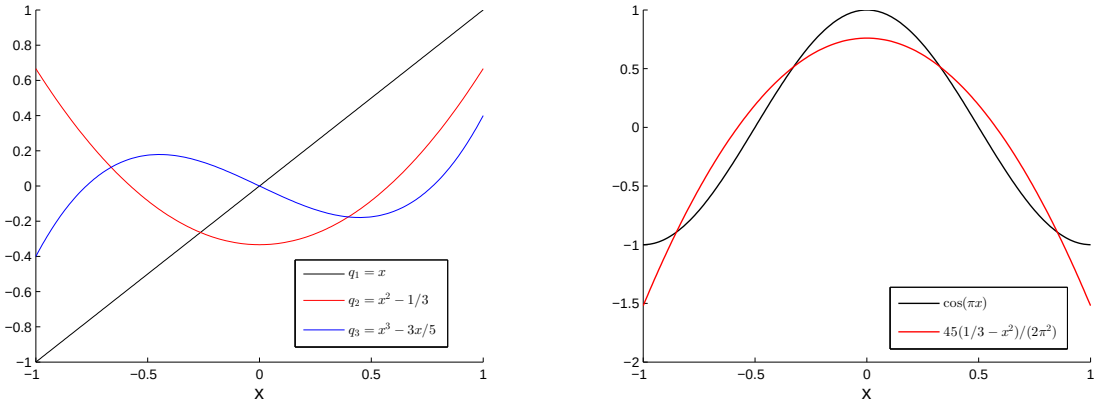



**Figure** 6.5. (A) The first three interesting Legendre polynomials. (B) The best quadratic approximation to $\cos(\pi x)$.

For example, with $f(x) = \cos(\pi x)$ and $p \in \mathcal{P}_2$ we find that

$$\begin{aligned}\int_{-1}^{1} (\cos(\pi x) - p(x))^2\, dx &= \int_{-1}^{1} (\cos(\pi x) - (a_0 + a_1 x + a_2(x^2 - 1/3))^2\, dx \\ &= 1 + (8/\pi^2)a_2 + 2a_0^2 + (2/3)a_1^2 + (8/45)a_2^2\end{aligned}$$

takes its minimum at $a_0 = a_1 = 0$ and $a_2 = -45/(2\pi^2)$. We illustrate the fit in Figure 6.5(B).

We now develop two other approaches to constructing these polynomials.

**Proposition** 6.5. Commencing from $q_{-1}(x) = 0$ and $q_0(x) = 1$ the Legendre polynomials obey the three term recurrence

$$q_{n+1}(x) = (x - \alpha_{n+1})q_n(x) - \beta_n q_{n-1}(x) \tag{6.27}$$

where

$$\alpha_{n+1} = \frac{\langle e_1 q_n, q_n\rangle}{\langle q_n, q_n\rangle} \quad \text{and} \quad \beta_n = \frac{\langle q_n, q_n\rangle}{\langle q_{n-1}, q_{n-1}\rangle}$$

Proof: As the $q_n$ are each monic it follows that $q_{n+1} - xq_n$ is of order no more than $n$ and so may be expressed as a linear combination of $q_0$ through $q_n$. Namely

$$q_{n+1}(x) - xq_n(x) = -\alpha_{n+1}q_n - \beta_n q_{n-1} + \sum_{m=0}^{n-2} a_m q_m(x). \tag{6.28}$$

On taking the inner product of each side of Eq. (6.28) with $q_n$ we find

$$\langle e_1 q_n, q_n\rangle = \alpha_{n+1}\langle q_n, q_n\rangle.$$

On taking the inner product of each side of Eq. (6.28) with $q_{n-1}$ we find

$$\langle e_1 q_n, q_{n-1}\rangle = \beta_n \langle q_{n-1}, q_{n-1}\rangle.$$

On taking the inner product of each side of Eq. (6.28) with $q_{n-2}$, we find $0 = a_{m-2}\langle q_{n-2}, q_{n-2}\rangle$ and so $a_{m-2} = 0$. Now moving through decreasing indicies this argument establishes that $a_m = 0$, $m = 0, \ldots, n-2$. It follows that Eq. (6.28) reveals Eq. (6.27). End of Proof.

We note that $\langle e_1 q_m, q_m \rangle = 0$ by oddness and so each $\alpha_n = 0$ while

$$\beta_n = \frac{n^2}{4n^2 - 1}$$

and so

$$q_{n+1}(x) = xq_n(x) - \frac{n^2}{4n^2 - 1} q_{n-1}(x).$$

There are a number of ways to normalize beyond monic. One standard choice is to ask that each polynomial obey $p_n(1) = 1$. We note that

$$q_n(1) = \frac{n}{2n-1} q_{n-1}(1)$$

so if

$$P_n(x) = \frac{(2n-1)!!}{n!} q_n(x) \tag{6.29}$$

(where !! denotes factorial through the odds) will obey $P_n(1) = 1$. What happens to the recurrence?

$$\frac{(n+1)!}{(2n+1)!!} P_{n+1}(x) = \frac{n!}{(2n-1)!!} xP_n(x) - \frac{n!n}{(2n+1)!!} P_{n-1}(x).$$

multiply through by $(2n+1)!!/n!$ to get

$$\boxed{(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x).} \tag{6.30}$$

This recurrence relation permits deeper study of the $P_n$. For example, it will yield an explicit formula for the **generating function**

$$G(x,t) \equiv \sum_{n=0}^{\infty} P_n(x) t^n \tag{6.31}$$

of the $P_n$. We begin our calculation of $G$ by noting that

$$G(1,t) \equiv \sum_{n=0}^{\infty} t^n = \frac{1}{1-t} \tag{6.32}$$

and then using (6.30) to write

$$\sum_{n=1}^{\infty} (nP_n(x) - (2n-1)xP_{n-1}(x) + (n-1)P_{n-2}(x)) t^n = 0. \tag{6.33}$$

The basic idea is to translate this expression into a differential equation for $G$. In what follows we will use $t$ is a subscript to denote differeniation with respect to $t$. With this we note that the first term in (6.33) is actually

$$\sum_{n=1}^{\infty} nP_n(x) t^n = t \sum_{n=1}^{\infty} nP_n(x) t^{n-1} = tG_t(x,t).$$

To get our hands on the second term we record

$$\sum_{n=1}^{\infty} (n-1)P_{n-1}(x) t^n = t^2 \sum_{n=1}^{\infty} (n-1)P_{n-1}(x) t^{n-2} = t^2 G_t(x,t)$$

and

$$\sum_{n=1}^{\infty} P_{n-1}(x)t^n = t\sum_{n=1}^{\infty} P_{n-1}(x)t^{n-1} = tG(x,t)$$

and so deduce

$$\begin{aligned} x\sum_{n=1}^{\infty}(2n-1)P_{n-1}(x)t^n &= 2x\sum_{n=1}^{\infty}(n-1)P_{n-1}(x)t^n + x\sum_{n=1}^{\infty}P_{n-1}(x)t^n \\ &= 2xt^2G_t(x,t) + xtG(x,t). \end{aligned}$$

Proceeding similarly, the last term in (6.33) becomes

$$\begin{aligned} \sum_{n=1}^{\infty}(n-1)P_{n-2}(x)t^n &= t^3\sum_{n=1}^{\infty}(n-2)P_{n-2}(x)t^{n-3} + t^2\sum_{n=1}^{\infty}P_{n-2}(x)t^{n-2} \\ &= t^3G_t(x,t) + t^2G(x,t). \end{aligned}$$

It follows that (6.33) is equivalent to $tG_t - 2xt^2G_t - xtG + t^3G_t + t^2G = 0$ or, after collecting terms,

$$(1-2xt+t^2)G_t + (t-x)G = 0. \tag{6.34}$$

To find a nonzero $G$ we separate the knowns from the unknowns in (6.34),

$$\frac{G_t}{G} = -\frac{1}{2}\frac{a_t}{a} \quad \text{where} \quad a(x,t) = 1-2xt+t^2.$$

We recognize each side to be a logarithmic derivative, i.e.,

$$(\log G)_t = -\frac{1}{2}(\log a)_t \quad \text{so, upon integration} \quad \log G = -(1/2)\log a + c,$$

for some constant $c$. On taking the exponential of each side we find $G(x,t) = a^{-1/2}(x,t)\exp(c)$. On setting $x=1$ and using (6.32) we find $c=0$ and so

$$\boxed{G(x,t) = \frac{1}{\sqrt{1-2xt+t^2}}.} \tag{6.35}$$

We will see in the exercises that this simple explicit generating function satisfies a beautiful partial differential equation that in turn implies that the $P_n$ themselves satisfy a family of ordinary differential equations.

We frequently work with a **weighted** inner product,

$$\langle f,g\rangle_w \equiv \int_{-1}^{1}\frac{f(x)g(x)}{\sqrt{1-x^2}}\,dx \tag{6.36}$$

This places much greater weight at the ends. Under the change of variable $x=\cos\phi$ this takes the more reminiscent form

$$\langle f,g\rangle_w \equiv \int_0^{\pi} f(\cos\phi)g(\cos\phi)\,d\phi. \tag{6.37}$$

So we denote by $c_n$ the orthogonalization of the standard basis with respect to Eq. (6.37) we find

$$c_0(x)=1, \quad c_1(x)=x, \quad c_2(x)=x^2-1/2, \quad \text{and} \quad c_3(x)=x^3-3x/4c_4(x)=x^4-x^2+1/8.$$

Regarding recurrence we note (by symmetry) again that $\alpha_{n+1} = 0$ while

$$\beta_1 = 1/2, \quad \text{and} \quad \beta_n = 1/4, \quad \text{for} \quad n > 1,$$

and so the $c_n$ obey

$$c_{n+1}(x) = xc_n(x) - (1/4)c_{n-1}(x), \quad \text{for} \quad n > 1,$$

Now $c_n(1) = (1/2)c_{n-1}(1)$ and so if we define

$$T_n(x) = 2^{n-1}c_n(x)$$

then we arrive at **Chebyshev polynomials** of the first kind

$$T_1(x) = 2x, \quad T_2(x) = 4x^2 - 1, \quad T_3(x) = 4x^3 - 3x \quad \text{and} \quad T_4(x) = 8x^4 - 8x^2 + 1$$

and they obey the recurrence

$$2^{1-(n+1)}T_{n+1}(x) = x2^{1-n}T_n(x) - (1/4)2^{1-(n-1)}T_{n-1}(x), \quad \text{for} \quad n > 1,$$

that is

$$\boxed{T_{n+1} = 2xT_n(x) - T_{n-1}(x).} \tag{6.38}$$

The slight change, $U_0(x) = 1$ and $U_1(x) = 2x$ and the same recurrence brings us to the Chebyshev polynomials of the second kind.

$$U_2(x) = 4x^2 - 1 \quad \text{and} \quad U_3(x) = 8x^3 - 4x.$$

We compute their generating function

$$G(x,t) \equiv \sum_{n=0}^{\infty} U_n(x)t^n$$

by unpacking

$$\sum_{n=1}^{\infty}(U_n(x) - 2xU_{n-1}(x) + U_{n-2}(x))t^n = 0$$

for

$$\sum_{n=1}^{\infty} U_{n-1}(x)t^n = t\sum_{n=1}^{\infty} U_{n-1}(x)t^{n-1} = tG(x,t)$$

and

$$\sum_{n=1}^{\infty} U_{n-2}(x)t^n = t^2\sum_{n=1}^{\infty} U_{n-2}(x)t^n = t^2G(x,t)$$

and so all togther we find $G - 1 - 2xtG + t^2G = 0$ or that is

$$\boxed{G(x,t) = \frac{1}{1 - 2xt + t^2}.} \tag{6.39}$$

We will meet these Chebyshev polynomials, and their generating function, when studying trees and cycles in our closing chapter on graph theory.

### 6.6. Detecting Integer Relations*

The previous section demonstrated the value of $QR$ for resolving overdetermined systems. In this section we will discuss its use in an underdetermined system that arises in a beautiful algorithm for detecting integer relations between real numbers. We denote the set of integers by $\mathbb{Z}$ and the lattice of integers in $\mathbb{R}^n$ by $\mathbb{Z}^n$. Given an $x \in \mathbb{R}^n$ we call a nonzero $z \in \mathbb{Z}^n$ an **integer relation** for $x$ if

$$z^T x = 0. \tag{6.40}$$

In the plane one can find such relations by merely computing the greatest common divisor of the scaled entries. For example, to find $z \in \mathbb{Z}^2$ such that $3.1z_1 + 0.04z_2 = 0$ we multiply by 100 to clear fractions and arrive at $310z_1 + 4z_2 = 0$. We then compute $g = \gcd(310, 4) = 2$ and note that

$$z_1 = 4/g = 2 \quad \text{and} \quad z_2 = -310/g = -155$$

provide an integer relation. The gcd is computed via a classic division algorithm that first appeared in Euclid. Attempts to generalize this to higher dimensions have only recently succeeded – with the resulting algorithm recognized as one of the ten best of the 20th century, and its application in the right hands has detected some amazing new patterns - opening for example new vistas on $\pi$. We will use it to uncover the integers in putative functional relations like

$$z_1 \sin(5\theta) + z_2 \sin(\theta) + z_3 \sin^3(\theta) + z_4 \sin^5(\theta) = 0. \tag{6.41}$$

The algorithm is known as PSLQ where PS stands for Partial Sums and LQ for the the Lower Triangular Orthogonal Decomposition, a transposed variant of QR. We have found it more convenient to present the method as PSQR.

First, regarding partial sums, we assume without loss that $\|x\| = 1$ and no $x_j = 0$ (otherwise there is an obvious integer relation), build

$$s_j^2 = \sum_{k=j}^{n} x_k^2.$$

and use these to populate the $(n-1) \times n$ upper triangular matrix

$$U_x = \begin{pmatrix} \frac{s_2}{s_1} & \frac{-x_2 x_1}{s_1 s_2} & \frac{-x_3 x_1}{s_1 s_2} & \cdots & \cdots & \frac{-x_n x_1}{s_1 s_2} \\ & \frac{s_3}{s_2} & \frac{-x_3 x_2}{s_2 s_3} & \cdots & \cdots & \frac{-x_n x_2}{s_2 s_3} \\ & & \ddots & \ddots & \vdots & \vdots \\ & & & \frac{s_{n-1}}{s_{n-2}} & \frac{-x_{n-1} x_{n-2}}{s_{n-1} s_{n-2}} & \frac{-x_n x_{n-2}}{s_{n-1} s_{n-2}} \\ & & & & \frac{s_n}{s_{n-1}} & \frac{-x_n x_{n-1}}{s_n s_{n-1}} \end{pmatrix}. \tag{6.42}$$

This matrix enjoys two lovely identities

$$\boxed{U_x U_x^T = I_{n-1} \quad \text{and} \quad U_x^T U_x = I_n - xx^T} \tag{6.43}$$

on which the entire algorithm depends. To establish the first we first examine the diagonal term

$$U_x(i,:)U_x(i,:)^T = (s_{i+1}/s_i)^2 + \frac{1}{(s_{i+1}s_i)^2} + x_i^2 \sum_{k=i+1}^{n} x_k^2 = \frac{s_{i+1}^2}{s_i^2} + \frac{x_i^2}{s_i^2} = 1.$$

Regarding the off–diagonal terms we suppose, without loss, $i < j$ and find

$$U_x(i,:)U_x(j,:)^T = -\frac{x_j x_i s_{j+1}}{s_i s_{i+1} s_j} + \frac{x_i x_j}{s_i s_{i+1} s_j s_{j+1}} \sum_{k=j+1}^{n} x_k^2 = -\frac{x_j x_i s_{j+1}}{s_i s_{i+1} s_j} + \frac{x_i x_j s_{j+1}}{s_i s_{i+1} s_j} = 0.$$

This proves the first identity in (6.43). The second may be proven in much the same way and is left as an exercise. An immediate consequence of the second is that

$$P_x \equiv U_x^T U_x \tag{6.44}$$

is orthogonal projection onto the orthogonal complement to the line through $x$, and

$$P_x z = z \quad \text{for any integer relation} \quad z \in \mathbb{Z}^n \text{ for } x. \tag{6.45}$$

With this preparation we may establish

**Proposition** 6.6. Suppose that $A \in \mathbb{Z}^{n\times n}$ is invertible and that $U_x A = QR$ where $Q \in \mathbb{R}^{(n-1)\times(n-1)}$ is orthogonal, $R \in \mathbb{R}^{(n-1)\times n}$ is upper triangular with no zeros on its diagonal, and $U_x$ is given in (6.42). If $z$ is an Integer Relation for $x$ then

$$1 \le \|z\| R_{max} \tag{6.46}$$

where $R_{max}$ is the magnitude of the largest diagonal element of $R$.

Proof: From $U_x A = QR$ we deduce $P_x A = U_x^T U_x A = U_x^T QR$. So if $z$ is an Integer Relation for $x$ then by (6.44) and (6.45) it follows that $z^T = z^T P_x$ and $z^T A = z^T P_x A = z^T (U_x^T Q) R$. Now, as $A$ is invertible $z^T A \neq 0$. Let $j$ be the least $j$ for which $z^T A_{:,j} \neq 0$, so $z^T A_{:,k} = 0$ for $k < j$. If $j = 1$ skip to (6.47). If $j > 1$ note that, as $R$ is upper triangular,

$$0 = z^T A_{:,1} = z^T (U_x^T Q) R_{:,1} = r_{1,1} z^T (U_x^T Q)_{:,1},$$

and, as $r_{1,1} \neq 0$ we conclude that $z^T (U_x^T Q)_{:,1} = 0$. Continuing in this fashion,

$$0 = z^T A_{:,2} = z^T (U_x^T Q) R_{:,2} = r_{1,2} z^T (U_x^T Q)_{:,1} + r_{2,2} z^T (U_x^T Q)_{:,2}.$$

In this way we see that $z^T (U_x^T Q)_{:,k} = 0$ for each $k < j$ and so

$$z^T A_{:,j} = r_{j,j} z^T (U_x^T Q)_{:,j}. \tag{6.47}$$

Finally, as the left side of (6.47) is a nonzero integer and each $\|(U_x^T Q)_{:,j}\| = 1$ we find

$$1 \le |z^T A_{:,j}| = |z^T (U_x^T Q)_{:,j} r_{j,j}| \le |r_{j,j}| \|z\| \|(U_x^T Q)_{:,j}\| = |r_{j,j}| \|z\| \le R_{max} \|z\|$$

as a consequence of the Cauchy–Schwarz inequality. End of Proof.

This result sets the stage for a tractable, constructive means for identifying an integer relation or for showing that one is improbable, i.e., of astronomically big norm. The idea is to successively choose the "free" integer matrix $A$ so to decrease $R_{max}$ (and so increase the size of any permissible integer relation) while monitoring the diagonal of $R$ for zeros which then reveal integer relations in the row(s) of $A^{-1}$.

The decrease in $R_{max}$ is achieved through a simple integer variation on Gaussian Elimination. It begins, given an upper triangular $R \in \mathbb{R}^{(n-1)\times n}$, by generating an upper triangular, unit diagonal, $D \in \mathbb{R}^{n\times n}$ such that $RD$ is diagonal and $\mathrm{diag}(R) = \mathrm{diag}(RD)$. We build this elimination matrix via successive construction of its super–diagonals

$$d_{i,i} = 1, \quad d_{i,i+1} = -r_{i,i+1}/r_{i,i}, \quad d_{i,i+2} = -(r_{i,i+1} d_{i+1,i+2} + r_{i,i+2} d_{i+2,i+2})/r_{i,i},$$

and in general

$$d_{i,i+j} = -\frac{1}{r_{i,i}} \sum_{k=1}^{j} r_{i,i+k} d_{i+k,i+j}. \tag{6.48}$$

For example, if

$$R = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \end{pmatrix} \quad \text{then} \quad D = \begin{pmatrix} 1 & -2 & -1/2 \\ 0 & 1 & -5/4 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad RD = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \end{pmatrix}.$$

As we will use $D$ to transform $A$ we must respect its integrality. This is accomplished by sequentially rounding to the nearest integer in our super-diagonal construction. That is we replace (6.48) with

$$d_{i,i+j} = \text{round}\left( -\frac{1}{r_{i,i}} \sum_{k=1}^{j} r_{i,i+k} d_{i+k,i+j} \right). \tag{6.49}$$

On application to the small example above,

$$D = \begin{pmatrix} 1 & -2 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad RD = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 1 \end{pmatrix}.$$

Although $RD$ is not necessarily diagonal, we can still prove

$$|(RD)_{i,j}| \leq |r_{j,j}|/2. \tag{6.50}$$

The construction of $D$ from $R$ via (6.49) is known as **Hermite Reduction** and as such we will denote it $D = H_{red}(R)$. We now have all of the pieces required of the

**Integer Relation Algorithm**

**Initialize.** Given $x \in \mathbb{R}^n$ build $U_x$ per (6.42). Set $U = U_x$ and $A = I_n$ and choose $w > 2/\sqrt{3}$.

**While** neither $x$ nor $\text{diag}(U)$ possess any zero elements

1. **Reduce:** Compute $D = H_{red}(U)$ and set

$$x = D^{-1}x, \quad U = UD \quad \text{and} \quad A = AD.$$

2. **Exchange:** Let $r$ be such that $w^j |u_{j,j}|$ is maximal for $j = r$. Let $\tilde{I}$ be the elementary perturbation of $I_n$ where columns $r$ and $r+1$ are swapped, and set

$$x = \tilde{I}x, \quad U = U\tilde{I} \quad \text{and} \quad A = A\tilde{I}.$$

   If $r = n-1$ then $U$ remains upper triangular and we return to top of while. Else,

3. **Fix:** Construct $F = I_{n-1}$ except the $2 \times 2$ block

$$\begin{pmatrix} f_{r,r} & f_{r,r+1} \\ f_{r+1,r} & f_{r+1,r+1} \end{pmatrix} = \frac{1}{(u_{r,r}^2 + u_{r+1,r}^2)^{1/2}} \begin{pmatrix} u_{r,r} & u_{r+1,r} \\ -u_{r+1,r} & u_{r,r} \end{pmatrix}.$$

   Set $U = FU$ and return to top of while.

**end of While**

We see from the exchange step that $w$ is a weight that can assist the motion of large diagonal elements down the diagonal and eventually off the diagonal and into column $n$. We have implemented this algorithm in `psqr`. For example, with $x = (1.5, 2.3, 3.2)^T$ and $\gamma = 3$ `psqr` returns

$$A^{-1} = \begin{pmatrix} 4 & -4 & 1 \\ 9 & -17 & 8 \\ 1 & -2 & 1 \end{pmatrix}$$

and we recognize two integer relations for $x$ in its first two rows.

### 6.7. Probabilistic and Statistical Interpretations*

The previous example offers entry into the fascinating world of probability theory. To say that our $j$th experiment was tainted with Gaussian noise (of zero mean and variance $\sigma^2$) is to say that

$$a_j x = b_j + \varepsilon_j \tag{6.51}$$

where $a_j$ is the $j$th row of $A$ and that $\varepsilon_j$ is drawn from a probability distribution with density

$$p(\varepsilon) \equiv \frac{1}{\sigma\sqrt{2\pi}} \exp(-\varepsilon^2/(2\sigma^2)). \tag{6.52}$$

This permits us to write the probability of observing $b_j$ given $a_j$ and the candidate $x$, as

$$p(b_j|a_j; x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-(a_j x - b_j)^2/(2\sigma^2)).$$

Next, as we expect the $j$th and $k$th experiments to be independent of one another (i.e., errors in one should not effect errors in the other) we can write the probability of observing $b_j$ and $b_k$ given $a_j$ and $a_k$ and the candidate $x$, as the product of the individual probabilities

$$p((b_j, b_k)|(a_j, a_k); x) = \frac{1}{\sigma^2 2\pi} \exp(-((a_j x - b_j)^2 - (a_k x - b_k)^2/(2\sigma^2)).$$

Combining now all of the experiments, we find

$$p(b|A; x) = \frac{1}{\sigma^m (2\pi)^{m/2}} \exp(-\|Ax - b\|^2/(2\sigma^2)).$$

This is often abbreviated as $L(x)$ and interpreted as the likelihood of $A$ producing $b$ given $x$. The **principle of maximum likelihood** is to choose $x$ to maximize this likelihood. This is clearly the $x$ that minimizes $\|Ax - b\|^2$.

This rationale generalizes naturally to the case where each experiment produces more than one measurement. In this case each $\varepsilon_j \in \mathbb{R}^n$ is drawn from the multivariate density

$$p(\varepsilon) \equiv \frac{1}{\sqrt{(2\pi)^n \det(C)}} \exp(-\varepsilon^T C^{-1} \varepsilon/2). \tag{6.53}$$

where $C \in \mathbb{R}^{n \times n}$ is the symmetric and positive definite **covariance matrix**. It follows that the probability of measuring $b_j \in \mathbb{R}^n$ given $a_j \in \mathbb{R}^{n \times m}$ and $x \in \mathbb{R}^m$ is

$$p(b_j|a_j; x) = \frac{1}{\sqrt{(2\pi)^n \det(C)}} \exp(-(a_j x - b_j)^T C^{-1} (a_j x - b_j)/2).$$

We would like to write this as a simple square. The way out is via the Cholesky factorization (recall (6.54))

$$C = LL^T$$

Finally, we note that positive definite matrices permit a simpler $LU$ factorization.

**Proposition** 6.7. **Cholesky Factorization** If $S$ is symmteric and positive definite then there exists a lower triangular matrix $L$, with positive diagonal elements, for which

$$S = LL^T. \tag{6.54}$$

Proof: From the symmtery of $S \in \mathbb{R}^{n\times n}$ we may begin with the representation

$$S = \begin{pmatrix} s_{11} & S_{21}^T \\ S_{21} & S_{22} \end{pmatrix}, \qquad s_{11} \in \mathbb{R}, \quad S_{21} \in \mathbb{R}^{n-1}, \quad S_{22} \in \mathbb{R}^{(n-1)\times(n-1)}.$$

As $S$ is positive definite we may conclude that $s_{11} > 0$ and

$$S_{22} - \frac{1}{s_{11}} S_{21} S_{21}^T \quad \text{is positive definite.} \tag{6.55}$$

To prove the first claim choose $x = (1, 0, \ldots, 0)^T$ and note that $0 < x^T S x = s_{11}$. To establish (6.55) write $x = (x_1, \tilde{x})^T$ with $x_1 \in \mathbb{R}$ and confirm that

$$x^T S x = s_{11} x_1^2 + 2 x_1 S_{21}^T \tilde{x} + \tilde{x}^T S_{22} \tilde{x}. \tag{6.56}$$

Next, show that you may choose $x_1$ such that

$$s_{11} x_1^2 + 2 x_1 S_{21}^T \tilde{x} - S_{21}^T S_{21} = 0 \tag{6.57}$$

and conclude that this choice confirms the claim in (6.55).

With these preliminaries we proceed to construct the factorization

$$\begin{pmatrix} s_{11} & S_{21}^T \\ S_{21} & S_{22} \end{pmatrix} = \begin{pmatrix} l_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix} \begin{pmatrix} l_{11} & L_{21}^T \\ 0 & L_{22}^T \end{pmatrix} = \begin{pmatrix} l_{11}^2 & l_{11} L_{21}^T \\ l_{11} L_{21} & L_{21} L_{21}^T + L_{22} L_{22}^T \end{pmatrix}.$$

Identifying terms we find

$$l_{11} = \sqrt{s}_{11}, \quad L_{21} = S_{21}/l_{11} \quad \text{and} \quad L_{22} L_{22}^T = S_{22} - S_{21} S_{21}^T / s_{11}. \tag{6.58}$$

The first two equalities are explicit and, thanks to $s_{11} > 0$, unambiguous. Regarding the third assignment in (6.58) we note that its right hand side is symmetric by inspection and positive definite by the argument following (6.57). As such, the third assignment in (6.58) is simply the Cholesky factorization of the $n-1$ dimensional matrix $S_{22} - S_{21} S_{21}^T / s_{11}$. Applying the above scheme to this will reduce our needs to the Cholesky factorization of an $n-2$ dimensional matrix. Continuing this process brings us the trivial one dimensional factorization. End of Proof.

To implement the algorithm at the center of the proof of (6.54) we simply build the columns of $L$ from longest to shortest. For example, we build the first column

$$\begin{pmatrix} 4 & 8 & 16 \\ 8 & 52 & 92 \\ 16 & 92 & 308 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ 4 & l_{22} & 0 \\ 8 & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} 2 & 4 & 8 \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{pmatrix}$$

by dividing the original first column by the square root of its first element. Proceeding on to the second column we find

$$\begin{pmatrix} 52 & 92 \\ 92 & 308 \end{pmatrix} - \begin{pmatrix} 4 \\ 8 \end{pmatrix} \begin{pmatrix} 4 & 8 \end{pmatrix} = \begin{pmatrix} l_{22} & 0 \\ l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} l_{22} & l_{32} \\ 0 & l_{33} \end{pmatrix}$$

that is,

$$\begin{pmatrix} 36 & 60 \\ 60 & 244 \end{pmatrix} = \begin{pmatrix} 6 & 0 \\ 10 & l_{33} \end{pmatrix} \begin{pmatrix} 6 & 10 \\ 0 & l_{33} \end{pmatrix}$$

and finally $l_{33}^2 = 244 - 100$ and so $l_{33} = 12$. All together,

$$\begin{pmatrix} 4 & 8 & 16 \\ 8 & 52 & 92 \\ 16 & 92 & 308 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ 4 & 6 & 0 \\ 8 & 10 & 12 \end{pmatrix} \begin{pmatrix} 2 & 4 & 8 \\ 0 & 6 & 10 \\ 0 & 0 & 12 \end{pmatrix}$$

Picking of the thread, for then (exercise)

$$p(b_j|a_j; x) = \frac{1}{\sqrt{(2\pi)^n \det(C)}} \exp(-\|L^{-1}(a_j x - b_j)\|^2/2).$$

and so using independence, after stacking and blocking

$$p(b|A; x) = \frac{1}{((2\pi)^n \det(C))^{m/2}} \exp(-\|L^{-1}(Ax - b)\|^2/2).$$

and so the $x$ that maximizes the likelihood is the $x$ that satisfies the **weighted** least squares problem

$$A^T C^{-1} A x = A^T C^{-1} b.$$

From the statistical point of view we call $E_{LS} \equiv (A^T A)^{-1} A^T$ the Least Squares Estimator. We note that $E_{LS} b$ is clearly linear in $b$ and we show here that if $\varepsilon_j$ is simply assumed to be independent and drawn from a distribution with mean zero and variance $\sigma^2$ then $E_{LS} b$ is the **Best Linear Unbiased Estimate**.

A linear estimator $E$ is called **unbiased** if the mean value of $Eb$ is $x$. But this is easy,

$$\text{mean}(Eb) = \text{mean}(EAx - E\varepsilon) = EAx + E\text{mean}(\varepsilon) = EAx$$

and so $E$ is unbiased iff $EA = I$. Please note that $E_{LS}$ is indeed unbiased. It follows that the variance of the unbiased estimate is simply

$$\text{var}(Eb) \equiv \text{mean}((Eb - x)(Eb - x)^T) = \text{mean}((E(Ax - \varepsilon) - x)(E(Ax - \varepsilon) - x)^T) = E\text{var}(\varepsilon)E^T. \quad (6.59)$$

From here we can now establish that $E_{LS} b$ is best in the sense that it has the least variance, in the sense of postivive definite matrices. To begin we write $E = E_{LS} + D$ and note that $EA = I$ implies $DA = 0$ and so $DE_{LS}^T = E_{LS} D = 0$. As such, in the single measurement case, i.e., $\text{var}(\varepsilon) = \sigma^2 I$, we find

$$\text{var}(Eb) = E\text{var}(\varepsilon)E^T = \sigma^2 EE^T = \sigma^2 (E_{LS} + D)(E_{LS} + D)^T = \text{var}(E_{LS} b) + \sigma^2 DD^T.$$

As $DD^T$ is a positive semidefinite we have shown that

$$y^T \text{var}(E_{LS} b) y \leq y^T \text{var}(Eb) y$$

for every vector $y$ and every Linear Unbiased Estimator, $E$.

In the multivariate case where $\text{var}(\varepsilon) = C$ we again invoke its Cholesky factorization, $C = LL^T$, to transform $b = Ax + (or -)\varepsilon$ to

$$\overline{b} = \overline{A}x + \overline{\varepsilon}, \quad \text{where} \quad \overline{b} = L^{-1}b,\ \overline{A} = L^{-1}A \quad \text{and} \quad \overline{\varepsilon} = L^{-1}\varepsilon.$$

This decorrelates the error terms for

$$\text{var}(\overline{\varepsilon}) = \text{var}(L^{-1}\varepsilon) = L^{-1}CL^{-T} = L^{-1}LL^TL^{-T} = I.$$

And so we may argue as above that

$$\text{var}(E\overline{b}) = E\text{var}(\overline{b})E^T = (E_{LS} + D)(E_{LS} + D)^T = \text{var}(E_{LS}\overline{b}) + DD^T$$

where

$$E_{LS} = (\overline{A}^T\overline{A})^{-1}\overline{A}^T = (A^TL_n^{-T}L_n^{-1}A)^{-1}A^TL_n^{-T}L_n^{-1} = (A^TC^{-1}A)^{-1}A^TC^{-1}$$

as above.

### 6.8. Autoregressive Models and Levinson's Algorithm*

We investigate a classic least squares problem from random signal theory. We suppose that $x \in \mathbb{R}^N$ is drawn from a zero–mean second order **stationary process**. That is,

$$\text{mean}(x_j) = 0 \quad \text{and} \quad \text{mean}(x_jx_{j+\ell}) \equiv c_\ell = c_{-\ell}. \tag{6.60}$$

We envision $x_j$ to be an observation at time $j$ and so interpret $\ell$ as the lag between observations. Hence each "mean" in (6.60) is to be interpreted as "average over observations." The stationary hypothesis states that the **covariance** between observations at different times depends only on the associated lag. As these are covariances of the same process it is common to call them **autocovariances**. .

The intuitive notion of stationary suggests that it may be reasonable to "predict" the value of $x_j$ from a linear combination of its past values. More precisely, we will fit the process to an **autoregressive model** of order $L$,

$$x_j = \sum_{\ell=1}^{L} a_\ell x_{j-\ell} \tag{6.61}$$

where $L$ is typically much less than $N$. We use (6.60) to arrive at a consistent and deterministic system for $a \in \mathbb{R}^L$. Multiplying each side of (6.61) by $x_{j+1}$ and taking means

$$\text{mean}(x_{j+1}x_j) = \sum_{\ell=1}^{L} a_\ell \text{mean}(x_{j+1}x_{j-\ell}), \quad \text{reveals} \quad c_1 = \sum_{\ell=1}^{L} a_\ell c_{\ell-1}, \tag{6.62}$$

while multiplying (6.61) by $x_{j+k}$ and taking means reveals

$$c_k = \sum_{\ell=1}^{L} a_\ell c_{\ell-k}. \tag{6.63}$$

All together we arrive at the system

$$C_La = c_{1:L} \tag{6.64}$$

where $c_{1:L} = (c_1, c_2, \ldots, c_L)^T$ and

$$C_L = \begin{pmatrix} c_0 & c_1 & \cdots & c_{L-2} & c_{L-1} \\ c_1 & \ddots & \ddots & \ddots & c_{L-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ c_{L-2} & \ddots & \ddots & \ddots & c_1 \\ c_{L-1} & c_{L-2} & \cdots & c_1 & c_0 \end{pmatrix}$$

The system (6.64) is known as the **Yule–Walker equations** for the autoregressive model, (6.61). We see that the stationarity condition, (6.60) has manifested itself in the condition that the matrix $C_L$ is constant along each diagonal. Such matrices are called **Toeplitz matrices**. They occur frequently in applications and enjoy a considerable theorrtical development (stay tuned).

In order to develop some intuition about the behavior of such models we simulate and analyze the first order process

$$x_j = a_1 x_{j-1} + \varepsilon_j, \tag{6.65}$$

where $\varepsilon_j$ is drawn from the Gaussian distribution with density (6.52). Squaring and taking means of each side reveals

$$c_0 = \text{mean}(x_j^2) = \text{mean}(a_1^2 x_{j-1}^2 + 2a_1 x_{j-1}\varepsilon_j + \varepsilon_j^2) = a_1^2 c_0 + 0 + \sigma^2,$$

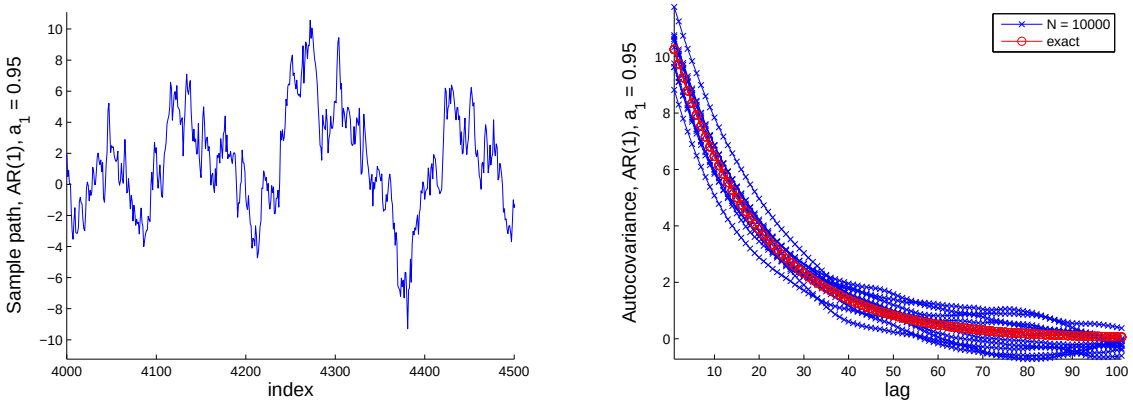
and so,

$$c_0 = \frac{\sigma^2}{1 - a_1^2}. \tag{6.66}$$

In a similar fashion,

$$c_k = \text{mean}(x_j x_{j+k}) = \text{mean}(x_j(a_1 x_{j+k-1} + \varepsilon_{j+k})) = a_1 c_{k-1} = a_1^2 c_{k-2} = \cdots = a_1^k c_0. \tag{6.67}$$

We illustrate in Figure 6.6 the processes and their autocovariances when $a_1 = \pm 0.95$ and $\sigma = 1$.



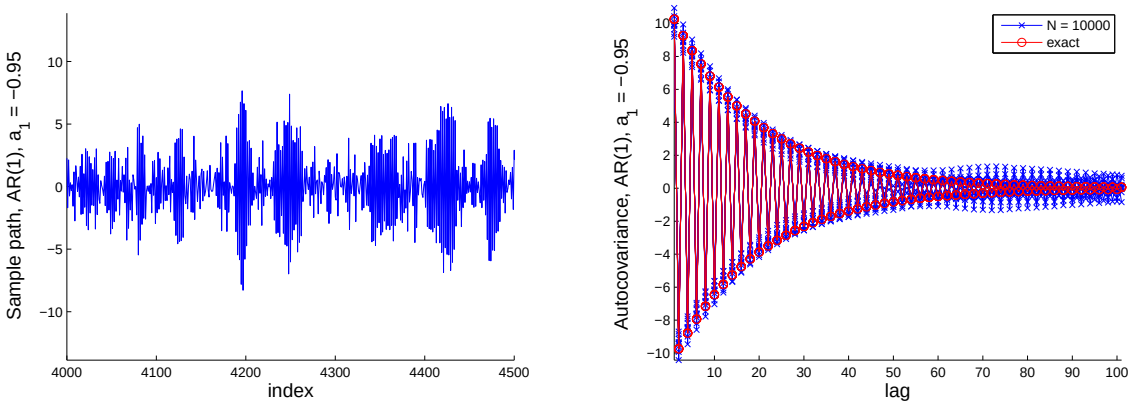



**Figure** 6.6 Sample paths and exact and empirical autocovariances for the first order model, (6.65), with $\sigma = 1$.

We next address the most parsimonious choice of model order, $L$. To distinguish the model orders we denote the solution to (6.61) at order $L$ by $a^{(L)}$ and the associated least squares error by

$$\rho(L) \equiv \frac{1}{N}\sum_{j=1}^{N} \text{mean}\left(\left(x_j - \sum_{\ell=1}^{L} a_\ell^{(L)} x_{j-\ell}\right)^2\right) = c_0 - c_{1:L}^T a^{(L)}.$$

The competing objectives are to render *both* the error, $\rho(L)$, and the model order, $L$, small. We accomplish this by minimizing the Relative Final Prediction Error

$$RFPE(L) = \frac{N+L+1}{N-L-1}\frac{\rho(L)}{c_0}. \tag{6.68}$$

The search for the minimum is eased by the fact that we may solve for $a^{(L)}$ via a fast and cheap update of $a^{(L-1)}$. The resulting algorithm was first discovered in this context by Norman Levinson and can be seen as a recursive inverse Cholesky factorization of $C_L$,

$$R_L^T C_L R_L = D_L, \quad \text{where} \quad D_L = \text{diag}(d_{1:L}) \quad \text{and} \quad R_L = \begin{pmatrix} R_{L-1} & r_{L-1} \\ 0 & 1 \end{pmatrix} \tag{6.69}$$

that hinges on the **persymmetry** of $C_L$, i.e.,

$$E_L C_L E_L = C_L$$

where $E_L = [e_L \; e_{L-1} \; \cdots \; e_1]$ is the **exchange matrix** obtained by exchanging, or reversing, the columns of the $L$-by-$L$ identity matrix. We construct the $L = 2$ factors by hand

$$R_2 = \begin{pmatrix} 1 & -c_1/c_0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad D_2 = \begin{pmatrix} c_0 & 0 \\ 0 & c_0 - c_1^2/c_0 \end{pmatrix}$$

and proceed to establish the general case.

**Proposition** 6.8. **Levinson's Algorithm**. Commencing from $r_1 = -c_1/c_0$, $d_2 = c_0 - c_1^2/c_0$ and $a^{(1)} = c_1/c_0$ and $k = 1$, we chart the recurrence relations

$$\begin{aligned} \gamma_{k+1} &= -(c_{1:k}^T r_k + c_{k+1})/d_{k+1} \\ a^{(k+1)} &= \begin{pmatrix} a^{(k)} \\ 0 \end{pmatrix} - \gamma_{k+1} \begin{pmatrix} r_k \\ 1 \end{pmatrix} \\ r_{k+1} &= \begin{pmatrix} 0 \\ r_k \end{pmatrix} + \gamma_{k+1} \begin{pmatrix} 1 \\ E_k r_k \end{pmatrix} \\ d_{k+2} &= (1 - \gamma_{k+1}^2) d_{k+1}. \end{aligned} \tag{6.70}$$

Proof: To set up the recursion we recall that $c_{1:k} = (c_1, c_2, \ldots, c_k)^T$ and express the growth of our Toeplitz matrix in terms of the associated exchange matrix:

$$C_{k+1} = \begin{pmatrix} C_k & E_k c_{1:k} \\ c_{1:k}^T E_k & c_0 \end{pmatrix} = \begin{pmatrix} c_0 & c_{1:k}^T \\ c_{1:k} & C_k \end{pmatrix} \tag{6.71}$$

and the growth of its associated Cholesky factor via

$$R_{k+1} = \begin{pmatrix} R_k & r_k \\ 0 & 1 \end{pmatrix}, \quad r_{k+1} = \begin{pmatrix} \gamma_{k+1} \\ s_k \end{pmatrix}. \tag{6.72}$$

To begin, we use the first representation of $C_{k+1}$ in (6.71) and equate last columns in $C_{k+1}R_{k+1} = R_{k+1}^{-T} D_{k+1}$, finding

$$C_k r_k + E_k c_{1:k} = 0 \tag{6.73}$$

and

$$c_{1:k}^T E_k r_k + c_0 = d_{k+1}. \tag{6.74}$$

The former becomes, on using persymmetry $C_k = E_k C_k E_k$ and multiplying across by $E_k^T$,

$$C_k E_k r_k + c_{1:k} = 0. \tag{6.75}$$

We next increment $k$ by 1 in (6.73), finding $C_{k+1} r_{k+1} + E_{k+1} c_{1:k+1} = 0$. We unpack this using (6.71) (2nd part) and (6.72) and find

$$\gamma_{k+1} c_0 + c_{1:k}^T s_k + c_{k+1} = 0 \tag{6.76}$$

and

$$\gamma_{k+1} c_{1:k} + C_k s_k + E_k c_{1:k} = 0. \tag{6.77}$$

Subtracting (6.73) from (6.77) and using (6.75) we arrive at

$$s_k = r_k + \gamma_{k+1} E_k r_k. \tag{6.78}$$

This completes the expression of

$$r_{k+1} = \begin{pmatrix} 0 \\ r_k \end{pmatrix} + \gamma_{k+1} \begin{pmatrix} 1 \\ E_k r_k \end{pmatrix}, \tag{6.79}$$

which is precisely the third statement in (6.70). Using (6.78) in (6.76) we find

$$c_0 \gamma_{k+1} + c_{1:k}^T (r_k + E_k r_k \gamma_{k+1}) + c_{k+1} = 0.$$

Applying (6.74) here reveals the first formula in (6.70). This formula, together with (6.74) (after incrementing $k$) and (6.79) yield

$$d_{k+2} = c_0 + c_{1:k+1}^T E_{k+1} r_{k+1} = c_0 + c_{1:k}^T E_k r_k + c_{1:k+1}^T \begin{pmatrix} r_k \\ 1 \end{pmatrix} \gamma_{k+1} = (1 - \gamma_{k+1}^2) d_{k+1},$$

the fourth piece of the stated recursion. Finally,

$$\begin{aligned} a^{(k+1)} &= R_{k+1} D_{k+1}^{-1} R_{k+1}^T c_{1:k+1} \\ &= \begin{pmatrix} R_k & r_k \\ 0 & 1 \end{pmatrix} \begin{pmatrix} D_k^{-1} & 0 \\ 0 & 1/d_{k+1} \end{pmatrix} \begin{pmatrix} R_k^T & 0 \\ r_k^T & 1 \end{pmatrix} c_{1:k+1} \\ &= \begin{pmatrix} R_k & r_k \\ 0 & 1 \end{pmatrix} \begin{pmatrix} D_k^{-1} R_k^T c_{1:k} \\ (r_k^T c_{1:k} + c_{k+1})/d_{k+1} \end{pmatrix} \\ &= \begin{pmatrix} R_k D_k^{-1} R_k^T c_{1:k} - \gamma_{k+1} r_k \\ -\gamma_{k+1} \end{pmatrix} = \begin{pmatrix} a^{(k)} \\ 0 \end{pmatrix} - \gamma_{k+1} \begin{pmatrix} r_k \\ 1 \end{pmatrix} \end{aligned}$$

fills in the last part of the recursion. End of Proof.

We illustrate its performance in identifying a 10th order model in Figure 6.7.

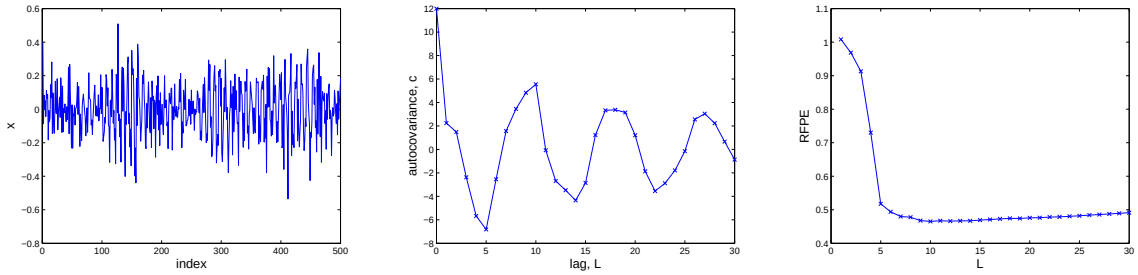



**Figure** 6.7 (A) A sample path of a 10th order process. (B) Its empirical autocovariance. (C) The relative final prediction error. `AkaLev.m`

### 6.9. Notes and Exercises

Our work on detecting integer patterns follows the excellent DH Bailey and Moll (2007). For the autoregressive approach see Brillinger (2001). The model criteria (6.68) is due to Akaike (1969). Our presentation of Levinson's Algorithm follows Ammar and Gragg (1987).

1. An elastic cable 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 cable's compliance, $c$, and original length, $L$, by least squares. In particular

   (i) Formulate the question as $Ax = b$, with numerical values in $A$ and $b$. What does $x$ signify?

   (ii) Solve the normal equations, $A^T Ax = A^T b$ by hand for $x$.

   (iii) Graph the three data points in the $(f, \ell)$ plane as well as the straight line fit corresponding to the $x$ found in (ii).

2. With regard to the example of §6.2 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 sufficient number of times (until the averages settle down) 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. Show that if $P$ is an othogonal projection onto a subspace of dimension $r$ then $\|P\|_F = \sqrt{r}$.

7. Show that if $P$ is a projection then so too are $P^T$ and $I - P$.

8. Show that the only invertible projection is the identity matrix.

9. (a) Show that if $P$ and $Q$ are projections then

$$(P - Q)^2 + (I - P - Q)^2 = I.$$

(b) Use (a) to show that if $P$ and $Q$ are orthogonal projections then $\|P - Q\| \leq 1$.

10. Not all projections are symmetric. Please confirm that

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

are projections. Sketch the column space of the first, and depict the oblique projection of $b = (1, 3)^T$ onto this space. How does your sketch differ from Figure 6.4?

11. Show that orthonormal matrices stretch no vector, i.e., if $Q^T Q = I$ then $\|Qx\| = \|x\|$ for all $x$.

12. Use the Gram–Schmidt procedure, by hand (don't even use a calculator, for they replace pregnant square roots with chaotic decimals), to compute orthonormal bases for the four fundamental subspaces of

$$A = \begin{pmatrix} 1 & 2 & 5 & 3 \\ 3 & 1 & 5 & 4 \\ 2 & -1 & 0 & 1 \\ 1 & 2 & 5 & 3 \end{pmatrix}. \tag{6.80}$$

13. Construct, by hand, the QR decomposition of

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

and explain its relationship to our first exercise.

14. Show that the generating function, (6.35), of the Legendre polynomials satisfies the partial differential equation

$$((1-x^2)G_x)_x + t(tG)_{tt} = 0, \tag{6.81}$$

and hence the $P_n$ obey the ordinary differential equation

$$((1-x^2)P_n'(x))' + n(n+1)P_n(x) = 0. \tag{6.82}$$

Hint: (6.81) is an explicit calculation. Its tedium is minimized via the symbolic toolbox in MATLAB. To get from (6.81) to (6.82) use (6.31).

15. Let us confirm that the density in Eq. (6.52) indeed obeys

$$\int_{-\infty}^{\infty} \exp(-t^2/(2\sigma^2))\, dt = \sigma\sqrt{2\pi} \tag{6.83}$$

Hint: Justify each step in

$$\left(\int_{-\infty}^{\infty} \exp(-t^2/(2\sigma^2))\, dt\right)^2 = \int_{-\infty}^{\infty} \exp(-t_1^2/(2\sigma^2))\, dt_1 \int_{-\infty}^{\infty} \exp(-t_2^2/(2\sigma^2))\, dt_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(-(t_1^2+t_2^2)/(2\sigma^2))\, dt_1 dt_2.$$
$$= \int_0^{2\pi}\int_0^{\infty} \exp(-r^2/(2\sigma^2)) r\, dr d\theta,$$

then notice that the final integrand is constant in $\theta$ and is proportional to the derivative of a clean function of $r$.

16. The $n$th moment of a function $f$ on $\mathbb{R}$ is

$$\mu_n(f) \equiv \int_{-\infty}^{\infty} x^n f(x)\, dx.$$

Please compute the moments of the Gaussian density. In particular, show that $\mu_n(p) = 0$ when $n$ is odd while for even $n$ that

$$\mu_{2m}(p) = (2m-1)!!\sigma^{2m}. \tag{6.84}$$

where $(2m-1)!! = (2m-1)(2m-3)\cdots(2m-(2m-1))$ denotes factorials through the odds. Hint: show that

$$\mu_{2m}(p) = -\sigma^2 \int_{-\infty}^{\infty} x^{2m-1} p'(x)\, dx$$

and then integrate by parts.

17. The (differential) entropy of a function $f$ on $\mathbb{R}$ is

$$S(f) \equiv -\int_{-\infty}^{\infty} f(x) \log\, f(x)\, dx.$$

Use the results of the previous exercise to compute the entropy of the gaussian:

$$S(p) = (1 + \log(2\pi) + \log(\sigma^2))/2. \tag{6.85}$$

18. Find the first three **Hermite polynomials** by orthonormalizing $\{1, x, x^2\}$ with respect to the inner product

$$\langle f, g\rangle_w \equiv \int_{-\infty}^{\infty} f(x)g(x)w(x)\, dx \quad \text{where} \quad w(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2).$$

Hint: Use (6.83) and the fact that $w(x) = w(-x)$.

19. Expand an arbitrary function, $f$, in terms of the Hermite polynomials of the previous exercise. In particular, show that if

$$f(x) = w(x) \sum_{n=0}^{\infty} f_n H_n(x) \quad \text{then} \quad f_n = \int_{-\infty}^{\infty} f(x) H_n(x)\, dx.$$

Establish the **Gram-Charlier Expansion:** if $f$ has zero mean and unit variance then

$$f(x) = w(x)(1 + \mu_3(f)H_3(x) + (\mu_4(f) - 3)H_4(x) + \cdots).$$

20. Establish the second identity in (6.43). Recall that $\|x\| = 1$ and then deduce, with little work, that $U_x x = 0$.

21. Confirm that (6.50) is a consequence of the fact that the largest rounding error is $1/2$. That is, $|t - \text{round}(t)| \leq 1/2$ for every $t \in \mathbb{R}$. Here is how to start in one concrete case

$$\begin{aligned}(RD)_{1,2} &= r_{1,1}d_{1,2} + r_{1,2} \\ &= r_{1,1}\text{round}(-r_{1,2}/r_{1,1}) + r_{1,2} \\ &= r_{1,1}\text{round}(-r_{1,2}/r_{1,1}) + r_{1,2} + r_{1,1}(-r_{1,2}/r_{1,1}) - r_{1,1}(-r_{1,2}/r_{1,1}) \\ &= r_{1,1}\left(\text{round}(-r_{1,2}/r_{1,1}) - (-r_{1,2}/r_{1,1})\right)\end{aligned}.$$

Finish this argument and then extend it to the rest of the super–diagonals.

22. Complete the integer relation (6.41) by invoking `psqr` at many random values of $\theta$.

23. Prove (6.55) by following the hints in (6.56)–(6.57).

24. Code Cholesky and contrast with `chol`.

25. The statistical modeling of §6.8 is precisely that needed to develop the finite **Wiener Filter**. Here we suppose we have measurements, $y_k$, of a true signal, $x_k$, contaminated by additive noise, $\varepsilon_k$, i.e.,

$$y_k = x_k + \varepsilon_k.$$

We suppose known the two autocovariances, $c_{yy}$ and $c_{xx}$, and the covariance, $c_{xy}$, and we seek the finite filter $a$ for which

$$\tilde{x}_k = \sum_{m=k-N}^{k} a_{k-m} y_m$$

minimizes $\text{mean}((x_k - \tilde{x}_k)^2)$. Show that the best $a$ obeys the Yule–Walker equations

$$\begin{pmatrix} c_{yy}(0) & c_{yy}(1) & \cdots & c_{yy}(N) \\ c_{yy}(1) & \ddots & \ddots & \ddots \\ \ddots & \ddots & \ddots & c_{yy}(1) \\ c_{yy}(N) & \ddots & c_{yy}(1) & c_{yy}(0) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_N \end{pmatrix} = \begin{pmatrix} c_{xy}(0) \\ c_{xy}(1) \\ \vdots \\ c_{xy}(N) \end{pmatrix}$$

26. Let us extend the Levinson Algorithm of §6.8 to the multivariate setting, where $X$ is drawn from an $n$–valued zero–mean second order stationary process. That is, each $X_j \in \mathbb{R}^n$,

$$\text{mean}(X_j) = 0 \quad \text{and} \quad C_\ell \equiv \text{mean}(X_j X_{j-\ell}^T) \tag{6.86}$$

and so $C_k \in \mathbb{R}^{n\times n}$. We fit $X$ to an autoregressive model of order $L$,

$$X_j = \sum_{\ell=1}^{L} A_\ell X_{j-\ell} \tag{6.87}$$

where each $A_\ell \in \mathbb{R}^{n\times n}$.

(i) By taking products and means of (6.87) when $L = 2$ derive the associated Yule–Walker equations

$$\begin{pmatrix} A_1^{(2)} & A_2^{(2)} \end{pmatrix} \begin{pmatrix} C_0 & C_1 \\ C_1^T & C_0 \end{pmatrix} = \begin{pmatrix} C_1 & C_2 \end{pmatrix}$$

and proceed to construct the explicit block Cholesky factorization

$$\begin{pmatrix} I & 0 \\ -C_1^T C_0^{-1} & I \end{pmatrix} \begin{pmatrix} C_0 & C_1 \\ C_1^T & C_0 \end{pmatrix} \begin{pmatrix} I & -C_0^{-1}C_1 \\ 0 & I \end{pmatrix} = \begin{pmatrix} C_0 & 0 \\ 0 & C_0 - C_1^T C_0^{-1} C_1 \end{pmatrix}.$$

(ii) In general, for a model of order $k$, derive

$$\begin{pmatrix} A_1^{(k)} & A_2^{(k)} & \cdots & A_k^{(k)} \end{pmatrix} \begin{pmatrix} C_0 & C_1 & \cdots & C_{k-1} \\ C_1^T & C_0 & \cdots & C_{k-2} \\ \vdots & \vdots & & \vdots \\ C_{k-1}^T & C_{k-2}^T & \cdots & C_0 \end{pmatrix} = \begin{pmatrix} C_1 & C_2 & \cdots & C_k \end{pmatrix}$$

and express it as

$$A^{(k)} \mathcal{C}_k = C_{1\to k}. \tag{6.88}$$

(iii) Define the block exchanger

$$\mathcal{E}_k = \begin{pmatrix} 0 & \cdots & 0 & I_n \\ 0 & \cdots & I_n & 0 \\ \vdots & \iddots & & \vdots \\ I_n & 0 & \cdots & 0 \end{pmatrix}$$

comprised of $k$ copies of $I_n$ along the antidiagonal and confirm the block persymmetry

$$\mathcal{E}_k \mathcal{C}_k = \tilde{\mathcal{C}}_k \mathcal{E}_k, \tag{6.89}$$

where

$$\tilde{\mathcal{C}}_k \equiv \begin{pmatrix} C_0 & C_1^T & \cdots & C_{k-1}^T \\ C_1 & C_0 & \cdots & C_{k-2}^T \\ \vdots & \vdots & & \vdots \\ C_{k-1} & C_{k-2} & \cdots & C_0 \end{pmatrix}.$$

(iv) It follows from (iii) that we should pursue the simultaneous Cholesky factorizations

$$\mathcal{R}_{k+1}^T \mathcal{C}_{k+1} \mathcal{R}_{k+1} = \mathcal{D}_{k+1} \quad \text{and} \quad \tilde{\mathcal{R}}_{k+1}^T \tilde{\mathcal{C}}_{k+1} \tilde{\mathcal{R}}_{k+1} = \tilde{\mathcal{D}}_{k+1}.$$

Adopt the conventions

$$\mathcal{R}_{k+1} = \begin{pmatrix} \mathcal{R}_k & R_k \\ 0 & I_n \end{pmatrix}, \quad R_{k+1} = \begin{pmatrix} \Gamma_{k+1} \\ S_k \end{pmatrix}, \quad \mathcal{D}_{k+1} = \begin{pmatrix} \mathcal{D}_k & 0 \\ 0 & D_{k+1} \end{pmatrix} \tag{6.90}$$

and

$$\tilde{\mathcal{R}}_{k+1} = \begin{pmatrix} \tilde{\mathcal{R}}_k & \tilde{R}_k \\ 0 & I_n \end{pmatrix}, \quad \tilde{R}_{k+1} = \begin{pmatrix} \tilde{\Gamma}_{k+1} \\ \tilde{S}_k \end{pmatrix}, \quad \tilde{\mathcal{D}}_{k+1} = \begin{pmatrix} \tilde{\mathcal{D}}_k & 0 \\ 0 & \tilde{D}_{k+1}, \end{pmatrix} \tag{6.91}$$

and derive, by arguing as in the proof of Prop. 6.8, the recursive scheme

$$\boxed{\begin{aligned} \Gamma_{k+1} &= -\tilde{D}_{k+1}^{-1}(C_{1\to k}R_k + C_{k+1}), \qquad \tilde{\Gamma}_{k+1} = -D_{k+1}^{-1}(C_{1\to k}^T\tilde{R}_k + C_{k+1}^T) \\ A^{(k+1)} &= \begin{pmatrix} A^{(k)} - \tilde{D}_{k+1}\Gamma_{k+1}D_{k+1}^{-1}R_k^T & -\tilde{D}_{k+1}\Gamma_{k+1}D_{k+1}^{-1} \end{pmatrix} \\ S_{k+1} &= R_k + \mathcal{E}_k\tilde{R}_k\Gamma_{k+1}, \qquad \tilde{S}_{k+1} = \tilde{R}_k + \mathcal{E}_kR_k\tilde{\Gamma}_{k+1} \\ D_{k+2} &= D_{k+1}(I_n - \tilde{\Gamma}_{k+1}\Gamma_{k+1}), \qquad \tilde{D}_{k+2} = \tilde{D}_{k+1}(I_n - \Gamma_{k+1}\tilde{\Gamma}_{k+1}). \end{aligned}} \tag{6.92}$$

Hint: exploit the representations

$$\mathcal{C}_{k+1} = \begin{pmatrix} \mathcal{C}_k & \mathcal{E}_kC_{1\downarrow k} \\ C_{1\to k}^T\mathcal{E}_k & C_0 \end{pmatrix} = \begin{pmatrix} C_0 & C_{1\to k} \\ C_{1\downarrow k}^T & \mathcal{C}_k \end{pmatrix} \tag{6.93}$$

and

$$\tilde{\mathcal{C}}_{k+1} = \begin{pmatrix} \tilde{\mathcal{C}}_k & \mathcal{E}_kC_{1\downarrow k}^T \\ C_{1\to k}\mathcal{E}_k & C_0 \end{pmatrix} = \begin{pmatrix} C_0 & C_{1\to k}^T \\ C_{1\downarrow k} & \tilde{\mathcal{C}}_k \end{pmatrix}. \tag{6.94}$$

# 7. Metabolic Networks

Metabolism, the conversion of food into energy, can be seen to function in three stages. In the first, fats, polysaccharides and proteins are broken down into fatty acids and glycerol, glucose and other sugars, and amino acids. In the second stage these metabolites are largely converted into acetyl units of acetyl CoA. In the third stage Acetyl CoA brings acetyl into the Citric Acid Cycle. With each turn of the cycle an acetyl group is oxidized and the associated flow of electrons is harnessed to generate ATP, life's principal currency of energy.

We derive and solve linear programming problems stemming from flux balance subject to thermodynamic constraints. We derive the Simplex Method, offer a geometric interpretation, apply it to the real problem of succinate production and close with an investigation of Elementary Flux Modes and Extremal Rays.

## 7.1. Flux Balance and Optimal Yield

In models of metabolism the concentration of individual metabolites is governed by the reactions in which they participate. Although metabolites undergo significant conversion we assume that the net *flux* of each metabolite is zero. In other words, nothing is lost (or gained) in its conversion. As there are typically many more reactions than metabolites (in *E. Coli* the ratio is ten to one) the modeler is typically faced with an underdetermined system of equilibrium equations. As such there is typically a large class of balanced flux distributions. Scientists and engineers have recently turned this freedom to their advantage by directing flow down pathways that deliver an optimal (or at least improved) yield of a desired product.

To illustrate this we consider the network of Figure 7.1 and suppose that we have a steady flow into metabolite 1 and wish to accumulate as much as possible of metabolite $y$.

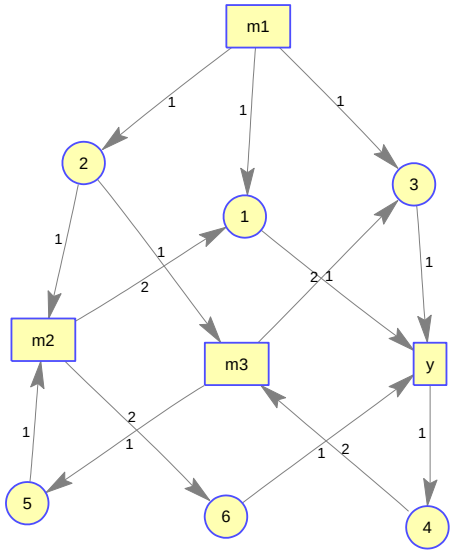



**Figure** 7.1. A metabolic network. `mstoich('ex1tab')`

There are many pathways from $m_1$ to $y$ and our task is to discover or design one that produces the greatest yield. Along the way we must obey flux balance and energy considerations. The former simply means that flux into an internal metabolite must balance flux out and that each of the fluxes has an energetically preferred direction.

More precisely, Figure 7.1 indicates that $m_1$ is supplied with rate one and then

$$\begin{aligned} m_1 + 2m_2 &\to y \quad \text{at rate } v_1 \\ m_1 &\to m_2 + m_3 \quad \text{at rate } v_2 \\ m_1 + 2m_3 &\to y \quad \text{at rate } v_3 \\ y &\to 2m_3 \quad \text{at rate } v_4 \\ m_3 &\to m_2 \quad \text{at rate } v_5 \\ 2m_2 &\to y \quad \text{at rate } v_6 \end{aligned} \tag{7.1}$$

To balance the flux of $m_j$ we balance its "total out" and "total in" rates. Its "total out" rate is the sum of the reaction rates in which it appears as a reactant (on the left), while its "total in" rate is the sum of the reaction rates in which it appears as a product (on the right). By this logic, the reaction scheme of (7.1) gives rise to this system of flux balances,

$$\begin{aligned} v_1 + v_2 + v_3 &= 1 \\ (2v_1 + 2v_6) - (v_2 + v_5) &= 0 \\ (2v_3 + v_5) - (v_2 + 2v_4) &= 0 \end{aligned} \tag{7.2}$$

Please note that we have balanced the flux of only the first 3 metabolites, for our goal is to in fact upset the balance of $y$ in our favor. It is now a simple manner to translate (7.2) into the matrix equation $Sv = f$ where

$$S = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 2 & -1 & 0 & 0 & -1 & 2 \\ 0 & -1 & 2 & -2 & 1 & 0 \end{pmatrix} \quad \text{and} \quad f = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \tag{7.3}$$

and the yield to

$$y = c^T v = v_1 + v_3 - v_4 + v_6. \tag{7.4}$$

While flux balance has been reduced to $Sv = f$ we note that each of our reactions is typically catalyzed by an enzyme with prescribed reactants and products. To say that we cannot interchange these is to say that the reaction is not reversible and this in turn implies that each reaction rate has a sign. For the sake of definiteness we suppose each $v_j \geq 0$. We may pose our design problem as

$$\boxed{\max_{v \in \mathcal{P}} c^T v \quad \text{where} \quad \mathcal{P} \equiv \{v \in \mathbb{R}^6 : Sv = f,\ v \geq 0\}.} \tag{7.5}$$

The maximization of a linear combination of variables subject to both linear and sign constraints is called a **Linear Programming** problem.

### 7.2. Linear Programming

To begin, we solve $Sv = f$ via row reduction. This marks $\{v_1, v_2, v_3\}$ as pivot variables and $\{v_4, v_5, v_6\}$ as free variables and so permits us to write the general solution as

$$\begin{aligned} v_1 &= (1/4) - (1/4)v_4 + (1/2)v_5 - (3/4)v_6 \\ v_2 &= (1/2) - (1/2)v_4 + (1/2)v_6 \\ v_3 &= (1/4) + (3/4)v_4 - (1/2)v_5 + (1/4)v_6. \end{aligned} \tag{7.6}$$

The associated yield is

$$y = v_1 + v_3 - v_4 + v_6 = 1/2 - (1/2)v_4 + (1/2)v_6. \tag{7.7}$$

The easiest way to guarantee that all rates are nonnegative is to set each of the free rates to zero. This gives us the simple starting solution

$$v^{(1)} = (1/4\ 1/2\ 1/4\ 0\ 0\ 0)^T \quad \text{yields} \quad y = 1/2. \tag{7.8}$$

To improve the yield we note (with respect to (7.7)) that increasing $v_6$ is the only good idea (for it is the only variable with a positive weight). But how far can we go? We note that, with $v_4 = v_5$ still zero that $v_1 \geq 0$ requires $v_6 \leq (1/3)$ (and that positivity of $v_2$ and $v_3$ provide no upper bounds on $v_6$). Hence we set $v_6 = 1/3$ and $v_4 = v_5 = 0$ in Eq. (7.10) and find

$$v^{(2)} = (0\ 2/3\ 1/3\ 0\ 0\ 1/3)^T \quad \text{yields} \quad y = 2/3. \tag{7.9}$$

We can interpret the setting of $v_1 = 0$ to imply that $\{v_1, v_2, v_3\}$ was not the best choice of pivot variables. Hence, we might be better served by a system that expresses the new pivot variables $\{v_2, v_3, v_6\}$ in terms of the new free variables $\{v_1, v_4, v_5\}$. On rearranging Eq. (7.10) we find that

$$\begin{aligned} v_6 &= (1/3) - (4/3)v_1 - (1/3)v_4 + (2/3)v_5 \\ v_2 &= (2/3) - (2/3)v_1 - (2/3)v_4 + (1/3)v_5 \\ v_3 &= (1/3) - (1/3)v_1 + (2/3)v_4 - (1/3)v_5 \end{aligned} \tag{7.10}$$

is such a system. As its yield is

$$y = v_1 + v_3 - v_4 + v_6 = (2/3) - v_1 - (2/3)v_4 + (1/3)v_5,$$

we naturally consider increasing $v_5$. Its only bound is $v_5 \leq 1$ and at this bound (with $v_1 = v_4 = 0$), we find

$$v^{(3)} = (0\ 1\ 0\ 0\ 1\ 1)^T \quad \text{yields} \quad y = 1. \tag{7.11}$$

Returning to Figure 7.1, as the supplied flow into $m_1$ was 1 our yield cannot exceed 1, and $v^{(3)}$ is an optimal pathway. As a check on our calculations we might wish to confirm this algebraically. More precisely, we express the new basic variables $\{v_2, v_5, v_6\}$ in terms of the new free variables $\{v_1, v_3, v_4\}$. On rearranging Eq. (7.10) we find

$$\begin{aligned} v_6 &= 1 - 2v_1 - 2v_3 + v_4 \\ v_2 &= 1 - v_1 - v_3 \\ v_5 &= 1 - v_1 - 3v_3 + 2v_4 \end{aligned} \tag{7.12}$$

is such is system. As its yield

$$y = v_1 + v_3 - v_4 + v_6 = 1 - v_1 - v_3,$$

has no rates with positive weights we conclude that indeed no further growth is possible.

### 7.3. The Simplex Method

We now attempt to formalize these steps. The step from (7.2) to (7.6) begins by row reduction of $Sv = f$ and subsequent identification of the **basic**, or pivot, variables indexed by $b$, and **nonbasic**,

or free, variables indexed by $n$. This permits us to express $Sv = f$ as $S_b v_b + S_n v_n = f$, and to solve for the basic in terms of the nonbasic

$$v_b = S_b^{-1}(f - S_n v_n) = v_b^* - S_b^{-1} S_n v_n \quad \text{where} \quad v_b^* = S_b^{-1} f. \tag{7.13}$$

Here $S_b$ corresponds to those columns of $S$ indexed by $b$. This permits us to express the yield

$$y = c^T v = c_b^T v_b + c_n^T v_n = c_b^T S_b^{-1} f + (c_n^T - c_b^T S_b^{-1} S_n) v_n.$$

To increase the yield we consider

$$w_i = \max w \quad \text{where} \quad w \equiv c_n^T - c_b^T S_b^{-1} S_n. \tag{7.14}$$

If $w_i \leq 0$ then there is no room for growth and we are done. If however $w_i > 0$ then we flex its associated free variable (with index $j = n_i$) to increase the yield while staying feasible. With $t$ denoting the value of the entering free variable we stay feasible (recalling Eq. (7.13)) so long as the new

$$v_b = v_b^* - t S_b^{-1} s_j \geq 0, \tag{7.15}$$

where $s_j$ is the $j$th column of $S$. As the yield increases with $t$ we should choose $t$ to be the largest value for which Eq. (7.15) holds. That largest value is

$$t_p = \min t \quad \text{where} \quad t \equiv v_b^* ./ (S_b^{-1} s_j) \tag{7.16}$$

With this choice of $t$ in Eq. (7.15) we note that the $p$th element of $v_b^* - t_p S_b^{-1} s_j$ is zero. As such the index $b_p$ moves from the basic to the nonbasic set while index $n_i$ moves in the opposite direction and the values of the basic variables are the associated nonzero values of $v_b^* - t_p S_b^{-1} s_j$ at the old indices together with $t_p$ at the new index. We have now derived the

**Simplex Algorithm**

1. Given basic, $b$, and nonbasic, $n$, index sets and the basic feasible solution $v_b^*$.
2. Compute $w_i$ in Eq. (7.14).
3. If $w_i \leq 0$ then $v_b^*$ produces the maximal yield, $c_b^T v_b^*$, and you may stop. Otherwise set $j = n_i$ and compute $t_p$ in Eq. (7.16).
4. Replace $b_p$ in $b$ with $n_i$ and replace $n_i$ in $n$ with $b_p$. The new $v_b^*$ is $t_p$ at the new index and the nonzero values of $v_b^* - t_p S_b^{-1} s_j$ at the old indices. Return to step 2.

The "cost" of each iteration of the Simplex Algorithm is dominated by the computation of $c^T S_b^{-1} S_n$ and $S_b^{-1} s_j$. We can accelerate the first by using the associative property of matrix multiplication to free us from the costly $S_b^{-1} S_n$. In particular, setting $u = S_b^{-T} c$ brings the desired $u^T S_n = c^T S_b^{-1} S_n$. We have implemented this in `simplex.m`

### 7.4. The Geometric Point of View*

Our approach so far has been algebraic and algorithmic. We complement these here by illustrating the beautiful fashion in which the Simplex Method moves from vertex to vertex of the admissible polyhedron,

$$\mathcal{P} \equiv \{v \in \mathbb{R}^n : Sv = b,\ v \geq 0\}.$$

We will do this first for the concrete problem of the first section and then attack the general problem using the theory of convex sets developed in §5.4.

For the $S$ and $b$ of (7.2) the space of $n = 6$ rates is beyond our visual comprehension. As $S$ has only 3 rows however we should be able to reduce our problem to one in three dimensions. We will do this with the help of $S^+$, the pseudo–inverse of $S$, and $E$, a 6-by-3 matrix whose columns comprise a basis for $\mathcal{N}(S)$. When put into action these yields

$$\mathcal{P} \equiv \{v = v_0 + Eu : u \in \mathcal{P}_{red}\}$$

where

$$\mathcal{P}_{red} = \{u \in \mathbb{R}^3 : Eu \geq -v_0\}, \quad v_0 = S^+ b = \begin{pmatrix} 1/4 \\ 1/2 \\ 1/4 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad E = \begin{pmatrix} -1/4 & 1/2 & -3/4 \\ -1/2 & 0 & 1/2 \\ 3/4 & -1/2 & 1/4 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

with associated yield

$$y = c^T v = 1/2 + 1/2(u_3 - u_1). \tag{7.17}$$

We write out $Eu \geq -v_0$

$$\begin{aligned} u_1 - 2u_2 + 3u_3 &\leq 1 \\ u_1 - u_3 &\leq 1 \\ -3u_1 + 2u_2 - u_3 &\leq 1 \end{aligned} \tag{7.18}$$

and note that as the yield, Eq. (7.17), is independent of $u_2$ it suffices to study this system in the $(u_1, u_3)$ plane.

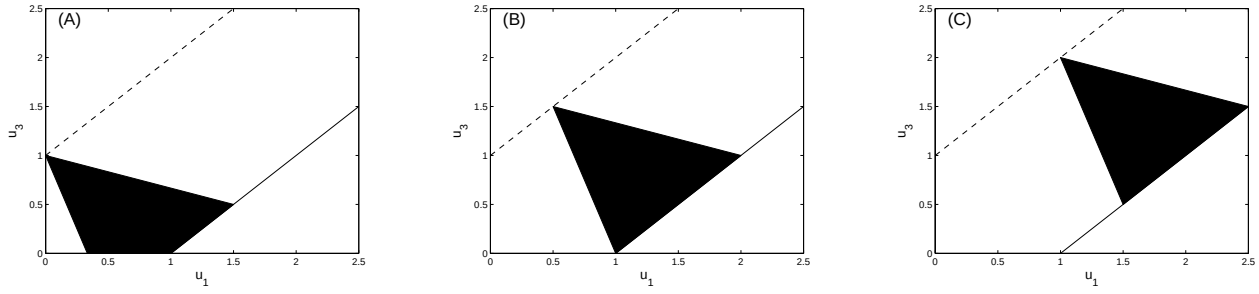



**Figure** 7.2. Projection of $\mathcal{P}_{red}$ onto the $(u_1, u_3)$ plane for three values of $u_2$. In each case the left face is the line $u_3 = 2u_2 - 3u_1 - 1$, the right face is the (solid) line $u_3 = u_1 - 1$ and the top face is $u_3 = (1 + 2u_2 - u_1)/3$. The yield, Eq. (7.17), is strictly a function of $u_3 - u_1$ and we have plotted $u_3 - u_1 = 1$ as a dotted line in each panel. (A) $u_2 = 1$ and bottom face is the line $u_3 = 0$. (B) $u_2 = 2$. (C) $u_2 = 3$.

## 7.5. Succinate Production*

$v_1$ : glu + pep → g6p + pyr

$v_2$ : g6p → 2g3p

$v_3$ : g3p + nadp → pep + nadh

$v_4$ : pep → pyr

$v_5$ : isocit +2nadp → succ +2nadh

$v_6$ : pyr + nadp → aCoA + nadh

$v_7$ : aCoA → ace

$v_8$ : aCoA + 2nadh → eth + 2nadp

$v_9$ : isocit → glyox + succ

$v_{10}$ : glyox + aCoA → mal

$v_{11}$ : oaa + nadh → mal + nadp

$v_{12}$ : mal → fum

$v_{13}$ : fum + nadh → succ + nadp

$v_{14}$ : aCoA + oaa → isocit

$v_{15}$ : pyr → form + aCoA

$v_{16}$ : form + nadp → cdh + nadh

$v_{17}$ : glu → g6p

$v_{18}$ : pyr → oaa

$v_{19}$ : pep → oaa

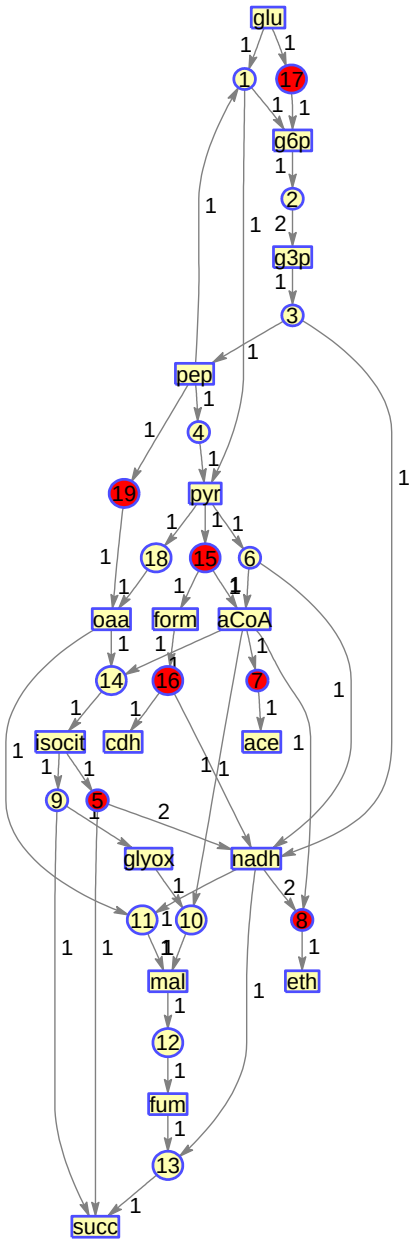



**Figure** 7.3 A set of 19 reactions among 16 metabolites from central metabolism and its associated network. Abbreviations: **glu**cose,

$$S = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
-1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\
0 & -2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
-1 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & -1 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & 0 & -2 & -1 & 0 & 2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0
\end{pmatrix} \tag{7.19}$$

where the supply is

$$f^T = (1 \ \texttt{zeros}(15,1))$$

and the yield is

$$y = v_5 + v_9 + v_{13} = c^T v \quad \text{where} \quad c^T = (0,0,0,0,1,0,0,0,1,0,0,0,1,0,0,0,0,0,0).$$

Our design problem is then

$$\boxed{\max \ c^T v \quad \text{subject to} \quad Sv = f, \quad 0 \le v \in \mathbb{R}^r}. \tag{7.20}$$

Proceeding as above, our $S$ is 16-by-19 with a three-dimensional nullspace and even identifying a starting point is problematic. We look for a systematic approach. Perhaps the simplest is to augment $S$ and $v$ to

$$\tilde{S} \equiv [S \ I_m] \quad \text{and} \quad \tilde{v}^T \equiv (\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_r, \tilde{v}_{r+1}, \tilde{v}_{r+2}, \ldots, \tilde{v}_{r+m})$$

and to consider the auxiliary problem

$$\boxed{\max \ p^T \tilde{v} \quad \text{subject to} \quad \tilde{S}\tilde{v} = f, \quad 0 \le \tilde{v} \in \mathbb{R}^{r+m}} \tag{7.21}$$

where $p$ is chosen to penalize the augmented variables,

$$p^T = (\texttt{zeros}(\texttt{r},1), \ -\texttt{ones}(\texttt{m},1)).$$

It follows that if the maximum value of the auxiliary problem is zero then each of the augmented variables must vanish and so we have a starting point for the real problem. Regarding a starting point for Problem (7.21) we have given ourselves so much elbow room that we simply choose our basic variables to be the auxiliary variables, i.e.,

$$\widehat{b} = (r+1, \ldots, r+m) \quad \text{and} \quad \tilde{v}^* = (\texttt{zeros}(\texttt{r},1); \ f). \tag{7.22}$$

Lets first test this approach on the small example. Our basic indices begin at $(7,8,9)$ and yield $-1$. The algorithm then exchanges 8 for 1 then 9 for 3 to bring $(7,1,3)$, all without increasing the yield. From here it enters the never ending cycle $(7,1,3) \to (7,2,3) \to (7,1,3)$. This is a well documented pathology for which there exist a number of remedies. Perhaps the simplest is just to

nudge $f$ a bit. To be precise we replace $f$ with $\tilde{f}$ where $\tilde{f}_j = f_j + 10^{-6j}$. With this **perturbation** our scheme again moves from $(7, 8, 9)$ to $(7, 1, 3)$ but then to $(2, 1, 3)$ and a yield of 0 and a starting vector precisely as $v^{(1)}$ in Eq. (7.8).

When applying this augmented and perturbed approach to the large problem we encounter yet another obstacle/nuance. We drive the yield to 0 but the basic indices include augmented variables! In particular

$$\widehat{b} = (2, 4, 3, 1, 6, 18, 5, 13, 12, 14, 10, 31, 32, 33, 11, 8)$$

The offenders being $31, 32$ and $33$. One solution is to replace them with proper columns of $S$, that is columns with indices in $n = \widehat{n} \cap \{1 : r\}$. Via,

**Replacement**

1. for each $k$ such that $r + k \in \widehat{b}$
2. solve $\widehat{S}_{\widehat{b}}^T r = \mathbf{e}_k$ where $\mathbf{e}_k$ is the $k$th column of $I_m$.
3. find a $j \in n$ such that $s_j^T r \neq 0$.
4. Exchange $r + k$ for $j$ in $\widehat{b}$.

At the optimal design

$$v = (7, 7, 14, 7, 0, 4, 0, 0, 2, 2, 8, 10, 10, 2, 0, 0, 0, 10, 0)^T/7$$

we achieve a succinate yield of $12/7$ and we produce no acetate, $CO_2$ and ethanol, in fact we do not direct any pyruvate to formate (the precursor of $CO_2$), and we do not split resources at the redundant pathways from glucose to g6p, from pep to oaa and from isocitrate to succinate.

### 7.6. Elementary Flux Modes and Extremal Rays*

We have so far specified specific source and sink metabolites. We now embark on the more difficult problem of achieving a minimal description of the balance of interior metabolites. Metabolically we seek all minimal sets of enzymes consistent with flux balance and thermodynamics. More precisely, given the stoichiometric matrix $S$ (restricted to $m$ internal metabolites participating in $d$ reactions) we denote by

$$\mathcal{C}(S) \equiv \{v \in \mathbb{R}^d : Sv = 0,\ v \geq 0\} \tag{7.23}$$

the set of all possible flux modes. A flux mode $v$ is said to be elementary if there is no other flux mode that uses a proper subset of the reactions of $v$. In symbols, define

$$z(v) = \{i : v_i = 0\}. \tag{7.24}$$

Then $v \in \mathcal{C}(S)$ is an **elementary flux mode** if there does not exist a $u \in \mathcal{C}(S)$ for which $z(u) \subset z(v)$.

Let's consider a concrete example. With respect to our first net, see Figure 7.1, the first metabolite, $m_1$, is sourced and the balance only $m_2$ and $m_3$ is governed by the stoichiometric matrix

$$S = \begin{pmatrix} 2 & -1 & 0 & 0 & -1 & 2 \\ 0 & -1 & 2 & -2 & 1 & 0 \end{pmatrix}. \tag{7.25}$$

The associated set of flux modes, $\mathcal{C}(S)$, is then the intersection of the positive orthant in $\mathbb{R}^6$ with the 4–dimensional subspace, $\mathcal{N}(S)$. For most, this is not an easy set to visualize. We will see that its elementary flux modes give us metabolically significant concrete basis vectors with which we may represent the entire set of flux modes. It turns out that the $S$ of Eq. (7.25) has the seven elementary flux modes, arranged as the columns of

$$R = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 2 & 2 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 2 & 1 & 2 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \end{pmatrix}. \tag{7.26}$$

Each column is indeed a flux mode, i.e., a nonnegative member of $\mathcal{N}(S)$. It is instructive to check that each is also elementary. Taking the first column we note that it uses only reactions 3 and 4. To see that it is elementary we must check that no flux mode could operate on only one of these reactions. This is clear both from the second row of $S$ as well as from Figure 7.1 – for reaction 3 or 4 alone would leave $m_3$ unbalanced.

To enumerate the elementary flux modes, as in Eq. (7.26), as well as to prove that they indeed constitute a kind of "basis" we translate our problem into one of enumerating the extremal rays in a pointed polyhedral cone.

To begin, from the $m \times d$ stoichiometric matrix $S$ we build the $(d + 2m) \times d$ matrix

$$A = \begin{pmatrix} I_d \\ S \\ -S \end{pmatrix} \tag{7.27}$$

and note that

$$\mathcal{C}(A) \equiv \{v \in \mathbb{R}^d : Av \geq 0\} \tag{7.28}$$

is precisely the set $\mathcal{C}(S)$ of flux modes in Eq. (7.23). There is a rich geometric language (and theory) associated to such objects. We call the members of $\mathcal{C}(A)$ **rays** because if $v \in \mathcal{C}(A)$ then so is $av$ for every $a \geq 0$. If $K$ is a set of row indices and $A_K$ denotes the associated rows of $A$ then

$$\mathcal{F}_K \equiv \{v \in \mathcal{C}(A) : A_K v = 0\} \tag{7.29}$$

is called a **face** of $\mathcal{C}(A)$. The **dimension** of a face is its number of linearly independent rays. A face of dimension 1 is called an **extreme ray** of $\mathcal{C}(A)$. Two extreme rays of $\mathcal{C}(A)$ are said to be **adjacent** if they span a two-dimensional face of $\mathcal{C}(A)$. Before considering the large stoichiometrically relevant $A$ lets pause to illustrate these definitions on the two small examples

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad A' = \begin{pmatrix} 1 & -1 & 0 \\ -8 & 0 & 5 \\ -1 & 5 & 0 \\ -1 & 0 & 5 \end{pmatrix} \tag{7.30}$$

Their associated cones, $\mathcal{C}(A)$ and $\mathcal{C}(A')$, lie in $\mathbb{R}^2$ and $\mathbb{R}^3$ respectively and so may be depicted as in Figure 7.4. The first cone is specially simple. It is a two dimensional face and

$$\mathcal{F}_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = r_1 \quad \text{and} \quad \mathcal{F}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} = r_2$$

are two adjacent extremal rays and every vector in $\mathcal{C}(A)$ is a positive linear combination of $r_1$ and $r_2$. The second example is considerably richer. Each row inequality $A_i'v \geq 0$ divides $\mathbb{R}^3$ into two half-spaces and the intersection of these half–spaces comprise $\mathcal{C}(A')$. The boundaries of these half–spaces are planes and these 4 planes intersect at the extremal rays, marked $r_1$ through $r_4$ in Figure 7.4. For example, $A_2'$ generates the top face and $A_3'$ generates the left face and so $r_3$ is the extremal ray $\mathcal{F}_{\{2,3\}}$. The full set of extremal rays appear in the columns of

$$R' = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0.2 & 0.2 \\ 0.5 & 1.6 & 1.6 & 0.5 \end{pmatrix} \tag{7.31}$$

and we observe, at least graphically, that any vector in $\mathcal{C}(A)$ is a positive linear combination of columns of $R$.

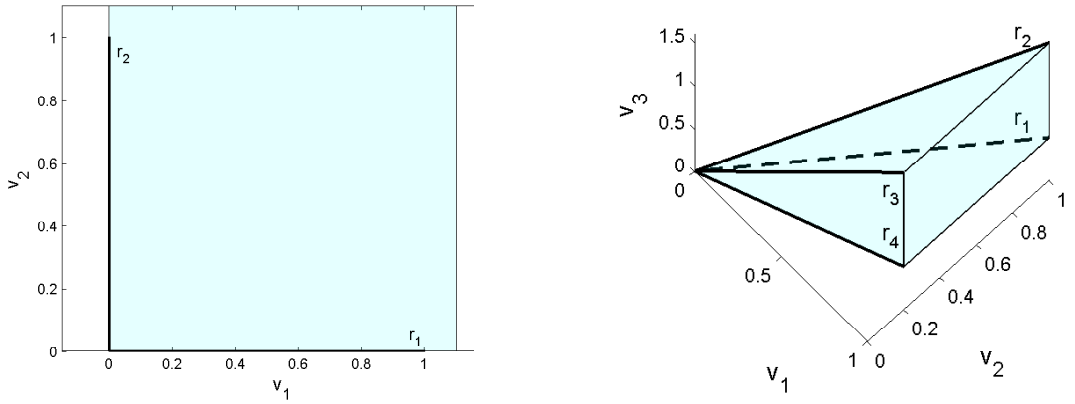



**Figure** 7.4. The cones (shaded regions) $\mathcal{C}(A)$ and $\mathcal{C}(A')$ for the two matrices specified in Eq. (7.30).

The fact that $A'$ and $R'$ independently fully determine the same cone has lead to the name **double description** for the pair $(A, R)$. In what follows we describe a procedure for building $R$ from $A$ for $A$ of the form Eq. (7.27).

For any vector $v \in \mathcal{C}(A)$ we define the zero set with respect to $A$ as the indices of the rows of $A$ that are orthogonal to $v$. That is

$$\zeta(v) = \{i : A_i v = 0\}. \tag{7.32}$$

This notation allows us to prove

**Proposition** 7.1. $\mathcal{F}_{\zeta(r)}$ is the smallest face in $\mathcal{C}(A)$ containing $r$.

Proof: As $A_{\zeta(r)}r = 0$ we see that $\mathcal{F}_{\zeta(r)}$ is a face containing $r$. If $\mathcal{F}_K$ is a face containing $r$ then $A_K r = 0$ and so $K \subset \zeta(r)$ and so $\mathcal{F}_{\zeta(r)} \subset \mathcal{F}_K$. End of Proof.

**Proposition** 7.2. If $r$ is a ray of $\mathcal{C}(A)$ and $\text{rank}(A_{\zeta(r)}) = d - k$ then
(a) $\text{rank}(A_{\zeta(r)\cup\{i\}}) = d - k + 1$ for each $i \notin \zeta(r)$.
(b) $\dim(\mathcal{F}_{\zeta(r)}) = k$.
(c) If $k \geq 2$ then $r$ is a nonnegative combination of two distinct rays $r_1$ and $r_2$ with $\text{rank}(A_{\zeta(r_i)}) > d - k$ for $i = 1, 2$.

Proof: (a) We must show that $A_i \notin \text{span}(A_{\zeta(r)})$. But this follows from $A_i r \neq 0$ and $A_{\zeta(r)} r = 0$.

(b) As $\dim(\mathcal{N}(A_{\zeta(r)})) = k$ it contains $k$ linearly independent vectors, say $r, v_2, v_3, \ldots v_k$. We now turn these into rays. Set $r_1 \equiv r$, $r_i \equiv r + a_i v_i$, $i = 2, \ldots k$. These vectors are linearly independent so long as each $a_i \neq 0$. If $v_i \in \mathcal{C}(A)$ set $a_i = 1$. Otherwise choose $a_i$ to guarantee $A_j r + a_i A_j v_i \geq 0$ for each $j$. Note that

$$0 < a_i \leq \min_j \{-A_j r / A_j v_i : A_j v_i < 0\}$$

suffices. The $k$ vectors $r_i$, $i = 1, \ldots, k$ are linearly independent rays in $\mathcal{F}_{\zeta(r)}$.

(c) If $k \geq 2$ then there exist $i$ and $j$ such that $\text{rank}(A_{\zeta(r)\cup\{i,j\}}) = \text{rank}(A_{\zeta(r)}) + 2$. Hence $A_j \notin \text{span}(A_{\zeta(r)\cup\{i\}})$ it follows from the Fundamental Theorem of Linear Algebra that $A_j$ is not orthogonal to $\mathcal{N}(A_{\zeta(r)\cup\{i\}})$ and so there exists $u_1 \in \mathcal{N}(A_{\zeta(r)\cup\{i\}})$ such that $A_j u_1 \neq 0$. Without loss we can assume that $A_j u_1 > 0$. By construction we also note that $A_i u_1 = 0$. By the same reasoning there exists a $u_2 \in \mathcal{N}(A_{\zeta(r)\cup\{j\}})$ such that $A_i u_2 < 0$ and $A_j u_2 = 0$.

Now $v \equiv u_1 + u_2$ satisfies $A_i v < 0$ and $A_j v > 0$. Let $r_1 \equiv r + a_1 v$ and $r_2 \equiv r - a_2 v$ with

$$a_1 = \min_m \{-A_m r / A_m v : A_m v < 0\} \quad \text{and} \quad a_2 = \min_n \{A_n r / A_n v : A_n v > 0\}. \tag{7.33}$$

By construction these are both strictly positive and the resulting $r_1$ and $r_2$ both lie in $\mathcal{F}_{\zeta(r)}$ and clearly

$$r = \frac{a_2}{a_1 + a_2} r_1 + \frac{a_1}{a_1 + a_2} r_2$$

is a positive combination of two rays. Finally, we denote by the $m_1$ and $n_2$ the idiocies at which the respective minima are attained in Eq. (7.33). As $A_{m_1} r_1 = A_{n_2} r_2 = 0$ it follows $\zeta(r)$ is strictly contained in $\zeta(r_1)$ and $\zeta(r_2)$ and so $\mathcal{N}(A_{\zeta(r_1)})$ and $\mathcal{N}(A_{\zeta(r_2)})$ both have dimensions strictly larger than $k$. End of Proof.

**Proposition** 7.3. Suppose that $r$ is a ray of $\mathcal{C}(A)$.
(a) If $r$ is a nonnegative combination of rays of $\mathcal{C}(A)$, say

$$r = \sum_j \lambda_j r_j, \quad \lambda_j > 0 \tag{7.34}$$

then each $r_j \in \mathcal{N}(A_{\zeta(r)})$.
(b) $r$ is extreme iff $\text{rank}(A_{\zeta(r)}) = d - 1$.
(c) $r$ is a nonnegative combination of extreme rays of $\mathcal{C}(A)$.

Proof: (a), We have already noted that $r \in \mathcal{N}(A_{\zeta(r)})$ hence applying $A_{\zeta(r)}$ to each side of Eq. (7.34) brings

$$0 = \sum_j \lambda_j A_{\zeta(r)} r_j.$$

but as each term in the sum is nonnegative it follows that each term is in fact zero.

(b), If $\text{rank}(A_{\zeta(r)}) = d-1$ , then $\mathcal{N}(A_{\zeta(r)}) = \{ar : a \in \mathbb{R}\}$ and so $\mathcal{F}_{\zeta(r)}$ is a one–dimensional face spanned by $r$. That is, $r$ is extreme. To prove the converse we prove its contrapositive. If $\text{rank}(A_{\zeta(r)}) < d-1$ then by Prop. 7.2(b) then $\dim(\mathcal{F}_{\zeta(r)}) > 1$ and so $r$ is not extremal.

(c), If $r$ is not extreme then $\dim(\mathcal{N}(A_{\zeta(r)})) = k > 1$ and so, by Prop. 7.2(c), $r = a_1 r_1 + a_2 r_2$ where $r_1$ and $r_2$ obey $\dim(\mathcal{N}(A_{\zeta(r_i)})) < \dim \mathcal{N}((A_{\zeta(r)}))$. If $k = 2$ we are done. If not apply this same reasoning to $r_1$ and $r_2$ (and their descendants) until one–dimensional null spaces are achieved. End of Proof.

Since every extreme ray is certainly necessary to generate $\mathcal{C}(A)$ we have the following

**Corollary** 7.4. Let $R$ be a minimal generating matrix of $\mathcal{C}(A)$. Then $R$ is the set of extreme rays of $\mathcal{C}(A)$.

Our final step is to establish a rank test for adjacency.

**Proposition** 7.5. Let $r_1$ and $r_2$ be distinct extremal rays of $\mathcal{C}(A)$. They are adjacent iff $\text{rank}(A_{\zeta(r_1)\cap\zeta(r_2)}) = d-2$.

Proof: Let $r_1$ and $r_2$ be distinct rays of $\mathcal{C}(A)$. Then $\mathcal{F}_{\zeta(r_1)\cap\zeta(r_2)}$ is the minimal face of $\mathcal{C}(A)$ containing $r_1$ and $r_2$. To prove the equivalence of (a) and (b) let $r_1$ and $r_2$ be extreme rays of $\mathcal{C}(A)$. Since $\mathcal{C}(A)$ is pointed, we have $\zeta(r_1) \neq \zeta(r_2)$.

We prove that (a) implies (b). If $r_1$ and $r_2$ are adjacent then they span a two–dimensional face. As $\mathcal{F}_{\zeta(r_1)\cap\zeta(r_2)}$ is the smallest such face it follows that $\dim(\mathcal{N}(A_{\zeta(r_1)\cap\zeta(r_2)})) \leq 2$. But, as $\mathcal{N}(A_{\zeta(r_1)\cap\zeta(r_2)})$ contains the two distinct one dimensional spaces, $\mathcal{N}(A_{\zeta(r_1)})$ and $\mathcal{N}(A_{\zeta(r_2)})$, it follows that $\dim(\mathcal{N}(A_{\zeta(r_1)\cap\zeta(r_2)})) \geq 2$.

We prove that (b) implies (a). If $\dim(\mathcal{N}(A_{\zeta(r_1)\cap\zeta(r_2)})) = 2$ then $r_1$ and $r_2$ generate $\mathcal{F}_{\zeta(r_1)\cap\zeta(r_2)}$, that is, each such $\mathbf{x}$ can be written $\mathbf{x} = a_1 r_1 + a_2 r_2$. To secure the signs of $a_1$ and $a_2$ note that if $i \in \zeta(r_2) \setminus \zeta(r_1)$ then

$$A_i \mathbf{x} = a_1 A_i r_1 + a_2 A_i r_2 = a_1 A_i r_1.$$

From the positivity of $A_i\mathbf{x}$ and $A_i r_1$ comes the positivity of $a_1$. To show that $a_2 > 0$ repeat this argument with $j \in \zeta(r_1) \setminus \zeta(r_2)$. As $r_1$ and $r_2$ span a two–dimensional face they are adjacent. End of Proof.

**Proposition** 7.6. Let $(A_K, R)$ be a DD pair such that $\text{rank}(A_K) = d$ and $R$ is minimal. Select a row index, $i$, of $A$ not in $K$. The new inequality $A_i v \geq 0$ partitions the columns of $R$ via

$$R^+ = \{r_j : A_i r_j > 0\}, \quad R^0 = \{r_j : A_i r_j = 0\} \quad \text{and} \quad R^- = \{r_j : A_i r_j < 0\}.$$

Retain the vectors in the first two sets for the updated

$$\tilde{R} = [R^+ \; R^0]$$

and whenever $r^+ \in R^+$ and $r^- \in R^-$ are adjacent append

$$\tilde{r} = (A_i r^+) r^- - (A_i r^-) r^+ \tag{7.35}$$

to $\tilde{R}$. Then $(A_{K\cup\{i\}}, \tilde{R})$ is a DD pair and $\tilde{R}$ is minimal.

Proof: We know from Prop. 7.3 that each extreme ray of $\mathcal{C}(A_K)$ must belong to $R$, and that only the extreme rays of $\mathcal{C}(A_{K\cup\{i\}})$ are necessary in $\tilde{R}$. For $r^{\pm} \in R^{\pm}$ set $W \equiv \zeta(r^+) \cap \zeta(r^-) \cap K$ and define $\tilde{r}$ as in Eq. (7.35). Note that $\zeta(\tilde{r}) \cap (K \cup \{i\}) = W \cup \{i\}$. If $r^{\pm}$ are adjacent extreme rays of $\mathcal{C}(A_K)$, then $\text{rank}(A_W) = d-2$. Then, by Prop. 7.2(a), $\text{rank}(A_{\zeta(\tilde{r})\cap(K\cup\{i\})}) = d-1$. Thus $\tilde{r}$ is an extreme ray of $\mathcal{C}(A_{K\cup\{i\}})$ and must belong to $\tilde{R}$.

Suppose $r^{\pm}$ are not adjacent extreme rays. If $\text{rank}(A_W) < d-2$ we get $\text{rank}(A_{\zeta(\tilde{r})\cap(K\cup\{i\})}) < d-1$ and so $\tilde{r}$ is not an extremal ray in $\mathcal{C}(A_{K\cup\{i\}})$. If rather $\text{rank}(A_W) = d-2$, then we know from Prop. 7.6 that $r^{\pm}$ cannot be both extreme rays of $\mathcal{C}(A_K)$. But they belong to a two dimensional face containing exactly two extreme rays of $\mathcal{C}(A_K)$ which thus belong to $R$: This adjacent pair will then produce a new ray equal to $\tilde{r}$, so $\tilde{r}$ is not necessary. Hence all new rays are extreme rays of $\mathcal{C}(A_{K\cup\{i\}})$ and $\tilde{R}$ is minimal. End of Proof.

I have coded this procedure in `efm.m`. When run on the $S$ of Eq. (7.25) we get the $R$ of Eq. (7.26) When run on the $S$ of Eq. (7.19) without rows 1 (glu), 14(CO2), 15(ace) and 16(eth) we get

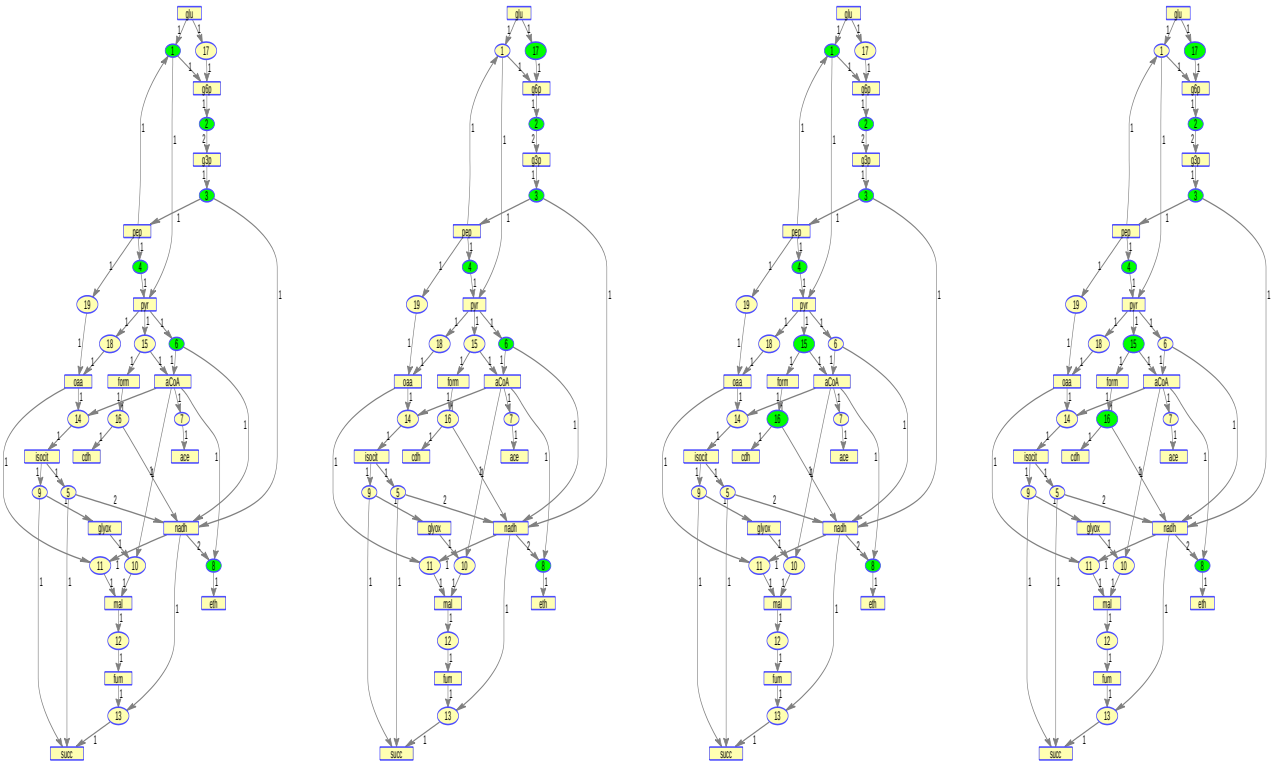


**Figure** 7.5 Elementary Flux Modes for ...

## 7.7. Notes and Exercises

For a general introduction to Metabolic Engineering see GN Stephanopoulos and Nielsen (1998). For Linear Programming we have followed Chvatal (1983). Our presentation of the Double Desription method for capturing Extremal Rays follows Fukuda and Prodon (1996).

1. Solve max $v_1$ subject to $v_1 + v_2 = 1$, $v_1 \geq 0$ and $v_2 \geq 0$ by graphing the constraints and seeing the largest $v_1$.

2. This leaves us then to solve the two linear systems

$$S_b z = s_j \quad \text{and} \quad S_b^T u = c.$$

We naturally use the lu factorization $S_b = LU$ to conclude that

$$z = U\backslash(L\backslash s_j) \quad \text{and} \quad u = L^T\backslash(U^T\backslash c).$$

3. Anaerobic Succinate network.

4. We consider the cycling reported in the last section for

$$\widehat{S} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 2 & -1 & 0 & 0 & -1 & 2 & 0 & 1 & 0 \\ 0 & -1 & 2 & -2 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad f = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \tag{7.36}$$

and $c^T = (0,0,0,0,0,0,-1,-1,-1)^T$ and

$$\widehat{b} = (7,8,9) \quad \text{and} \quad \widehat{v}_{\widehat{b}}^{\,*} = f \quad \text{and} \quad \widehat{S}_{\widehat{b}} = I_3.$$

(a) Execute one iteration of the Simplex Algorithm by hand. In particular, show that Eq. (7.14) brings

$$w = (3,-1,3,-2,0,2,1,1,1).$$

and argue why column 1 will now become basic. To find the departing column show that Eq. (7.15) asks for the least nonnegative $t$ for which

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} - t \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix} \geq 0, \tag{7.37}$$

and so conclude that $t = 0$ is the desired value. As this is forced by the second element in Eq. (7.37) conclude that the second index of $\widehat{b}$ must leave and conclude that $\widehat{b} = (7,1,9)$. Argue that as $t = 0$ this step will not increase the yield.

(b) Now reflect on our remedy. In particular, note that if the column on the left in Eq. (7.37) was actually $(1+\varepsilon;\varepsilon^2;\varepsilon^3)$ for some small $\varepsilon$ then the least nonnegative $t$ would be $\varepsilon^2/2$ and that the yield increase accordingly.

5. Lets establish the validity of each of the **Replacement Algorithm**.

(a) If $S$ is full rank then there exists a nonbasic column of $S$ which is not orthogonal to $r$.

(b) Show that after each replacement that $\widehat{S}_{\widehat{b}}$ remains invertible. Hint: Let $S_1$ and $S_2$ denote $\widehat{S}_{\widehat{b}}$ before and after replacement of $\mathbf{e}_k$ with $s_j$. Prove that $S_2 = S_1\mathbf{E}$ where $\mathbf{E}$ is the identity matrix except for its $k$th column is $S_1^{-1}s_j$. As $S_1$ is invertible prove that $S_2$ is invertible if $\mathbf{E}$ is invertible. Argue that $\mathbf{E}$ is invertible if the $k$th element of $S_1^{-1}s_j$ is nonzero. Finally argue that this element is precisely $s_j^T r$, which was shown in (a) above to be nonzero.

6. This device of augmentation also applies nicely to the larger class of problems with linear **inequalities**. For example, the inequality $v_1 + v_2 \leq f_1$ can be handled by introducing a new variable, $v_3$, and requiring that both $v_1 + v_2 + v_3 = f_1$ and $v_3 \geq 0$.

(a) Use this idea to transform max $v_1$ subject to $v_1 + v_2 \leq 1$, $v_1 - v_2 \leq 0$, $v_1 \geq 0$ and $v_2 \geq 0$ into an equality constrained problem. Note that Eq. (7.22) is indeed a basic feasible solution. Proceed to solve it by hand by the simplex method. Confirm your answer by graphically solving the original problem.

(b) Adapt `TwoPhaseSimplex.m` to solve the mixed problem

$$\max c^T v, \quad \text{subject to} \quad Sv = b, \quad Av \leq \mathbf{d}, \quad v \geq 0.$$

# 8. Dynamical Systems

We shift our focus here from linear systems of *algebraic* equations to linear systems of *differential* equations in the quest to understand the dynamics of electrical, mechanical and metabolic networks. Although the dynamical phenomena that we will be able to explain are indeed substantially richer than the equilibria that occupied our earlier study we will see, miraculously, that the requisite mathematical techniques remain essentially *algebraic* in nature.

Our first, and easiest, task will be to show, with regard to building models, that the methodology of the Strang Quartet extends naturally to linear dynamical systems. Next, we will argue that the Laplace Transform takes the differential equation $x'(t) = Bx(t)$, where $B \in \mathbb{R}^{n \times n}$, into an algebraic equation whose easy solution may pass through the Inverse Laplace Transform, and back to $x(t)$, once one has mastered the eigenvalue problem, $Bv = \lambda v$. In this chapter we *motivate*, via many varied examples, rather than *master*, the eigenvalue problem. These examples provide a physical appreciation for the importance of eigenvalues to dynamics – largely through their role in the matrix exponential, $\exp(Bt)$. This appreciation, it is hoped, will whet the appetite for mastery of the eigenproblem and sustain the reader through two supporting chapters on the calculus of functions of a complex variable.

In addition to our analytical approach through the Laplace Transform and eigenvalue problem we also pursue approximate, or numerical, means to solving $x'(t) = Bx(t)$. On discretizing time this equation becomes *immediately* algebraic and so very simple, in light of our last 7 chapters, to code and analyze. We carry this out for electrical and mechanical networks and show that this method extends naturally to the nonlinear systems, $x'(t) = F(x(t))$, that appear in modeling the dynamics of metabolic networks.

## 8.1. Dynamics of Electrical Networks

In Chapter 2 we modeled and obtained the neuron's response to a *steady* (constant) stimulus. In reality, neurons receive and integrate an ongoing barage of *transient* stimuli. These time varying stimuli engage time sensitive properties of the cell membrane. In particular, beginning with the single compartment model in Figure 8.1, as the cell membrane separates charge it produces a capacitative current

$$y_2(t) = C_m \frac{dx(t)}{dt} \tag{8.1}$$

whenever the charge moves. Here $x(t)$ is the transmembrane potential at time $t$ and $C_m = A_S c$ is the whole cell capacitance, where $A_S$ is the surface area of the cell membrane, in $cm^2$, and $c$ is the native capacitance of the cell membrane, typically 1 $\mu F/cm^2$ (micro Farad per square centimeter). Balancing currents in the single compartment neuron of Figure 8.1 brings the **differential equation**

$$C_m x'(t) + x(t)/R_m = i_0(t) \tag{8.2}$$

in response to the transient current stimulus, $i_0(t)$.

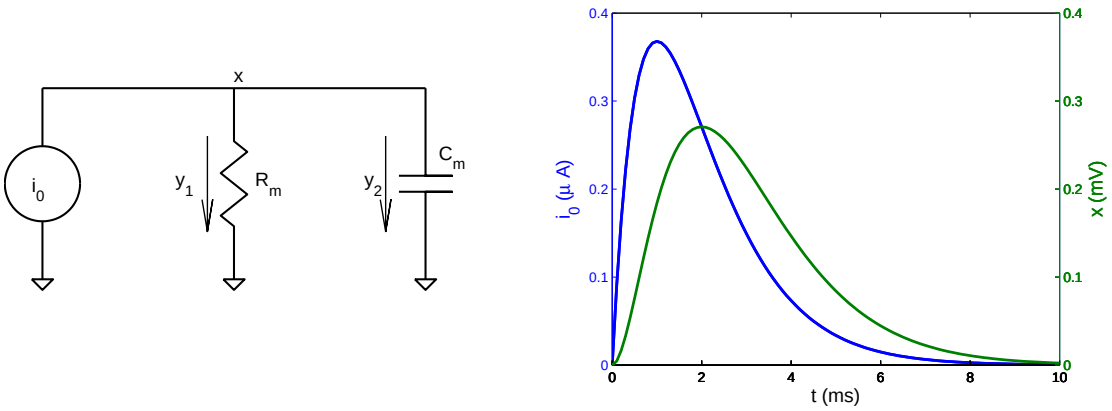



**Figure** 8.1. (A) A single compartment neuron. (B) In the clean case where $R_m = 1\ k\Omega$ and $C_m = 1\mu F$ and (B) The stimulus $i_0(t) = t\exp(-t)\ \mu A$ and response $x(t) = (t^2/2)\exp(-t)\ mV$ in the case where $x(0) = 0$ and $R_m = 1\ k\Omega$ and $C_m = 1\mu F$.

The easiest differential equations to solve are the ones that are pure derivatives. The key step is to see in (8.2) the product rule for differentiation in the guise

$$(x(t)\exp(t))' = (x'(t) + x(t))\exp(t).$$

For, in the clean setting where $R_m = C_m = 1$ and $i_0(t) = t\exp(-t)$, this permits us to express (8.2) as

$$(x(t)\exp(t))' = t.$$

As the left is a pure derivative we may integrate both sides and find

$$x(t)\exp(t) - x(0) = t^2/2, \quad \text{i.e.,} \quad x(t) = \exp(-t)x(0) + \exp(-t)t^2/2.$$

We have plotted this in Figure 8.1(B). We have made a clean choice of resistance and capacitance values. In general, (8.2) looks like

$$x'(t) + x(t)/\tau = f(t), \quad \text{where} \quad \tau = R_mC_m \quad \text{and} \quad f(t) = i_0(t)/C_m. \tag{8.3}$$

where $\tau = R_mC_m$ and $f(t) = i_0(t)/C_m$. On multiplying both sides of (8.3) by $\exp(t/\tau)$ it takes the form

$$(x(t)\exp(t/\tau))' = \exp(t/\tau)f(t).$$

On integrating each side from $t = 0$ to $t = T$ the Fundamental Theorem of Calculus yields

$$x(T)\exp(T/\tau) - x(0) = \int_0^T \exp(t/\tau)f(t)\,dt,$$

which after moving $x(0)$ the other side, and multiplying through by $\exp(-T/\tau)$ brings the final, explicit representation

$$\boxed{x(T) = \exp(-T/\tau)x(0) + \int_0^T \exp((t-T)/\tau)f(t)\,dt.} \tag{8.4}$$

You might wish to evaluate this expression for a variety of stimuli, e.g., step functions and sinusoids.

We now move on to the multicompartment circuit (neuron) and show that the strategy of the Strang Quartet from Chapter 2 offers a principaled path to larger models and that the explicit solution, (8.4), to the scalar problem has a natural generalization in terms of the matrix exponential. In order that the details not obscure the ideas we proceed gently and consider first the two compartment model of Figure 8.2.

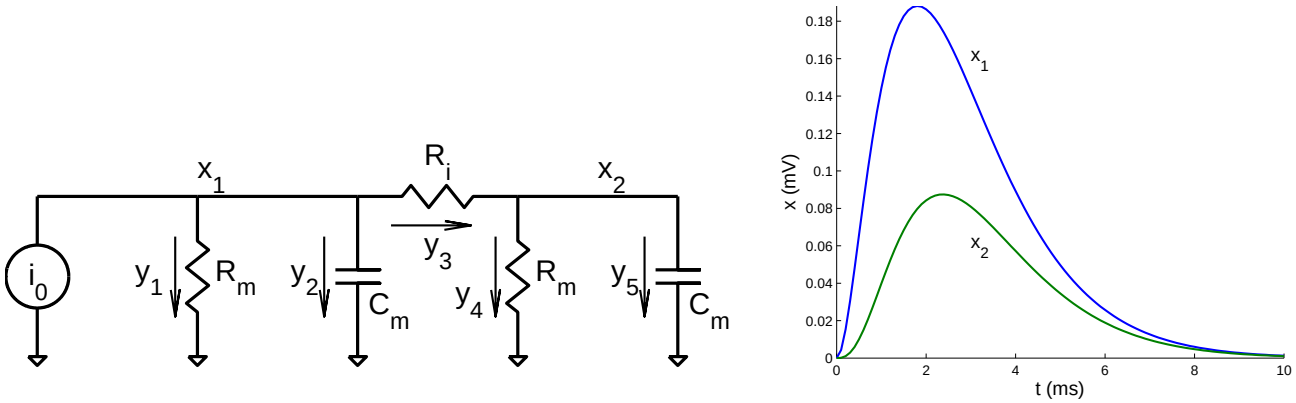



**Figure** 8.2. (A) A two compartment RC model of a neuron. (B) Its response, when $R_i = R_m = C_m = 1$, to $i_0(t) = t\exp(-t)$, as established in (8.23).

With $N$ compartments each compartment has length $\ell/N$ and radius $a$ and so capacitance $C_m = 2\pi a(\ell/N)c$. We ask now how the static Strang Quartet of Chapter 2 should be augmented. Regarding (S1') we proceed as before. The voltage drops are

$$e_1 = x_1, \quad e_2 = x_1, \quad e_3 = x_1 - x_2, \quad e_4 = x_2, \quad e_5 = x_2,$$

and so

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

In (S2) we must now augment Ohm's law with voltage–current law obeyed by a capacitor, namely (8.1). This yields,

$$y_1 = C_m e_1', \quad y_2 = e_2/R_m, \quad y_3 = e_3/R_i, \quad y_4 = C_m e_4', \quad y_5 = e_5/R_m$$

or, in matrix terms,

$$y = Ge + Ce'$$

where $G = \text{diag}(0\ 1/R_m\ 1/R_i\ 0\ 1/R_m)$ and $C = \text{diag}(C_m\ 0\ 0\ C_m\ 0)$ 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. That is,

$$i_0 - y_1 - y_2 - y_3 = 0 \quad \text{and} \quad y_3 - y_4 - y_5 = 0,$$

or, in matrix terms,

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

Step (S4) remains one of assembling, hence

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

becomes

$$\boxed{A^T C A x'(t) + A^T G A x(t) = f(t).} \tag{8.5}$$

where

$$A^T C A = \begin{pmatrix} C_m & 0 \\ 0 & C_m \end{pmatrix} \quad \text{and} \quad A^T G A = \begin{pmatrix} 1/R_i + 1/R_m & -1/R_i \\ -1/R_i & 1/R_i + 1/R_m \end{pmatrix}. \tag{8.6}$$

Contrasting the two-compartment model, (8.5), with the one-compartment model, (8.2), we note that the former can be written

$$\boxed{x'(t) = Bx(t) + f(t) \quad \text{where} \quad B = -A^T G A / C_m \quad \text{and} \quad f(t) = (i_0(t)/C_m \; 0)^T.} \tag{8.7}$$

Our goal, over this and the next 4 chapters, is to understand the sense in which the solution of the scalar problem generalizes to

$$\boxed{x(T) = \exp(BT)x(0) + \int_0^T \exp(B(T-t))f(t)\,dt.} \tag{8.8}$$

More precisely, we will be looking for effective ways to compute and understand the **matrix exponential**

$$\boxed{\exp(Bt) \equiv \sum_{k=0}^{\infty} \frac{(Bt)^k}{k!}.} \tag{8.9}$$

There are two classes of matrices, nilpotents and projections, for which this sum may be evaluated by inspection. See Exer. 8.2 for details. The Spectral Theorem of Chapter 11 will state that every matrix can be written as a weighted sum of projections and nilpotents, where the weights are eigenvalues of $B$.

### 8.2. Analytical Methods

In attempting to understand the purported solution, (8.8)–(8.9), to the matrix problem our first idea is to hew close to the scalar case. There we simply multiplied by the right scaler exponential and integrated. In the matrix case, unsure of the "right" scaler exponential we begin with a "variable" scaler, $-s$. In particular, multiplying each side of (8.7) by $\exp(-st)$ brings

$$x'(t)\exp(-st) = Bx(t)\exp(-st) + g(t)\exp(-st). \tag{8.10}$$

The next step in our complete solution of the scalar problem was one of integration. We start on the left in (8.10) and note that

$$\int_0^T x'(t)\exp(-st)\,dt = x(t)\exp(-st)\Big|_{t=0}^{t=T} + s\int_0^T x(t)\exp(-st)\,dt, \tag{8.11}$$

upon integrating by parts. This has left us with "mixed" terms in the sense that the desired $x$ appears both alone and under the integral sign. We can purify this mix by letting $T \to \infty$ in (8.11).

More precisely, under the physically plausible assumption that $x(t)\exp(-st) \to 0$ as $t \to \infty$ we deduce from (8.11) that

$$\int_0^\infty x'(t)\exp(-st)\,dt = -x(0) + s\int_0^\infty x(t)\exp(-st)\,dt. \tag{8.12}$$

It follows that if we integrate both sides of (8.10) over all time then

$$s\int_0^\infty x(t)\exp(-st)\,dt - x(0) = B\int_0^\infty x(t)\exp(-st)\,dt + \int_0^\infty g(t)\exp(-st)\,dt. \tag{8.13}$$

The integral transforms of the known $g$ and unknown $x$ are called the **Laplace Transforms** of $g$ and $x$ and are written

$$G(s) \equiv \int_0^\infty g(t)\exp(-st)\,dt \quad \text{and} \quad X(s) \equiv \int_0^\infty x(t)\exp(-st)\,dt \tag{8.14}$$

to stress their dependence on the variable $s$. With this definition we see that (8.13) becomes $sX(s) - x(0) = BX(s) + G(s)$, or after simple rearrangement

$$(sI - B)X(s) = x(0) + G(s). \tag{8.15}$$

From this we see that if $s$ is such that $(sI - B)$ is invertible then

$$X(s) = (sI - B)^{-1}(x(0) + G(s)), \tag{8.16}$$

delivers the Laplace transform of the response in terms of the Laplace transform of the stimulus.

We return to the 2-compartment neuron and assemble the players when

$$B = \begin{pmatrix} -2 & 1 \\ 1 & -2 \end{pmatrix} \quad \text{and} \quad g(t) = \begin{pmatrix} t\exp(-t) \\ 0 \end{pmatrix}. \tag{8.17}$$

The Laplace transform of $g$ is

$$G(s) = \int_0^\infty \exp(-st)\begin{pmatrix} t\exp(-t) \\ 0 \end{pmatrix} dt = \frac{1}{(s+1)^2}\begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{8.18}$$

and the inverse of $(sI - B)$ may be computed from the general formula

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}. \tag{8.19}$$

In particular,

$$(sI - B)^{-1} = \frac{1}{(s+1)(s+3)}\begin{pmatrix} s+2 & 1 \\ 1 & s+2 \end{pmatrix}. \tag{8.20}$$

On substitution of $G(s)$ and $(sI - B)^{-1}$ into (8.16) we find

$$X(s) = \frac{1}{(s+1)^3(s+3)}\begin{pmatrix} s+2 \\ 1 \end{pmatrix}. \tag{8.21}$$

It remains to invert the Laplace transform and recover $x(t)$ from $X(s)$. As $X(s)$ was built from integration in $t$ of $x(t)\exp(-st)$ we may expect that $x(t)$ can be recovered by integration in $s$ of

$X(s)\exp(st)$. With the right definition of "integration in $s$" this is indeed the case. In particular, the **Inverse Laplace Transform** of $X$ is

$$x(t) = \frac{1}{2\pi i}\int_C X(s)\exp(st)\,ds, \tag{8.22}$$

where $i = \sqrt{-1}$ and $C$ is a closed curve in the complex plane that encircles all of the **poles** of $X$. The poles of $X$ are those values of $s$ for which $|X(s)| = \infty$. The poles of the $X$ in (8.21) are at $s = -1$ and $s = -3$. Integration in the complex plane is so fundamental for understanding $\exp(Bt)$ that we will devote the next two chapters to it. There we will see that Cauchy's Integral Formula draws

$$\begin{aligned} x_1(t) &= \frac{t^2 + t - 1/2}{4}\exp(-t) + \frac{1}{8}\exp(-3t), \\ x_2(t) &= \frac{t^2 - t + 1/2}{4}\exp(-t) - \frac{1}{8}\exp(-3t) \end{aligned} \tag{8.23}$$

from (8.22). We have plotted these responses in Figure 8.2. Contrasting (8.23) and (8.20) we see that the two rates of exponential decay in both $x_1$ and $x_2$ coincide with the two poles of $(sI - B)^{-1}$. This matrix, and its poles, are of such central importance to both dynamics and higher linear algebra that they have been named. In particular, we call $(sI - B)^{-1}$ the **resolvent** of $B$ and we call the poles of the resolvent the **eigenvalues** of $B$. The meaning of this German–English hybrid is better gleaned from its Spanish equivalent, *autovalor*. This suggests that poles of the resolvent of $B$ are auto– or self–values of $B$. To see where this notion of self–value arises, note that if $\lambda$ is a pole of the resolvent then $(\lambda I - B)$ has no inverse. In this case the columns of $(\lambda I - B)$ are linearly dependent and, equivalently, $(\lambda I - B)$ has a nontrivial null space. The latter implies that there exists and nonzero vector $x$ such that $(\lambda I - B)x = 0$. On rearranging we find

$$\boxed{Bx = \lambda x,} \tag{8.24}$$

and finally arrive at the etymology of *self*. In particular, in (8.24) there are no exogenous stimuli or initial conditions for $B$ to respond to. The $x$ and $\lambda$ are therefore solely reflections of $B$ itself. Moreover, as $Bx$ is simply a scalar multiple of $x$ we call this $x$ a self–vector and the associated $\lambda$ and self–value of $B$. With the words now unpacked we will revert to their common usage, eigenvalue and eigenvector.

To take a concrete case we find the eigenvectors of the $B$ in (8.17) associated with the eigenvalues $\lambda_1 = -1$ and $\lambda_2 = -3$. If $x_1 \in \mathcal{N}(\lambda_1 I - B)$ then

$$\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} x_1(1) \\ x_1(2) \end{pmatrix} \quad \text{hence} \quad x_1 = a\begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{8.25}$$

for any $a \in \mathbb{R}$. Similarly, If $x_2 \in \mathcal{N}(\lambda_2 I - B)$ then

$$\begin{pmatrix} -1 & -1 \\ -1 & -1 \end{pmatrix}\begin{pmatrix} x_2(1) \\ x_2(2) \end{pmatrix} \quad \text{hence} \quad x_2 = a\begin{pmatrix} 1 \\ -1 \end{pmatrix} \tag{8.26}$$

for any $a \in \mathbb{R}$. To visualize the general two-by-two case, and to see how rare it is for matrix to simply scale a vector, I recommend that you invoke `eigshow` in MATLAB.

The connection between the mysterious representation (8.22) and the less mysterious representation (8.8) will stem from writing the latter as a convolution and then noting that the Laplace

Transform is especially well–suited to convolutions. To be precise we define the **convolution** of two functions $f$ and $g$, defined for $t > 0$, to be

$$\boxed{(f \star g)(t) = \int_0^t f(t-y)g(y)\,dy.} \tag{8.27}$$

We will now prove that the Laplace Transform of the convolution of $f$ and $g$ is simply the product of their Laplace Transforms.

**Proposition** 8.1 If $h(t) = (f \star g)(t)$ then $H(s) = F(s)G(s)$.

Proof: On taking the Laplace Transform of each side of Eq. (8.27) we find

$$\begin{aligned} H(s) &= \int_0^\infty \int_0^t f(t-y)g(y)\,dy\,\mathrm{e}^{-st}\,dt \\ &= \int_0^\infty g(y)\mathrm{e}^{-sy} \int_0^\infty f(t-y)\mathrm{e}^{-s(t-y)}\,dt\,dy, \quad \text{as } g(y) = 0 \text{ for } y < 0 \text{ and } f(t-y) = 0 \text{ for } y > t \\ &= \int_0^\infty g(y)\mathrm{e}^{-sy} \int_0^\infty f(r)\mathrm{e}^{-sr}\,dr\,dy, \quad \text{using } r = t - y \text{ and } f(r) = 0 \text{ for } r < 0 \\ &= F(s)G(s), \end{aligned}$$

as claimed. End of Proof.

On reconciling this result with (8.8) and (8.16) we arrive at the conclusion that $(sI - B)^{-1}$ is the Laplace Transform of $\exp(Bt)$. Or, put the other way round, $\exp(Bt)$ is the inverse Laplace Transform of the resolvent. That is

$$\boxed{\exp(Bt) = \frac{1}{2\pi i}\int_C (sI - B)^{-1}\exp(st)\,ds} \tag{8.28}$$

where $C$ encloses all of the eigenvalues of $B$. With $B$ and $(sI - B)^{-1}$ as given, (8.17) and (8.20), for our 2-compartment neuron Cauchy's Integral Theorem will reveal the concrete

$$\exp(Bt) = \frac{\exp(-t)}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \frac{\exp(-3t)}{2}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \tag{8.29}$$

It is no accident that these two matrices are orthogonal projectors of $\mathbb{R}^2$ onto the respective subspaces spanned by the eigenvectors, $x_1$ and $x_2$ from (8.25)–(8.26).

On substition of (8.29) into (8.8) with $f(t) = [t\exp(-t)\ 0]^T$ we find

$$\begin{aligned} x(t) &= \int_0^t \exp(B(t-y))f(y)\,dy \\ &= \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}\int_0^t \exp(y-t)y\exp(-y)\,dy + \begin{pmatrix} 1/2 \\ -1/2 \end{pmatrix}\int_0^t \exp(3(y-t))y\exp(-y)\,dy \\ &= \exp(-t)\begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}\int_0^t y\,dy + \exp(-3t)\begin{pmatrix} 1/2 \\ -1/2 \end{pmatrix}\int_0^t \exp(2y)y\,dy \end{aligned}$$

and so indeed recover (8.23).

### 8.3. Numerical Methods

Where in the previous section we tackled the derivative in (8.7) via an integral transform we pursue in this section a much simpler strategy, namely, integrate the derivative exactly and approximate the integral of the right hand side as a simple sum. More precisely, one chooses a small "time step," $\varepsilon$, and replaces the differential equation (8.7), for the function $x(t)$, by a difference equation for the vector $x_\varepsilon = [x(0)\ x(\varepsilon)\ x(2\varepsilon)\ \cdots\ x((N-1)\varepsilon)]$, by integrating (8.7) over one time step

$$\int_{(n-1)\varepsilon}^{n\varepsilon} x'(t)\,dt = x(n\varepsilon) - x((n-1)\varepsilon) = x_\varepsilon(n+1) - x_\varepsilon(n) = \int_{(n-1)\varepsilon}^{n\varepsilon} Bx(t) + g(t)\,dt,$$

and then approximating the integral via a sum involving only values of the integrand at one or both endpoints, via one of three common choices

$$\int_{(n-1)\varepsilon}^{n\varepsilon} f(t)\,dt \approx \varepsilon \begin{cases} f((n-1)\varepsilon) & \text{Left} \\ f(n\varepsilon) & \text{Right} \\ (f((n-1)\varepsilon) + f(n\varepsilon))/2 & \text{Average} \end{cases} \tag{8.30}$$

The Left choice leads to $x_\varepsilon(n+1) - x_\varepsilon(n) = \varepsilon B x_\varepsilon(n) + \varepsilon g((n-1)\varepsilon)$, that is

$$\boxed{\text{Explicit Euler:}\quad x_\varepsilon(n+1) = (I + \varepsilon B)x_\varepsilon(n) + \varepsilon g((n-1)\varepsilon).} \tag{8.31}$$

The Right choice leads to $x_\varepsilon(n+1) - x_\varepsilon(n) = \varepsilon B x_\varepsilon(n+1) + \varepsilon g(n\varepsilon)$, that is

$$\boxed{\text{Implicit Euler:}\quad (I - \varepsilon B)x_\varepsilon(n+1) = x_\varepsilon(n) + \varepsilon g(n\varepsilon).} \tag{8.32}$$

The Average choice leads to $x_\varepsilon(n+1) - x_\varepsilon(n) = \varepsilon B(x_\varepsilon(n+1) + x_\varepsilon(n))/2 + \varepsilon(g(n\varepsilon) + g((n-1)\varepsilon))/2$, that is

$$\boxed{\text{Trapezoid:}\quad (I - (\varepsilon/2)B)x_\varepsilon(n+1) = (I + (\varepsilon/2)B)x_\varepsilon(n) + \varepsilon(g(n\varepsilon) + g((n-1)\varepsilon))/2.} \tag{8.33}$$

Each of these provide means to march through time by computing the next element of $x_\varepsilon$ in terms of its current value. We will implement and study the Implicit method here and the other two methods in the exercises.

Regarding implementation we find

$$x_\varepsilon(n+1) = (I - \varepsilon B)\backslash(x_\varepsilon(n) + \varepsilon g(n\varepsilon)), \tag{8.34}$$

and code this in `cab2.m` for the circuit of Figure 8.2.

```
% cab2.m  Implicit Euler on the 2-compartment neuron
B = [-2 1;1 -2];
eps = 0.1;
N = ceil(10/eps);
[L,U] = lu(eye(2)-eps*B);
x = zeros(2,N);
t = zeros(1,N);
for n=2:N,
    t(n) = (n-1)*eps;
    g = [t(n)*exp(-t(n)); 0];
    x(:,n) = L\(U\(x(:,n-1) + eps*g));
end
plot(t,x)
```

We compare the performance of this procedure with the exact solution in Figure 8.3. We have chosen large values of $\varepsilon$ for ease of illustration. You may wish to confirm that the approximation "looks" much better at much smaller $\varepsilon$.

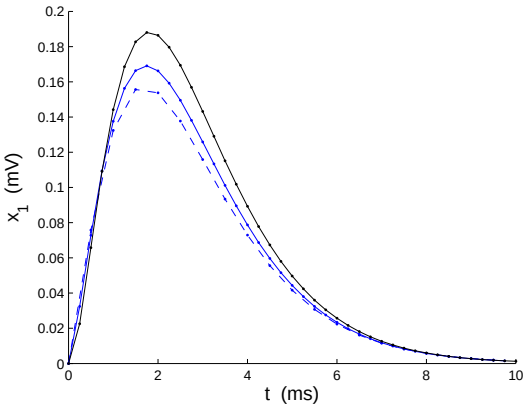



**Figure** 8.3. We contrast the exact solution, $x_1(t)$, in (8.23), in black at times $t = (0.25)n$, with solutions delivered by `cab2.` with $\varepsilon = 0.5$ (dashed blue) and $\varepsilon = 0.25$ (solid blue).

In order to see, mathemticallly, that the Implicit Euler solution, $x_\varepsilon$, indeed approaches the exact solution, $x$, as $\varepsilon \to 0$ lets first suppose that the stimulus, $g$, is zero. In this case, (8.34) may be written

$$x_\varepsilon(n) = ((I - \varepsilon B)^{-1})^n x(0). \tag{8.35}$$

Now, for a fixed time $t$ we suppose that $\varepsilon = t/n$ and ask whether

$$x(t) = \lim_{n\to\infty} ((I - (t/n)B)^{-1})^n x(0). \tag{8.36}$$

We will here establish the truth of (8.36) for scalar $B$ and then return to the matrix case in Chapter 11. If $B \in \mathbb{R}$ and $n > |Bt|$ then

$$\begin{aligned} ((I - (t/n)B)^{-1})^n &= \left(\frac{1}{1 - Bt/n}\right)^n \\ &= \exp((\log(1/(1 - Bt/n)))^n) \\ &= \exp(n \log(1/(1 - Bt/n))) \\ &= \exp(-n \log(1 - Bt/n)) \\ &= \exp(-n(-Bt/n + O(1/n^2))) \\ &\to \exp(Bt) \quad \text{as } n \to \infty. \end{aligned} \tag{8.37}$$

The second equality follows from exp being the inverse function of log, i.e., $x = \exp(\log(x))$. The remaining steps follow from basic properties of log: $(\log(x))^n = n \log(x)$, $\log(1/x) = -\log(x)$ and $\log(1 - x) = -x + O(x^2)$ where $O(x^2)$ indicates terms that go to zero at least as fast as $x^2$ when $x \to 0$.

Returning to the question posed in (8.36), we have shown that $x_{t/n}(n) \to \exp(Bt)x(0)$ as $n \to \infty$, when $B$ is scalar and the stimulus is zero. Comparing this with our analytical findings, (8.8), we conclude that the approximate solution, $x_{t/n}(n)$, computed by the Implicit Euler Method, indeed converges to the solution of the dynamical system $x'(t) = Bx(t)$.

At this point we have developed analytical and numerical approaches to differential equations of the form $x'(t) = Bx(t) + g(t)$. In §8.5 we will investigate the extent to which our analytical

approach provide insight into the behavior of solutions to differential equations of the form

$$x'(t) = F(t, x(t)), \tag{8.38}$$

where $F$ may be, in general, a nonlinear function of each of its arguments. We begin the numerical integration of (8.38) by applying our Trapezoidal scheme

$$x_\varepsilon(n+1) - x_\varepsilon(n) = \varepsilon\{F(n\varepsilon, x_\varepsilon(n)) + F((n+1)\varepsilon, x_\varepsilon(n+1))\}/2. \tag{8.39}$$

The difficulty in marching from $x_\varepsilon(n)$ to $x_\varepsilon(n+1)$ is that the latter now appears "inside" the nonlinear $F$ as the last term in (8.39). There are many ways to work our way out from under this nonlinearity. One popular scheme, known as **Heun's Method**, is to approximate this latter $x_\varepsilon(n+1)$ by with its Explicit Euler update from $x_\varepsilon(n)$. That is, to replace (8.39) with

$$\boxed{x_\varepsilon(n+1) = x_\varepsilon(n) + \varepsilon\{F(n\varepsilon, x_\varepsilon(n)) + F((n+1)\varepsilon, x_\varepsilon(n) + \varepsilon F((n+1)\varepsilon, x_\varepsilon(n)))\}/2.} \tag{8.40}$$

We note that this is fully explicit but much more accurate than Explicit Euler. To code this we need only specify the function $F$, the initial values $x(0)$, the final time, $T$, and the timestep, $\varepsilon$.

```
function [t,x] = heun(F,x0,T,eps)
N = ceil(T/eps);
m = length(x0);
x = zeros(m,N);
x(:,1) = x0;
t = zeros(1,N);
for n=2:N
    t(n) = eps*(n-1);
    F1 = F(t(n-1),x(:,n-1));
    F2 = F(t(n),x(:,n-1));
    F3 = F(t(n),x(:,n-1) + eps*F2);
    x(:,n) = x(:,n-1) + eps*(F1 + F3)/2;
end
return
```

We test this on a Goodwin Oscillator

$$\begin{aligned} x_1'(t) &= \frac{1}{1+x_2(t)} - 1/2 \\ x_2'(t) &= x_1(t) - 1 \end{aligned} \tag{8.41}$$

with initial values $x(0) = [1; 2]$, final time $T = 20$ and timestep $\varepsilon = 0.01$ via

```
>> [t,x] = heun(@goodwin,[1;2],20,0.01);
>> plot(t,x(1,:),t,x(2,:))
```

where `goodwin` is encoded as

```
function dx = goodwin(t,x)
dx(1,1) = 1/(1+x(2)) - 1/2;
dx(2,1) = x(1) - 1;
return
```

We have plotted both $x_1$ and $x_2$ against $t$ and against one another in Figure 8.4.

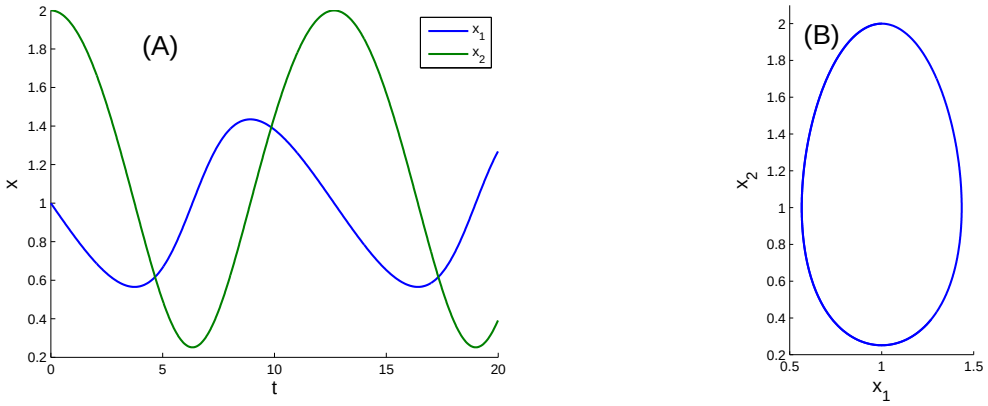



**Figure** 8.4. Solutions of the Goodwin Oscillator, (8.41), via Heun's Method. (A) The two elements of $x$ *vs.* time. (B) The trajectory $(x_1(t), x_2(t))$ produces a closed curve as $t$ increases.

The closed curve in the $(x_1, x_2)$ plane in Figure 8.4(B) encircles the steady state solution $x_1 = x_2 = 1$ to (8.41). This curve is in fact uniquely prescibed by the choice of initial data. You are encouraged to experiment with both smaller and larger values.

### 8.4. Dynamics of Mechanical Networks

Regarding the mechanical networks of Chapter 3, we may move from the equilibrium equations,

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

for the displacement $x$ due to a constant force, $f$, 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. That is,

$$Mx''(t) + Sx(t) = f(t), \qquad x(0) = x_0, \quad x'(0) = v_0, \tag{8.42}$$

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 (8.42) takes the form

$$u_2' = x'' = -M^{-1} S u_1 + M^{-1} f(t).$$

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

$$u' = Bu + g, \quad u(0) = u_0 \tag{8.43}$$

where

$$B = \begin{pmatrix} 0 & I \\ -M^{-1}S & 0 \end{pmatrix}, \quad g = \begin{pmatrix} 0 \\ M^{-1} f \end{pmatrix}, \quad u_0 = \begin{pmatrix} x_0 \\ v_0 \end{pmatrix}. \tag{8.44}$$

**Example 1:** As our first example of (8.44) we consider, see Figure 8.5, a single mass, of mass $m$, suspended by a single spring, of stiffness $k$. In this case

$$B = \begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix} \tag{8.45}$$

where $\omega^2 = k/m$.

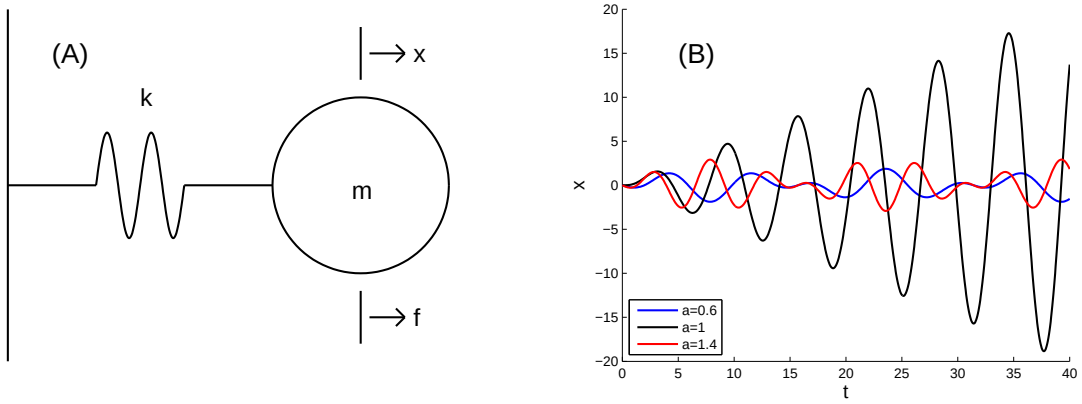



**Figure** 8.5. (A) A single mass suspended from a single spring. (B) The response of the mass to a driving force $\sin(at)$ as derived in (8.49) and (8.50).

Via Gauss–Jordan or (8.19) we find that

$$(sI - B)^{-1} = \frac{1}{s^2 + \omega^2} \begin{pmatrix} s & 1 \\ -\omega^2 & s \end{pmatrix} \tag{8.46}$$

has poles at $\pm i\omega$ where $i \equiv \sqrt{-1}$ is the imaginary unit. Invoking (8.28) (will) then bring

$$\exp(Bt) = \frac{\exp(i\omega t)}{2i\omega} \begin{pmatrix} i\omega & 1 \\ -\omega^2 & i\omega \end{pmatrix} - \frac{\exp(-i\omega t)}{2i\omega} \begin{pmatrix} -i\omega & 1 \\ -\omega^2 & -i\omega \end{pmatrix} = \begin{pmatrix} \cos(\omega t) & \sin(\omega t)/\omega \\ -\omega \sin(\omega t) & \cos(\omega t) \end{pmatrix} \tag{8.47}$$

Thanks to the beautiful identities

$$\cos(\omega t) = \frac{\exp(i\omega t) + \exp(-i\omega t)}{2} \quad \text{and} \quad \sin(\omega t) = \frac{\exp(i\omega t) - \exp(-i\omega t)}{2i} \tag{8.48}$$

to be proven in the next chapter. In the unloaded case, $f = 0$, and no intial velocity, $v_0 = 0$, it follows that the displacement is simply $x(t) = x_0 \cos(\omega t)$. Conversely, with no initial displacement or velocity but with driving force $f(t) = \sin(at)$ we find

$$x(t) = \frac{1}{m\omega} \int_0^t \sin(\omega y) \sin(a(t-y))\, dy = \frac{\sin(\omega t) - \sin(at)}{2m\omega(a - \omega)} + \frac{\sin(\omega t) - \sin(at)}{2m\omega(a + \omega)}. \tag{8.49}$$

From this we see that as the driving frequency, $a$, approaches $\omega$ the displacement approaches

$$x(t) = \frac{\sin(\omega t)}{2m\omega^2} - \frac{t \cos(\omega t)}{2m\omega}. \tag{8.50}$$

As the amplitude of this displacement grows with time we speak of $\omega$ as the **resonant frequency** of our system. It is no accident that the resonant frequency is the imaginary part of the eigenvalue of the associated $B$ matrix, (8.45). We have illustrated this phenomenon in Figure 8.5(B).

**Example 2:** For our second example we consider the chain of 2 masses in Figure 8.6(A). If each node has mass $m$ and each spring has stiffness $k$ then

$$M^{-1}S = \omega^2 \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \quad \text{and hence} \quad B = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -2\omega^2 & \omega^2 & 0 & 0 \\ \omega^2 & -2\omega^2 & 0 & 0 \end{pmatrix}$$

and so,

$$(sI-B)^{-1} = \frac{1}{(s^2+\omega^2)(s^2+3\omega^2)} \begin{pmatrix} s(s^2+2\omega^2) & s\omega^2 & s^2+2\omega^2 & \omega^2 \\ s\omega^2 & s(s^2+2\omega^2) & \omega^2 & s^2+2\omega^2 \\ -\omega^2(2s^2+3\omega^2) & s^2\omega^2 & s(s^2+2\omega^2) & s\omega^2 \\ s^2\omega^2 & -\omega^2(2s^2+3\omega^2) & s\omega^2 & s(s^2+2\omega^2) \end{pmatrix}.$$

We did not compute this by hand via Gauss–Jordan but rather invoked the symbolic toolbox in MATLAB. In particular,

```
>> syms s
>> inv(s*eye(4)-B)
```

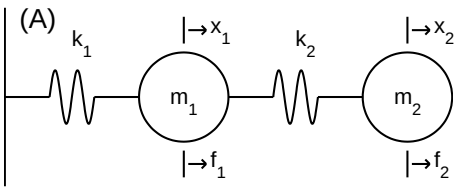
does the job. We see that this resolvent has poles at $\pm i\omega$ and $\pm i\omega\sqrt{3}$ and that they appear explicitly in the associated matrix exponential

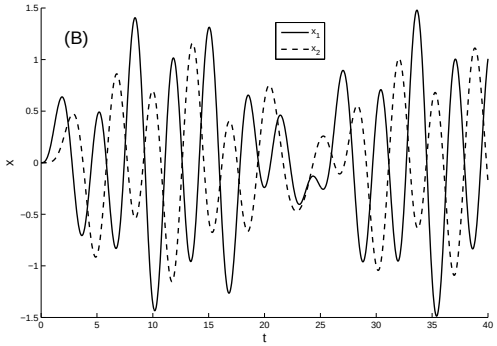
$$\exp(Bt) = \frac{1}{2}\begin{pmatrix} \cos(\omega t) & \cos(\omega t) & \sin(\omega t)/\omega & \sin(\omega t)/\omega \\ \cos(\omega t) & \cos(\omega t) & \sin(\omega t)/\omega & \sin(\omega t)/\omega \\ -\omega\sin(\omega t) & -\omega\sin(\omega t) & \cos(\omega t) & \cos(\omega t) \\ -\omega\sin(\omega t) & -\omega\sin(\omega t) & \cos(\omega t) & \cos(\omega t) \end{pmatrix}$$
$$+\frac{1}{2}\begin{pmatrix} \cos(\omega\sqrt{3}t) & -\cos(\omega\sqrt{3}t) & \sin(\omega\sqrt{3}t)/(\omega\sqrt{3}) & -\sin(\omega\sqrt{3}t)/(\omega\sqrt{3}) \\ -\cos(\omega\sqrt{3}t) & \cos(\omega\sqrt{3}t) & -\sin(\omega\sqrt{3}t)/(\omega\sqrt{3}) & \sin(\omega\sqrt{3}t)/(\omega\sqrt{3}) \\ -\omega\sqrt{3}\sin(\omega\sqrt{3}t) & \omega\sqrt{3}\sin(\omega\sqrt{3}t) & \cos(\omega\sqrt{3}t) & -\cos(\omega\sqrt{3}t) \\ \omega\sqrt{3}\sin(\omega\sqrt{3}t) & -\omega\sqrt{3}\sin(\omega\sqrt{3}t) & -\cos(\omega\sqrt{3}t) & \cos(\omega\sqrt{3}t) \end{pmatrix}$$

Hence, if $f(t) = \sin(at)[u_1\ u_2]^T$ then

$$\begin{aligned} x_1(t) &= \frac{((a/\omega)\sin(\omega t) - \sin(at))(u_1+u_2)}{2(a^2-\omega^2)} + \frac{((a/\sqrt{3}\omega)\sin(\sqrt{3}\omega t) - \sin(at))(u_1-u_2)}{2(a^2-3\omega^2)} \\ x_2(t) &= \frac{((a/\omega)\sin(\omega t) - \sin(at))(u_1+u_2)}{2(a^2-\omega^2)} - \frac{((a/\sqrt{3}\omega)\sin(\sqrt{3}\omega t) - \sin(at))(u_1-u_2)}{2(a^2-3\omega^2)}. \end{aligned} \tag{8.51}$$

Contrasting this response with that of the single mass system, (8.49), we see that the two–mass system has two distinct **modes of vibration**. Namely, if $u_1 = u_2$ then $x_1(t) = x_2(t)$ while if $u_1 = -u_2$ then $x_1(t) = -x_2(t)$. In general the response is a mixture of these two modes, see, e.g., Figure 8.6(B), where $a = 2$ and $u = [1\ 0]^T$. With this same $u$ we illustrate in Figure 8.6(C and D) resonance at $a = \omega$ and $a = \sqrt{3}\omega$. As in the one–mass example, the resonant frequencies of the mechanical network are precisely the imaginary parts of the eigenvalues of the associated $B$ matrix.



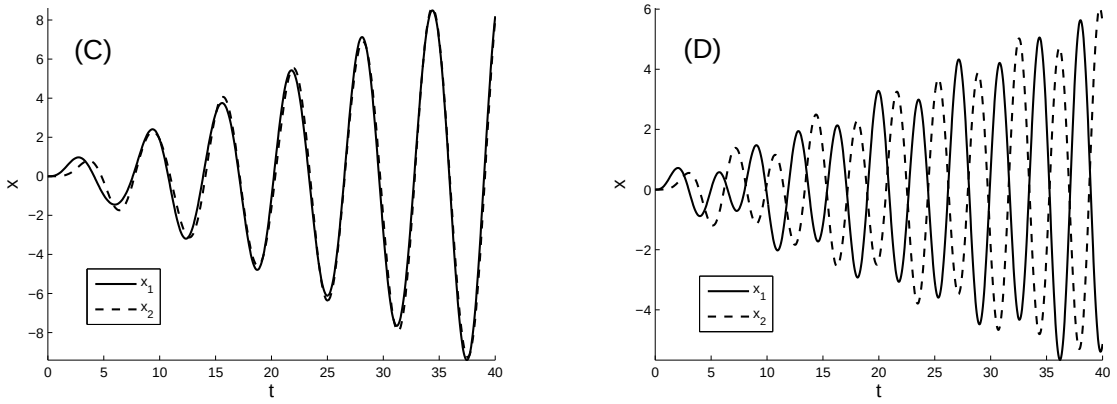



**Figure** 8.6. (A) The two–mass, two–spring network. (B) The displacements of the 2 masses, with $k/m = 1 = \omega^2$ and driving force $f(t) = \sin(at)[1\ 0]^T$ and $a = 2$. (C) The displacements for the same conditions as in (B) except $a = 1$. (D) The displacements for the same conditions as in (B) except $a = \sqrt{3}$. `ms2.m`

We note that Implicit Euler is just as useful here as in the circuit case. In particular, changing the system and stimulus in `cab2.m` brings

```
% chain2.m
S = [2 -1;-1 2];
B = [zeros(2) eye(2); -S zeros(2)];
eps = 0.0001;
N = ceil(40/eps);
[L,U] = lu(eye(2)-eps*B);
x = zeros(4,N);
t = zeros(1,N);
for n=2:N,
    t(n) = (n-1)*eps;
    g = [0; 0; sin(2*t(n)); 0];
    x(:,n) = L\(U\(x(:,n-1) + eps*g));
end
plot(t,x(1:2,:))
```

This code indeed replicates Figure 8.6(B), although with significantly greater effort (as measured by the timestep $\varepsilon$) than that required of `cab2.m`.

### 8.5. Dynamics of Metabolic Networks*

We recall the steady state state flux balance $Sv = f$ is the rest state of

$$m'(t) = Sv - f \tag{8.52}$$

where to "close" this system we must express $v$ as a function of $m$. This field is highly complex and there is nothing as clean as the early phases of the Strang Quartet, i.e., we may not expect $v = DS^T m$ for some diagonal matrix $D$. Rather each $v_j(m)$ is typically a distinct nonlinear function of the associated metabolite concentrations.

To begin we consider the interaction of two metabolites via

$$\begin{aligned} &\to m_1 \quad \text{at rate } v_0 \\ m_1 &\to m_2 \quad \text{at rate } v_1 \\ m_1 &\to \quad \text{at rate } v_2 \\ m_2 &\to \quad \text{at rate } v_3 \end{aligned}$$

where the feed flux, $v_0$, is fixed and the two decay fluxes are linear

$$v_2 = 5m_2 \quad \text{and} \quad v_3 = m_1,$$

while the flux from $m_1$ to $m_2$ obeys the nonlinear law

$$v_1 = m_1(1 + m_2^3).$$

In terms of the general case, (8.52), this produces the concrete nonlinear system

$$\begin{aligned} m_1'(t) &= v_0 - m_1(1 + m_2^3) - m_1 \\ m_2'(t) &= m_1(1 + m_2^3) - 5m_2. \end{aligned} \tag{8.53}$$

Our first goal is to demonstrate (numerically) that solutions to this system behave in three very distinct ways, depending on the size of the input, $v_0$. We use Heun's method from §8.3 to solve (8.53), from two choices of initial data and for three choices of $v_0$, and present our findings in Figure 8.7.

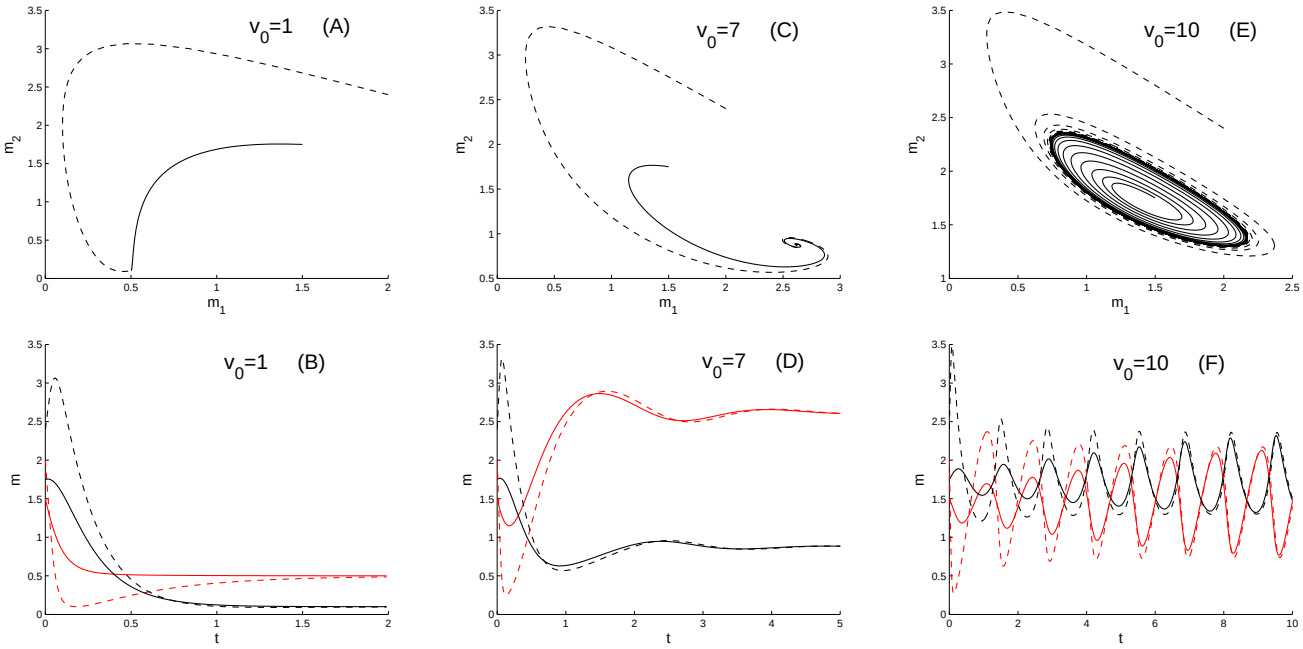



**Figure** 8.7. Behavior of the nonlinear metabolic network, (8.53), at three distinct input fluxes: $v_0 = 1,\ 7$ and 10. In the top row we plot $(m_1(t), m_2(t))$ commencing from initial metabolite levels $(1.5, 1.75)$ (solid) and $(2, 2.4)$ (dashed). In the bottom row we plot the associated $(t, m_1(t))$ (black) and $(t, m_2(t))$ (red) with the solid/dash convention as above.

In panels (A) and (B) of Figure 8.7, with $v_0 = 1$, we see exponential decay to a steady state. A solution of (8.53) is called a **steady state** if it does not depend on $t$. For the situation illustrated in

panels (A) and (B) we call its steady state and **stable node**. It is stable in the sense that nearby initial conditions are attracted to it while the word node is used to distinguish it from the spiral, which comes next.

In panels (C) and (D) of Figure 8.7, with $v_0 = 7$, we see exponential decaying oscillation to a steady state and we call this state a **stable spiral** because (with an eye on panel (C)) neighboring initial conditions spiral into it.

In panels (E) and (F) of Figure 8.7, with $v_0 = 10$, we see exponential decay (from without) and growth (from within) to an oscillatory, in fact periodic, state. The growth from within is actually away from steady state that we hence call an **unstable spiral**. This periodic state is called (with an eye to panel (E)) a **stable limit cycle** because nearby initial states cycle into to it over time.

In order to see that these modes of behavior, and associated labels, are not merely artifacts of our particular system of equations or choice of initial conditions we need a more systematic approach. Although we can not find an exact solution for such systems, we can use our analytical tools to study its solution in the neighborhood of its steady state solution. To begin, let us write (8.53) more succinctly as

$$m'(t) = F(m(t)). \tag{8.54}$$

We call a vector $\overline{m}$ a **steady state** of (8.54) when $F(\overline{m}) = 0$. The steady state of (8.53) for example obeys

$$\overline{m}_1 = \frac{5\overline{m}_2}{1+\overline{m}_2^3} \quad \text{and} \quad 5\overline{m}_2^4 - v_0\overline{m}_2^3 + 10\overline{m}_2 - v_0 = 0. \tag{8.55}$$

If we choose initial data close to $\overline{m}$ it seems reasonable that $m$ should remain close to $\overline{m}$ for small time, and that during this time the function $F$ might be well approximated by the first term in its Taylor expansion. More precisely, we suppose that

$$m(t) \approx \overline{m} + \varepsilon p(t) \quad \text{and} \quad F(\overline{m} + \varepsilon p(t)) \approx F(\overline{m}) + \varepsilon \nabla F(\overline{m}) p \tag{8.56}$$

where $\varepsilon$ is small, $p$ is the "perturbation" of $\overline{m}$, and $\nabla F(\overline{m})$, the **gradient** of $F$ at $\overline{m}$, is the matrix of partial derivatives

$$\nabla F(\overline{m}) = \begin{pmatrix} \frac{\partial F_1}{\partial m_1}(\overline{m}) & \frac{\partial F_1}{\partial m_2}(\overline{m}) \\ \frac{\partial F_2}{\partial m_1}(\overline{m}) & \frac{\partial F_2}{\partial m_2}(\overline{m}) \end{pmatrix}.$$

In the case of (8.53) we have $F_1(m) = v_0 - 2m_1 - m_1 m_2^3$ and $F_2(m) = m_1(1+m_2^3) - 5m_2$ and so

$$\nabla F(\overline{m}) = \begin{pmatrix} -(2+\overline{m}_2^3) & -3\overline{m}_1\overline{m}_2^2 \\ 1+\overline{m}_2^3 & 3\overline{m}_1\overline{m}_2^2 - 5 \end{pmatrix}.$$

On substituting our approximations, (8.56), into (8.54) we find $m' = (\overline{m} + \varepsilon p)' = \varepsilon \nabla F(\overline{m}) p$, which upon dividing by $\varepsilon$, yields a linear system of differential equations for $p$,

$$p' = Bp, \qquad B \equiv \nabla F(\overline{m}), \tag{8.57}$$

that is amenable to our analytical tools. In the case of our example system, (8.53), this $B$ matrix is

$$B = \begin{pmatrix} -(2+a) & -15a/(1+a) \\ 1+a & 15a/(1+a) - 5 \end{pmatrix} \quad \text{where} \quad a = \overline{m}_2^3$$

The behavior of this system is completely determined by the poles of

$$(sI - B)^{-1} = \frac{1}{(1+a)s^2 + (a^2 - 7a + 7)s + 5a^2 + 10} \begin{pmatrix} (1+a)s - 10a + 5 & -15a \\ (1+a)^2 & (1+a)(2+a+s) \end{pmatrix}.$$

That is, by

$$\lambda_\pm(a) \equiv \frac{7a - a^2 - 7 \pm \sqrt{D(a)}}{2(a+1)}, \quad \text{where} \quad D(a) = a^4 - 34a^3 + 43a^2 - 138a + 9. \tag{8.58}$$

We will first determine how these poles travel with $a$ and then translate this, via $a = \overline{m}_2^3$ and (8.55), into conclusions that help us to better understand, and eventually transcend, Figure 8.7.

As $a$ increases from zero we note $\lambda_\pm(a)$ are real (and negative) until $a$ reaches the first real zero of $D(a)$ at $a_1 = 0.0665$. These pole are visible in Figure 8.8(A) as the two pair of (black) real poles in the vicinity of $\lambda = -3$. As $a$ approaches $a_1$ from below $\lambda_+(a)$ and $\lambda_-(a)$ coincide and then split into a complex conjugate pair (blue) with negative real part, until $a$ reaches the first root of $7a - a^2 - 7$ at $a_2 = 1.2087$. As $a$ increases beyond $a_2$ the real part of $\lambda_\pm$ becomes positive (red) until $a$ reaches the second root of $7a - a^2 - 7$ at $a_3 = 5.7913$. After this point the poles are again a complex conjugate pair with negative real part (blue again), until $a$ reaches the second (and final) real root of $D(a)$ at $a_4 = 32.8176$. For $a$ beyond $a_4$ the poles remain distinct real and negative (black again) in the vicinity of $\lambda = -12$.

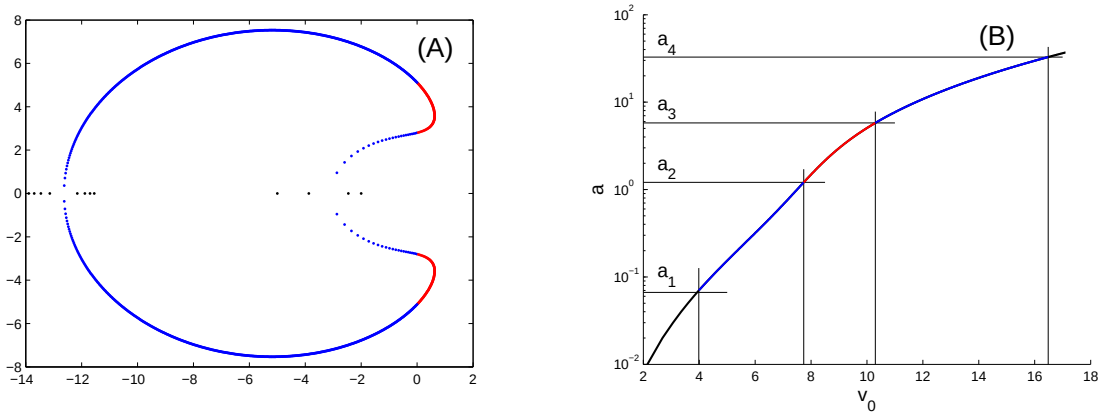



**Figure** 8.8. (A) Tracking the poles, $\lambda_\pm(a)$, as $a$ increases from 0 to 33 in steps of 0.05. As $a$ increases to $a_1$ the black pair of poles collide near $\lambda = -3$ and turn blue while they remain in the left half plane, $\{z \in \mathbb{C} : \Re z < 0\}$, until $a = a_2$ after which we paint the poles red. They re-enter the left half plane at $a = a_3$ and we again color the poles blue until they collide once again when $a = a_4$ near $\lambda = -12$ after which they turn black and real. `metaeigtrack.m` (B) The plot of $a$ *vs.* $v_0$ derived from $a = \overline{m}^3$ and $v_0 = 5\overline{m}^{1/3}(a+2)/(a+1)$ per (8.55). The four transition points in $a$ map onto four transition points for $v_0$. `metav0a.m`

If the behavior of the full nonlinear system, (8.54), for initial data in the vicinity of a steady state, $\overline{m}$, is indeed captured by the associated linear system (8.57) then we may glean all from the matrix exponential $\exp(Bt)$. As the latter is governed by the scalar exponentials, $\exp(\lambda_\pm)$, we see that states near $\overline{m}$ are attracted to $\overline{m}$ if the real part of $\lambda_\pm$ is negative and are repelled if the real part is positive. If the associated imaginary part is nonzero then attraction/repulsion will be oscillatory. With this we may now reconcile Figures 8.7 and 8.8. In particular, from Figure 8.8 we learn that that $\lambda_\pm$ are real and negative for $v_0 < 4$. Hence $\exp(\lambda_\pm t)$ are both decaying exponentials, in agreement with the stable node exhibited in Figure 8.7(A–B).

Returning to Figure 8.8 we see for $4 < v_0 < 7.9$ that $\lambda_\pm$ are nonreal but have negative real part and hence solutions are decaying oscillations, in agreement with the stable spiral exhibited in Figure 8.7(C–D).

Next, from Figure 8.8 we see for $7.9 < v_0 < 10.2$ that $\lambda_\pm$ are nonreal but have postive real part and hence solutions are growing oscillations, in agreement with the unstable spiral exhibited in Figure 8.7(E–F).

Our analysis now permits us to make predictions about regions of $v_0$ untested in the empirical solutions of Figure 8.7. In particular, Figure 8.8, predicts that the steady state is a stable spiral when $10.2 < 0 < 16.2$ and a stable node after that.

### 8.6. Notes and Exercises

In §8.5 we suggested that the behavior of solutions to the nonlinear system $m'(t) = F(m(t))$ near a steady state, $\overline{m}$, could be predicted by the behavior of the associated linear system, $p'(t) = \nabla F(\overline{m})p$, and we supported our suggestion with a two dimensional example. We offer a 3 dimensional example in Exer. 8.14. This prediction is better known as the Hartman–Grobman Theorem. Please see Perko (1991) for a precise statement and proof.

Our presentation of the Goodwin Oscillator follows Gonze and Abou-Jaoud (2013).

Our presentation of Metabolic Networks follows Heinrich et al. (1977).

1. For matrices with special powers we may sum the Taylor series, (8.9), by hand and arrive at the matrix exponential. We considered a few easy examples in Chapter 1. In particular lets return to (1.35).

   (a) Show that if

   $$B = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \quad \text{then} \quad B^n = \begin{pmatrix} 1 & 2n \\ 0 & 1 \end{pmatrix} \tag{8.59}$$

   for $n = 0, 1, 2, \ldots$.

   (b) Use (a) to show that

   $$\exp(Bt) = \begin{pmatrix} \sum_{n=0}^{\infty} t^n/n! & \sum_{n=0}^{\infty} 2nt^n/n! \\ 0 & \sum_{n=0}^{\infty} t^n/n! \end{pmatrix} = \exp(t) \begin{pmatrix} 1 & 2t \\ 0 & 1 \end{pmatrix}.$$

   Carefully explain your confirmation of the off diagonal term.

2. This previous example is rewarding but rare. We seek larger classes of matrices for which the sums in (8.9) may be expressed.

   (a) The easiest case is the nilpotent case. Perhaps you ran into these matrices in §4.4. We call $B$ **nilpotent** if $B^m = 0$ for some positive integer $m$. In this case (8.9) reduces to the finite sum

   $$\exp(Bt) = \sum_{k=0}^{m-1} \frac{(Bt)^k}{k!}. \tag{8.60}$$

   Show that

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

   is nilpotent and use (8.60) to express its matrix exponential.

   (b) The next easiest class is the class of projection matrices. For if $P^2 = P$ then $P^k = P$ for integer $k \geq 1$. Show that if $P^2 = P$ then

   $$\exp(Pt) = I - P + \exp(t)P. \tag{8.61}$$

(c) The next easiest class is the class of weighted sums of independent projection matrices. For example, show that if

$$B = \lambda_1 P_1 + \lambda_2 P_2 \quad \text{where} \quad P_1^2 = P_1,\ P_2^2 = P_2,\ P_1 P_2 = P_2 P_1 = 0,\ P_1 + P_2 = I$$

and $\lambda_1$ and $\lambda_2$ are scalars then

$$\exp(Bt) = \exp(\lambda_1 t) P_1 + \exp(\lambda_2 t) P_2. \tag{8.62}$$

Show that (8.29) is an example of this class by identifying the $\lambda_j$ and $P_j$.

3. Confirm, by hand, the following Laplace transforms

   (a) If $u(t) = \exp(t)$ then $U(s) = 1/(s-1)$.

   (b) If $u(t) = t\exp(-t)$ then $U(s) = 1/(s+1)^2$.

   (c) If $u(t) = \sin(t)$ then $U(s) = 1/(s^2+1)$. Hint: Integrate by parts twice.

4. Confirm that the $B$ matrix in (8.17) and its matrix exponential in (8.29) obey $(\exp(Bt))' = B\exp(Bt)$.

5. Our initial brush with the eigenvalues of the matrix $B$ defined them as the poles of the resolvent, $(sI - B)^{-1}$. In other words, as those values of $s$ at which $(sI - B)$ is not invertible. Recalling our work in §3.2 it follows that the eigenvalues of $B$ are precisely those $s$ for which $sI - B$ has a zero pivot. Recalling the definition of the determinant, Definition 3.2, we note that $\det(sI - B)$ is either plus or minus the product of these pivots and hence $s$ is an eigenvalue of $B$ when $\det(sI - B)$=0. Let us put this into practice in the 2–by–2 case where

$$B = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

   (a) Show that $\det(sI - B) = s^2 - (a+d)s + ad - bc$.

   (b) Show that (a) can be re-expressed as $\det(sI - B) = s^2 - \mathrm{tr}(B)s + \det(B)$ and hence that the eigenvalues of $B$ are

$$\lambda_\pm = \frac{\mathrm{tr}\, B \pm \sqrt{(\mathrm{tr}\, B)^2 - 4\det\, B}}{2}. \tag{8.63}$$

   (c) The expression in (b) permits us to classify the eigenvalues (and therefore the behavior of the associated dynamical system) by where the point $(\det(B), \mathrm{tr}(B))$ lies in the $(d, t)$ plane.

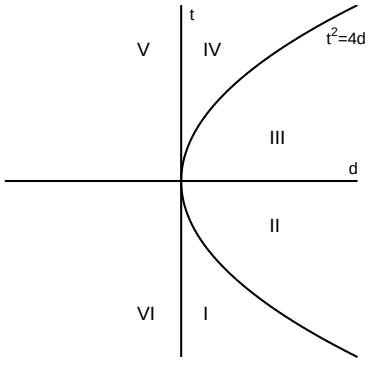


**Figure** 8.9. The coordinate axes and the curve $t^2 = 4d$ partition the $(d, t)$ plane into 5 regions that correspond to qualitatively distinct dymanics.

Please show that

In region I, $\lambda_\pm$ are real and negative.

In region II, $\lambda_\pm$ are nonreal but in LHP

In region III, $\lambda_\pm$ are nonreal in RHP

in region IV, $\lambda_\pm$ are real and positive

In regions V, $\lambda_\pm$ are real and of different signs.

6. In order to visualize the convolution, (8.27), of two functions let us see that the convolution of two blocks is a trapezoid. In particular, suppose that $g$ and $f$ are the step functions $f(y) = \mathbb{1}_{(0,1)}(y)$ and $g(y) = \mathbb{1}_{(2,4)}(y)$. Show that $(f \star g)(t)$ is the length of the overlap of the intervals $(t-1, t)$ and $(2, 4)$, as in Figure 8.10, and so

$$(f \star g)(t) = \begin{cases} 0 & \text{if} \quad t \leq 2 \\ t-2 & \text{if} \quad 2 \leq t \leq 3 \\ 1 & \text{if} \quad 3 \leq t \leq 4 \\ 5-t & \text{if} \quad 4 \leq t \leq 5 \\ 0 & \text{if} \quad 5 \leq t. \end{cases}$$

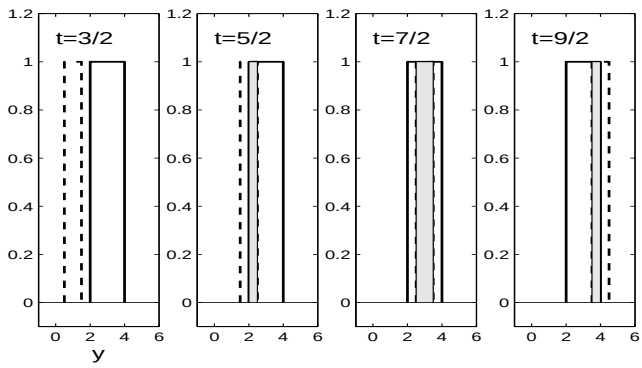



**Figure** 8.10. The integrand $f(t-y)g(y)$ for increasing values of $t$. The convolution at each $t$ is the area of the shaded region, of overlap of $g$ and a shifted copy of $f$.

7. Regarding the errors produced in the integral approximations (8.30), show

   (a) If $f$ is constant then the three choices agree and are exact.

   (b) If $f$ is linear, e.g., $f(t) = 1 + t$, then the three choices give three distinct values and that the average choice is exact.

8. We return to the Implicit Euler scheme and show that $x_{t/n}(n) \to x(t)$ even with a stimulus.

   (a) If $g \neq 0$ then argue that (8.35) takes the form

$$x_\varepsilon(n) = ((I - \varepsilon B)^{-1})^n \left( x(0) + \varepsilon \sum_{k=1}^{n} (I - \varepsilon B)^{k-1} g(k\varepsilon) \right).$$

Again setting $\varepsilon = t/n$ we arrive at

$$x_{t/n}(n) = ((I - (t/n)B)^{-1})^n \left( x(0) + (t/n) \sum_{k=1}^{j} (I - tB/n)^{k-1} g(kt/n) \right).$$

(b) Argue as in (8.37) that the coefficients of the $g$ terms above obey

$$(I - (t/n)B)^{k-1} = \exp(-B(kt/n)) + O(1/n), \tag{8.64}$$

and so arrive at

$$x_{t/n}(n) = ((I - (t/n)B)^{-1})^n \left( x(0) + \frac{t}{n} \sum_{k=1}^{n} \exp(-Btk/n) g(tk/n) + O(1/n) \frac{t}{n} \sum_{k=1}^{n} g(tk/n) \right).$$

(c) Explain why

$$\lim_{n\to\infty} (t/n) \sum_{k=1}^{n} g(tk/n) = \int_0^t g(y)\, dy$$

and

$$\lim_{n\to\infty} (t/n) \sum_{k=1}^{n} \exp(-Btk/n) g(tk/n) = \int_0^t \exp(-By) g(y)\, dy,$$

and finally that $x_{t/n}(n) \to x(t)$ where $x$ is the solution of (8.8).

9. With resistors, capacitors and opamps we can design filters. Consider the circuit of Figure 8.11

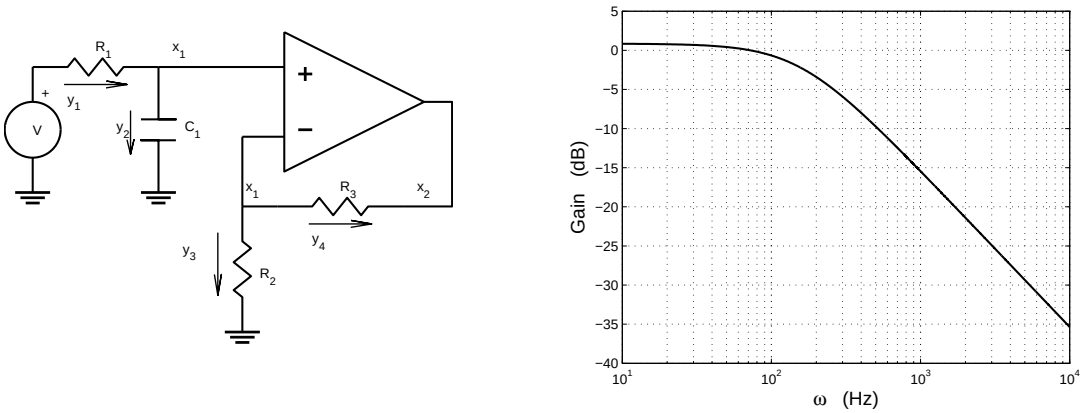



**Figure** 8.11. A low pass filter circuit and its gain *vs.* frequency.

(a) From the two current balances, $y_1 - y_2 = 0$ and $y_3 + y_4 = 0$ derive the pair of equations

$$R_1 C_1 x_1'(t) = v(t) - x_1(t), \qquad x_2(t) = (1 + R_3/R_2) x_1(t). \tag{8.65}$$

(b) Note that $\tau \equiv R_1 C_1$ has units of time and use Eq. (8.4) to conclude that

$$x_1(t) = \exp(-t/\tau) x_1(0) + \frac{1}{\tau} \int_0^t \exp((y - t)/\tau) v(y)\, dy.$$

(c) Suppose that $x_1(0) = 0$ and $v(y) = \sin(2\pi\omega y)$ and conclude that

$$x_1(t) = \frac{\sin(2\pi\omega t) + 2\pi\omega\tau\{\exp(-t/\tau) - \cos(2\pi\omega t)\}}{1 + (2\pi\omega\tau)^2}. \tag{8.66}$$

Note that as the numerator is of order $\omega$ while the denominator is of order $\omega^2$ that the magnitude of $x_1(t)$ will decrease with increasing frequency. As such we call the circuit of Figure 8.11 a low pass filter.

(d) Use Eq. (8.14) to take the Laplace Transform of Eq. (8.65) (with again $x_1(0) = 0$) and show that

$$X_2(s) = H(s)V(s) \quad \text{where} \quad H(s) = \frac{1 + R_3/R_2}{1 + \tau s}.$$

We speak of $H$ as the **transfer function**. Graph its associated

$$\text{Gain}(\omega) \equiv 20 \log_{10} |H(2\pi i\omega)|,$$

as in Figure 8.11(B) for the concrete choice

$$R_1 = 98.8\,k\Omega, \quad R_2 = 978\,k\Omega, \quad R_3 = 100.6\,k\Omega \quad \text{and} \quad C_1 = 10.4\,nF.$$

10. 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.

11. Find the eigenvectors, by hand, of the $B$ matrix, (8.45), associated with the single vibrating mass. The eigenvalues are $\pm i\omega$ and so you must find the two null spaces $\mathcal{N}(\pm\omega I - B)$.

12. The restoring force of a viscous damper is proportional to the velocity of the attached mass. In the notation of Figure 8.12(A) this reads $y_d(t) = dx'(t)$ and so force balance takes the form

$$mx''(t) + dx'(t) + kx(t) = f(t).$$

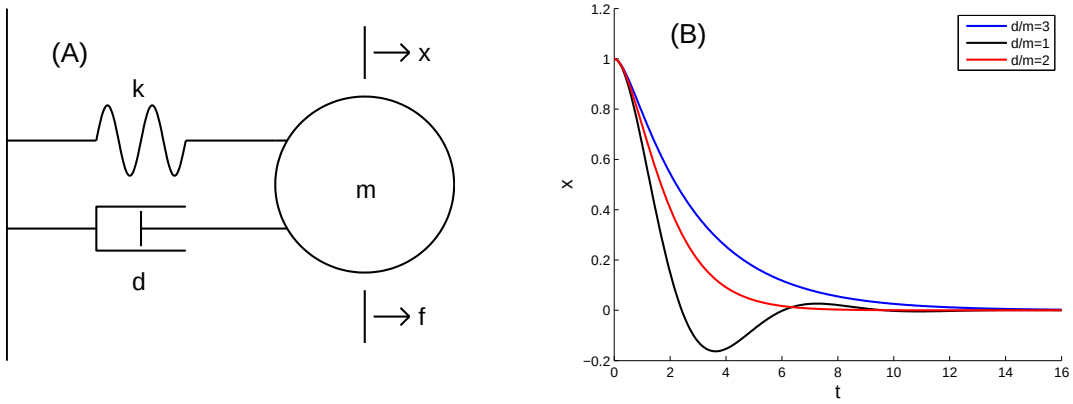



**Figure** 8.12. (A) One mass attached to a spring and a damper. (B) The response of the system to the initial disturbance $x(0) = 1$, $x'(0) = 0$ when $k/m$ and $d/m$ is either the under–, over–, or critically damped regime.

(a) Express this as a first order system, $u' = Bu + g$.

(b) Show that

$$(sI - B)^{-1} = \frac{1}{s^2 + (d/m)s + (k/m)} \begin{pmatrix} s + d/m & 1 \\ -k/m & s \end{pmatrix} \tag{8.67}$$

(c) Show that the poles of $(sI - B)^{-1}$ are at

$$\lambda_\pm \equiv \frac{-d \pm \sqrt{d^2 - 4km}}{2m}.$$

Note that these are real and distinct when $d^2 > 4km$, real and coincident when $d^2 = 4km$, and nonreal and distinct when $d^2 < 4km$. We will see that these three scenarios lead to distinctly different responses.

(d) Show that if $d^2 \neq 4km$ then

$$\exp(Bt) = \frac{\exp(\lambda_+ t)}{\lambda_+ - \lambda_-} \begin{pmatrix} \lambda_+ + d/m & 1 \\ -k/m & \lambda_+ \end{pmatrix} + \frac{\exp(\lambda_- t)}{\lambda_- - \lambda_+} \begin{pmatrix} \lambda_- + d/m & 1 \\ -k/m & \lambda_- \end{pmatrix}$$

satisfies $(\exp(Bt))' = B\exp(Bt)$.

(e) Use $u(t) = \exp(Bt)u(0)$ to show that if $u(0) = [1\ 0]^T$ then

$$\begin{aligned} x(t) &= \frac{\exp(\lambda_+ t)(\lambda_+ + d/m)}{\lambda_+ - \lambda_-} + \frac{\exp(\lambda_- t)(\lambda_- + d/m)}{\lambda_- - \lambda_+} \\ &= \exp(-dt/(2m))(\cosh(\delta t/(2m)) + d\sinh(\delta t/(2m))/\delta) \end{aligned} \tag{8.68}$$

where $\delta \equiv \sqrt{d^2 - 4km}$.

(f) If $d^2 > 4km$ then $\delta > 0$ and the system is called **overdamped** because the $x$ in (8.68) decays to zero without oscillation. For example, if $k/m = 1$ and $d/m = 3$ then

$$x(t) = \exp(-3t/2)(\cosh(\sqrt{5}t/2) + 3\sinh(\sqrt{5}t/2)/\sqrt{5})$$

(g) If $d^2 < 4km$ then $\delta = i2m\omega_d$ where $\omega_d \equiv \sqrt{4km - d^2}/2m$ and the system is called **underdamped** because the $x$ in (8.68) oscillates as it decays to zero. To understand this we unpack

$$\cosh(\delta t/(2m)) = \cosh(i\omega_d t) = \frac{\exp(i\omega_d t) + \exp(-i\omega_d t)}{2} = \cos(\omega_d t)$$

and similarly for sinh and so arrive at

$$x(t) = \exp(-dt/(2m))(\cos(\omega_d t) + d\sin(\omega_d t)/\delta).$$

Show that the natural frequency, $\omega_d$, of the underdamped system is less that the natural frequency, $\omega = \sqrt{k/m}$, of the undamped system. For example, if $k/m = 1$ and $d/m = 1$ then

$$x(t) = \exp(-t/2)(\cos(\sqrt{3}t/2) + \sin(\sqrt{3}t/2)/\sqrt{3}).$$

(h) Finally, if $d^2 = 4km$ we call the system **critically damped**. Rather than working all the way back through a new $\exp(Bt)$, simply let $\delta \to 0$ in (8.68) to arrive at

$$x(t) = \exp(-dt/(2m))(1 + dt/(2m)).$$

For example, if $k/m = 1$ and $d/m = 2$ then

$$x(t) = \exp(-t)(1 + t).$$

13. Find the steady state, $\overline{x}$, of the Goodwin Oscillator (8.41), evaluate (by hand) the associated gradient $B = \nabla F(\overline{x})$, and show (by hand) that the eigenvalues of $B$ are purely imaginary. In this situation we declare $\overline{x}$ **neutrally stable** and call it a **center**.

14. Our Goodwin Oscillator, (8.41), is a stripped down version of his original system

$$\begin{aligned} x_1'(t) &= k_1 \frac{K^n}{K^n + x_3^n(t)} - k_2 x_1(t) \\ x_2'(t) &= k_3 x_1(t) - k_4 x_2(t) \\ x_3'(t) &= k_5 x_2(t) - k_6 x_3(t) \end{aligned} \tag{8.69}$$

Let us set the coefficients

$$k_1 = k_3 = k_5 = K = 1 \quad \text{and} \quad k_2 = k_4 = k_6 = 1/10 \tag{8.70}$$

and investigate the stability of steady state solutions of (8.69) as $n$ varies through the small positive integers.

(a) At steady state show that $\overline{x}_3^{n+1} + \overline{x}_3 = 1000$ and that $\overline{x}_1$ and $\overline{x}_2$ are known multiples of $\overline{x}_3$.

(b) Show that the associated gradient is of the form

$$B = \nabla F(\overline{x}) = \begin{pmatrix} -1/10 & 0 & -\alpha^3 \\ 1 & -1/10 & 0 \\ 0 & 1 & -1/10 \end{pmatrix}$$

and express $\alpha$ in terms of $n$ and $\overline{x}_3$.

(c) Show that the eigenvalues of $B$ are

$$-\alpha - 1/10 \quad \text{and} \quad \alpha \frac{1 \pm i\sqrt{3}}{2} - 1/10$$

and hence that they lie in the left half plane so long as $\alpha \leq 1/5$.

(d) Use (a), (b) and (c) to show that the eigenvalues of $B$ lie in the left half plane so long as $n \leq 8$.

(e) Check the predictions of (d) by solving (8.69), with the parameter set (8.70), via Heun's method with $n = 6$ and $n = 12$ and so reproduce Figure 8.13

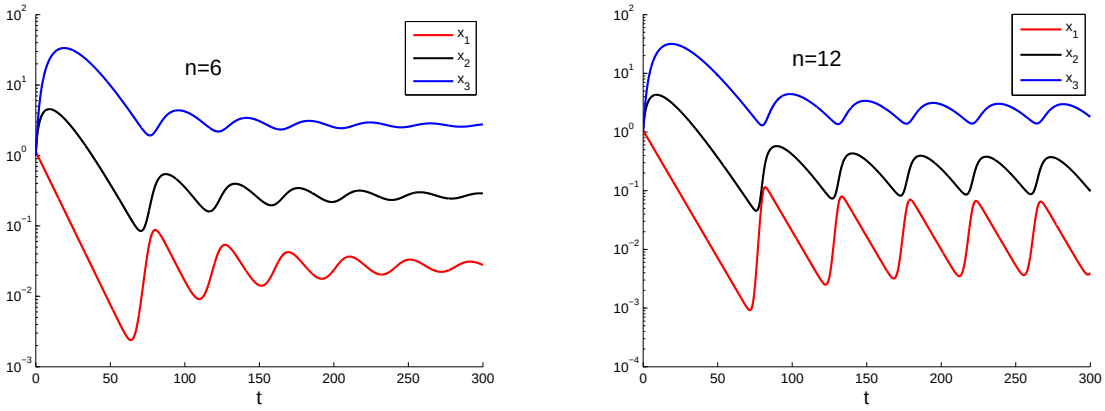



**Figure** 8.13. Numerical solution of (8.69)–(8.70) from initial data $x_1(0) = x_2(0) = x_3(0) = 1$ at $n = 6$ and $n = 12$. Discuss these solutions in light of the predictions made in part (d). `good3dyn.m`

15. Check the predictions made at the end of §8.5 by solving (8.53) via Heun's method and showing that $v_0 = 15$ gives rise to a stable spiral and $v_0 = 18$ yields a stable node.

# 9. Complex Numbers, Functions and Derivatives

In this Chapter we investigate the algebra and geometry of the complex plane, $\mathbb{C}$, and begin the study of the calculus of functions from $\mathbb{C}$ to $\mathbb{C}$. Although there may be much that is new about this chapter – the basic tool of Partial Fraction Expansion is elementary and perhaps fondly remembered from real calculus. Our intent here is to prepare the way for the complex integration required to make sense of the resolvent, eigenvalue problem and the inverse Laplace transform and their role in understanding dynamics – though we pause to develop the basics of Fourier Series and Transforms and its application to Time Series.

### 9.1. Complex Numbers

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

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

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

$$z = |z|(\cos\theta + i\sin\theta), \tag{9.1}$$

where $|z|$ is the magnitude of $z$ and $\theta$ is the angle that it make with the positive real axis. We have illustrated these various representations in Figure 9.1.

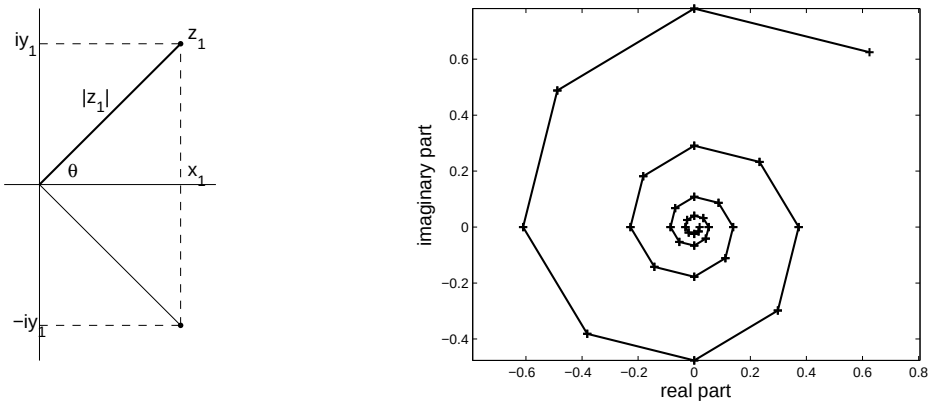



**Figure** 9.1. (A) An illustration of the complex number $z_1 = x_1 + iy_1 = |z_1|(\cos\theta + i\sin\theta)$. (B) The trajectory of $((1+i)/1.6)^n$ as $n$ grows 1 to 32.

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

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

As a result, for integer values of $n$, we find

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

This formula dictates that taking powers of a complex number simultaneously scales its magnitude and rotates its argument. The spiral in Figure 9.1(B) offers a concrete illustration.

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 inner product of two complex vectors requires, however, some care. 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

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

Now norm **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 norm of a complex vector via

$$\|z\| \equiv \sqrt{\overline{z}^T z}. \tag{9.2}$$

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 double symbol, conjugate transpose, occurs so often – for both vectors and matrices – that it has been contracted to a single symbol. Namely

$$\boxed{Z^* \equiv \overline{Z}^T, \quad Z \in \mathbb{C}^{m\times n}.} \tag{9.3}$$

The notion of magnitude in (9.2) 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

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

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 $(i/2)^n \to 0$, for given an $\varepsilon$ we note that if $n > N = \log_2(1/\varepsilon)$ then $|(i/2)^n| < \varepsilon$. Similarly, the series

$$\sum_{n=1}^{\infty} z_n$$

is said to converge to the number $Z$ if the sequence of partial sums

$$Z_n \equiv \sum_{m=1}^{n} z_m$$

converge to $Z$. We shall make use of

**Proposition** 9.1. If $z_j \to z$ and

$$c_n \equiv \frac{1}{n}\sum_{j=1}^{n} z_j$$

then $c_n \to z$.

Proof: Given $\varepsilon > 0$ as $z_j \to z$ there exists an $N > 0$ such that $|z_j - z| < \varepsilon/2$ when $j \geq N$. Now, for $n > N$ write

$$\begin{aligned} |z - c_n| = \left| z - \frac{1}{n}\sum_{m=1}^{n} z_m \right| = \left| \frac{1}{n}\sum_{m=1}^{n} (z - z_m) \right| &\leq \frac{1}{n}\sum_{m=1}^{n} |z - z_m| \\ &= \frac{1}{n}\sum_{m=1}^{N} |z - z_m| + \frac{1}{n}\sum_{m=N+1}^{n} |z - z_m| \\ &\leq \frac{\varepsilon}{2} + \frac{1}{n}\sum_{m=1}^{N} |z - z_m| \end{aligned}$$

and so $|z - c_n| < \varepsilon$ when $n > N$ and

$$n > \frac{2}{\varepsilon}\sum_{m=1}^{N} |z - z_m|.$$

End of Proof.

### 9.2. 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

$$f(z) = (z - \lambda_1)^{o_1}(z - \lambda_2)^{o_2} \cdots (z - \lambda_h)^{o_h}. \tag{9.4}$$

Each $\lambda_j$ is a **root** of $f$ of **order** $o_j$. Here $h$ is the number of **distinct** roots of $f$. In the previous chapter we observed the appearance of ratios of polynomials, or so called **rational** functions, when taking Laplace transforms, (8.18), and evaluating resolvents, (8.20). Suppose

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

is rational, that $f$ is of degree at most $m-1$ while $g$ is of degree $m$ with $m$ distinct roots $\{\lambda_1, \ldots, \lambda_m\}$. Our central task is to arrive at multiple, and complementary means of computing the $r_j$ in the **Partial Fraction Expansion**

$$r(z) = \sum_{j=1}^{m} \frac{r_j}{z - \lambda_j} \tag{9.5}$$

of $r$. Our first approach is a direct one, that you may recall from calculus. We uncover the $r_j$ by first multiplying each side of (9.5) by $(z - \lambda_j)$ and then setting $z = \lambda_j$. For example, if

$$\frac{1}{z^2 + 1} = \frac{r_1}{z + i} + \frac{r_2}{z - i} \tag{9.6}$$

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 (9.6) by $(z - i)$ and then set $z = i$. So doing we find $r_2 = -i/2$, and so

$$\frac{1}{z^2 + 1} = \frac{i/2}{z + i} + \frac{-i/2}{z - i}. \tag{9.7}$$

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

$$r_j = (z - \lambda_j) r(z)\big|_{z=\lambda_j}. \tag{9.8}$$

You should be able to use this to confirm that

$$\frac{z}{z^2 + 1} = \frac{1/2}{z + i} + \frac{1/2}{z - i}. \tag{9.9}$$

We now have the tools to compute the partial fraction expansion of the resolvent

$$(zI - B)^{-1} = \frac{1}{z^2 + 1} \begin{pmatrix} z & 1 \\ -1 & z \end{pmatrix} \tag{9.10}$$

of the matrix, (8.45), associated with the vibration of a single mass, when $k = m$. In particular, (9.7) and (9.9) allow us to write (9.10) as

$$(zI - B)^{-1} = \frac{1}{z + i} \begin{pmatrix} 1/2 & i/2 \\ -i/2 & 1/2 \end{pmatrix} + \frac{1}{z - i} \begin{pmatrix} 1/2 & -i/2 \\ i/2 & 1/2 \end{pmatrix}. \tag{9.11}$$

We recognize these coefficient matrices as precisely those that appear in (8.47).

In Chapter 8 we were confronted with the complex exponential when considering the Laplace Transform. By analogy to the real exponential we define

$$\boxed{\exp(z) \equiv \sum_{n=0}^{\infty} \frac{z^n}{n!}} \tag{9.12}$$

and find that, for $\theta \in \mathbb{R}$,

$$\begin{aligned} \exp(i\theta) &= 1 + i\theta + (i\theta)^2/2 + (i\theta)^3/3! + (i\theta)^4/4! + \cdots \\ &= (1 - \theta^2/2 + \theta^4/4! - \cdots) + i(\theta - \theta^3/3! + \theta^5/5! - \cdots) \\ &= \cos\theta + i\sin\theta. \end{aligned} \tag{9.13}$$

This should hopefully clear up any mystery remaining from (8.48). With these observations, the polar form is now simply $z = |z|\exp(i\theta)$. One may just as easily verify that

$$\cos\theta = \frac{\exp(i\theta) + \exp(-i\theta)}{2} \quad \text{and} \quad \sin\theta = \frac{\exp(i\theta) - \exp(-i\theta)}{2i}.$$

These suggest the definitions, for complex $z$, of

$$\boxed{\cos z \equiv \frac{\exp(iz) + \exp(-iz)}{2}} \quad \text{and} \quad \boxed{\sin z \equiv \frac{\exp(iz) - \exp(-iz)}{2i}.} \tag{9.14}$$

As in the real case the exponential enjoys the property that

$$\exp(z_1 + z_2) = \exp(z_1)\exp(z_2)$$

and in particular

$$\exp(x + iy) = \exp(x)\exp(iy) = \exp(x)(\cos y + i\sin y).$$

In order to visualize these complex functions we plot in Figure 9.2 their transformation of a regular grid.

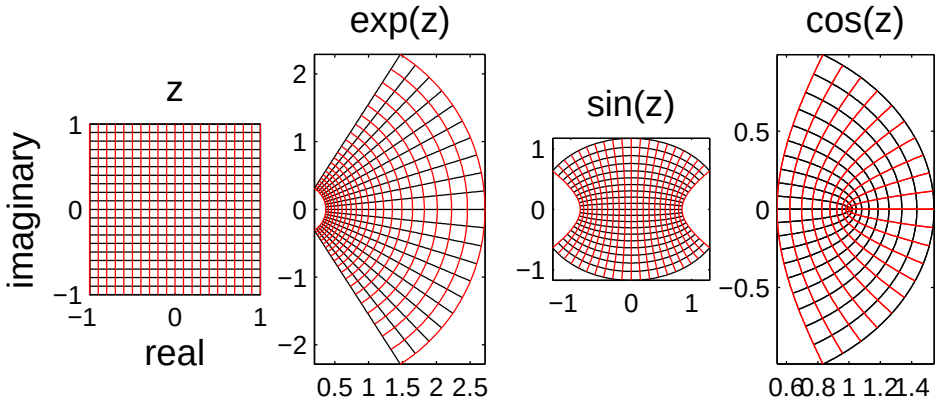



**Figure** 9.2. The deformation of horizontal (black) segments and vertical (red) segments by exp, sin and cos.

### 9.3. Complex Differentiation and the First Residue Theorem

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.

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

**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** 9.2. 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**

$$\boxed{\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \quad \text{and} \quad \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}.} \tag{9.15}$$

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

$$z+2 = r_{1,1}(z+1) + r_{1,2}. \tag{9.16}$$

On setting $z=-1$ this gives $r_{1,2}=1$. With $r_{1,2}$ computed (9.16) 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

$$(z+2)^2 = r_{1,1}(z+1)^2 + r_{1,2}(z+1) + r_{1,3}. \tag{9.17}$$

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 (9.17) is to hold for **all** $z$. We exploit this by taking two derivatives, with respect to $z$, of (9.17). 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** 9.3. **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 $\{o_1, \ldots, o_h\}$, may be expanded in partial fractions

$$r(z) = \sum_{j=1}^{h} \sum_{k=1}^{o_j} \frac{r_{j,k}}{(z - \lambda_j)^k} \tag{9.18}$$

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,

$$r_{j,k} = \lim_{z \to \lambda_j} \frac{1}{(o_j - k)!} \frac{d^{o_j - k}}{dz^{o_j - k}} \{(z - \lambda_j)^{o_j} r(z)\}. \tag{9.19}$$

This result permits us to compute the partial fraction of the resolvent, recall (8.67), of the critically damped single mass, $k = m = 1$ and $d = 2$,

$$B = \begin{pmatrix} 0 & 1 \\ -1 & -2 \end{pmatrix} \quad \text{and} \quad (sI - B)^{-1} = \frac{1}{(s+1)^2} \begin{pmatrix} s+2 & 1 \\ -1 & s \end{pmatrix}. \tag{9.20}$$

The required expansions,

$$\frac{s}{(s+1)^2} = \frac{1}{s+1} - \frac{1}{(s+1)^2}$$
$$\frac{s+2}{(s+1)^2} = \frac{1}{s+1} + \frac{1}{(s+1)^2}$$

were constructed en route to Prop. 9.3. It follows that the resolvent in (9.20) may be written

$$(sI - B)^{-1} = \frac{1}{s+1} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \frac{1}{(s+1)^2} \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}. \tag{9.21}$$

### 9.4. Möbius Transformations and Discrete Dynamics*

We study a simple, and yet amazingly rich, class of rational functions, named after August Möbius. A **Möbius** transformation is a function of the form

$$\mu(z) = \frac{az + b}{cz + d} \tag{9.22}$$

for fixed, complex numbers $a,\ b,\ c$ and $d$. On differentiating $\mu$ with respect to $z$ we find

$$\mu'(z) = \frac{ad - bc}{(cz + d)^2}, \tag{9.23}$$

and so $\mu$ is nonconstant when $ad - bc \neq 0$. As multiplication of $a,\ b,\ c$ and $d$ by a common (nonzero) factor leads to the same Möbius transformation we adopt the convention (for the remainder of this section)

$$ad - bc = 1. \tag{9.24}$$

This condition in fact allows us to extend $\mu$ to an invertible mapping of the extended complex plane, $\mathbb{C}_\infty \equiv \mathbb{C} \cup \infty$. In particular, if $c = 0$ then set $\mu(\infty) = \infty$, while if $c \neq 0$ then (9.24) permits us to

unambiguously express $\mu(-d/c)$:

$$\mu(-d/c) = \frac{b - ad/c}{d - cd/c} = \frac{bc - ad}{dc - cd} = \frac{-1}{0} \equiv \infty.$$

Conversely,

$$\mu(\infty) = \frac{a\infty + b}{c\infty + d} = \frac{a + b/\infty}{c + d/\infty} = \frac{a + 0}{c + 0} = a/c.$$

We next observe that if $\mu_1$ and $\mu_2$ are Möbius Transformations then their composition

$$\mu_3(z) \equiv \mu_1(\mu_2(z)) = \frac{a_1\mu_2(z) + b_1}{c_1\mu_2(z) + d_1} = \frac{b_1 + a_1(a_2z + b_2)/(c_2z + d_2)}{d_1 + c_1(a_2z + b_2)/(c_2z + d_2)} = \frac{(a_1a_2 + b_1c_2)z + (a_1b_2 + b_1d_2)}{(c_1a_2 + c_2d_1)z + (c_1b_2 + d_1d_2)}.$$

is another Möbius Transformation. Moreover, the coefficients of the composition correspond precisely to the multiplication of the two associated 2-by-2 matrices,

$$\begin{pmatrix} a_3 & b_3 \\ c_3 & d_3 \end{pmatrix} = \begin{pmatrix} a_1a_2 + b_1c_2 & a_1b_2 + b_1d_2 \\ c_1a_2 + c_2d_1 & c_1b_2 + d_1d_2 \end{pmatrix} = \begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix} \begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix}.$$

As the inverse of $\mu$ is that function for which $\mu^{-1}(\mu(z)) = z$ for all $z$ it follows that the coefficients of $\mu^{-1}$ are precisely the elements of

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}. \tag{9.25}$$

In what follows we first show that each Möbius Transformation takes a circle or a line to a circle or a line.

The general equation for the circle and the line in the $(x, y)$ plane is

$$A(x^2 + y^2) + b_1x + b_2y + C = 0 \tag{9.26}$$

where each constant is real. In order to make this more suitable to our Möbius tranformations we invert $z = x + iy$ and $\overline{z} = x - iy$ for

$$x = (z + \overline{z})/2 \quad \text{and} \quad y = i(\overline{z} - z)/2.$$

These, together with $x^2 + y^2 = z\overline{z}$ and $B \equiv (b_1 - ib_2)/2$, permit us to express (9.26) as

$$Az\overline{z} + Bz + \overline{B}\overline{z} + C = 0. \tag{9.27}$$

Now if we set $w = \mu(z)$ then, recalling (9.25),

$$z = \mu^{-1}(w) = \frac{dw - b}{a - cw} \quad \text{and} \quad \overline{z} = \frac{\overline{d}\overline{w} - \overline{b}}{\overline{a} - \overline{cw}}.$$

On using these expressions for $z$ and $\overline{z}$ in (9.27), and clearing fractions, we find

$$A(dw - b)(\overline{d}\overline{w} - \overline{b}) + B(dw - b)(\overline{a} - \overline{cw}) + \overline{B}(\overline{d}\overline{w} - \overline{b})(a - cw) + C(a - cw)(\overline{a} - \overline{cw}) = 0.$$

Collecting terms brings

$$\alpha w\overline{w} + \beta w + \overline{\beta}\overline{w} + \gamma = 0 \tag{9.28}$$

where

$$\alpha = A|d|^2 - 2\Re(Bd\overline{c}) + C|c|^2, \quad \beta = -Ad\overline{b} + Bd\overline{a} + \overline{B}\overline{b}c - Cc\overline{a}, \quad \gamma = A|b|^2 - 2\Re(Bb\overline{a}) + C|a|^2.$$

As $w = \mu(z)$ obeys (9.28) with real $\alpha$ and $\gamma$ we have proven

**Proposition** 9.4. Each Möbius transformation maps circles and lines onto circles and lines.

Our interest here is in classifying Möbius Transformations via the nature of their orbits. Namely by the properties of

$$z_n = \mu(z_{n-1}) \tag{9.29}$$

as $n \to \infty$. As in the continuous case we begin with the steady–state solutions. In this case, if $z_n \to z_*$ then (9.29) implies that

$$z_* = \mu(z_*).$$

We call such a $z_*$ a **Fixed Point** of $\mu$.

As $\mu(\infty) = a/c$ we note that $\mu$ fixes $\infty$ iff $c = 0$. When $c = 0$ we note that $d = 1/a$ and that $\mu(z) = z$ requires $az + b = z/a$. This equation forks two ways; if $a = \pm 1$ then $z = \infty$ is the only fixed point, while if $a \neq \pm 1$ then $z = ab/(1 - a^2)$ is the second fixed point of $\mu$.

In the case that $c \neq 0$, the fixed point condition $\mu(z) = z$ is simply the quadratic equation $cz^2 + (d - a)z - b = 0$. Its roots,

$$z_\pm = \frac{a - d \pm \sqrt{(\mathrm{tr}M)^2 - 4}}{2c}, \tag{9.30}$$

where $\mathrm{tr}M = a + d$ and $M$ is the 2-by-2 matrix associated with $\mu$, are distinct when $(\mathrm{tr}M)^2 \neq 4$. These observations in fact establish

**Proposition** 9.5. Suppose $\mu$ is a Möbius Transformation with matrix representation $M$. If $(\mathrm{tr}M)^2 \neq 4$ then $\mu$ has two fixed points. If $(\mathrm{tr}M)^2 = 4$ and $\mu$ is not the identity then $\mu$ has one fixed point.

As in the case of continuous dynamics we expect these fixed points (steady–states) to be either attracting or repelling or the center of oscillations.

To begin, for $c \neq 0$ and $(\mathrm{tr}\, M)^2 \neq 4$, $\mu$ has two distinct finite roots, $z_\pm$, per (9.30). These roots are mapped to 0 and $\infty$ by the explicit Möbius Transformation

$$\sigma(z) \equiv \frac{z - z_+}{z - z_-}. \tag{9.31}$$

As the inverse of $\sigma$ therefore maps 0 and $\infty$ to $z_\pm$ it follows that

$$\nu(z) \equiv \sigma(\mu(\sigma^{-1}(z))) \tag{9.32}$$

fixes 0 and $\infty$ and is therefore of a very simple form. To see this we note that its associated matrix

$$N = SMS^{-1} = \begin{pmatrix} 1 & -z_+ \\ 1 & -z_- \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \frac{1}{z_+ - z_-} \begin{pmatrix} -z_- & z_+ \\ -1 & 1 \end{pmatrix} = \begin{pmatrix} \lambda & 0 \\ 0 & 1/\lambda \end{pmatrix} \tag{9.33}$$

is **diagonal**, where

$$\lambda = \frac{\mathrm{tr}M + \sqrt{(\mathrm{tr}\, M)^2 - 4}}{2}. \tag{9.34}$$

It follows that $\nu$ is merely multiplication

$$\nu(z) = \frac{\lambda z + 0}{0 + 1/\lambda} = \lambda^2 z. \tag{9.35}$$

As such its orbits, $z_n = \nu(z_{n-1}) = \lambda^{2n} z_0$, grow, decay or oscillate as $|\lambda|$ is respectively greater than, less than, or equal to one. To see how this facilitates study of the orbits of $\mu$ we simply invert (9.32) to go from

$$z_n = \mu(z_{n-1}) \quad \text{to} \quad z_n = \sigma^{-1}(\nu(\sigma(z_{n-1}))) = \sigma^{-1}(\lambda^{2n}\sigma(z_0)). \tag{9.36}$$

Hence, the orbits of $\mu$ are determined solely by powers of $\lambda^2$.

Although the situation when $c = 0$ and $a \neq \pm 1$ is really just a special case of the analysis above, it is worth spelling it out. In this case the fixed points are $z_+ = ab/(1 - a^2)$ and $z_- = \infty$ and so we replace the fixed point shifter, (9.31), with the even simpler

$$\sigma(z) \equiv z - z_+.$$

It follows, as above, that $\nu(z) \equiv \sigma(\mu(\sigma^{-1}(z)))$ fixes 0 and $\infty$ and is therefore merely multiplication. In particular, as its associated matrix

$$N = SMS^{-1} = \begin{pmatrix} 1 & -z_+ \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ 0 & 1/a \end{pmatrix} \begin{pmatrix} 1 & z_+ \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a & 0 \\ 0 & 1/a \end{pmatrix},$$

it follows that $\nu(z) = a^2 z$ and hence the dynamics of $\mu$ are completely determined by $a$.

We consider then its polar expression $\lambda^2 = r\exp(i\theta)$ and consider the three distinct cases.

The Möbius Transformation $\mu$ is called **hyperbolic** when $\lambda^2 = r > 0$. It follows that if $z_0 \neq z_\pm$ and $r > 1$ then $z_n \to \sigma^{-1}(\infty) = z_-$ and hence $z_-$ is attracting while $z_+$ is repelling. Similarly, if $r < 1$ then $z_n \to \sigma^{-1}(0) = z_+$ and so now $z_+$ is attracting while $z_-$ is repelling. Given (9.34) we should be able to state necessary and sufficient conditions on $\mathrm{tr}\, M$ for $\lambda^2 = r$. From (9.33) it follows that $\mathrm{tr}\, M = \mathrm{tr}\, N = \lambda + 1/\lambda$. Squaring this expression brings

$$(\mathrm{tr}\, M)^2 = 2 + \lambda^2 + 1/\lambda^2. \tag{9.37}$$

Now, if $\lambda^2 = r > 0$ then $(\mathrm{tr}, M)^2$ is real and bounded below by 4 (for $r + 1/r \geq 2$ when $r > 0$). Hence, if $\mu$ is hyperbolic then $\mathrm{tr}\, M \in \mathbb{R}$ and $|\mathrm{tr}\, M| > 2$. We illustrate this case by plotting orbits of

$$\mu_H(z) = \frac{z+1}{z+2} \tag{9.38}$$

in Figure 9.3(A).

The Möbius Transformation $\mu$ is called **elliptic** when $\lambda^2 = \exp(i\theta)$. In this case we note that $\lambda^{2n}\sigma(z_0) = \exp(2ni\theta)\sigma(z_0)$ does not converge but that each iterate lies on the circle of radius $|\sigma(z_0)|$. As $\sigma^{-1}$ takes circles to circles it follows that each $z_n$ lies on a circle (centered at $z_+$ or $z_-$). If $\mu$ is elliptic then (9.37) requires that

$$(\mathrm{tr}\, M)^2 = 2 + \exp(i\theta) + \exp(-i\theta) = 2 + 2\cos(\theta), \tag{9.39}$$

from which we deduce that $\operatorname{tr} M \in \mathbb{R}$ and $|\operatorname{tr} M| < 2$. We have illustrated this case by plotting orbits of

$$\mu_E(z) = \frac{z/2 - 1/2}{z + 1} \tag{9.40}$$

in Figure 9.3(C).

The Möbius Transformation $\mu$ is called **loxodromic** when $\lambda^2 = r\exp(i\theta)$ where $r > 0$, $r \neq 1$ and $\theta \neq 2k\pi$. In this case (9.37) reads

$$(\operatorname{tr} M)^2 = 2 + (r + 1/r)\cos(\theta) + i(r - 1/r)\sin(\theta).$$

The right hand side is not real unless $\theta = \pi$. In this case we find $(\operatorname{tr} M)^2 = 2 - (r + 1/r) < 0$ (as $r + 1/r > 2$). Hence, if $\mu$ is loxodromic then $\operatorname{tr} M \notin \mathbb{R}$. We have illustrated this case by plotting orbits of

$$\mu_L(z) = \frac{(1 - 5i)z + 1}{z + 1 - 5i} \tag{9.41}$$

in Figure 9.3(C).

Finally, if $\operatorname{tr} M = \pm$ then, by (9.30), $\mu$ has the single fixed point $(a - d)/(2c)$. We use

$$\sigma(z) \equiv \frac{1/c}{z - (a - d)/(2c)}$$

to send it to $\infty$ so that $\nu(z) \equiv \sigma(\mu(\sigma^{-1}(z)))$ fixes $\infty$. The associated matrix

$$N = SMS^{-1} = \begin{pmatrix} 0 & 1/c \\ 1 & (d-a)/(2c) \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} (a-d)/2 & 1 \\ c & 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

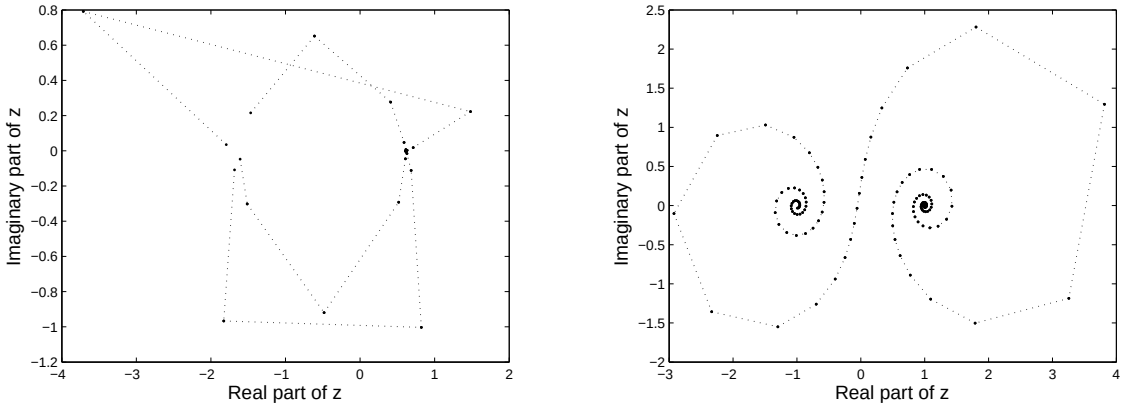
is what we called a **Jordan Block** in §4.4. It follows that $\nu(z) = z + 1$ and that the dynamics of $\mu$ come down to

$$z_n = \mu(z_{n-1}) = \sigma^{-1}(\nu(\sigma(z_{n-1}))) = \sigma^{-1}(\sigma(z_0) + n) \to \sigma^{-1}(\infty) = (a - d)/(2c).$$

This shows that the lone fixed point is attracting. Möbius transformations with $\operatorname{tr} M = \pm 2$ are deemed **parabolic**. We have illustrated this case by plotting orbits of

$$\mu_P(z) = \frac{z}{z + 1} \tag{9.42}$$

in Figure 9.3(D).



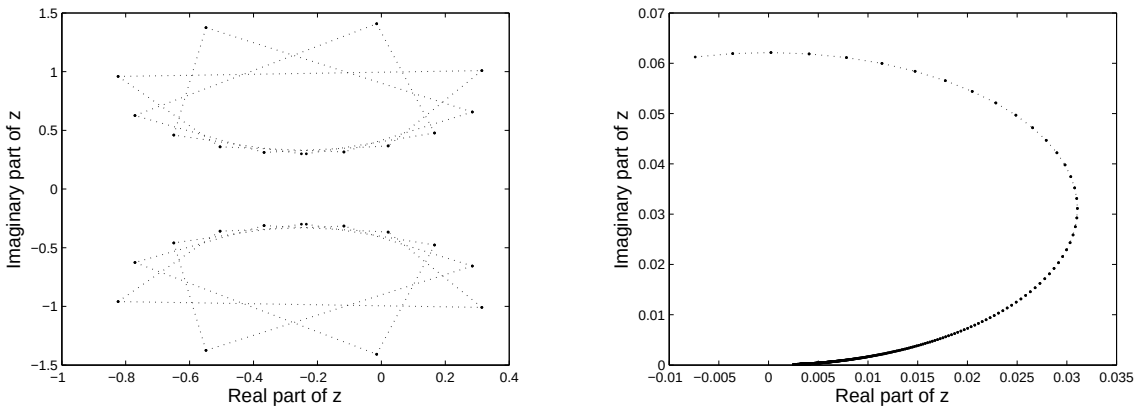



**Figure** 9.3. Orbits of the four types of Möbius Transformations.

To summarize, we have established necessary conditions on the trace of $M$, the matrix associated with a Möbius transformation, $\mu$, for $\mu$ to be of 1 of 4 types. As these conditions are mutually exclusive they must in fact be sufficient. As a result, we have established the fundamental classification of Möbius transformations.

**Proposition** 9.6. The Möbius transformation $\mu$ is
Hyperbolic iff $\operatorname{tr} M \in \mathbb{R}$ and $|\operatorname{tr} M| > 2$,
Elliptic iff $\operatorname{tr} M \in \mathbb{R}$ and $|\operatorname{tr} M| < 2$,
Parabolic iff $\operatorname{tr} M \pm 2$,
Loxodromic iff $\operatorname{tr} M \notin \mathbb{R}$.

We will return to this classification in our work in Chapters 14 and 15 on matrix groups and representation theory. For now, we wish to emphasize that we have also, en route, classified 2-by-2 matrices, with determinant equal to one. To make this precise we need to recall that when $N = SMS^{-1}$ we call $S$ a **similarity transform** and say that $N$ and $M$ are **similar**.

To summarize, we have shown that if $(\operatorname{tr} M)^2 = 4$ then $M$ is similar to

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \text{else } M \text{ is similar to} \quad \begin{pmatrix} \lambda & 0 \\ 0 & 1/\lambda \end{pmatrix}. \tag{9.43}$$

This condition $(\operatorname{tr} M)^2 = 4$ can be seen as a degeneracy condition. Our main spectral result will say that each nondegenerate matrix is similar to a diagonal matrix, and that each degenerate matrix is similar to a diagonal matrix plus a nilpotent matrix.

### 9.5. Fourier Series and Transforms*

The complex exponential is fundamental to the two central "transforms" of modern science, those named after Fourier and Laplace. We pause here to develop the key properties of Fourier Series and Transforms.

We observed in §8.4 that a mechanical system with 2 degrees of freedom vibrates at 2 characteristic frequencies. One of the most common applications of Fourier tools is to the problem of spectral analysis, i.e., the determination of the spectral or frequency components of a signal.

Most everything follows from the fact that $\exp(2\pi imt)$ is "orthogonal" to $\exp(2\pi int)$ in the sense

that, for integers $m$ and $n$,

$$\int_0^1 \exp(2\pi int)\exp(-2\pi imt)\,dt = \int_0^1 \exp(2\pi i(n-m)t)\,dt = \begin{cases} 0 & \text{if } m \neq n, \\ 1 & \text{if } m = n. \end{cases} \tag{9.44}$$

The key idea is now to use these orthonormal exponentials as a basis for a class of functions defined on $0 \leq t \leq 1$. More precisely, given a function $f$ we develop it in a **Fourier Series**

$$f(t) = \sum_{n=-\infty}^{\infty} \hat{f}(n)\exp(2\pi int). \tag{9.45}$$

To determine the Fourier coefficients, $\hat{f}(m)$, multiply each side of (9.45) by $\exp(-2\pi imt)$ then integrate and invoke (9.44):

$$\hat{f}(m) = \int_0^1 f(t)\exp(-2\pi imt)\,dt. \tag{9.46}$$

For example, if

$$f(t) = t \quad \text{then} \quad \hat{f}(0) = \frac{1}{2} \quad \text{and} \quad \hat{f}(n) = \frac{i}{2\pi n}, \quad |n| > 0, \tag{9.47}$$

while if

$$f(t) = t(1-t) \quad \text{then} \quad \hat{f}(0) = \frac{1}{6} \quad \text{and} \quad \hat{f}(n) = \frac{-1}{2(n\pi)^2}, \quad |n| > 0. \tag{9.48}$$

We speak of $\hat{f}(m)\exp(2\pi imt)$ as the projection of $f$ onto $\exp(2\pi imt)$ and so interpret $\hat{f}(m)$ as the "amount" of $f$ at frequency $m$. Regarding the sense of "negative frequency" we note that if $f$ is real then $\hat{f}(-n) = \overline{\hat{f}(n)}$ and so (9.45) takes the form

$$f(t) = \hat{f}_0 + 2\sum_{n=1}^{\infty} \Re\{\hat{f}(n)\exp(-2\pi int)\}. \tag{9.49}$$

This in turn suggests that we write

$$\hat{f}(m) = \int_0^1 f(t)\exp(-2\pi imt)\,dt = \int_0^1 f(t)(\cos(2\pi mt) - i\sin(2\pi mt))\,dt = 2(\hat{f}_c(m) - i\hat{f}_s(m)) \tag{9.50}$$

where

$$\hat{f}_c(m) = \frac{1}{2}\int_0^1 f(t)\cos(2\pi mt)\,dt \quad \text{and} \quad \hat{f}_s(m) = \frac{1}{2}\int_0^1 f(t)\sin(2\pi mt)\,dt, \quad m = 1, 2, \ldots$$

In which case (9.49) becomes

$$f(t) = \hat{f}_0 + \sum_{n=1}^{\infty} \hat{f}_c(n)\cos(2\pi nt) + \hat{f}_s(n)\sin(2\pi nt). \tag{9.51}$$

Returning to our two examples, we find

$$t = \frac{1}{2} - \sum_{n=1}^{\infty} \frac{\sin(2\pi nt)}{n\pi} \quad \text{and} \quad (1-t)t = \frac{1}{6} - \sum_{n=1}^{\infty} \frac{\cos(2\pi nt)}{n^2\pi^2}. \tag{9.52}$$

It is very instructive to visually interrogate the convergence of these sums by graphing the partial sums

$$S_N(t) \equiv \hat{f}_0 + \sum_{n=1}^{N} \hat{f}_c(n)\cos(2\pi nt) + \hat{f}_s(n)\sin(2\pi nt) \tag{9.53}$$

for increasing values of $N$.

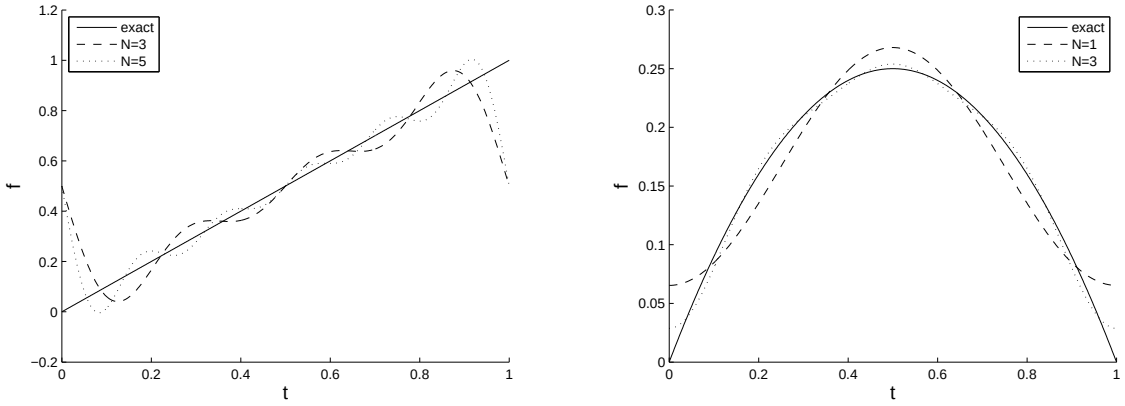



**Figure** 9.4 (**A**) The exact $f(t) = t$ and two of its low frequency (9.53) Fourier approximants. (**B**) The exact $f(t) = (1-t)t$ and two of its low frequency (9.53) Fourier approximants. (`fftexa.m`)

In practice we are most often confronted not with an analytical expression of a function on the unit interval but rather with $N$ samples over an interval of duration $T$:

$$f_N(m) \equiv f(mdt), \quad dt \equiv T/N, \quad m = 0, \ldots, N-1.$$

We now attempt to develop $f_N$ in a discrete Fourier series of the form

$$f_N(m) = \frac{1}{N}\sum_{n=0}^{N-1} \hat{f}_N(n)\exp(2\pi inm/N). \tag{9.54}$$

On defining

$$w_N \equiv \exp(2\pi i/N)$$

we note that Eq. (9.54) takes the very simple form,

$$\frac{1}{N}\sum_{n=0}^{N-1} w_N^{mn}\hat{f}_N(n) = f_N(m). \tag{9.55}$$

This in turn may be written as the matrix equation

$$\frac{1}{N}F_N\hat{f}_N = f_N \tag{9.56}$$

where, noting that $w_N^N = 1$,

$$F_N = \begin{pmatrix} 1 & 1 & 1 & \cdot & 1 \\ 1 & w_N & w_N^2 & \cdot & w_N^{N-1} \\ 1 & w_N^2 & w_N^4 & \cdot & w_N^{2(N-1)} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & w_N^{N-1} & w_N^{2(N-1)} & \cdot & w_N^{(N-1)^2} \end{pmatrix}. \tag{9.57}$$

We now exploit this very special structure of $F_N$ and arrive at an elegant solution to Eq. (9.56). To begin we examine the $jk$ element of $F_N^* F_N$, i.e., row $j$ of $F_N^*$ (the conjugate transpose of $F_N$) times column $k$ of $F_N$,

$$(F_N^* F_N)_{jk} = 1 \cdot 1 + \overline{w}_N^{j-1} w_N^{k-1} + \overline{w}_N^{2(j-1)} w_N^{2(k-1)} + \cdots + \overline{w}_N^{(N-1)(j-1)} w_N^{(N-1)(k-1)}.$$

If $j = k$ then $\overline{w}_N^{m(j-1)} w_N^{m(k-1)} = \exp(-2\pi i m(j-1)) \exp(2\pi i m(j-1)) = 1$ for each $m$ and $(F_N^* F_N)_{jj} = N$. If $j \neq k$ we let $z = \overline{w}_N^{(j-1)} w_N^{(k-1)}$ and find the finite geometric series

$$(F_N^* F_N)_{jk} = 1 + z + z^2 + \cdots + z^{N-1} = \frac{1 - z^N}{1 - z} = 0. \tag{9.58}$$

The middle equality is justified in Exer. 9.3 and the final equality stems from $z^N = 1$. Gathering the above computations, we have shown that

$$F_N^* F_N = NI \quad \text{and so} \quad F_N^{-1} = \frac{1}{N} F_N^* \quad \text{and} \quad \hat{f}_N = F_N^* f_N \tag{9.59}$$

is the solution to Eq. (9.56). We speak of $\hat{f}_N$ as the **Discrete Fourier Transform** (DFT) of $f_N$ and note the latter equation in (9.59) may be expressed in component form as

$$\hat{f}_N(m) = \sum_{n=0}^{N-1} \overline{w}_N^{mn} f_N(n) = \sum_{n=0}^{N-1} \exp(-2\pi i (m/T)(ndt)) f(ndt). \tag{9.60}$$

As such we interpret $m/T$ as the associated discrete frequencies. It also follows from Eq. (9.60) that if $f_N$ is real then

$$\hat{f}_N(N/2 + j) = \overline{\hat{f}}_N(N/2 - j), \quad j = 1, 2, \ldots, N/2 - 1, \tag{9.61}$$

and as such only the first $1 + N/2$ frequencies

$$\omega_m = m/T, \quad m = 0, \ldots, N/2,$$

carry information. This observation in fact leads to a fast method, known as the `fft`, for implementing the Discrete Fourier Transform. The basic idea is that a DFT of order $N$ can be implemented by 2 DFTs of order $N/2$.

We illustrate it, in Figure 9.5, by "recovering" the driving frequency and two frequencies of vibration of the 2–mass system Figure 8.6(A) from knowledge of the displacement of the first mass. In particular, with a driving frequency of $a/2\pi$ at the first mass we use `chain2.m` to compute the first displacement up to time $T = 40$ at $N = 400000$ points. The code

```
fx1 = fft(x1);
omega = (0:N-1)/T;
plot(omega,abs(fx1)/max(abs(fx1)),'k')
```

then peaks at the frequencies present in $x_1$.

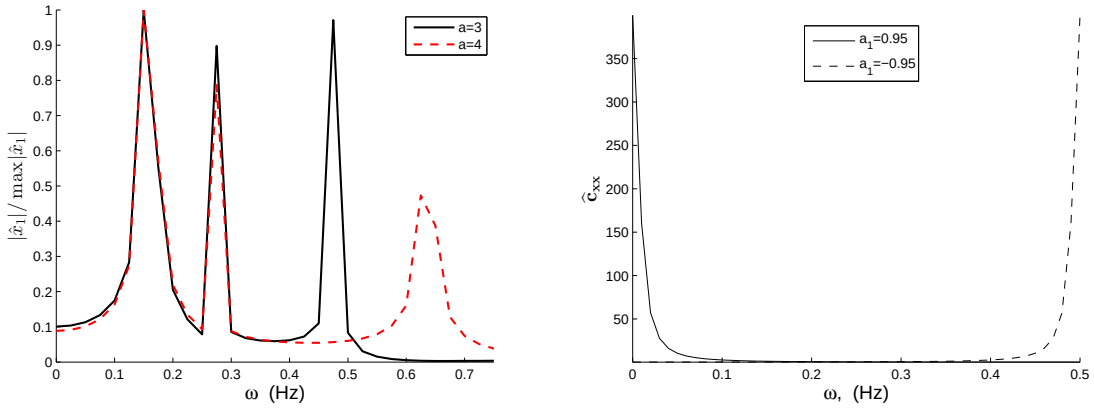



**Figure** 9.5 (A) Detecting the frequencies present in the vibration of a two mass system when the first mass is driven at frequency $a/(2\pi)$ where $a = 3$ and $a = 4$. The two curves coincide at the system's natural, or resonant, frequencies $1/(2\pi)$ and $\sqrt{3}/(2\pi)$, while the black curve also peaks at the driving frequency $3/(2\pi)$ and the red curve also peaks at the driving frequency $4/(2\pi)$. (`chainfreq.m`) (B) The power spectra of $x_{j+1} - a_1 x_j = \varepsilon_j$ with $a_1 = \pm 0.95$.

### 9.6. The Power Spectra of Stationary Processes*

Given two discrete stationary processes, $x$ and $y$, we search for the "filter" $a$ for which the Fourier Transform of the autocovariance of $y - a \star x$ is minimal at every frequency.

We consider a Stationary Process, $x$, of §6.8 and define its **Power Spectrum** to be the Fourier Transform of its autocovariance:

$$\hat{c}_{xx}(\omega) \equiv \sum_{k=-\infty}^{\infty} c_{xx}(k) \exp(-2\pi i k dt \omega) \quad c_{xx}(k) \equiv \text{mean}(x_j x_{j+k}) \tag{9.62}$$

where $dt$ is the time step between samples of $x$.

If $x$ is the AR(1) process of (6.65) and $dt = 1$ $s$ then, recalling (6.66)–(6.67),

$$\begin{aligned} \hat{c}_{xx}(\omega) &= c_0 + c_0 \sum_{k=1}^{\infty} a_1^k \{\exp(2\pi i k\omega) + \exp(-2\pi i k\omega)\} \\ &= c_0 + c_0 \frac{a_1 \exp(2\pi i\omega)}{1 - a_1 \exp(2\pi i\omega)} + c_0 \frac{a_1 \exp(-2\pi i\omega)}{1 - a_1 \exp(-2\pi i\omega)} \\ &= c_0 \frac{1 - a_1^2}{1 - 2a_1 \cos(2\pi\omega) + a_1^2} = \frac{\sigma^2}{1 - 2a_1 \cos(2\pi\omega) + a_1^2} \end{aligned} \tag{9.63}$$

We illustrate this in Figure 9.5(B).

For our next class we consider moving averages

$$x_n = \sum_{j=-\infty}^{\infty} a_j \xi_{n+j}, \quad \text{mean}(\xi_m \xi_n) = \delta_{mn}. \tag{9.64}$$

We argue that the Fourier Transform of its autocovariance

$$c_{xx}(n) \equiv \text{mean}(x_{m+n} x_m) = \sum_{j=-\infty}^{\infty} a_{j-n} a_j$$

can be represented in terms of the Fourier Transform of $a$:

$$\hat{a}(\omega) \equiv \sum_{k=-\infty}^{\infty} a_k \exp(2\pi i k\omega),$$

for

$$a_k = \int_{-1/2}^{1/2} \hat{a}(\omega) \exp(-2\pi i k\omega)\, d\omega$$

and so

$$\begin{aligned}\int_{-1/2}^{1/2} \hat{c}_{xx}(\omega) \exp(2\pi i n\omega)\, d\omega &= \sum_{j=-\infty}^{\infty} a_{j-n}\overline{a}_j = \sum_{j=-\infty}^{\infty} \overline{a}_j \int_{-1/2}^{1/2} \hat{a}(\omega) \exp(2\pi i(n-j)\omega)\, d\omega \\ &= \int_{-1/2}^{1/2} \hat{a}(\omega) \exp(2\pi i n\omega) \sum_{j=-\infty}^{\infty} \overline{a}_j \exp(-2\pi i j\omega)\, d\omega \\ &= \int_{-1/2}^{1/2} \hat{a}(\omega) \exp(2\pi i n\omega)\overline{\hat{a}(\omega)}\, d\omega = \int_{-1/2}^{1/2} |\hat{a}(\omega)|^2 \exp(2\pi i n\omega)\, d\omega\end{aligned}$$

and so if $x$ is a moving average of the form (9.64) then

$$\hat{c}_{xx}(\omega) = |\hat{a}(\omega)|^2.$$

More generally, we consider the discrete convolution of two time series,

$$\boxed{(a \star x)_j \equiv \sum_p a_p x_{j-p}} \tag{9.65}$$

and establish

**Proposition** 9.7. If $y = a \star x$ then
(i) $\hat{y}(\omega) = \hat{a}(\omega)\hat{x}(\omega)$.
(ii) If $x$ is stationary then $c_{xy} = a \star c_{xx}$ and $c_{yy} = (E_\infty a) \star (a \star c_{xx})$.
(iii) If $x$ is stationary then $\hat{c}_{xy}(\omega) = \hat{a}(\omega)\hat{c}_{xx}(\omega)$ and $\hat{c}_{yy}(\omega) = |\hat{a}(\omega)|^2 \hat{c}_{xx}(\omega)$.

Proof: Regarding (i) we find

$$\begin{aligned}\hat{y}(\omega) &= \sum_k (a \star x)_k \exp(-2\pi i k dt\omega) \\ &= \sum_k \sum_p a_p x_{k-p} \exp(-2\pi i(k-p+p)dt\omega) \\ &= \sum_p a_p \exp(-2\pi i p dt\omega) \sum_k x_{k-p} \exp(-2\pi i(k-p)dt\omega) = \hat{a}(\omega)\hat{x}(\omega).\end{aligned}$$

Regarding (ii) we find

$$\begin{aligned}c_{xy}(k) &= \text{mean}(x_j(a \star x)_{j+k}) = \text{mean}(x_j \sum_p a_p x_{j+k-p}) \\ &= \sum_p a_p \text{mean}(x_j x_{j+k-p}) = \sum_p a_p c_{xx}(k-p) = (a \star c_{xx})_k,\end{aligned}$$

and

$$c_{yy}(k) = \text{mean}((a \star x)_j (a \star x)_{j+k}) = \text{mean}(\sum_p a_p x_{j-p} \sum_q a_q x_{j+k-q})$$
$$= \sum_p a_p \text{mean}(x_{j-p} \sum_q a_q x_{j+k-q}) = \sum_p a_p \sum_q a_q \text{mean}(x_{j-p} x_{j+k-q})$$
$$= \sum_p a_p \sum_q a_q c_{xx}(k+p-q) = \sum_p a_p (a \star c_{xx})_{k+p} = ((E_\infty a) \star (a \star c_{xx}))_k.$$

Finally (iii) follows from (i) and (ii) on noting that $\widehat{E_\infty a} = \overline{\hat{a}}$. End of Proof.

We apply this to the AR(1) process, $x_j - a_1 x_{j-1} = \varepsilon_j$, so $\hat{a}(\omega) = 1 - a_1 \exp(-2\pi i\omega)$ and so

$$\hat{c}_{\varepsilon\varepsilon}(\omega) = (1 - 2a_1 \cos(2\pi\omega) + a_1^2)\hat{c}_{xx}(\omega).$$

As $\hat{c}_{\varepsilon\varepsilon}(\omega) = \sigma^2$ we have arrived at (9.63) by different means.

We have seen three examples of the $\hat{c}_{xx}(\omega) \geq 0$. We prove

**Proposition** 9.8. If

$$\sum_{k=-\infty}^{\infty} |c_{xx}(k)| < \infty$$

then

$$\hat{c}_{xx}(\omega) = \lim_{N\to\infty} \frac{1}{N} \text{mean} \left| \sum_{n=0}^{N-1} x_n \exp(-2n\pi i\omega) \right|^2.$$

Proof: We define

$$\hat{x}_N(\omega) \equiv \sum_{n=0}^{N-1} x_n \exp(-2n\pi i\omega)$$

and note that

$$\text{mean}|\hat{x}_N(\omega)|^2 = \text{mean} \left( \sum_{n=0}^{N-1} x_n \exp(-2n\pi i\omega) \sum_{m=0}^{N-1} x_m \exp(2m\pi i\omega) \right)$$
$$= \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} c_{xx}(m-n) \exp(2(m-n)\pi i\omega)$$
$$= \sum_{n=1-N}^{N-1} (N - |n|) c_{xx}(n) \exp(-2n\pi i\omega)$$

where the final equality follows from the lovely identity (1.43). It follows that

$$\lim_{N\to\infty} \frac{1}{N} \text{mean}|\hat{x}_N(\omega)|^2 = \lim_{N\to\infty} \sum_{n=1-N}^{N-1} (1 - |n|/N) c_{xx}(n) \exp(-2n\pi i\omega)$$
$$= \sum_{n=-\infty}^{\infty} c_{xx}(n) \exp(-2n\pi i\omega)$$

thanks to Exer. 9.4. End of Proof.

We next define the **coherence** of two processes to be the normalized cross–spectrum

$$R_{xy}(\omega) \equiv \frac{\hat{c}_{xy}(\omega)}{\sqrt{\hat{c}_{xx}(\omega)\hat{c}_{yy}(\omega)}} \tag{9.66}$$

and note that Prop. 9.8 asserts that if $y = a \star x$ then $R_{xy}(\omega) = \hat{a}(\omega)/|\hat{a}(\omega)|$. As a simple example, if $y_k = x_{k+d}$ is a delayed copy of $x$ then $y = a \star x$ with $a_j = 0$ save $a_{-d} = 1$ in which case $R_{xy}(\omega) = \hat{a}(\omega) = \exp(2d\pi i\omega)$.

Given two processes, $x$ and $y$, we now search for the filter $a$ for which $y$ is closest to $a \star x$ in the sense that the error

$$\varepsilon = y - (a \star x) \tag{9.67}$$

has minimal power spectrum at each frequency. Arguing as in the proof of Prop. 9.7, the autocovariance of each side of (9.67) spells

$$c_{\varepsilon\varepsilon} = c_{yy} - (E_\infty a) \star c_{xy} - a \star c_{yx} + (E_\infty a) \star (a \star c_{xx}).$$

and so the associated error Spectrum is

$$\begin{aligned} \hat{c}_{\varepsilon\varepsilon}(\omega) &= \hat{c}_{yy}(\omega) - \overline{\hat{a}}\hat{c}_{xy} - \hat{a}(\omega)\overline{\hat{c}}_{xy}(\omega) + |\hat{a}(\omega)|^2\hat{c}_{xx}(\omega) \\ &= \hat{c}_{xx}|\hat{a} - \hat{c}_{xy}/\hat{c}_{xx}|^2 + \hat{c}_{yy}(1 - |R_{xy}|^2). \end{aligned} \tag{9.68}$$

To minimize this we choose

$$\hat{a}(\omega) = \hat{c}_{xy}(\omega)/\hat{c}_{xx}(\omega). \tag{9.69}$$

In which case $\hat{c}_{\varepsilon\varepsilon}(\omega) = \hat{c}_{yy}(\omega)(1 - |R_{xy}(\omega)|^2)$. As the $a$ that results from (9.69) requires knowledge of covariances at an infinite number of lags it is known as the filter with *Infinite Impulse Response*, by contrast with the Finite Impulse Response of Exer. 6.25. For real-time filters the obstacle is not access to the infinite past but rather access to the future. To design best filters that require no knowledge of the future will require deeper knowledge of residues.

### 9.7. Notes and Exercises

For a more thorough introduction to Complex Variables see Levinson and Redheffer (1970). Our presentation of Möbius transformations follows Ford (1957). For a more spectacular view consult **?**ndra. For a more thorough study of the Fourier Analysis of Time Series see Brillinger (2001).

1. Please show that $|\exp(x + iy)| = \exp(x)$.

2. Show that if $|z| = 1$ is not a root of unity then its powers are dense on the unit circle.

3. 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 (9.58) from this result.

4. Deduce from Prop. 9.1 that if

$$Z = \lim_{N\to\infty} \sum_{n=1}^{N} z_n \quad \text{then} \quad Z = \lim_{N\to\infty} \sum_{n=1}^{N} (1 - n/N) z_n.$$

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

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

7. The beautiful $\cos(\theta) = (\exp(i\theta) + \exp(-i\theta))/2$ plays a fundamental role in analyzing the state of the random walker. This walker begins, at time $t = 0$, at position $x = 0$ and steps to a neighboring integer, $\pm 1$, at time $t = 1$, with equal probability, $1/2$. From that position he again steps to a neighboring integer with equal probability, $1/2$, and so on. We denote the probability of being at position $n$ at time $t$ by $P(x(t) = n)$.

   (a) Show that $P(x(1) = 1) = 1/2$ and that $P(x(1) = 1)$ is the coefficient of $\exp(i\theta)$ in $\cos(\theta)$.

   (b) Show that $P(x(2) = 0) = 1/2$, $P(x(2) = 2) = 1/4$ and $P(x(2) = -2) = 1/4$ and argue that $P(x(2) = n)$ is the coefficient of $\exp(in\theta)$ in $\cos^2(\theta)$.

   (c) Generalize your argument in (b) to conclude that $P(x(t) = n)$ is the coefficient of $\exp(in\theta)$ in $\cos^t(\theta)$.

   (d) Use (c) to deduce that

$$P(x(t) = n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos^t(\theta) \exp(-in\theta)\, d\theta.$$

   (e) Lets next define $F(t)$ to be the probability that $t$ is the first time that $x(t) = 0$. Please confirm that

$$P(x(t) = 0) = \sum_{j=0}^{t-1} P(x(j) = 0) F(t - j).$$

   Now sum this from $t = 1$ to $t = \infty$ and conclude that

$$\mathcal{P} - 1 = \mathcal{P}\mathcal{F} \quad \text{where} \quad \mathcal{P} = \sum_{t=1}^{\infty} P(x(t) = 0) \quad \text{and} \quad \mathcal{F} = \sum_{t=1}^{\infty} F(t).$$

   Explain why $\mathcal{F}$ is the probability that the walker *ever* returns to the origin.

   (f) Use (e) to show that $\mathcal{F} = 1 - 1/\mathcal{P}$ and (d) to show that

$$\mathcal{P} = \sum_{t=1}^{\infty} P(x(t) = 0) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{1 - \cos(\theta)}\, d\theta. \tag{9.70}$$

   (g) To evaluate this integral please confirm that $\cos(\theta)$ lies above the chord $1 - 2\theta/\pi$ when $0 \le \theta \le \pi/2$ to conclude that $1 - \cos(\theta) \le (2/\pi)\theta$ there. Conclude from this that $\mathcal{P} = \infty$ and that the walker is therefore assured, by (f), of returning home.

8. Given a complex matrix $Z$ define its "real part" to be $X \equiv (Z + Z^*)/2$ and its "imaginary part" to be $Y = (Z - Z^*)/(2i)$. Recall that $Z^*$ denotes its conjugate transpose, (9.3) Show that $Z = X + iY$ and $X^* = X$ and $Y^* = Y$.

9. Suppose $Z \in \mathbb{C}^{n\times n}$. Show that if $\mathrm{tr}(ZH) = 0$ for every $H \in \mathbb{C}^{n\times n}$ for which $H = H^*$ then $Z = 0$. Hint: use the previous exercise to write $Z^* = X - iY$ and note that $ZZ^* = ZX - iZY$.

10. Use (9.59) to derive **Parseval's Theorem**

$$\|\hat{f}_N\|^2 = N\|f_N\|^2. \tag{9.71}$$

11. Suppose $N$ is even and use (9.60) and $w_N^2 = w_{N/2}$ to arrive at

$$\hat{f}_N(m) = \sum_{n=0}^{N/2-1} \overline{w}_{N/2}^{mn} f_N(2n) + \overline{w}_N^m \sum_{n=0}^{N/2-1} \overline{w}_{N/2}^{mn} f_N(2n+1), \quad m = 0, 1, \ldots, N-1.$$

From here use $w_N^{m+N/2} = -w_N^m$ to establish

$$\begin{aligned} \hat{f}_N(m) &= (F_{N/2} f_{N_e})(m) + \overline{w}_N^m (F_{N/2} f_{N_o})(m) \\ \hat{f}_N(m+N/2) &= (F_{N/2} f_{N_e})(m) - \overline{w}_N^m (F_{N/2} f_{N_o})(m) \end{aligned} \quad m = 0, 1, \ldots, N/2-1,$$

where $f_{N_e}$ denotes the even elements and $f_{N_o}$ denotes the odd elements of $f_N$. Compute the savings.

12. Lets build a multivariate generalization of Prop. 9.1. For $X_j \in \mathbb{R}^n$ and $A_p \in \mathbb{R}^{n\times n}$ define

$$Y_j = (A \star X)_j \equiv \sum_p A_p X_{j-p} \tag{9.72}$$

and establish

(i) $\hat{Y}(\omega) = \hat{A}(\omega)\hat{X}(\omega)$.

(ii) If $X$ is stationary then $C_{YY} = A \star E_\infty(C_{XX} \star A^T)$ and $\hat{C}_{YY}(\omega) = \hat{A}(\omega)\hat{C}_{XX}(\omega)\hat{A}^*(\omega)$.

13. Verify that $\sin z$ and $\cos z$ satisfy the Cauchy-Riemann equations (9.15) and use Prop. 9.2.1 to evaluate their derivatives.

14. Use the Cauchy–Riemann equations to conclude that the real and imaginary parts of a differentiable function are each **harmonic**. That is, if $f(x,y) = u(x,y) + iv(x,y)$ then

$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = 0 \quad \text{and} \quad \frac{\partial^2 v(x,y)}{\partial x^2} + \frac{\partial^2 v(x,y)}{\partial y^2} = 0$$

for the points $x + iy$ where $f$ is smooth. Confirm that the real and imaginary parts of $\sin z$ and $\cos z$ are indeed harmonic. (perhaps reverse these subparts)

15. We call $\Psi(x,t) = \exp(2\pi i(px - Et)/h)$ a wave in space $x$, and time $t$. We accordingly speak of $h/p$ as its wavelength and $E/h$ as its frequency. In Figure 9.6 we have plotted its real part over a region of space-time for particular values of $p$, $E$ and $h$.

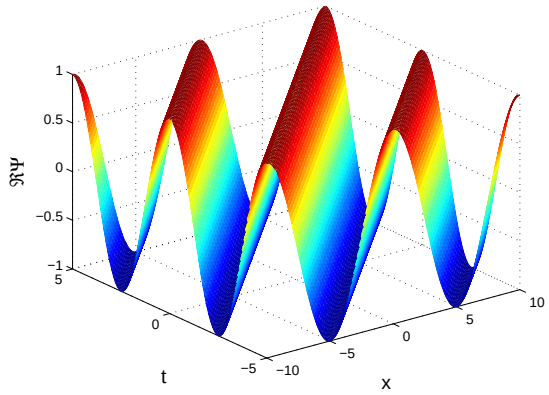



**Figure** 9.6 A plot of $\Re\Psi$ with $p = 1$, $E = 2$ and $h = 10$. `psiwave.m`

Show that $\Psi$ obeys **Schrödinger's equation**

$$-\frac{h^2}{8\pi^2 m}\frac{\partial^2\Psi(x,t)}{\partial x^2} + V\Psi(x,t) = \frac{ih}{2\pi}\frac{\partial\Psi(x,t)}{\partial t} \tag{9.73}$$

when $E = V + p^2/(2m)$. In this interpretation, $h$ is Planck's constant, $m$ is mass, $p$ is momentum, $V$ is potential energy and $E$ is total (kinetic plus potential) energy.

16. Submit a MATLAB diary documenting your calculation, via the `symbolic toolbox`, of the partial fraction expansion of the resolvent of

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

You should achieve

$$(sI - B)^{-1} = \frac{1}{s-(2+\sqrt{2})}\frac{1}{4}\begin{pmatrix} 1 & -\sqrt{2} & 1 \\ -\sqrt{2} & 2 & -\sqrt{2} \\ 1 & -\sqrt{2} & 1 \end{pmatrix} + \frac{1}{s-2}\frac{1}{2}\begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix} + \frac{1}{s-(2-\sqrt{2})}\frac{1}{4}\begin{pmatrix} 1 & \sqrt{2} & 1 \\ \sqrt{2} & 2 & \sqrt{2} \\ 1 & \sqrt{2} & 1 \end{pmatrix}.$$

17. If $A \neq 0$ and $|B|^2 > AC$ show that (9.27) is equivalent to $|z - z_0| = r$. Find $z_0$ and show that $r > 0$. Us this to show that $\mu(z) = 1/(z+2)$ takes the circle centered at zero of radius one to the circle centered at $2/3$ of radius $1/3$.

18. Consider a Möbius transformation, $\mu$, and its associated matrix $M$. Show that $z = [z_1; z_2]$ is an eigenvector of $M$ iff $z_1/z_2$ is a fixed point of $\mu$.

# 10. Complex Integration

Our goal here is to develop the main results of complex integration theory and to apply these to partial fraction expansion of the resolvent of a matrix and to the Inverse Laplace Transform of a rational function. With these tools and applications we will have placed the dynamics work in §8.2 on a solid foundation.

These tools will also permit us to solve the challenging Causal Wiener Filter problem set in the previous chapter, and to develop the tools needed for a careful study of the eigenvalue perturbation problem.

### 10.1. Cauchy's Theorem

We will 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,

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

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

$$\boxed{\int_C f(z)\, dz \equiv \int_{t_1}^{t_2} f(z(t))z'(t)\, dt.} \tag{10.1}$$

For example, if $C = \{t + it : 0 \leq t \leq 1\}$ 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 $\{\exp(it) : 0 \leq t \leq 2\pi\}$ then

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

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

$$\int_C z^{-1}\, dz = \int_0^{2\pi} \exp(-it)i\exp(it)\, dt = 2\pi i.$$

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

$$C(a, \rho) \equiv \{a + \rho\exp(it) : 0 \leq t \leq 2\pi\}, \tag{10.2}$$

the circle of radius $\rho$ centered at $a$. We find

$$\begin{aligned}\int_{C(a,\rho)} (z-a)^m\,dz &= \int_0^{2\pi} (a+\rho\exp(it)-a)^m \rho i\exp(it)\,dt \\ &= i\rho^{m+1}\int_0^{2\pi} \exp(i(m+1)t)\,dt \\ &= i\rho^{m+1}\int_0^{2\pi} \{\cos((m+1)t)+i\sin((m+1)t)\}\,dt \\ &= \begin{cases} 2\pi i & \text{if } m=-1, \\ 0 & \text{otherwise,}\end{cases}\end{aligned} \tag{10.3}$$

*regardless* of the size of $\rho$!

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

$$\int_C f(z)\,dz = \int_C f(x+iy)(dx+idy) = \int_C f(x+iy)\,dx + i\int_C f(x+iy)\,dy. \tag{10.4}$$

This representation is specially suited to one of the cornerstones of multivariable calculus. In particular, Green's Theorem permits us to express our contour integral, Eq. (10.4), as an especially convenient integral over space.

**Proposition** 10.1. **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

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

Proof: We suppose that $C$ encloses the lens shaped region like that of Figure 10.1 with the dual description

$$\begin{aligned} C_{in} &= \{(x,y) : y_1 \le y \le y_2,\ x_a(y) \le x \le x_b(y)\} \\ &= \{(x,y) : x_1 \le x \le x_2,\ y_a(x) \le y \le y_b(x)\}. \end{aligned}$$

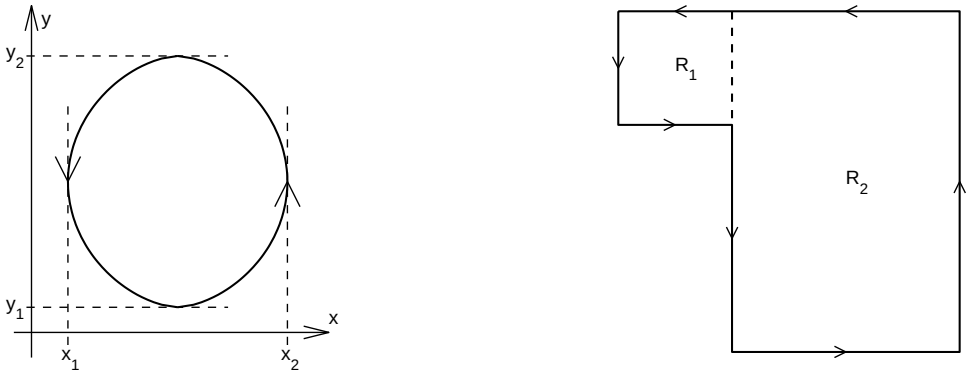



**Figure** 10.1. (A) A lens shaped region. (B) A union of lenses.

The advantage of such a region is that it permits us to express integrals over $(x, y)$ as sequential, or iterated, integrals over $x$ and $y$ individually. When our integrand is a partial derivative with respect to $x$ or $y$ we may then invoke the fundamental theorem of calculus, relegating the final integral to the boundary. More precisely

$$\begin{aligned}\iint_{C_{in}} \frac{\partial N(x,y)}{\partial x}\,dxdy &= \int_{y_1}^{y_2}\int_{x_a(y)}^{x_b(y)} \frac{\partial N(x,y)}{\partial x}\,dxdy \\ &= \int_{y_1}^{y_2} (N(x_b(y),y) - N(x_a(y),y))\,dy = \int_C N(x,y)\,dy\end{aligned}$$

where the second equality is a direct application of the fundamental theorem of calculus. To see the third equality use Figure 10.1 as a guide and note that integrating up along $x_b$ agrees with the orientation of $C$ while integrating up along $x_a$ is counter to $C$'s orientation, but as this is prefaced with a minus sign we arrive at integration of $N$ round the full contour. Similarly

$$\begin{aligned}\iint_{C_{in}} \frac{\partial M(x,y)}{\partial y}\,dxdy &= \int_{x_1}^{x_2}\int_{y_a(x)}^{y_b(x)} \frac{\partial M(x,y)}{\partial y}\,dydx \\ &= \int_{x_1}^{x_2} (M(x,y_b(x)) - M(x,y_a(y)))\,dy = -\int_C M(x,y)\,dx\end{aligned}$$

Regarding the final equality, note that integrating rightward along $y_b$ is counter to $C$'s orientation while integrating rightward along $y_a$ conforms to $C$'s orientation.

To complete the proof we return to a general smooth closed curve $C$ and express $C_{in}$ as a union of lens shaped regions. Take for example the region of Figure 10.1(B) where $C_{in} = R_1 \cup R_2$. Now

$$\begin{aligned}\iint_{C_{in}} \frac{\partial N(x,y)}{\partial x}\,dxdy &= \iint_{R_1} \frac{\partial N(x,y)}{\partial x}\,dxdy + \iint_{R_2} \frac{\partial N(x,y)}{\partial x}\,dxdy \\ &= \int_{C_1} N(x,y)\,dy + \int_{C_2} N(x,y)\,dy\end{aligned} \tag{10.5}$$

where $C_j$ is the counterclockwise contour enclosing $R_j$. As $C_1$ and $C_2$ have opposite orientations along their common (dashed) segment these contributions to Eq. (10.5) cancel and we find

$$\iint_{C_{in}} \frac{\partial N(x,y)}{\partial x}\,dxdy = \int_C N(x,y)\,dy.$$

The same argument applies to the $M$ integral. It also applies to all finite unions of lens shaped regions. End of Proof.

Applying this proposition to Eq. (10.4), 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** 10.2. **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 (10.3), 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 $m$th degree polynomial with $h$ distinct zeros at $\{\lambda_j\}_{j=1}^h$ with respective multiplicities of $\{p_j\}_{j=1}^h$ we found that

$$r(z) = \sum_{j=1}^{h}\sum_{k=1}^{p_j} \frac{r_{j,k}}{(z-\lambda_j)^k}. \tag{10.6}$$

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

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

In (10.3) we found that each, save the first, of the integrals under the sum is in fact zero. Hence

$$\int_{C_j} r(z)\,dz = 2\pi i r_{j,1}. \tag{10.7}$$

With $r_{j,1}$ in hand, say from (9.19) or `residue`, one may view (10.7) 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,

$$\int_{C_j} r(z)(z-\lambda_j)^{k-1}\,dz = 2\pi i r_{j,k}. \tag{10.8}$$

It is a simple, but highly important, matter to extend this representation to a matrix of rational functions. More precisely, if $R(z) \equiv (zI - B)^{-1}$ is the resolvent associated with $B$ then (10.6) and (10.8) state that

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

where

$$R_{j,k} = \frac{1}{2\pi i}\int_{C_j} R(z)(z-\lambda_j)^{k-1}\,dz. \tag{10.9}$$

Lets consider these in the concrete setting of critically damped single mass. The resolvent in that case, recall (9.20), can then be expressed

$$(sI-B)^{-1} = \frac{1}{(s+1)^2}\begin{pmatrix} s+2 & 1 \\ -1 & s \end{pmatrix} = \frac{R_{1,1}}{s+1} + \frac{R_{1,2}}{(s+1)^2} \tag{10.10}$$

where

$$R_{1,1} = \frac{1}{2\pi i}\begin{pmatrix} \int_{C_1} \frac{z+2}{(z+1)^2}\,dz & \int_{C_1} \frac{1}{(z+1)^2}\,dz \\ \int_{C_1} \frac{-1}{(z+1)^2}\,dz & \int_{C_1} \frac{z}{(z+1)^2}\,dz \end{pmatrix} \quad \text{and} \quad R_{1,2} = \frac{1}{2\pi i}\begin{pmatrix} \int_{C_1} \frac{z+2}{z+1}\,dz & \int_{C_1} \frac{1}{z+1}\,dz \\ \int_{C_1} \frac{-1}{z+1}\,dz & \int_{C_1} \frac{z}{z+1}\,dz \end{pmatrix} \tag{10.11}$$

and $C_1$ encloses the pole $z = -1$. The off-diagonal terms of both matrices may be computed directly from our bare–handed result, (10.3). Evaluation of the diagonal terms will follow from the theory built in the next section.

### 10.2. The Second Residue Theorem

After (10.7) and (10.8) 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** 10.3. 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.

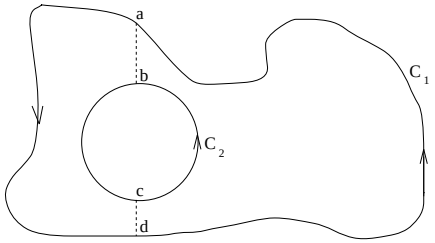



**Figure** 10.2. The Curve Replacement Lemma.

As an example, recalling (10.6) and (10.7), we may express the integral of a rational function around a curve that encircles **all** of its poles as a sum of residues

$$\int_C r(z)\,dz = \sum_{j=1}^{h}\sum_{k=1}^{p_j}\int_{C_j} \frac{r_{j,k}}{(z-\lambda_j)^k}\,dz = 2\pi i \sum_{j=1}^{h} r_{j,1}. \tag{10.12}$$

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 $\rho$ and so proceeds to let $\rho \to 0$, 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 $\rho$ 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 (in fact it tends to $f'(a)$) as the region of integration shrinks to a point, the integral must tend to zero. We have just proven

**Proposition** 10.4. **Cauchy's Integral Formula.** If $f$ is differentiable on and in the closed curve $C$ then

$$f(a) = \frac{1}{2\pi i}\int_C \frac{f(z)}{z-a}\,dz \tag{10.13}$$

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 of (10.13) is a perfectly smooth function of $a$. Hence, differentiating each side, we find

$$f'(a) = \frac{df(a)}{da} = \frac{1}{2\pi i}\int_C \frac{d}{da}\frac{f(z)}{z-a}\,dz = \frac{1}{2\pi i}\int_C \frac{f(z)}{(z-a)^2}\,dz \tag{10.14}$$

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

$$\frac{d^n f}{da^n}(a) = \frac{n!}{2\pi i}\int_C \frac{f(z)}{(z-a)^{1+n}}\,dz \tag{10.15}$$

for each $a$ lying inside $C$. The upshot is that once $f$ is shown to be differentiable it must, in fact, be infinitely differentiable. Regarding concrete examples, the diagonal terms in (10.11), where $C_1$ is a circle centered at $z=-1$, may be evaluated by (10.13) and (10.15) respectively. In particular,

$$\frac{1}{2\pi i}\int_C \frac{s}{s+1} = -1 \quad \text{and} \quad \frac{1}{2\pi i}\int_C \frac{s}{(s+1)} = 1. \tag{10.16}$$

On substiution into (10.11) we find

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

which on substitution in (10.10) delivers

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

in complete agreement with (9.21).

As a second example let us consider

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

where $f$ is differentiable on and in $C$ and $C$ encircles both $\lambda_1$ and $\lambda_2$. 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 $\lambda_j$ now lies in only $C_j$. As $f(z)/(z-\lambda_2)$ is well behaved in $C_1$ we may use (10.13) 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 (10.14) to conclude that

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

These calculations can be read as a concrete instance of

**Proposition** 10.5. **The Second Residue Theorem.** If $g$ is a polynomial with roots $\{\lambda_j\}_{j=1}^h$ of degree $\{p_j\}_{j=1}^h$ and $C$ is a closed curve encircling each of the $\lambda_j$ and $f$ is differentiable on and in $C$ then

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

where

$$\operatorname{res}(f/g, \lambda_j) = \lim_{z\to\lambda_j} \frac{1}{(p_j-1)!}\frac{d^{p_j-1}}{dz^{p_j-1}}\left((z-\lambda_j)^{p_j}\frac{f(z)}{g(z)}\right) \tag{10.17}$$

is called the **residue** of $f/g$ at $\lambda_j$ by extension of (9.19).

The generality of this statement, and the notation required to specify the residue, should not obscure the fact that it permits us to compute important integrals by merely *evaluating the good parts at the bad places.* The bad places are of course the poles of the integrand and the good part of the integrand is what remains after multiplying by the offending factor, and perhaps taking a few derivatives. For example, the integrands in

$$\frac{1}{2\pi i}\int_{C(0,1)} \frac{\exp(z)}{z^p}\,dz, \quad p = 1, 2, \ldots$$

have $z = 0$ ia a pole of order $p$ and so we must take $p-1$ derivatives of the good part, $\exp(z)$, at $z = 0$, and finally divide by $(p-1)!$. Hence,

$$\frac{1}{2\pi i}\int_{C(0,1)} \frac{\exp(z)}{z^p}\,dz = \frac{1}{(p-1)!}.$$

One of the most useful applications of the Second Residue Theorem is the formula for the inverse Laplace transform of a rational function.

### 10.3. The Inverse Laplace Transform and Return to Dynamics

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

$$\boxed{(\mathcal{L}^{-1}r)(t) \equiv \frac{1}{2\pi i}\int_C r(z)\exp(zt)\,dz} \tag{10.18}$$

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

$$(\mathcal{L}^{-1}r)(t) = \sum_{j=1}^{h} \operatorname{res}(r(z)\exp(zt), \lambda_j). \tag{10.19}$$

Let us put this lovely formula to the test. We take our examples from dynamical systems of Chapter 8. According to (10.19) the inverse Laplace Transform of

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

is simply the residue of $r(z)\exp(zt)$ at $z = -1$, i.e.,

$$\operatorname{res}(r(z)\exp(zt), -1) = \lim_{z\to -1}\frac{d}{dz}\exp(zt) = t\exp(-t).$$

This closes the circle on the example begun in §8.3 and continued in Exer. 8.1.

For our next example we return to (10.7) and take the inverse Laplace transform of the constituents of the resolvent

$$(zI - B)^{-1} = \frac{1}{z^2+1}\begin{pmatrix} z & 1 \\ -1 & z\end{pmatrix} \quad \text{of} \quad B = \begin{pmatrix} 0 & 1 \\ -1 & 0\end{pmatrix}.$$

In particular, we compute

$$\mathcal{L}^{-1}\frac{z}{z^2+1} = \left.\frac{z\exp(zt)(z+i)}{z^2+1}\right|_{z=-i} + \left.\frac{z\exp(zt)(z-i)}{z^2+1}\right|_{z=i} = \exp(-it)/2 + \exp(it/2) = \cos(t)$$

and

$$\mathcal{L}^{-1}\frac{1}{z^2+1} = \left.\frac{\exp(zt)(z+i)}{z^2+1}\right|_{z=-i} + \left.\frac{\exp(zt)(z-i)}{z^2+1}\right|_{z=i} = i\exp(-it)/2 - i\exp(it)/2 = \sin(t)$$

and recall the mysterious (8.28) en route to the lovely

$$\exp(Bt) = \mathcal{L}^{-1}(sI - B)^{-1} = \begin{pmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t)\end{pmatrix}.$$

You may wish to confirm that this matrix indeed obeys $(\exp(Bt))' = B\exp(Bt)$.

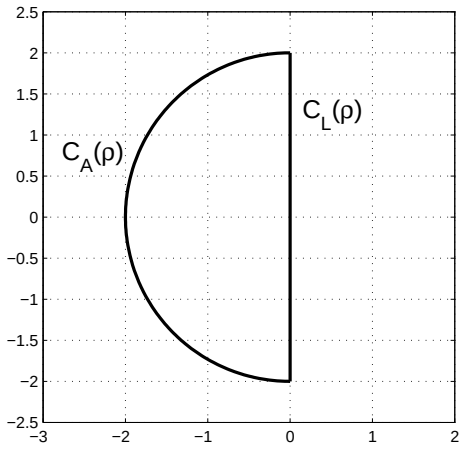



**Figure** 10.3 An illustration of the contour in (10.23) when $\rho = 2$.

The curve replacement lemma of course gives us considerable freedom in our choice of the curve $C$ used to define the inverse Laplace transform, (10.18). 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, see Figure 10.3,

$$C = C_L(\rho) \cup C_A(\rho), \tag{10.20}$$

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

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

where $\rho$ is chosen large enough to encircle the poles of $r$. With this concrete choice, (10.18) takes the form

$$\begin{aligned}(\mathcal{L}^{-1}r)(t) &= \frac{1}{2\pi i}\int_{C_L} r(z)\exp(zt)\,dz + \frac{1}{2\pi i}\int_{C_A} r(z)\exp(zt)\,dz \\ &= \frac{1}{2\pi}\int_{-\rho}^{\rho} r(i\omega)\exp(i\omega t)\,d\omega + \frac{\rho}{2\pi}\int_{\pi/2}^{3\pi/2} r(\rho\exp(i\theta))\exp(\rho\exp(i\theta)t)\exp(i\theta)\,d\theta.\end{aligned} \tag{10.21}$$

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

$$\boxed{(\mathcal{L}^{-1}r)(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} r(i\omega)\exp(i\omega t)\,d\omega,} \tag{10.22}$$

the conventional definition of the inverse Laplace transform.

### 10.4. The Inverse Fourier Transform and the Causal Wiener Filter*

Our interest is conditions on $f$ that assure causality, i.e.,

$$\int_0^1 f(\omega)\exp(2k\pi i\omega)\,d\omega = 0 \qquad \forall\ k = -1, -2, \ldots. \tag{10.23}$$

Suppose $f$ is a rational function of $\exp(2\pi i\omega)$, i.e.,

$$f(\omega) = \frac{p(\exp(2\pi i\omega))}{q(\exp(2\pi i\omega))}$$

where $p$ and $q$ are polynomials. The change of variable $z = \exp(2\pi i\omega)$ will reduce (10.23) to a residue problem. In particular, as $dz = 2\pi i z d\omega$ we find

$$\begin{aligned}\int_0^1 \frac{p(\exp(2\pi i\omega))}{q(\exp(2\pi i\omega))} \exp(-2k\pi i\omega)\, d\omega &= \frac{1}{2\pi i}\int_{C_1} \frac{p(z)}{q(z)} z^{-k-1}\, dz \\ &= \sum_{j=1}^{h}\sum_{m=1}^{m_j} \frac{r_{j,m}}{2\pi i}\int_{C_1} \frac{dz}{(z-\lambda_j)^m z^{k+1}}\end{aligned}$$

where the last equality follows from the partial fraction expansion of $p/q$. We now note that if $|\lambda_j| < 1$ then the residues at $z = 0$ and $z = \lambda_j$ are equal and opposite

$$\begin{aligned}\frac{1}{2\pi i}\int_{C_1} (z-\lambda)^{-m} z^{-k-1}\, dz &= \frac{1}{(m-1)!}\frac{d^{m-1}}{dz^{m-1}} z^{-k-1}\Big|_{z=\lambda} + \frac{1}{k!}\frac{d^k}{dz^k}(z-\lambda)^{-m}\Big|_{z=0} \\ &= \frac{(m-1+k)!}{(m-1)!k!}(-1)^{m-1} z^{-m-k}\Big|_{z=\lambda} + \frac{(m-1+k)!}{k!(m-1)!}(-1)^k (z-\lambda)^{-m-k}\Big|_{z=0} \\ &= \frac{(m-1+k)!}{k!(m-1)!}\{(-1)^{m-1} + (-1)^m\}\lambda^{-m-k} = 0.\end{aligned}$$

This provides the clue to revising the IIR filter (9.69) $\hat{a}(\omega) = \hat{c}_{xy}(\omega)/\hat{c}_{xx}(\omega)$. The obvious guess of simply keeping the good part is too naive. Instead we establish the

**Proposition** 10.6. If $f(\omega) > 0$ and integrable and is a rational function of $\exp(2\pi i\omega)$ then $f$ may be factored as

$$f(\omega) = L(\exp(2\pi i\omega))\overline{L(\exp(2\pi i\omega))} \tag{10.24}$$

where all poles of $L$ lie strictly inside the unit disk.

Proof: To begin

$$f(\omega) = c\frac{\prod_{j=1}^{n}(\exp(2\pi i\omega) - \mu_j)}{\prod_{j=1}^{d}(\exp(2\pi i\omega) - \lambda_j)}$$

where $c \neq 0$ and no $\mu_j$ nor $\lambda_j$ is zero. Also, no $\lambda_j$ has magnitude 1, for this would make $f$ nonintegrable. And no $\mu_j$ has magnitude 1, for $f$ never vanishes.

As $f$ is real we equate conjugates and find

$$\begin{aligned}c\frac{\prod_{j=1}^{n}(\exp(2\pi i\omega) - \mu_j)}{\prod_{j=1}^{d}(\exp(2\pi i\omega) - \lambda_j)} &= \overline{c}\frac{\prod_{j=1}^{n}(\exp(-2\pi i\omega) - \overline{\mu}_j)}{\prod_{j=1}^{d}(\exp(-2\pi i\omega) - \overline{\lambda}_j)} \\ &= \overline{c}\exp(2\pi i(d-n)\omega)\frac{\prod_{j=1}^{n}(1/\overline{\mu}_j - \exp(2\pi i\omega))\overline{\mu}_j}{\prod_{j=1}^{d}(1/\overline{\lambda}_j - \exp(2\pi i\omega))\overline{\lambda}_j}\end{aligned}$$

and so for pole $\lambda_j$ we find that $1/\overline{\lambda}_j$ is also a pole. As such we may partition the poles into those lying within the unit circle $\{\lambda_j^+ : j = 1, \ldots, d/2\}$ and their reflections across the unit circle

$\{1/\overline{\lambda}_j^+ : j = 1, \ldots, d/2\}$. Similar, we partition the zeros into those lying within the unit circle $\{\mu_j^+ : j = 1, \ldots, n/2\}$ and their reflections across the unit circle $\{1/\overline{\mu}_j^+ : j = 1, \ldots, n/2\}$. With this we may define

$$L(\omega) \equiv \frac{\prod_{j=1}^{n/2}(\exp(2\pi i\omega) - \mu_j^+)}{\prod_{j=1}^{d/2}(\exp(2\pi i\omega) - \lambda_j^+)},$$

note that its zeros and poles lie strictly inside the unit circle and compute

$$\begin{aligned} L(\omega)\overline{L(\omega)} &= \frac{\prod_{j=1}^{n/2}(\exp(2\pi i\omega) - \mu_j^+)(\exp(-2\pi i\omega) - \overline{\mu}_j^+)}{\prod_{j=1}^{d/2}(\exp(2\pi i\omega) - \lambda_j^+)(\exp(-2\pi i\omega) - \overline{\lambda}_j^+)} \\ &= \exp(\pi i(d-n)\omega)\frac{\prod_{j=1}^{n/2}(\exp(2\pi i\omega) - \mu_j^+)(1/\overline{\mu}_j^+ - \exp(2\pi i\omega))\overline{\mu}_j^+}{\prod_{j=1}^{d/2}(\exp(2\pi i\omega) - \lambda_j^+)(1/\overline{\lambda}_j^+ - \exp(2\pi i\omega))\overline{\lambda}_j^+} \\ &= \exp(\pi i(d-n)\omega)(-1)^{(n-d)/2}\frac{\prod_{j=1}^{n/2}\mu_j^+}{\prod_{j=1}^{d/2}\lambda_j^+}\frac{f(\omega)}{c}. \end{aligned}$$

As both the left hand side and $f$ are nonnegative we can take magnitudes of both sides and conclude that

$$L(\omega)\overline{L(\omega)} = \rho f(\omega) \quad \text{where} \quad \rho = \left|\frac{\prod_{j=1}^{n/2}\mu_j^+}{c\prod_{j=1}^{d/2}\lambda_j^+}\right|$$

On setting $L = L/\sqrt{\rho}$ we conclude (10.24). End of Proof.

We assume that $c_{xx}$ is rational and exploit its factorization in the the associated error spectrum, recall (9.68), takes the form

$$\begin{aligned} \hat{c}_{\varepsilon\varepsilon}(\omega) &= \hat{c}_{yy}(\omega) - \overline{\hat{a}}(\omega)\hat{c}_{xy}(\omega) - \hat{a}(\omega)\overline{\hat{c}}_{xy}(\omega) + |\hat{a}(\omega)|^2 L(\exp(2\pi i\omega))\overline{L(\exp(2\pi i\omega))} \\ &= \hat{c}_{yy}(\omega) + |\hat{a}(\omega)L(\exp(2\pi i\omega)) - \hat{c}_{xy}/\overline{L(\exp(2\pi i\omega))}|^2 - |\hat{c}_{xy}(\omega)|^2/c_{xx}(\omega) \end{aligned}$$

Now, as we wish to determine the best causal $a$ we expect $\hat{a}$ to have all of its poles within $C_1$. As the poles of $L$ are also there then the best we can do is to match the causal part of scaled cross-spectrum. More precisely, we suppose that $\hat{c}_{xy}$ is also rational then use partial fractions to write

$$\frac{\hat{c}_{xy}}{\overline{L(\exp(2\pi i\omega))}} = \left\{\frac{\hat{c}_{xy}}{\overline{L(\exp(2\pi i\omega))}}\right\}_+ + \left\{\frac{\hat{c}_{xy}}{\overline{L(\exp(2\pi i\omega))}}\right\}_-$$

as a sum of terms with poles within and without (respectively) of $C_1$. With this notation we may solve the causal Wiener filter for rational spectra via

$$\hat{a}(\omega) = \frac{1}{L(\exp(2\pi i\omega))}\left\{\frac{\hat{c}_{xy}}{\overline{L(\exp(2\pi i\omega))}}\right\}_+. \tag{10.25}$$

### 10.5. Further Applications of the Second Residue Theorem*

We show how to structure an integrand so that the Second Residue Theorem returns the number of zeros of a pair of functions within a chosen region. This in turn leads to a beautiful comparison

theorem that permits us to equate the number of zeros of a "hard" function with that of an "easy" function. This in turn permits us to prove the Fundamental Theorem of Algebra, i.e., the statement that every $n$th order polynomial has $n$ zeros. And finally, by a very similar argument we prove a useful perturbation result. Namely, we show that the zeros of a polynomial don't move drastically when one perturbs the polynomial coefficients. We begin with the zero counter.

**Proposition** 10.7. **The Argument Principle.** If $r = f/g$ is the ratio of two differentiable functions on and in the simple closed curve $C$ and neither $f$ nor $g$ have zeros on $C$ then

$$\frac{1}{2\pi i}\int_C \frac{r'(z)}{r(z)}\,dz = Z(f,C) - Z(g,C), \tag{10.26}$$

where $Z(f,C)$ is the number (counting multiplicity) of zeros of $f$ in $C$.

Proof: From $r(z) = f(z)/g(z)$ comes

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

and so each pole of $r'/r$ is a zero of either $f$ or $g$. We take these up separately.

If $\lambda$ is a zero of $f$ of order $k$ then $r(z) = (z-\lambda)^k q(z)$ where $q(\lambda) \neq 0$. It follows that $r'(z) = k(z-\lambda)^{k-1}q(z) + (z-\lambda)^k q'(z)$ and so

$$\frac{r'(z)}{r(z)} = \frac{k}{z-\lambda} + \frac{q'(z)}{q(z)}.$$

As the latter term is well behaved at $z = \lambda$ it follows that $\mathrm{res}(r'/r, \lambda) = k$.

If $\mu$ is a zero of $g$ of order $m$ then $r(z) = (z-\mu)^{-m}p(z)$ where $p(\mu) \neq 0$. It follows that $r'(z) = -m(z-\mu)^{-m-1}p(z) + (z-\mu)^{-m}p'(z)$ and so

$$\frac{r'(z)}{r(z)} = \frac{-m}{z-\mu} + \frac{p'(z)}{p(z)}.$$

As the latter term is well behaved at $z = \mu$ it follows that $\mathrm{res}(r'/r, \mu) = -m$.

Combining these two residue calculations, the Second Residue Theorem delivers (10.26). End of Proof.

From this we establish the very useful comparison principle.

**Proposition** 10.8. **Rouché's Theorem.** If $f$ and $g$ are two complex differentiable functions on and in the simple closed curve $C$ and

$$|f(z) - g(z)| < |g(z)| \qquad \forall\ z \in C \tag{10.27}$$

then $f$ and $g$ have the same number of zeros in $C$.

Proof: We define $r \equiv f/g$ and deduce from (10.27) that $r$ has neither zeros nor poles on $C$. As such we may read from the Argument Principle that

$$\frac{1}{2\pi i}\int_C \frac{r'(z)}{r(z)}\,dz = Z(f,C) - Z(g,C),$$

and so it remains to show that this integral is zero. To wit, we note that

$$F(t) = \frac{1}{2\pi i}\int_C \frac{r'(z)}{r(z)+t}\,dz = Z(f+tg, C) - Z(g, C)$$

is both integer valued and continuous in $t$. This implies that $F(t)$ is constant. This constant reveals itself on noting that

$$|F(t)| \leq \frac{|C|}{2\pi}\frac{\max\{|r'(z)| : z \in C\}}{t - \max\{|r(z)| : z \in C\}}.$$

This implies that $F(t) \to 0$ as $t \to \infty$. Hence, $0 = F(0) = Z(f, C) - Z(g, C)$ as claimed. End of Proof.

This has many subtle applications, perhaps the simplest being

**Proposition** 10.9. **Fundamental Theorem of Algebra.** If $f(z)$ is a polynomial of degree $n$ the $f$ has precisely $n$ zeros.

Proof: From $f(z) = f_0 + f_1 z + \cdots + f_n z^n$ we construct $g(z) = f_n z^n$ and $C = C(0, R)$ where

$$R = 1 + \frac{1}{|f_n|}\sum_{j=0}^{n-1} |f_j|.$$

It now follows from the triangle inequality that for $z \in C$,

$$|f(z) - g(z)| = |f_0 + f_1 z + \cdots + f_{n-1}z^{n-1}| \leq R^{n-1}\sum_{j=0}^{n-1}|f_j| < |f_n|R^n = |f_n z^n| = |g(z)|.$$

It now follows from Rouché's Theorem that $f$ and $g$ have the same number of zeros in $C$. As $g$ has $n$ zeros there then so too does $f$. End of Proof.

By a very similar argument we can show that the roots of polynomial are continuous functions of its coefficients. We will use this result to build a quantitative perturbation theory in Chapter 12.

**Proposition** 10.10. Suppose $f(\varepsilon, z) = f_0(\varepsilon) + f_1(\varepsilon)z + f_2(\varepsilon)z^2 + \cdots + f_n(\varepsilon)z^n$ where each $f_j$ is a continuous complex function of the complex parameter $\varepsilon$ in some ball about $\varepsilon = 0$. If $\lambda$ is a zero of order $k$ of $z \mapsto f(0, z)$ then there exists an $\rho > 0$ and $\varepsilon_0 > 0$ such that $z \mapsto f(\varepsilon, z)$ has precisely $k$ zeros in $C(\lambda, \rho)$ when $|\varepsilon| < \varepsilon_0$.

Proof: Pick $\rho > 0$ so that no other zeros of $z \mapsto f(0, z)$ lie in $C = C(\lambda, \rho)$ and record

$$F = \min_{z\in C}|f(0, z)| \quad \text{and} \quad R = 1 + \max_{z\in C}|z|.$$

As each $f_j$ is continuous we may choose an $\varepsilon_0$ small enough to guarantee that

$$\max_j |f_j(\varepsilon) - f_j(0)| < \frac{F}{(n+1)R^n}, \qquad \forall\ |\varepsilon| < \varepsilon_0.$$

This permits us to establish the bound

$$|f(\varepsilon, z) - f(0, z)| \leq \sum_{j=0}^{n} |f_j(\varepsilon) - f_j(0)||z^j| < F \leq |f(0, z)|,$$

and so conclude from Rouché's Theorem that $z \mapsto f(\varepsilon, z)$ has the same number of zeros in $C$ as $z \mapsto f(0, z)$ for each $|\varepsilon| < \varepsilon_0$. End of Proof.

### 10.6. Notes and Exercises

We have followed Levinson and Redheffer (1970) throughout, except for Doob (1990) for the Wiener Filter.

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 (10.3), 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. Choose $C$ in the Cauchy Integral Formula to be the circle of radius $\rho$ centered about $a$ defined in (10.2) and deduce from (10.13) the beautiful Mean Value Theorem

$$f(a) = \frac{1}{2\pi}\int_0^{2\pi} f(a + \rho\exp(it))\,dt. \tag{10.28}$$

   Confirm this for $f(z) = z^m$, where $m$ is a positive integer, by computing both sides of (10.28) by hand for arbitrary $a$ and $\rho$.

4. Use (10.19) 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

$$x'(t) + x(t) = \exp(-t)\sin t, \qquad x(0) = 0.$$

   Hint: Take the Laplace transform of each side.

6. Evaluate all expressions in (10.20) in MATLAB's symbolic toolbox via `syms`, `diff` and `subs` and confirm that the final result jibes with (***.

7. Let us check the limit we declared in going from (10.21) to (10.22). First show that

$$|\exp(\rho\exp(i\theta)t)| = \exp(\rho t\cos\theta).$$

   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

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

and finally argue why

$$\max_\theta|r(\rho\exp(i\theta))|\to 0$$

as $\rho\to\infty$.

# 11. The Eigenvalue Problem

Eigenvalues appeared naturally in our discussion of dynamics in Chapter 8. After a two–chapter tour of the necessary tools from Complex Analysis we are now prepared to fully understand eigenvalues, and therefore dynamics.

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

$$B = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} \tag{11.1}$$

then the Gauss–Jordan method delivers

$$(sI - B)^{-1} = \frac{1}{(s-1)^2(s-2)} \begin{pmatrix} (s-1)(s-2) & s-2 & 0 \\ 0 & (s-1)(s-2) & 0 \\ 0 & 0 & (s-1)^2 \end{pmatrix} \tag{11.2}$$

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 $j$th **eigenspace** and call each of its nonzero members a $j$th **eigenvector**. With respect to the $B$ in (11.1) we note that

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{11.3}$$

respectively span $\mathcal{N}(\lambda_1 I - B)$ and $\mathcal{N}(\lambda_2 I - B)$. That $B \in \mathbb{R}^{3\times 3}$ but possesses only 2 linearly independent eigenvectors 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 $\lambda_1$ 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. We achieve that in the first two sections, arriving at the **Spectral Representation**, the most significant and general Proposition of the second third of this text.

In the subsequent sections we provide alternate expressions, with examples, of this general result in a number of important instances. This yields the Schur and Jordan Forms and several points of view on the matrix exponential and permits the presentation and solution of the problem of ranking web pages.

### 11.1. 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

$$\frac{1}{s-b}. \tag{11.4}$$

This function is a scaled version of the even simpler function $1/(1-z)$. This latter function satisfies (recall the $n$-term geometric series, Exer. 9.3)

$$\frac{1}{1-z} = 1 + z + z^2 + \cdots + z^{n-1} + \frac{z^n}{1-z} \tag{11.5}$$

for each positive integer $n$. Furthermore, if $|z| < 1$ then $z^n \to 0$ as $n \to \infty$ and so (11.5) becomes, in the limit,

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

the full geometric series. Returning to (11.4) 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,

$$(sI-B)^{-1} = s^{-1}(I-B/s)^{-1} = \frac{1}{s} + \frac{B}{s^2} + \cdots + \frac{B^{n-1}}{s^n} + \frac{B^n}{s^n}(sI-B)^{-1}, \tag{11.6}$$

and hence, so long as $s$ is larger than any element of $B$, e.g., if $|s| > \|B\|_F$ where the latter is defined in (1.17), we find

$$(sI-B)^{-1} = s^{-1}\sum_{n=0}^{\infty}(B/s)^n. \tag{11.7}$$

Although (11.7) is indeed a formula for the resolvent you may, regarding computation, not find it any more attractive than the Gauss-Jordan method. We view (11.7) however as an analytical rather than computational tool. More precisely, it facilitates the computation of integrals of the resolvent. For example, if $C_\rho$ is the circle of radius $\rho$ centered at the origin and $\rho > \|B\|$ then

$$\int_{C_\rho}(sI-B)^{-1}\,ds = \sum_{n=0}^{\infty} B^n \int_{C_\rho} s^{-1-n}\,ds = 2\pi i I. \tag{11.8}$$

Lets check this on the concrete resolvent in (11.2). For $\rho > 2$ we find indeed that

$$\frac{1}{2\pi i}\int_{C_\rho}(sI-B)^{-1}\,ds = \frac{1}{2\pi i}\begin{pmatrix} \int_{C_\rho}\frac{ds}{s-1} & \int_{C_\rho}\frac{ds}{(s-1)^2} & 0 \\ 0 & \int_{C_\rho}\frac{ds}{s-1} & 0 \\ 0 & 0 & \int_{C_\rho}\frac{ds}{s-2}\end{pmatrix} = \begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}. \tag{11.9}$$

This result is essential to our study of the eigenvalue problem. As are the two **resolvent identities**. Regarding the first, with $R(s) \equiv (sI-B)^{-1}$, 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

$$\boxed{R(s_2) - R(s_1) = (s_1-s_2)R(s_2)R(s_1).} \tag{11.10}$$

The second identity is simply a rewriting of

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

namely,

$$\boxed{BR(s) = R(s)B = sR(s) - I.} \tag{11.11}$$

You may wish to confirm that our concrete resolvent, (11.2), indeed obeys these identities.

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 $\{\mu_j\}_{j=1}^h$, and recall that (10.9) produced

$$R(s) = \sum_{j=1}^{h} \sum_{k=1}^{\mu_j} \frac{R_{j,k}}{(s-\lambda_j)^k} \quad \text{where} \quad R_{j,k} = \frac{1}{2\pi i} \int_{C_j} R(z)(z-\lambda_j)^{k-1}\, dz. \tag{11.12}$$

In the previous 2 chapters we took a fairly pedestrian, element-wise approach and evaluated these as matrices of integrals. For example, the resolvent in (11.2) has coefficients

$$R_{1,1} = \frac{1}{2\pi i} \begin{pmatrix} \int_{C_1} \frac{ds}{s-1} & \int_{C_1} \frac{ds}{(s-1)^2} & 0 \\ 0 & \int_{C_1} \frac{ds}{s-1} & 0 \\ 0 & 0 & \int_{C_1} \frac{ds}{s-2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{11.13}$$

and similarly

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

As noted already, these matrices enjoy some amazing properties, e.g.,

$$R_{1,1}^2 = R_{1,1}, \quad R_{2,1}^2 = R_{2,1}, \quad R_{1,1}R_{2,1} = 0, \quad \text{and} \quad R_{1,2}^2 = 0. \tag{11.14}$$

To establish that such structure inheres to the coefficients in the partial fraction expansion of the resolvent of *every* matrix we must step back from an element-wise focus on the $R_{j,k}$ and instead view them as integrals of matrices. To begin, lets show that each $R_{j,1}$ is a projection.

**Proposition** 11.1. $R_{j,1}^2 = R_{j,1}$.

Proof: Recall that the $C_j$ appearing in (11.12) is any circle about $\lambda_j$ that neither touches nor encircles any other root. Suppose, as in Figure 11.1, that $C_j$ and $C_j'$ are two such circles and $C_j'$ encloses $C_j$. Now, by the curve replacement lemma,

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

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, (11.10), in moving from the second to the third line. In moving from the fourth to the fifth we used only

$$\int_{C_j'}\frac{1}{w-z}\,dw = 2\pi i \quad \text{and} \quad \int_{C_j}\frac{1}{w-z}\,dz = 0. \tag{11.15}$$

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

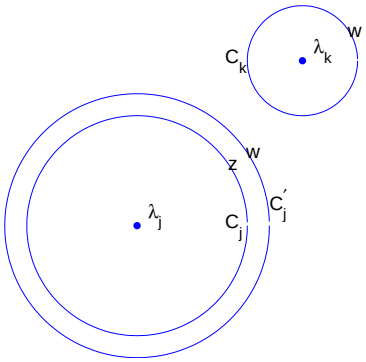



**Figure** 11.1. The curves that figure in Propositions 11.1–11.3.

Recalling Definition 6.1 that matrices that equal their squares are projections we adopt the abbreviation

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

With respect to the proof that $P_jP_k = 0$ when $j \neq k$, the calculation runs along the same lines. The difference comes in (11.15) where, regarding Figure 11.1, as $C_j$ lies completely outside of $C_k$, *both* integrals are zero. Hence,

**Proposition** 11.2. If $j \neq k$ then $P_jP_k = 0$.

Along the same lines we define

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

and prove that its lower powers indeed coincide with the subsequent $R_{j,k}$ and that its higher powers vanish entirely.

**Proposition** 11.3. If $1 \le k \le \mu_j - 1$ then $D_j^k = R_{j,k+1}$. $D_j^{\mu_j} = 0$.

Proof: For $k$ and $\ell$ greater than or equal to one and $C_j$ and $C_j'$ as in Figure 11.1,

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

because

$$\int_{C_j'} \frac{(w-\lambda_j)^\ell}{w-z}\,dw = 2\pi i(z-\lambda_j)^\ell \quad \text{and} \quad \int_{C_j} \frac{(z-\lambda_j)^k}{w-z}\,dz = 0. \tag{11.16}$$

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 $k = \mu_j$ we find

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

by Cauchy's Theorem. End of Proof.

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

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

for the integrand then has a term proportional to $1/(z-\lambda_j)$, which we know, by (10.3), leaves a nonzero residue. When some power of a matrix vanishes we call the matrix **nilpotent**.

With this we have now arrived at a much richer specification of the generic expansion (11.12), namely

$$\boxed{R(z) = \sum_{j=1}^{h}\left\{\frac{1}{z-\lambda_j}P_j + \sum_{k=1}^{\mu_j-1}\frac{1}{(z-\lambda_j)^{k+1}}D_j^k\right\},} \tag{11.17}$$

along with verification of a number of the properties enjoyed by the **eigenprojections**, $P_j$, and **eigennilpotents**, $D_j$.

### 11.2. 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, (11.11), to $P_j$. More precisely, we note that (11.11) implies that

$$\begin{aligned} BP_j = P_jB &= \frac{1}{2\pi i}\int_{C_j} (zR(z) - I)\,dz \\ &= \frac{1}{2\pi i}\int_{C_j} zR(z)\,dz \\ &= \frac{1}{2\pi i}\int_{C_j} R(z)(z-\lambda_j)\,dz + \frac{\lambda_j}{2\pi i}\int_{C_j} R(z)\,dz \\ &= D_j + \lambda_j P_j, \end{aligned} \tag{11.18}$$

where the second equality is due to Cauchy's Theorem and the third arises from adding and subtracting $\lambda_j R(z)$. Summing (11.18) over $j$ we find

$$B\sum_{j=1}^{h} P_j = \sum_{j=1}^{h} \lambda_j P_j + \sum_{j=1}^{h} D_j. \tag{11.19}$$

We can go one step further, namely the evaluation of the first sum. This stems from (11.8) where we integrated $R(s)$ over a circle $C_\rho$ 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 (11.8) we find

$$\sum_{j=1}^{h} P_j = I, \tag{11.20}$$

and so (11.19) takes the form

$$\boxed{B = \sum_{j=1}^{h} \lambda_j P_j + \sum_{j=1}^{h} D_j.} \tag{11.21}$$

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 (11.18) as

$$(B - \lambda_j I)P_j = D_j \tag{11.22}$$

and then raise each side to the $k$th power to arrive at

$$(B - \lambda_j I)^k P_j = D_j^k, \tag{11.23}$$

where we've used the fact that $P_j^2 = P_j$ and $BP_j = P_jB$. With $k = \mu_j$ in (11.23) we arrive at the lovely

$$\boxed{(B - \lambda_j I)^{\mu_j} P_j = 0.} \tag{11.24}$$

For this reason we call the range of $P_j$ the $j$th **generalized eigenspace**, call each of its nonzero members a $j$th **generalized eigenvector**. The completion of a basis of eigenvectors to a basis of generalized eigenvectors will follow from the following nesting property.

**Proposition** 11.4. $\mathcal{N}((\lambda_j I - B)^k) \subset \mathcal{R}(P_j)$ for $k = 1, \ldots, \mu_j$, and $\mathcal{N}((B - \lambda_j I)^{\mu_j}) = \mathcal{R}(P_j)$.

Proof: The key is the second resolvent identity, (11.11). For starting with $k = 1$ we find that if $e \in \mathcal{N}(B - \lambda_j)$ then $Be = \lambda_j e$ and (11.11) reveals that $R(s)Be = \lambda_j R(s)e = sR(s)e - e$, which upon simple rearrangement brings

$$R(s)e = \frac{1}{s - \lambda_j} e. \tag{11.25}$$

and so

$$P_j e = \frac{1}{2\pi i} \int_{C_j} R(s)e \, ds = \frac{1}{2\pi i} \int_{C_j} \frac{1}{s - \lambda_j} e \, ds = e.$$

Regarding $k = 2$ we note that if $x \in \mathcal{N}((B - \lambda_j I)^2)$ then $Bx = \lambda_j x + e$ for some $e \in \mathcal{N}(B - \lambda_j I)$. The second resolvent identity applied to $x$ now reveals $R(s)Bx = \lambda_j R(s)x + R(s)e = sR(s)x - x$, and upon rearrangement and recalling (11.25)

$$R(s)x = \frac{1}{(s - \lambda_j)^2} e + \frac{1}{s - \lambda_j} x. \tag{11.26}$$

Upon integrating this around $C_j$ we find indeed that $P_j x = x$. If $y \in \mathcal{N}((B - \lambda_j I)^3)$ then $By = \lambda_j y + x$ for some $x \in \mathcal{N}((B - \lambda_j I)^2)$. The second resolvent identity applied to $y$ now reveals $R(s)By = \lambda_j R(s)y + R(s)x = sR(s)y - y$, and upon rearrangement and recalling (11.26)

$$R(s)y = \frac{1}{(s - \lambda_j)^3} e + \frac{1}{(s - \lambda_j)^2} x + \frac{1}{s - \lambda_j} y,$$

and integrating this around $C_j$ brings $P_j y = y$. The pattern is now clear. Finally, if $x \in \mathcal{R}(P_j)$ then (11.24) reveals that $(B - \lambda_j I)^{\mu_j} x = 0$, i.e., $x \in \mathcal{N}((B - \lambda_j I)^{\mu_j})$. End of Proof.

This result suggests that the pole order, $\mu_j$, does not tell the full story. The paucity of eigenvectors associated with $\lambda_j$ is instead signified by the difference between the **geometric multiplicity**,

$$n_j \equiv \dim \mathcal{N}(B - \lambda_j I) \tag{11.27}$$

and the **algebraic multiplicity**

$$m_j \equiv \dim \mathcal{R}(P_j). \tag{11.28}$$

The previous proposition establishes that $n_j \leq m_j$.

With regard to the example, (11.1), with which we began the chapter we note that $n_1 = 1 < 2 = m_1$ and that the two eigenvectors in (11.3) may be completed by the second column of the associated $P_1$ (see (11.13)). We will see that there is a canonical means of completing the eigenvectors. However, prior to developing that we look at the much easier, and in fact typical, case where each $m_j = n_j$.

### 11.3. Diagonalization of a Semisimple Matrix

If $\mu_j = 1$ then we call $\lambda_j$ **semisimple**. If *each* $\mu_j = 1$ we call $B$ semisimple. Our first observation is that each nilpotent vanishes in the semisimple case, i.e.,

$$B = \sum_{j=1}^{h} \lambda_j P_j. \tag{11.29}$$

Our first objective is to construct a concrete alternative to this beautiful but perhaps overly concise representation.

As each $\mu_j = 1$ it follows from Prop. 11.4 that the algebraic and geometric multiplicities coincide, i.e., $m_j = n_j$.

Let us now show that its columns are linearly independent. Suppose $e_j \in \mathcal{R}(P_j)$ but that

$$e_k = \sum_{j \neq k} a_j e_j.$$

As we may write this as

$$P_k e_k = \sum_{j \neq k} a_j P_j e_j \quad \text{it follows that} \quad e_k = P_k P_k e_k = \sum_{j \neq k}^{h} a_j P_k P_j e_j = 0.$$

It then follows from (11.20) that these multiplicities sum to the ambient dimension. i.e.,

$$\sum_{j=1}^{h} n_j = n. \tag{11.30}$$

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

$$B e_{j,k} = \lambda_j e_{j,k},$$

and so

$$BE = E\Lambda \quad \text{where} \quad E = [E_1\ E_2\ \cdots E_h] \tag{11.31}$$

and $\Lambda$ is the diagonal matrix of eigenvalues,

$$\Lambda = \text{diag}(\lambda_1 \texttt{ones}(n_1, 1)\ \lambda_2 \texttt{ones}(n_2, 1)\ \cdots \lambda_h \texttt{ones}(n_h, 1)).$$

It follows from (11.30) that $E$ is square.

As such, it is invertible and so we have established

**Proposition** 11.5. If $B$ is semisimple then there exists an invertible matrix, $E$, of eigenvectors of $B$, and a diagonal matrix, $\Lambda$, of eigenvalues of $B$ (repeated according to their geometric multiplicities), such that

$$\boxed{B = E\Lambda E^{-1} \quad \text{and} \quad \Lambda = E^{-1} B E.} \tag{11.32}$$

In this sense we say that $E$ **diagonalizes** $B$. Let us work out a few examples and a number of striking consequences. The resolvent of the rotation matrix

$$B = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \tag{11.33}$$

has the partial fraction expansion

$$(sI - B)^{-1} = \frac{1}{s-i}\begin{pmatrix} 1/2 & -i/2 \\ i/2 & 1/2 \end{pmatrix} + \frac{1}{s+i}\begin{pmatrix} 1/2 & i/2 \\ -i/2 & 1/2 \end{pmatrix}$$

We recognize that the two eigenvalues are simple poles of the resolvent, and note by inspection 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. We lay these into $E = [e_1, e_2]$ and confirm that

$$E^{-1}BE = \Lambda = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}$$

as claimed. For our second example, the resolvent of the Clement matrix

$$B = \begin{pmatrix} 0 & 1 & 0 \\ 2 & 0 & 2 \\ 0 & 1 & 0 \end{pmatrix} \tag{11.34}$$

has the partial fraction expansion

$$(sI - B)^{-1} = \frac{1/4}{s+2}\begin{pmatrix} 1 & -1 & 1 \\ -2 & 2 & -2 \\ 1 & -1 & 1 \end{pmatrix} + \frac{1/2}{s}\begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix} + \frac{1/4}{s-2}\begin{pmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 1 & 1 & 1 \end{pmatrix}$$

and we read off the eigenvectors from the first columns of the respective projections

$$e_1 = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}, \quad e_2 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \quad \text{and} \quad e_3 = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}.$$

We lay these into $E = [e_1, e_2, e_3]$ and confirm that

$$E^{-1}BE = \Lambda = \begin{pmatrix} -2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Regarding functional consequences of our two representations

$$B = E\Lambda E^{-1} = \sum_{j=1}^{h} \lambda_j P_j, \tag{11.35}$$

we note, regarding the first and using only $E^{-1}E = I$, that

$$B^2 = BB = E\Lambda E^{-1} E\Lambda E^{-1} = E\Lambda^2 E^{-1},$$

and so, taking higher powers

$$\boxed{B^k = E\Lambda^k E^{-1}.} \tag{11.36}$$

This is worth boxing because $B^k$ is in general expensive to compute and difficult to interpret, while $\Lambda^k$ is the diagonal matrix of the $k$th powers of the eigenvalues of $B$. Regarding the second representation in (11.35) we note, using $P_iP_j = \delta_{ij}P_j$ that

$$B^2 = \left(\sum_{j=1}^{h} \lambda_j P_j\right)\left(\sum_{i=1}^{h} \lambda_i P_i\right) = \sum_{j=1}^{h} \lambda_j^2 P_j,$$

and so, for higher powers,

$$\boxed{B^k = \sum_{j=1}^{h} \lambda_j^k P_j.} \tag{11.37}$$

As each of these representations reduces powers of the matrix to powers of (scalar) eigenvalues we may take familiar sums of these scalar powers and arrive at new functions of the original matrix. In particular, recalling the power series definition, (9.12),

$$\boxed{\exp(Bt) \equiv \sum_{k=0}^{\infty} \frac{1}{k!}(Bt)^k}. \tag{11.38}$$

In light of (11.36) this becomes

$$\boxed{\exp(Bt) = \sum_{k=1}^{\infty} \frac{1}{k!} E(\Lambda t)^k E^{-1} = E\left(\sum_{k=1}^{\infty} (\Lambda t)^k/k!\right) E^{-1} = E\exp(\Lambda t)E^{-1}} \tag{11.39}$$

where $\exp(\Lambda t)$ is the diagonal matrix

$$\exp(\Lambda t) = \mathrm{diag}(\exp(\lambda_1 t)\mathtt{ones}(n_1,1)\ \exp(\lambda_2 t)\mathtt{ones}(n_2,1)\ \cdots \exp(\lambda_h t)\mathtt{ones}(n_h,1)).$$

In like fashion we draw from (11.37) the representation

$$\boxed{\exp(Bt) = \sum_{j=1}^{h} \exp(\lambda_j t) P_j.} \tag{11.40}$$

As the matrix exponential is of such fundamental importance to the dynamics of linear systems we pause to develop (in fact complete) yet a third representation. It begins from the partial fraction expansion of the resolvent of a semisimple matrix

$$(zI - B)^{-1} = \sum_{j=1}^{h} \frac{P_j}{z - \lambda_j}. \tag{11.41}$$

Recalling our discussion of the Backward Euler Method in §8.3 we now 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 (11.41) may be written

$$(I - B/z)^{-1} = \sum_{j=1}^{h} \frac{zP_j}{z - \lambda_j} = \sum_{j=1}^{h} \frac{P_j}{1 - \lambda_j/z} \tag{11.42}$$

If we now set $z = k/t$, where $k$ is a positive integer, and use $P_i P_j = \delta_{ij} P_j$ we arrive first at

$$(I - (t/k)B)^{-k} = \sum_{j=1}^{h} \frac{P_j}{(1 - (t/k)\lambda_j)^k} \tag{11.43}$$

and so, in the limit find

$$\boxed{\lim_{k\to\infty} (I - (t/k)B)^{-k} = \sum_{j=1}^{h} \exp(\lambda_j t) P_j = \exp(Bt).} \tag{11.44}$$

Although each of these representations have stemmed naturally from their scalar formulations we should still check that they do indeed resolve the dynamics problem. We carry this out for the representation (11.40).

$$\begin{aligned} \frac{d}{dt}\exp(Bt) &= \frac{d}{dt}\sum_{j=1}^{h} \exp(\lambda_j t) P_j = \sum_{j=1}^{h} \frac{d}{dt}\exp(\lambda_j t) P_j = \sum_{j=1}^{h} \lambda_j \exp(\lambda_j t) P_j \\ &= \sum_{i=1}^{h} \lambda_i P_i \sum_{j=1}^{h} \exp(\lambda_j t) P_j = B\exp(Bt). \end{aligned}$$

For the rotation matrix, (11.33), representation (11.40) yields

$$\exp(Bt) = \exp(it)\begin{pmatrix} 1/2 & -i/2 \\ i/2 & 1/2 \end{pmatrix} + \exp(-it)\begin{pmatrix} 1/2 & i/2 \\ -i/2 & 1/2 \end{pmatrix} = \begin{pmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{pmatrix}. \tag{11.45}$$

While for the Clement matrix, (11.34), we find

$$\begin{aligned} \exp(Bt) &= \frac{\exp(-2t)}{4}\begin{pmatrix} 1 & -1 & 1 \\ -2 & 2 & -2 \\ 1 & -1 & 1 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix} + \frac{\exp(2t)}{4}\begin{pmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 1 & 1 & 1 \end{pmatrix} \\ &= \frac{1}{2}\begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \cosh(2t) & \sinh(2t) & \cosh(2t) \\ 2\sinh(2t) & 2\cosh(2t) & 2\sinh(2t) \\ \cosh(2t) & \sinh(2t) & \cosh(2t) \end{pmatrix} \end{aligned} \tag{11.46}$$

### 11.4. The Schur Form and the QR Algorithm*

In general, $B$ is nonsemisimple and there is at least one $\mu_j > 1$, and so there are too few eigenvectors with which to construct a diagonalizing similarity transformation. In this case however there still exist *triangularizing* similarity transformations. The diagonal of the resulting triangular matrix will indeed be comprised of the eigenvalues of $B$. There are two standard ways of developing this triangularization; the Schur form is simple to construct but typically delivers a full upper triangle while the Jordan form is difficult to construct but confines the upper triangle to a single super–diagonal. We start with the former.

**Proposition** 11.6. For $B \in \mathbb{C}^{n\times n}$ there exists a unitary $Q \in \mathbb{C}^{n\times n}$ and an upper triangular $U \in \mathbb{C}^{n\times n}$ (the **Schur form** of $B$) such that

$$Q^* B Q = U. \tag{11.47}$$

Proof: The proof is by induction on $n$. For $n = 1$ the proof is trivial since $B$ is a scalar in this case.

Now assume that for any $B_{n-1} \in \mathbb{C}^{(n-1)\times(n-1)}$ there exists a unitary $Q_{n-1} \in \mathbb{C}^{(n-1)\times(n-1)}$ such that $Q_{n-1}^* B_{n-1} Q_{n-1}$ is upper triangular.

Now suppose $Bx_1 = \lambda_1 x_1$. Suppose $\|x_1\| = 1$ and denote by $\{x_2, \ldots, x_n\}$ an orthonormal basis for the orthogonal complement of $x_1$. The matrix $X = [x_1, x_2, \ldots, x_n]$ is therefore unitary and

$$X^* B X = \begin{pmatrix} \lambda_1 & y^* \\ 0 & B_{n-1} \end{pmatrix}$$

where $y \in \mathbb{C}^{n-1}$ and $B_{n-1} \in \mathbb{C}^{(n-1)\times(n-1)}$. From the induction hypothesis there exists a unitary $Q_{n-1} \in \mathbb{C}^{(n-1)\times(n-1)}$ such that $Q_{n-1}^* B_{n-1} Q_{n-1}$ is upper triangular. Choose

$$Q = X \begin{pmatrix} 1 & 0 \\ 0 & Q_{n-1} \end{pmatrix}$$

and note that $Q$ is unitary and

$$\begin{aligned} Q^* B Q &= \begin{pmatrix} 1 & 0 \\ 0 & Q_{n-1}^* \end{pmatrix} X^* B X \begin{pmatrix} 1 & 0 \\ 0 & Q_{n-1} \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & Q_{n-1}^* \end{pmatrix} \begin{pmatrix} \lambda_1 & y^* \\ 0 & B_{n-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & Q_{n-1} \end{pmatrix} \\ &= \begin{pmatrix} \lambda_1 & y^* Q_{n-1} \\ 0 & Q_{n-1}^* B_{n-1} Q_{n-1} \end{pmatrix} \end{aligned}$$

where the last matrix is an upper triangular matrix because $Q_{n-1}^* B_{n-1} Q_{n-1}$ is upper triangular. End of Proof.

Regarding examples, the Schur form is typically achieved by an efficient implementation of a very simple iteration. Namely, compute the QR factorization of $B$, reverse the factors and repeat. In symbols,

$$B_1 = B, \quad [Q_1, R_1] = \mathrm{qr}(B_1), \quad B_2 = R_1 Q_1, \quad [Q_2, R_2] = \mathrm{qr}(B_2), \quad \ldots \tag{11.48}$$

and we repeat until $B_k$ is triangular. This procedure is called the QR method for determining the eigenvalues of $B$. It is far from easy to see why these iterates should become triangular. It is however easy to see that they are each similar to one another, for

$$B_{k+1} = R_k Q_k = Q_k^T Q_k R_k Q_k = Q_k^T B_k Q_k.$$

Lets run through these steps for

$$B = \begin{pmatrix} 16 & 12 & 8 & 4 \\ -4 & 2 & -7 & -6 \\ 2 & 4 & 16 & 8 \\ -1 & -2 & -3 & 6 \end{pmatrix}. \tag{11.49}$$

Its resolvent is

$$(sI - B)^{-1} = \frac{1}{s-10} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + \frac{1}{(s-10)^2} \begin{pmatrix} 6 & 12 & 8 & 4 \\ -4 & -8 & -7 & -6 \\ 2 & 4 & 6 & 8 \\ -1 & -2 & -3 & -4 \end{pmatrix}$$

and so $\lambda_1 = 10$ is the sole eigenvalue. For practical purposes we have terminated (11.48) when the Eigenerror, $\max|\mathrm{diag}(B_k) - 10|$, was less than 0.01. This required 1000 iterations, see Figure 11.2, and resulted in the Schur form

$$B_{1000} = \begin{pmatrix} 10.0100 & 0.0298 & -10.4400 & -20.0570 \\ 0.0000 & 10.0000 & 3.4933 & -7.1777 \\ 0.0000 & 0.0000 & 10.0000 & -0.0205 \\ 0.0000 & 0.0000 & -0.0000 & 9.9900 \end{pmatrix}.$$

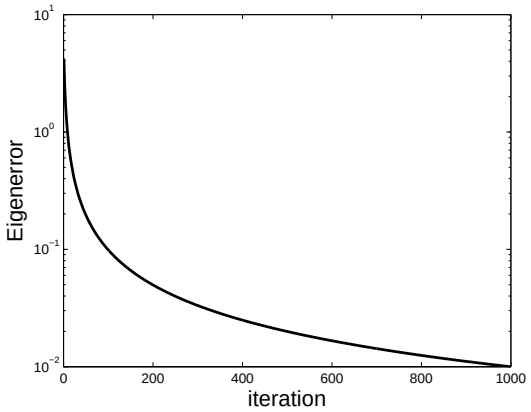



**Figure** 11.2. The convergence of the naive QR algorithm, (11.48), to a Schur form of (11.49).

### 11.5. The Jordan Canonical Form*

To the insufficient number of eigenvectors we add carefully constructed generalized eigenvectors from $\mathcal{R}(P_j)$. As this space is filled out, recall Prop. 11.4, by null spaces of powers of $(B - \lambda_j I)$, we will construct Jordan bases for the $\mathcal{R}(P_j)$ along the lines used in the Jordan decomposition of nilpotents in §4.4.

As a warm up lets restrict ourselves to matrices $B$ with a single distinct nonsemisimple eigenvalue, i.e., $h = 1$ and $p_1 > 1$. We first record the dimensions

$$d_j = \dim \mathcal{N}((B - \lambda_1 I)^j), \quad j = 1, \ldots, p_1,$$

and proceed to

**Step 1':** construct a basis $\{v_{p_1}^1, v_{p_1}^2, \cdots, v_{p_1}^{c_{p_1}}\}$ for $\mathcal{N}((B - \lambda_1 I)^{p_1})$ mod $\mathcal{N}((B - \lambda_1 I)^{p_1 - 1})$, where $v_{p_1}^k$ is the coordinate vector for the $k$th pivot column of $(B - \lambda_1 I)^{p_1 - 1}$.

**Step 2:** Note that $(B - \lambda_1 I)v_{p_1}^j \in \mathcal{N}((B - \lambda_1 I)^{p_1 - 1})$ and that $\{(B - \lambda_1 I)v_{p_1}^j : j = 1, \ldots, c_{p_1}\}$ is linearly independent mod $\mathcal{N}((B - \lambda_1 I)^{p_1 - 2})$. We complete this to a basis by appending $\{v_{p_1-1}^1, \ldots, v_{p_1-1}^{c_{p_1-1}}\}$.

**Step 3:** Repeat Step 2, with $p_1$ replaced by $p_1 - 1$.

On completion we arrange the basis vectors by increasing block size

$$\begin{aligned} X = \{&v_1^1, \ldots, v_1^{c_1}, (B - \lambda_1 I)v_2^1, v_2^1, \ldots, (B - \lambda_1 I)v_2^{c_2}, v_2^{c_2}, \ldots, \\ &(B - \lambda_1 I)^{i-1} v_i^1, \ldots, v_i^1, \ldots, (B - \lambda_1 I)^{i-1} v_i^{c_i}, , \ldots, v_i^{c_i}, \ldots\}, \end{aligned}$$

and observe the chain/block structure inherent in the columns of $X$. Namely, for the 1-blocks

$$Bx_j = \lambda_1 x_j, \quad j = 1, \ldots, c_1,$$

and for 2-blocks,

$$Bx_j = \lambda_1 x_j \quad \text{and} \quad Bx_{j+1} = \lambda_1 x_{j+1} + x_j, \quad j = c_1 + 1, \ldots, c_1 + c_2$$

and for 3-blocks,

$$Bx_j = \lambda_1 x_j, \quad Bx_{j+1} = \lambda_1 x_{j+1} + x_j, \quad Bx_{j+2} = \lambda_1 x_{j+2} + x_{j+1}, \quad j = c_1 + 2c_2, \ldots, c_1 + 2c_2 + c_3$$

and so on. It follows that $BX = XJ$ where $J$ is the block diagonal matrix begining with $c_1$ zeros, then $c_2$ blocks of size 2, then $c_3$ blocks of size 3 up through $c_m$ blocks of size $m$. Each block has $\lambda_1$ along its diagonal and ones along its superdiagonal. The chain numbers, $c_j$, are determined by the null space dimensions, $d_j$, precisely as in (4.14).

Let's take the concrete example of (11.49). The associated null spaces have dimensions $d_1 = 2$ and $d_2 = 4$ and so we expect to build 2 chains of length two. Regarding the pivot columns of $(B - \lambda_1 I)$ we note, either directly or via (11.22), that $(B - \lambda_1 I) = D_1$. As columns 1 and 3 are pivot columns of $D_1$ it follows from Step 1' in that

$$v_2^1 = (1\ 0\ 0\ 0)^T \quad \text{and} \quad v_2^2 = (0\ 0\ 1\ 0)^T$$

comprise a basis for $\mathcal{N}((B - \lambda_1 I)^2)$ mod $\mathcal{N}(B - \lambda_1 I)$. We then build

$$X = [(B - \lambda_1 I)v_2^1\ v_2^1\ (B - \lambda_1 I)v_2^2\ v_2^2] = \begin{pmatrix} 6 & 1 & 8 & 0 \\ -4 & 0 & -7 & 0 \\ 2 & 0 & 6 & 1 \\ -1 & 0 & -3 & 0 \end{pmatrix} \tag{11.50}$$

and find that

$$X^{-1}BX = J = \begin{pmatrix} 10 & 1 & 0 & 0 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 10 & 1 \\ 0 & 0 & 0 & 10 \end{pmatrix},$$

as predicted.

In the general case we merely repeat this procedure for each eigenvalue. The basis vectors for the distinct $\lambda_j$ and $\lambda_k$ are independent from another because $P_j P_k = 0$.

**Proposition** 11.7, If $B \in \mathbb{C}^{n\times n}$ then $B$ is similar to a Jordan matrix $J \in \mathbb{C}^{n\times n}$. In particular, there exists an $X$ (comprised of generalized eigenvectors of $B$) for which

$$\boxed{X^{-1}BX = J.} \tag{11.51}$$

The structure of $J$ is determined by the eigenvalues of $B$, denoted $\lambda_1, \lambda_2, \ldots, \lambda_h$, their associated orders as poles of the resolvent of $B$, denoted $p_1, p_2, \ldots, p_h$, and finally by their associated nullities $d_{j,k} = \dim \mathcal{N}((B - \lambda_k)^j)$ for $k = 1, \ldots, h$ and $j = 1, \ldots, p_k$. In particular, for each $k = 1, \ldots, h$, there will be $d_{1,k}$ Jordan blocks with $\lambda_k$ on the diagonal. Among these there will be $c_{j,k}$ blocks of size $j$, where $j = 1, \ldots, p_k$ and $c_{:,k} = S_{p_k} d_{:,k}$ where $S_{p_k}$ is the hanging chain matrix of (4.14). The sum of the sizes of all Jordan blocks associated with $\lambda_k$ is $m_k = \dim \mathcal{R}(P_k)$, the algebraic multiplicity of $\lambda_k$.

Let us construct the Jordan form of

$$B = \begin{pmatrix} 3 & 2 & 3 & 5 & -7 & 4 & -9 \\ 1 & 4 & 4 & 6 & -7 & 4 & -10 \\ 1 & 1 & 5 & 5 & -6 & 4 & -8 \\ 1 & 3 & 5 & 11 & -11 & 7 & -15 \\ 1 & 3 & 6 & 10 & -10 & 8 & -17 \\ 1 & 1 & 3 & 6 & -7 & 7 & -9 \\ 1 & 1 & 2 & 4 & -5 & 3 & -4 \end{pmatrix}. \tag{11.52}$$

Its resolvent is

$$(sI - B)^{-1} = \frac{1}{s-1}P_1 + \frac{1}{s-2}P_2 + \frac{1}{s-3}P_3 + \frac{1}{(s-2)^2}D_2 + \frac{1}{(s-3)^2}D_3 + \frac{1}{(s-3)^3}D_3^2,$$

where the eigenprojections, eigennilpotents and their nonzero powers are

$$P_1 = \begin{pmatrix} -1 & -2 & -3 & -5 & 7 & -4 & 9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -2 & -3 & -5 & 7 & -4 & 9 \\ -1 & -2 & -3 & -5 & 7 & -4 & 9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad P_2 = \begin{pmatrix} 2 & 2 & 3 & 5 & -7 & 4 & -9 \\ 0 & 0 & -1 & -2 & 2 & -1 & 3 \\ 0 & 0 & 0 & -1 & 1 & -1 & 2 \\ 1 & 1 & 2 & 3 & -4 & 2 & -5 \\ 1 & 1 & 1 & 2 & -3 & 2 & -4 \\ 0 & 0 & -1 & -2 & 2 & -1 & 3 \\ 0 & 0 & 0 & -1 & 1 & -1 & 2 \end{pmatrix}$$

and

$$D_2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 2 & 3 & -4 & 2 & -5 \\ 1 & 1 & 2 & 3 & -4 & 2 & -5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 2 & 3 & -4 & 2 & -5 \\ 1 & 1 & 2 & 3 & -4 & 2 & -5 \end{pmatrix}, \quad P_3 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 2 & -2 & 1 & -3 \\ 0 & 0 & 1 & 1 & -1 & 1 & -2 \\ 0 & 1 & 1 & 3 & -3 & 2 & -4 \\ 0 & 1 & 2 & 3 & -3 & 2 & -5 \\ 0 & 0 & 1 & 2 & -2 & 2 & -3 \\ 0 & 0 & 0 & 1 & -1 & 1 & -1 \end{pmatrix}$$

and

$$D_3 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & -1 & 1 & -2 \\ 0 & 0 & 0 & 1 & -1 & 1 & -1 \\ 0 & 0 & 1 & 1 & -1 & 1 & -2 \\ 0 & 0 & 1 & 2 & -2 & 2 & -3 \\ 0 & 0 & 0 & 1 & -1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad D_3^2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 1 & -1 \\ 0 & 0 & 0 & 1 & -1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

corresponding to the three eigenvalues, $\lambda_j = j$, and their respective pole orders, $\mu_j = j$. The Jordan Form is fully determined by the sequence of dimensions

$$d_{1,1} = \dim \mathcal{N}(B - I) = 1,$$
$$d_{1,2} = \dim \mathcal{N}(B - 2I) = 2, \quad d_{2,2} = \dim \mathcal{N}((B - 2I)^2) = 3,$$
$$d_{1,3} = \dim \mathcal{N}(B - 3I) = 1, \quad d_{2,3} = \dim \mathcal{N}((B - 3I)^2) = 2, \quad d_{3,3} = \dim \mathcal{N}((B - 3I)^3) = 3.$$

In particular, $d_{1,1} = 1$ specifies that there is one Jordan block associated with $\lambda_1$. The size of the block is $p_1 = 1$. $d_{1,2} = 2$ specifies that there are two Jordan blocks associated with $\lambda_2$. Their sizes are $d_{2,2} - d_{1,2} = 1$ and $p_2 = 2$. $d_{1,3} = 1$ specifies that there is one Jordan block associated with $\lambda_3$. The size of the block is $p_3 = 3$. As a result $B$ is similar to

$$J = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3 \end{pmatrix} \tag{11.53}$$

and the transformation matrix is $X = [X_1, X_2, X_3]$ where

$$X_1 = v^1_{1,1}, \quad X_2 = \{v^1_{1,2}, (B-\lambda_2 I)v^1_{2,2}, v^1_{2,2}\} \quad \text{and} \quad X_3 = \{(B-\lambda_3 I)^2 v^1_{3,3}, (B-\lambda_3 I)v^1_{3,3}, v^1_{3,3}\} \tag{11.54}$$

Here $v^1_{1,1}$ is a basis for $\mathcal{N}(B-\lambda_1 I)$,

$$v^1_{1,1} = (1\ 0\ 0\ 1\ 1\ 0\ 0)^T,$$

$v^1_{2,2}$ is a basis for $\mathcal{N}((B-\lambda_2 I)^2)$ mod $\mathcal{N}(B-\lambda_2 I)$, and $v^1_{3,3}$ is a basis for $\mathcal{N}((B-\lambda_3 I)^3)$ mod $\mathcal{N}((B-\lambda_3 I)^2)$. Recalling Step 1' and $(B-\lambda_2 I)P_2 = D_2$ we obtain $v^1_{2,2} = P_2 e_2$ where $e_2$ is the coordinate vector of the pivot column of $D_2$. As column 1 is the pivot column of $D_2$ we find

$$v^1_{2,2} = (2\ 0\ 0\ 1\ 1\ 0\ 0)^T \quad \text{and} \quad (B-\lambda_2 I)v^1_{2,2} = (0\ 1\ 1\ 0\ 0\ 1\ 1)^T.$$

These two lie, by construction in $\mathcal{N}(B-\lambda_2 I)^2$. We recognize that $v^1_{1,2} = (1\ 0\ 1\ 1\ 0\ 0\ 1)^T$ completes the basis of $\mathcal{N}(B-\lambda_2 I)^2$.

Similarly, as $(B-\lambda_3 I)^2 P_3 = D_3^2$ we obtain $v^1_{3,3} = P_3 e_3$ where $e_3$ is the coordinate vector of the pivot column of $D_3^2$. As column 4 is the pivot column of $D_3^2$ we find

$$v^1_{3,3} = (0\ 2\ 1\ 3\ 3\ 2\ 1)^T, \quad (B-\lambda_3 I)v^1_{3,3} = (0\ 1\ 1\ 1\ 2\ 1\ 0)^T \quad \text{and} \quad (B-\lambda_3 I)^2 v^1_{3,3} = (0\ 1\ 0\ 1\ 1\ 0\ 0)^T.$$

These vectors arranged, per (11.54), indeed deliver $X^{-1}BX = J$ for the $B$ and $J$ of (11.52) and (11.53).

As each matrix is similar to a triangular matrix with its eigenvalues, repeated by their algebraic multiplicity, along the diagonal, we may use the product formulas for the determinant and trace to establish

**Corollary** 11.8. In the language of Prop. 11.6,

$$\det(B-zI) = \prod_{j=1}^{h} (\lambda_j - z)^{m_j} \quad \text{and} \quad \operatorname{tr}(B) = \sum_{j=1}^{h} m_j \lambda_j. \tag{11.55}$$

The first statement in (11.55) finally brings us to a concise formulation of the characteristic polynomial,

$$\chi_B(z) \equiv \det(B-zI) = \prod_{j=1}^{h} (\lambda_j - z)^{m_j} \tag{11.56}$$

first encountered in §8.2 during our introduction to dynamics.

We close with two applications of a generalized form of Cauchy's Integral Formula. If $C$ is a closed curve that strictly contains the eigenvalues of $B$ and $f$ is differentiable on and in $C$ then

$$\begin{aligned} f(B) &= \frac{1}{2\pi i} \int_C (zI-B)^{-1} f(z)\, dz \\ &= \frac{1}{2\pi i} \sum_{j=1}^{h} \int_{C_j} \left( \frac{f(z)}{z-\lambda_j} P_j + \sum_{k=1}^{m_j-1} \frac{f(z)}{(z-\lambda_j)^{k+1}} D_j^k \right) dz \\ &= \sum_{j=1}^{h} \left( f(\lambda_j) P_j + \sum_{k=1}^{m_j-1} \frac{1}{k!} \frac{d^k f(\lambda_j)}{dz^k} D_j^k \right). \end{aligned} \tag{11.57}$$

As $d^k\chi_B(\lambda_j)/dz^k = 0$ for each $j$ and $0 \le k < m_j$ we have proven the Cayley–Hamilton Theorem,

**Proposition** 11.9. **Cayley–Hamilton Theorem**. Each matrix satisfies its own characteristic polynomial, i.e., $\chi_B(B) = 0$.

Our second application of (11.57) is the representation of $\exp(Bt)$ as the Laplace Transform of the resolvent

$$\exp(Bt) = \frac{1}{2\pi i}\int_C (zI - B)^{-1}\exp(zt)\,dz = \sum_{j=1}^{h}\exp(\lambda_j t)\left(P_j + \sum_{k=1}^{m_j-1}\frac{t^k}{k!}D_j^k\right). \tag{11.58}$$

With this representation we return to the example (11.49) and compute

$$\exp(Bt) = \exp(10t)(I + tD_1) = \exp(10t)\begin{pmatrix} 1+6t & 12t & 8t & 4t \\ -4t & 1-8t & -7t & -6t \\ 2t & 4t & 1+6t & 8t \\ -t & -2t & -3t & 1-4t \end{pmatrix}.$$

### 11.6. Positive Matrices and the PageRank Algorithm*

The Jordan form, Eq. (11.51), is canonical in the sense that it is the simplest spectral form that an arbitrary square matrix can take. Of course if we work over a class of more structured matrices we may hope that this structure is in some sense inherited by the associated eigenvalues and eigenvectors. This is in fact the theme for the remaining chapters of this book. We kick it off here with a focus on the dramatic implications of the spectra of positive matrices.

One key to the rapid search through a large and diverse set of documents is to have an efficient ordering, or ranking, of their utility or reliability. Rather than applying some external measure, and so setting oneself up as the arbiter of truth and utility, the best search tools use an internal measure and rank documents based upon the number of times they are referenced by other high ranking documents. This circular definition is resolved with a little linear algebra. To see this lets begin with the five documents in Figure 11.3.

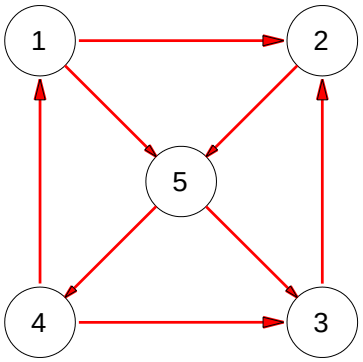



**Figure** 11.3. Among the five documents, document 1 cites documents 2 and 5, document 2 cites only document 5, and so on.

We quantify the citations depicted in Figure 11.3 by scaling the impact of a citation by the total number of citations made by that document. Hence documents 5 and 2 each receive one half of the

rank of document 1 while document 5 receives all of the rank of document 2. If $x_j$ denotes the rank of document $j$ we may then interpret Figure 11.3 to say

$$\begin{aligned} x_1 &= x_4/2 \\ x_2 &= x_1/2 + x_3 \\ x_3 &= x_4/2 + x_5/2 \\ x_4 &= x_5/2 \\ x_5 &= x_1/2 + x_2. \end{aligned}$$

This is best captured as

$$Ax = x. \tag{11.59}$$

where $A$ is the citation matrix

$$A = \begin{pmatrix} 0 & 0 & 0 & 1/2 & 0 \\ 1/2 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0 & 1/2 \\ 1/2 & 1 & 0 & 0 & 0 \end{pmatrix} \tag{11.60}$$

associated with Figure 11.3. The self-referential nature of our preliminary definition of rank is echoed in Eq. (11.59) – which states that the rank vector is a "self" (i.e., "eigen") vector of the citation matrix. This however comes with a strong twist – $x$ is not just any eigenvector of $A$, it is the eigenvector associated with the eigenvalue 1. Does every citation matrix have 1 as an eigenvalue? If so is it paired with a unique (positive) eigenvector? If so, how can we compute it effectively (when faced with tens of millions of documents)? With a slightly more robust definition we will see that each of these questions can be answered in the affirmative. Before working out the general theory lets finish off this example by noting that 1 is a simple eigenvalue of (11.60) and that it has the positive eigenvector

$$x = (2,\ 7,\ 6,\ 4,\ 8) \tag{11.61}$$

We glean from this $x$ that the documents in Figure 11.3 should be ranked, in order of decreasing importance as 5, 3, 2, 4, 1. This should coarsely jibe with your intuition, the top three documents each receive 2 citations while the bottom two each receive only one. That document 5 is clearly better than document 2 which is clearly more important than document 3 can only be discerned by consider all of the documents simultaneously – and this is precisely what is done in (11.59).

In order to see that not all document graphs produce viable, in the sense of (11.59), citation matrices we need only reverse two links in Figure 11.3.

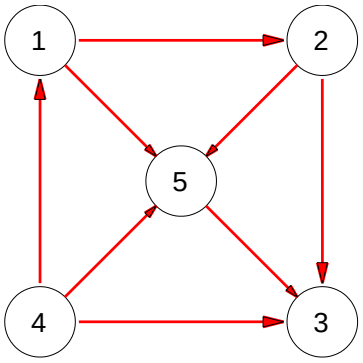


**Figure** 11.4. A reversal, with respect to Figure 11.3, of the citations between documents 2 and 3 and documents 4 and 5.

The citation matrix for Figure 11.4

$$A = \begin{pmatrix} 0 & 0 & 0 & 1/3 & 0 \\ 1/2 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/3 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 1/3 & 0 \end{pmatrix} \tag{11.62}$$

has no chance of satisfying (11.59), for ALL of its eigenvalues are zero – in fact it is nilpotent!, $A^5 = 0$. Of course this graph is "degenerate" in the sense that document 3 makes no citations and document 4 receives no citations. The most unbiased way to rectify this is to have every document cite every other document but with weights that do not swamp those established by the original citation matrix. This is accomplished by instead studying

$$B = (1-t)A + tK \tag{11.63}$$

where $t$ is a tuning parameter and $K$ is the kumbaya matrix

$$K = \frac{1}{n} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} \tag{11.64}$$

where each of the $n$ documents shares one nenth of its rank with every other document, including itself. If we apply this strategy to the $A$ in (11.62) with $t = 0.15$ we find that $B$ has one simple positive eigenvalue, $\lambda_1 \approx 0.56$, and a positive eigenvector that leads to the reasonable ranking (3, 5, 2, 1, 4). That this $\lambda_1$ falls short of 1 is a reflection of the fact that the dangling document, 3, contributes to a column sum of $t$ in $B$ while all other column sums are 1.

We consider $B \in \mathbb{R}^{n\times n}$ that are positive in the sense that each and every element of $B$ is positive, i.e., strictly greater than 0. Let us write $B > 0$ to mean that every element of $B$ is strictly greater than 0 and $x \geq 0$ to mean that every element of $x$ is greater than or equal to 0. It will be useful to make use of the vector and matrix norms

$$\|x\|_1 = \sum_{i=1}^{n} x_i \quad \text{and} \quad \|B\|_1 = \sum_{i=1}^{n}\sum_{j=1}^{n} b_{i,j}$$

and the inequality

$$\|Bx\|_1 \leq \|B\|_1 \|x\|_1.$$

On hunting for a positive eigenvalue of $B$ it make sense to search among those numbers in

$$L \equiv \{\lambda \geq 0 : \exists\ x \geq 0 \text{ such that } Bx \geq \lambda x\}.$$

We note that $L$ is nonempty, for $0 \in L$, and bounded above, for if $Bx \geq \lambda x$ then $\|B\|_1\|x\|_1 \geq \|Bx\|_1 \geq \lambda\|x\|_1$ and so $\lambda \leq \|B\|_1$. It therefore makes sense to consider

$$\lambda^+ = \sup_{\lambda\in L} \lambda. \tag{11.65}$$

**Proposition** 11.10. **Perron's Theorem.** If each element of $B \in \mathbb{R}^{n\times n}$ is positive then the $\lambda^+$ in (11.65) is an eigenvalue of $B$. It is positive, its geometric and algebraic multiplicities are each one and it has an associated positive eigenvector.

Proof: Choose $\lambda_j \in L$ such that $\lambda_j \to \lambda^+$ and denote by $x_j \geq 0$ a vector for which $Bx_j \geq \lambda_j x_j$ and $\|x_j\|_1 = 1$. As the $x_j$ comprise a bounded sequence it follows from Proposition 1.3 that they possess a subsequence, $\{x_{j_k}\}$ that converges to a limit $x^+$. We note that $x^+ \geq 0$ and $\|x^+\|_1 = 1$ and $Bx^+ \geq \lambda^+ x^+$. To see that $\lambda^+ > 0$ set

$$c_j = \min_i b_{i,j}$$

and note that $Be_j \geq c_j e_j$. So

$$\lambda^+ \geq \max_j c_j.$$

If strict inequality holds in $Bx^+ \geq \lambda^+ x^+$ at say the first element then

$$(Bx^+)_1 - \lambda^+ x_1^+ = d_1 > 0$$

then setting

$$y \equiv x^+ + (d_1/\lambda^+)e_1 \quad \text{and} \quad \varepsilon \equiv \frac{d_1}{\lambda^+}\frac{\min_i b_{i,1}}{\max_i y_i}$$

produces $By \geq (\varepsilon + \lambda^+)y$ contrary to the maximality of $\lambda^+$. Hence $\lambda^+$ and $x^+$ are an eigenpair for $B$. Now $Bx^+ = \lambda^+ x^+$ together with $B > 0$ and $\lambda^+ > 0$ imply that $x^+ > 0$.

To see that $\lambda^+$ is the largest eigenvalue of $B$ suppose that $By = \lambda y$ and $|\lambda| \geq \lambda^+$. We denote by $|y|$ the vector of magnitudes of elements of $y$. From $By = \lambda y$ and $B > 0$ we glean

$$B|y| \geq |By| = |\lambda y| = |\lambda||y| \tag{11.66}$$

and so conclude (as $|y|$ is a contender in $L$) that $|\lambda| \leq \lambda^+$. Together with the opening supposition of the clause it follows that $|\lambda| = \lambda^+$ and that equality must hold in (11.66) which implies that $|y|$ is a second eigenvector of $B$ with eigenvalue $\lambda^+$. It follows that $x^+ - \varepsilon|y|$ is also such an eigenvector. If $x^+$ and $|y|$ are not colinear then there exists a smallest positive value of $\varepsilon$ such that one (or more) elements of $x^+ - \varepsilon|y|$ are zero while the remainder are positive. But this would contradict

$$B(x^+ - \varepsilon|y|) = \lambda^+(x^+ - \varepsilon|y|)$$

at those elements, hence $|y|$ and $x^+$ are colinear and so $\lambda^+$ is geometrically simple.

If its algebraic multiplicity exceeds one then, by Proposition 11.4, there exists a vector $w$ such

$$(B - \lambda^+ I)^k w = 0 \quad \text{and} \quad (B - \lambda^+ I)^{k-1} w \neq 0$$

for some $k \geq 2$. This states that $(B - \lambda^+ I)^{k-1} w$ is an eigenvector associated with $\lambda^+$ and so is a multiple of $x^+$. Without loss we can assume this multiple is one, hence

$$x^+ = (B - \lambda^+ I)^{k-1} w.$$

If we then set $z = (B - \lambda^+ I)^{k-2} w$ then $(B - \lambda^+ I)z = x^+$. This now reads

$$Bz = \lambda^+ z + x^+ > \lambda^+ z$$

which leads to $B|z| > \lambda^+|z|$ in contradiction of the definition of $\lambda^+$. End of Proof.

This now guarantees the existence of a unique rank vector for every generalized citation matrix, (11.63). We leave it as a useful exercise to prove that if every column sum of $B$ is one then $\lambda^+ = 1$.

Regarding an effective means of estimating $x^+$ we note that our spectral representation takes on a beautiful form when $\lambda^+ = 1$. Namely

$$B = P^+ + \sum_{j=2}^{h} (\lambda_j P_j + D_j), \tag{11.67}$$

where $P^+$ is projection onto the desired $x^+$. As each $|\lambda_j| < 1$ and each $D_j$ is nilpotent and $P^+ P_j = P^+ D_j = 0$ it follows that

$$P^+ = \lim_{k \to \infty} B^k. \tag{11.68}$$

In practice this suggests the very simple algorithm: let $e$ be the vector of all ones and compute

$$Be,\ BBe,\ BBBe,\ \ldots$$

until convergence to $x^+$.

### 11.7. Notes and Exercises

We have followed Kato (1980). For more on the pagerank problem please consult Bryan and Leise (2006).

1. Use (3.32) to build the partial fraction expansion of the resolvent of the rank one matrix, $uv^T$, where $u$ and $v$ lie in $\mathbb{R}^n$.

   (a) Show that if $u^T v = 0$ then $uv^T$ is nilpotent and

   $$(sI - uv^T)^{-1} = \frac{1}{s} I + \frac{1}{s^2} uv^T.$$

   (b) Show that if $u^T v \neq 0$ then

   $$(sI - uv^T)^{-1} = \frac{1}{s} \left( I - \frac{uv^T}{u^T v} \right) + \frac{1}{s - u^T v} \frac{uv^T}{u^T v}.$$

2. Argue as in Prop. 11.1 that if $j \neq k$ then $D_j P_k = P_j D_k = 0$.

3. Argue from (11.22) and $P_j^2 = P_j$ that $D_j P_j = P_j D_j = D_j$. Use the latter to prove that $\mathcal{R}(D_j) \subset \mathcal{R}(P_j)$. From this result deduce that $\mu_j$, the multiplicity of $\lambda_j$ as a pole of the resolvent, can not exceed $m_j$, the dimension of $\mathcal{R}(P_j)$.

4. Show that $A$ and $A^T$ have the same eigenvalues and same multiplicities.

5. Use Corollary 4.14 and Eq. (11.17) and Eq. (11.56) to establish

   $$\operatorname{tr} (zI - B)^{-1} = \sum_{j=1}^{h} \frac{m_j}{z - \lambda_j} = \frac{\chi_B'(z)}{\chi_B(z)}. \tag{11.69}$$

6. Let us consider a semisimple matrix with a multiple eigenvalue,

$$B = \begin{pmatrix} 4 & 0 & 0 \\ -1 & 2 & -3 \\ 2 & 4 & 10 \end{pmatrix}$$

(i) Find the partial fraction expansion of its resolvent. (ii) From the projections in (i) extract three linearly independent eigenvectors and use these to diagonalize $B$.

7. Starting from the representation (11.57) use the second resolvent identity to confirm that $(\exp(Bt))' = B\exp(Bt)$.

8. Note that $-B$ has the same spectral representation as $B$ (except for a change in sign in the eigenvalues). Construct $\exp(-Bt)$ and show that it is the inverse of $\exp(Bt)$.

9. Show that $\exp(B^T t) = (\exp(Bt))^T$.

10. Regarding (11.45), note that $\exp(Bt)^T \exp(Bt) = I$. Show that this is a direct consequence of $B^T = -B$.

11. Show that $\exp((A+B)t) = \exp(At)\exp(Bt)$ if and only if $AB = BA$.

12. We use the Schur form to diagonalize normal matrices. A matrix $B \in \mathbb{C}^{n\times n}$ is said to be **normal** when $BB^* = B^*B$. This is a natural extension of the class of Hermitian matrices.

    (i) Use (11.47) to write $B = QUQ^*$ where $Q$ is unitary and $U$ is upper triangular. Taking adjoints of both sides conclude that $B^* = QU^*Q^*$.

    (ii) Use (i) to show that if $B$ is normal then so is $U$.

    (iii) Show that the only normal triangular matrices are diagonal and conclude from (ii) that if $B$ is normal then there exists a unitary $Q$ and **diagonal** $\Lambda$ such that $Q^*BQ = \Lambda$.

    (iv) Finally establish the converse: If there exists a unitary $Q$ and diagonal $\Lambda$ such that $B = Q\Lambda Q^*$ then $B$ is normal.

13. The Frobenius norm of $B \in \mathbb{C}^{n\times n}$ is $\|B\|_F \equiv \sqrt{\mathrm{tr}(BB^*)}$. Use the Schur Form to establish

$$\sum_{j=1}^{n} |\lambda_j|^2 \le \|B\|_F^2, \tag{11.70}$$

    where the $\lambda_j$ are the eigenvalues of $B$. Show that equality holds in Eq. (11.70) if and only if $B$ is normal.

14. (C-K Li) Given $A \in \mathbb{C}^{n\times n}$ we study its numerical range:

$$W(B) \equiv \{x^*Bx : x \in \mathbb{C}^n,\ \|x\| = 1\}. \tag{11.71}$$

    (i) Use the result of the previous exercise to prove that if $B$ is normal and $n = 2$ then $W(B)$ is the line segment in $\mathbb{C}$ between the two eigenvalues of $B$.

    (ii) Show that for every $B$ and scalar $a$ and $b$ that $W(aB - bI) = aW(B) - b$.

(iii) Given a nonnormal $B$ and $n = 2$ set $C = B - (\mathrm{tr}B)I/2$. Show that $\mathrm{tr}C = 0$ and conclude that the two eigenvalues of $C$ are equal and opposite, say $\pm\lambda$. If $\lambda = 0$ then use the Schur form to argue that $C$ is unitarily similar to

$$\begin{pmatrix} 0 & u \\ 0 & 0 \end{pmatrix}$$

and conclude that $W(C) = \{ux_1x_2 : x_1, x_2 \in \mathbb{C},\ |x_1|^2 + |x_2|^2 = 1\}$ is the circle about the origin of radius $|u|$. Conclude from (ii) that $W(B)$ is the circle of radius $u$ centered at $(\mathrm{tr}B)/2$.

(iv) Finally, if in part (iii) $\lambda \neq 0$ set $D = C/\lambda$ and use the Schur form to conclude that $D$ is unitarily similar to

$$E = \begin{pmatrix} 1 & 2c \\ 0 & -1 \end{pmatrix}$$

with $c > 0$.

15. Discuss Hessenberg and QR.

16. Use Cor. 11.8 to conclude that

$$\det \exp(B) = \exp(\mathrm{tr}(B)).$$

17. We have seen a natural means for attaching a polynomial to a matrix. We often have the need to go the other way round. Suppose

$$a(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots a_1x + a_0$$

and consider the **companion matrix**

$$C(a) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & & & & \vdots \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{pmatrix}$$

and prove that

$$\det(xI - C(a)) = a(x).$$

Hint: Row reduce $xI - C(a)$ and compute the product of its pivots. Note: This is how MATLAB computes roots of polynomials.

18. We will now use companion matrices and the Cayley–Hamilton Theorem to derive Newton's Identities between the roots $x_j$ and the coefficients of polynomial $p$, via the sums of powers

$$s_k = \sum_{j=1}^{n} x_j^k.$$

(i) Prove that $s_k = \mathrm{tr}(C(p)^k)$.

(ii) Suppose $k > n$ and use Cayley-Hamilton to conclude that $\mathrm{tr}(C^{k-n}p(C)) = 0$ then expand this identity to arrive at the high powered Newton Identity

$$s_k + a_{n-1}s_{k-1} + \cdots + a_0 = 0 \quad (k > n). \tag{11.72}$$

(iii) We can use $p(C) = 0$ to write $p(xI) = p(xI) - p(C)$. Expand and rearrange the right hand side to arrive at

$$p(x) = (x-C)[x^{n-1}+(C+a_{n-1})x^{n-2}+(C^2+a_{n-1}C+a_{n-2})x^{n-3}+\cdots+(C^{n-1}+a_{n-1}C^{n-2}+\cdots+a_1)].$$

(iv) Use part (iii) to arrive at

$$p(x)\mathrm{tr}(x-C)^{-1} = nx^{n-1} + \mathrm{tr}(C+a_{n-1})x^{n-2}+ \\ \mathrm{tr}(C^2+a_{n-1}C+a_{n-2})x^{n-3}+\cdots+\mathrm{tr}(C^{n-1}+a_{n-1}C^{n-2}+\cdots+a_1)$$

(v) Use part (iv) and Eq. (11.69) to conclude that

$$s_1x^{n-2}+(s_2+a_{n-1}s_1)x^{n-3}+\cdots+(s_{n-1}+a_{n-1}s_{n-2}+\cdots+a_2s_1) = -a_{n-1}x^{n-2}-2a_{n-2}x^{n-3}-\cdots-2a_2x-a_1.$$

Now identify coefficients of like powers and conclude that

$$s_k + a_{n-1}s_{k-1} + \cdots + a_{n-k+1}s_1 = -ka_{n-k} \quad (1 \le k \le n) \tag{11.73}$$

19. 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

$$B = F_4\mathrm{diag}(F_4^*B(:,1))F_4^*/4.$$

Why must we divide by 4? Now check the analogous formula on a circulant matrix in $\mathbb{R}^{5\times5}$ of your choice. Submit a marked-up diary of your computations.

20. Let us return to Exer. 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.

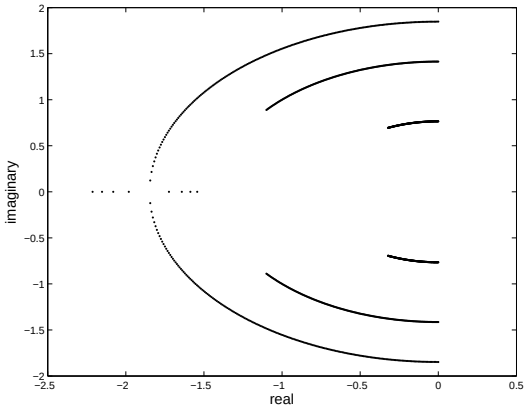


**Figure** 11.5. Trajectories of eigenvalues of the damped chain as the damping increased.

(ii) Argue that if $[u; v]^T$ is an eigenvector of $B$ with eigenvalue $\lambda$ 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 11.5 in light of your analysis.

21. Regarding Perron's Theorem, Prop. 11.10, please show that if every column sum of $B$ is one then $\lambda^+ = 1$.

22. Suppose that $B > 0$ and that $\lambda^+$ is its associated Perron eigenvalue. Show that the resolvent $(\lambda I - B)^{-1} > 0$ if and only if $\lambda > \lambda^+$. Hint: Eq. (11.7).

# 12. The Hermitian Eigenvalue Problem

In the previous chapter we arrived at the Spectral Representation of a general complex square matrix, and interpreted it as diagonal decomposition in the case that the matrix was semisimple. That condition was fairly technical and far from easy to check and so we devote an entire chapter to a very large, very important class of semisimple matrices, namely matrices that coincide with their conjugate transposes. For short, we call such matrices **Hermitian**. Note that real Hermitian matrices are symmetric.

We show that if $B = B^*$ then each eigenvalue, $\lambda_j$, is real, each eigenprojection, $P_j$, is Hermitian (and therefore orthogonal) and each eigennilpotent, $D_j$, vanishes. We saw a concrete example of this in Exer. 9.16.

We next show how to construct an orthonormal basis for each $\mathcal{R}(P_j)$, how to characterize the $\lambda_j$ as extrema of Rayleigh quotients and how to estimate the $\lambda_j$ via powers of $B$. We also study how eigenvalues move when the underlying matrix is perturbed and we close with applications to Molecular Orbital Theory and the optimal damping of mechanical networks.

### 12.1. The Spectral Representation

We establish the three key properties, one at a time.

**Proposition** 12.1. If $B = B^*$ then the eigenvalues of $B$ are real.

Proof: We suppose that $\lambda$ and $x$ comprise an eigenpair of $B$, i.e., $Bx = \lambda x$. On taking the conjugate transpose of each side we find

$$Bx = \lambda x \quad \text{and} \quad x^*B = \overline{\lambda}x^*$$

We now multiply the first by $x^*$ from the left and the second by $x$ from the right and find

$$x^*Bx = \lambda\|x\|^2 \quad \text{and} \quad x^*Bx = \overline{\lambda}\|x\|^2.$$

It follows that $\lambda\|x\|^2 = \overline{\lambda}\|x\|^2$ and, as $\|x\| > 0$, that $\lambda = \overline{\lambda}$. End of Proof.

In order to move onto the eigen–projections and –nilpotents it will help to recall that the conjugate transpose commutes with inversion. More precisely, the conjugate transpose of each side of the identity $A^*(A^*)^{-1} = I$ reveals $((A^*)^{-1})^*A = I$. That is $((A^*)^{-1})^* = A^{-1}$. Now the conjugate transpose of each side of this reveals

$$(A^*)^{-1} = (A^{-1})^*. \tag{12.1}$$

With this we can establish

**Proposition** 12.2. If $B$ is Hermitian then each eigenprojection, $P_j$, and each eigennilpotent, $D_j$, is Hermitian.

Proof: From (12.1) we learn that

$$\{(sI - B)^{-1}\}^* = \{(sI - B)^*\}^{-1} = (\overline{s}I - B)^{-1}$$

Next, as each eigenvalue is real we may choose each curve $C_j$ to be a circle centered on the real line. Hence

$$\begin{aligned} P_j^* &= \left( \frac{1}{2\pi i} \int_{C_j} (sI - B)^{-1}\, ds \right)^* = \frac{-1}{2\pi i} \int_{C_j} (\overline{s}I - B)^{-1}\, ds \\ &= \frac{1}{2\pi i} \int_{C_j} (sI - B)^{-1}\, ds = P_j, \end{aligned}$$

because integration of $\overline{s}$ merely reverses the curve's orientation. By the same token, we find that each $D_j^* = D_j$. End of Proof.

The next result will banish the nilpotent component.

**Proposition** 12.3. The zero matrix is the only Hermitian nilpotent matrix.

Proof: Suppose that $D = D^*$ 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 $x \in \mathbb{C}^n$ then

$$\begin{aligned} \|D^{m-1}x\|^2 &= x^*(D^{m-1})^* D^{m-1}x \\ &= x^* D^{m-1} D^{m-1} x \\ &= x^* D^{m-2} D^m x \\ &= 0. \end{aligned}$$

As $D^{m-1}x = 0$ for every $x$ it follows (recall Exer. 4.11) 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** 12.4. If $B$ is Hermitian then

$$\boxed{B = \sum_{j=1}^{h} \lambda_j P_j} \tag{12.2}$$

where the $\lambda_j$ are real and the $P_j$ are orthogonal projections that sum to the identity and whose pairwise products vanish.

The bulk of the symmetric matrices constructed in Chapters 2 and 3 were also positive definite, i.e., they obeyed

$$x^* B x > 0, \qquad \forall\ x \in \mathbb{C}^n.$$

It follows that the eigenvalues of such a matrix are positive and this in turn permits us to construct its **square root**,

**Proposition** 12.5. If $B \in \mathbb{R}^{n\times n}$ is Hermitian and positive definite with eigenvalues, $\lambda_j$, and eigenprojections, $P_j$, then

$$B^{1/2} \equiv \sum_{j=1}^{h} \sqrt{\lambda_j} P_j$$

is Hermitian and positive definite and obeys $B^{1/2}B^{1/2} = B$.

In applications one often is confronted with the so-called **generalized eigenproblem**

$$Su = \lambda Mu \tag{12.3}$$

where $S$ and $M$ are symmetric and $M$ is positive definite. For example, the dynamics of the chain Figure 3.1, with uniform stiffness, $k_1 = k_2 = k_3 = k_4 = 1$, but nonuniform mass, $m_j = j$ requires the solution of (12.3) with

$$S = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \quad \text{and} \quad M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

The eigenvalues in the this case are the poles of the generalized resolvent $(S - \lambda M)^{-1}$ or the zeros of the associated $\det(S - \lambda M)$. In this case

$$\lambda_1 = \frac{4 + \sqrt{10}}{3},\ \lambda_2 = 1,\ \lambda_3 = \frac{4 - \sqrt{10}}{3},$$

and the $j$th eigenvector spans $\mathcal{N}(S - \lambda_j M)$. Rather than producing an associated generalized spectral representation based on this generalized resolvent we instead note that Prop. 12.5 permits us to write (12.3) in standard form. Namely, from

$$Su = \lambda M^{1/2} M^{1/2} u$$

we multiply on the left by $M^{-1/2}$ and set $q = M^{1/2}u$ and arrive at the standard symmetric eigenproblem,

$$Bq = \lambda q \quad \text{where} \quad B = M^{-1/2} S M^{-1/2}. \tag{12.4}$$

For the example above

$$B = \begin{pmatrix} 2 & -1/\sqrt{2} & 0 \\ -1/\sqrt{2} & 1 & -1/\sqrt{6} \\ 0 & -1/\sqrt{6} & 2/3 \end{pmatrix}.$$

### 12.2. Orthonormal Diagonalization of Hermitian Matrices

It follows, as in §11.3, that each Hermitian matrix, $B$, has a full set of eigenvectors that may be used to diagonalize $B$, as in (11.32). As our $P_j$ are in fact *orthogonal* projections we may now argue that we may diagonalize $B$ with an orthonormal transformation. If $x \in \mathcal{R}(P_j)$ and $y \in \mathcal{R}(P_k)$ then

$$x^* y = (P_j x)^* P_k y = x^* P_j^* P_k y = x^* P_j P_k y = 0.$$

and hence the individual eigenbases

$$E_j \equiv [e_{j,1}\ e_{j,2}\ \ldots\ e_{j,n_j}]$$

are orthogonal to one another. In the simple case where each $n_j = 1$ we will have a full set of orthogonal eigenvectors. For example, we return to Exer. 9.16 and note that eigenvectors may be read from the columns of the associated projections. In particular

$$e_1 = \begin{pmatrix} 1 \\ -\sqrt{2} \\ 1 \end{pmatrix}, \qquad e_2 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \quad \text{and} \quad e_3 = \begin{pmatrix} 1 \\ \sqrt{2} \\ 1 \end{pmatrix}$$

are eigenvectors of

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

associated with the eigenvalues $\lambda_1 = 2+\sqrt{2}$, $\lambda_2 = 2$ and $\lambda_3 = 2-\sqrt{2}$. The $e_j$ are clearly orthogonal to one another, and upon simple normalization, i.e., $q_1 = e_1/2$, $q_2 = e_2/\sqrt{2}$ and $q_3 = e_3/2$ we note that the matrix

$$Q = [q_1, q_2, q_3] = \begin{pmatrix} 1/2 & 1/\sqrt{2} & 1/2 \\ -1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 1/2 & -1/\sqrt{2} & 1/2 \end{pmatrix}$$

enjoys both the lovely property $Q^*Q = I$ and

$$Q^*BQ = \Lambda = \begin{pmatrix} 2+\sqrt{2} & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2-\sqrt{2} \end{pmatrix},$$

the diagonal matrix of eigenvalues. A matrix $Q$ for which $Q^*Q = I$ is called **unitary**. A real unitary matrix is called orthogonal. A unitary square matrix is easily inverted, namely $Q^{-1} = Q^*$. We have in fact proven, and illustrated, that every Hermitian matrix with simple eigenvalues may be diagonalized by a unitary matrix of eigenvectors.

So let us move toward the general case with the loop matrix

$$B = \begin{pmatrix} 2 & -1 & 0 & -1 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ -1 & 0 & -1 & 2 \end{pmatrix} \tag{12.5}$$

and recognize in its resolvent

$$(sI - B)^{-1} = \frac{1}{s}P_1 + \frac{1}{s-2}P_2 + \frac{1}{s-4}P_3,$$

where

$$P_1 = \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \quad P_2 = \frac{1}{2}\begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix} \quad \text{and} \quad P_3 = \frac{1}{4}\begin{pmatrix} 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \end{pmatrix}$$

and eigenvalue $\lambda_1 = 0$ with a 1–dimensional eigenspace, an eigenvalue $\lambda_2 = 2$ with a 2–dimensional eigenspace, and an eigenvalue $\lambda_3 = 4$ with a 1–dimensional eigenspace. One easy way of seeing the dimensions is via $n_j = m_j = \mathrm{tr}(P_j)$. In any event, we proceed as above to read eigenvectors from the first $n_j$ columns of $P_j$. In particular, $E = [E_1, E_2, E_3]$ where

$$E_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \qquad E_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 0 \\ 0 & -1 \end{pmatrix} \quad \text{and} \quad E_3 = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}.$$

As expected the $E_j^T E_k = 0$ when $j \neq k$. As a perhaps lucky bonus we recognize that the two columns of $E_2$ are also orthogonal to one another. As a result the normalized collection

$$Q = [E_1/2, E_2/\sqrt{2}, E_3/2]$$

is an orthonormal diagonalizer of the $B$ in (12.5). It remains only to show that we can always engineer such luck.

By constructing an orthogonal basis $\{q_{j,k} : 1 \leq k \leq n_j\}$ for each $\mathcal{R}(P_j)$, collecting the basis vectors in $Q_j = [q_{j,1}\ q_{j,2}\ \cdots\ q_{j,n_j}]$ and assembling the $Q_j$ into a single square unitary matrix $Q = [Q_1\ \cdots\ Q_h]$, we arrive at

$$\boxed{B = Q\Lambda Q^* \quad \text{and} \quad Q^*BQ = \Lambda} \tag{12.6}$$

where $\Lambda$ is the diagonal matrix of eigenvalues of $B$, ordered as in the construction of $Q$, and repeated by their multiplicity.

Finally, let us return to the generalized eigenproblem, (12.3). Recall that $\lambda$ and $u$ are an eigenpair for $(S, M)$ if and only if $\lambda$ and $q = M^{1/2}u$ are an eigenpair for $B = M^{-1/2}SM^{-1/2}$. The orthonormality of $Q = (q_1, q_2, \ldots, q_n)$ then implies that the matrix of $(S, M)$ eigenvectors, $U = (u_1, u_2, \ldots, u_n)$, obeys

$$I = Q^*Q = (M^{1/2}U^*)M^{1/2}U = U^*MU, \tag{12.7}$$

i.e., $U$ is unitary with weight matrix $M$.

### 12.3. Perturbation Theory*

It is often the case in practice that one can solve the "base" instantiation of a problem but wishes to know how the solution changes if one perturbs the problem. In our setting we ask how the eigenvalues change when we change the underlying matrix. More precisely, given Hermitian $B$ and $C$ and a real $\varepsilon$ we attempt to expand the eigenvalues of $B + \varepsilon C$ in a Taylor series in $\varepsilon$. To lessen the notation (and with an eye to future applications) we develop the theory for the greatest eigenvalue

$$\lambda_1(B + \varepsilon C) = \lambda_1(B) + \varepsilon d_1 + \varepsilon^2 d_2/2 + \cdots$$

where our calculus suggests that $d_k$ should denote the $k$th derivative of $\lambda_1$ at $B$ in the direction $C$. First some examples. If $C = I$ and $Bq_1 = \lambda_1 q_1$ then $(B + \varepsilon I)q_1 = (\lambda_1 + \varepsilon)q_1$ and so

$$\lambda_1(B + \varepsilon I) = \lambda_1(B) + \varepsilon$$

and so clearly $d_1 = 1$ and the rest of $d_k = 0$. Similarly, if $C = B$ then

$$\lambda_1(B + \varepsilon B) = (1 + \varepsilon)\lambda_1(B) = \lambda_1(B) + \varepsilon\lambda_1(B)$$

and so $d_1 = \lambda_1(B)$ and again the remaining $d_k = 0$. We could generalize this to those $C$ that are multiples of $I$ or $B$, but this still remains too sparse a sampling of the direction space. Our third example exposes the key obstacle. The eigenvalues of

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \varepsilon \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{12.8}$$

are $1 + \varepsilon$ and $1 - \varepsilon$ which appear as banal as our two previous examples – until we order them. For if $\lambda_1$ is the largest and $\lambda_2$ the smallest we find

$$\lambda_1(B + \varepsilon C) = 1 + |\varepsilon| \quad \text{and} \quad \lambda_2(B + \varepsilon C) = 1 - |\varepsilon|. \tag{12.9}$$

Hence, a diagonal perturbation of a diagonal matrix can produce eigenvalues that do not vary smoothly with $\varepsilon$. The culprit here is the multiplicity of $\lambda_1(I)$, for as it is 2 we find that it splits when $I$ is perturbed. Although there are means to accommodate such splitting we will assume here that $\lambda_1$ of the unperturbed matrix is simple.

To resolve the general case we return to our source and develop the resolvent

$$R(s,\varepsilon) = (sI - (B + \varepsilon C))^{-1}$$

in a Taylor series in $\varepsilon$. Namely,

$$\begin{aligned} R(s,\varepsilon) &= ((sI - B) - \varepsilon C(sI - B)^{-1}(sI - B))^{-1} \\ &= (sI - B)^{-1}(I - \varepsilon C(sI - B)^{-1})^{-1} \\ &= R(s,0)(I - \varepsilon CR(s,0))^{-1} \\ &= R(s,0)\sum_{j=0}^{\infty} \varepsilon^j (CR(s,0))^j \\ &= R(s,0) + \varepsilon R(s,0)CR(s,0) + O(\varepsilon^2). \end{aligned}$$

Regarding the associated eigenprojection we note that $\lambda_1(B + \varepsilon C)$ is a pole of $s \mapsto R(s,\varepsilon)$. and a zero of the characteristic polynomial, $s \mapsto \chi_B(s,\varepsilon) = \det(sI - (B + \varepsilon C))$. This is a polynomial in $s$ with coefficients that are polynomials in $\varepsilon$. It follows from Prop. 10.9 that there exists a small circle, $C_1$, about $\lambda_1(B)$ that both includes $\lambda_1(B + \varepsilon C)$ and excludes $\lambda_j(B + \varepsilon C)$ for $j > 1$ and all $\varepsilon$ in some interval about 0. As a result, the perturbed eigenprojection is

$$\begin{aligned} P_1(B + \varepsilon C) &= \frac{1}{2\pi i}\int_{C_1} R(s,\varepsilon)\, ds \\ &= \frac{1}{2\pi i}\int_{C_1} (R(s,0) + \varepsilon R(s,0)CR(s,0) + O(\varepsilon^2))\, ds \\ &= P_1(B) + \varepsilon \frac{1}{2\pi i}\int_{C_1} R(s,0)CR(s,0)\, ds + O(\varepsilon^2). \end{aligned}$$

We now turn to the partial fraction expansion of $R(s,0)$ to achieve

$$\begin{aligned} \frac{1}{2\pi i}\int_{C_1} R(s,0)CR(s,0)\, ds &= \frac{1}{2\pi i}\int_{C_1} \sum_{j=1}^{h} \frac{P_j(B)}{s - \lambda_j(B)} C \sum_{k=1}^{h} \frac{P_k(B)}{s - \lambda_k(B)}\, ds \\ &= \frac{1}{2\pi i}\int_{C_1} \frac{P_1(B)}{s - \lambda_1(B)} C \sum_{k=2}^{h} \frac{P_k(B)}{s - \lambda_k(B)} + \sum_{j=2}^{h} \frac{P_j(B)}{s - \lambda_j(B)} C \frac{P_1(B)}{s - \lambda_1(B)}\, ds \\ &= P_1(B) C \sum_{k=2}^{h} \frac{P_k(B)}{\lambda_1(B) - \lambda_k(B)} + \sum_{k=2}^{h} \frac{P_k(B)}{\lambda_1(B) - \lambda_k(B)} C P_1(B). \end{aligned}$$

It follows that

$$P_1(B + \varepsilon C) = P_1(B) + \varepsilon(P_1(B)CS_1 + S_1CP_1(B)) + O(\varepsilon^2) \tag{12.10}$$

where

$$S_1 \equiv \sum_{k=2}^{h} \frac{P_k(B)}{\lambda_1(B) - \lambda_k(B)},$$

is the so–called reduced resolvent. In order to derive the perturbation series for $\lambda_1(B+\varepsilon C)$ we take the trace of each side of

$$(B+\varepsilon C)P_1(B+\varepsilon C) = \lambda_1(B+\varepsilon C)P_1(B+\varepsilon C).$$

Namely

$$\mathrm{tr}((B+\varepsilon C)P_1(B+\varepsilon C)) = \lambda_1(B+\varepsilon C)\mathrm{tr}(P_1(B+\varepsilon C)) = \lambda_1(B+\varepsilon C). \tag{12.11}$$

where we have used the fact that $\lambda_1(B+\varepsilon C)$ is simple and that the trace of a projection is its rank. We now develop the left side of (12.11).

$$\begin{aligned}(B+\varepsilon C)P_1(B+\varepsilon C) &= (B+\varepsilon C)(P_1(B)+\varepsilon(P_1(B)CS_1+S_1CP_1(B))+O(\varepsilon^2)) \\ &= \lambda_1(B)P_1(B)+\varepsilon(CP_1(B)+B(P_1(B)CS_1+S_1CP_1(B)))+O(\varepsilon^2)\end{aligned} \tag{12.12}$$

We note that $P_1(B)S_1 = S_1P_1(B) = 0$ will cause the $S_1$ terms in (12.12) to vanish upon taking the trace. In particular

$$\mathrm{tr}(BP_1(B)CS_1) = \mathrm{tr}(BP_1(B)S_1C) = 0 \quad \text{and} \quad \mathrm{tr}(BS_1CP_1(B)) = \mathrm{tr}(BS_1P_1(B)C) = 0.$$

Hence, the trace of (12.12) reveals

$$\lambda_1(B+\varepsilon C) = \mathrm{tr}((B+\varepsilon C)P_1(B+\varepsilon C)) = \lambda_1(B)+\varepsilon\mathrm{tr}(CP_1(B))+O(\varepsilon^2). \tag{12.13}$$

or, on recalling that $P_1(B) = q_1q_1^*$ is just projection onto the first eigenvector, that

**Proposition** 12.6. If $B$ and $C$ are Hermitian and $\lambda_1(B)$, the greatest eigenvalue of $B$, is simple then

$$\boxed{\lambda_1(B+\varepsilon C) = \lambda_1(B)+q_1^*Cq_1\varepsilon+O(\varepsilon^2).} \tag{12.14}$$

where $q_1 = q_1(B)$ is the associated eigenvector of $B$.

As a very simple example, you might wish to confirm that the largest eigenvalue of

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

behaves like $3+\varepsilon+O(\varepsilon^2)$. We will develop much richer examples in sections and exercises to come. One such example will require an understanding the generalized perturbation problem

$$(S+\varepsilon X)u = \lambda Mu$$

where $M$ is Hermitian and positive definite. As with (12.3) we may return this to the standard perturbation problem

$$(B+\varepsilon C)q = \lambda q$$

via $q = M^{1/2}u$, $B = M^{-1/2}SM^{-1/2}$ and $C = M^{-1/2}XM^{-1/2}$. In this case the key term in the perturbation expansion is

$$q_1^*Cq_1 = (M^{1/2}u_1)^*M^{-1/2}XM^{-1/2}M^{1/2}u_1 = u_1^*Xu_1.$$

As a result,

**Proposition** 12.7. If $S,\ M$ and $X$ are Hermitian and $M$ is positive definite and $\lambda_1(S,M)$, the greatest eigenvalue of $(S,M)$, is simple then

$$\boxed{\lambda_1(S+\varepsilon X, M) = \lambda_1(S,M) + u_1^* X u_1 \varepsilon + O(\varepsilon^2).} \tag{12.15}$$

where $u_1 = u_1(S,M)$ is the associated eigenvector of $(S,M)$.

### 12.4. Rayleigh's Principle and the Power Method*

As the eigenvalues of a Hermitian matrix, $B \in \mathbb{R}^{n\times n}$, are real we may order them from high to low,

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n. \tag{12.16}$$

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

$$q_1,\ q_2,\ \ldots,\ q_n \tag{12.17}$$

and note that each $x \in \mathbb{R}^n$ enjoys the expansion

$$x = (x^*q_1)q_1 + (x^*q_2)q_2 + \cdots + (x^*q_n)q_n. \tag{12.18}$$

Applying $B$ to each side we find

$$Bx = (x^*q_1)\lambda_1 q_1 + (x^*q_2)\lambda_2 q_2 + \cdots + (x^*q_n)\lambda_n q_n. \tag{12.19}$$

Now taking the inner product of (12.18) and (12.19) we find

$$\begin{aligned} x^*Bx &= (x^*q_1)^2\lambda_1 + (x^*q_2)^2\lambda_2 + \cdots + (x^*q_n)^2\lambda_n \\ &\leq \lambda_1\{(x^*q_1)^2 + (x^*q_2)^2 + \cdots + (x^*q_n)^2\} \\ &= \lambda_1 x^*x. \end{aligned}$$

That is, $x^*Bx \leq \lambda_1 x^*x$ for every $x \in \mathbb{R}^n$. This, together with the fact that $q_1^*Bq_1 = \lambda_1 q_1^*q_1$ establishes

**Proposition** 12.8. **Rayleigh's Principle.** If $B$ is Hermitian then its largest eigenvalue is

$$\lambda_1 = \max_{x\neq 0} \frac{x^*Bx}{x^*x}$$

and the maximum is attained on the line through the associated eigenvector, $q_1$.

As Rayleigh's Principle identifies only the principle "direction" it is often presumed that the maximum is taken over unit vectors, i.e., Rayleigh's Principle is written

$$\lambda_1 = \max_{x^*x=1} x^*Bx. \tag{12.20}$$

As unit vectors in the real plane are specially easy to write, namely $x = x(\theta) = (\cos(\theta), \sin(\theta))^T$, we may visualize the Rayleigh quotient for real 2–by–2 matrices. For example, if

$$B = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \tag{12.21}$$

then $x(\theta)^T Bx(\theta) = 2 - \sin(2\theta)$ clearly assumes its maximum of 3 at $\theta = \pi/4$ and its minimum of 1 when $\theta = 3\pi/4$. Confirm that 3 and 1 are the eigenvalues of $B$ and that their associated eigendirections are $\pi/4$ and $3\pi/4$. With little extra work we can prove

**Proposition** 12.9. **Minimax Principle.** If $B$ is Hermitian with eigenvalues and eigenvectors in (12.16) and (12.17) then

$$\lambda_k = \max_{\dim(W)=k} \min_{0\neq x\in W} \frac{x^* Bx}{x^* x}, \tag{12.22}$$

where the $W$ that appear in the maximum are subspaces of $\mathbb{C}^n$.

Proof: Given a subspace $W \subset \mathbb{C}^n$ we define

$$\lambda(W) \equiv \min_{0\neq x\in W} \frac{x^* Bx}{x^* x},$$

and proceed to show that $\lambda_k \geq \lambda(W)$ when $\dim(W) = k$ and that there exists a subspace, $W_k$, of dimension $k$ for which $\lambda_k \leq \lambda(W_k)$. Now, as $\dim(W) = k$ and $\dim(\mathrm{sp}\{q_k, q_{k+1}, \ldots, q_n\}) = n-k+1$ it follows that their intersection has dimension at least 1. If $x$ is a nonzero vector in their intersection then

$$x = \sum_{j=k}^{n} (x^* q_j) q_j \quad \text{and} \quad x^* Bx = \sum_{j=k}^{n} x^* (x^* q_j) Bq_j = \sum_{j=k}^{n} (x^* q_j)^2 \lambda_j \leq \lambda_k \sum_{j=k}^{n} (x^* q_j)^2 = \lambda_k x^* x$$

and so

$$\lambda(W) \leq \frac{x^* Bx}{x^* x} = \lambda_k.$$

Conversely, *for every* $x \in W_k = \mathrm{sp}\{q_1, q_2, \ldots, q_k\}$ we find

$$x^* Bx = \sum_{j=1}^{k} x^* (x^* q_j) Bq_j = \sum_{j=1}^{k} (x^* q_j)^2 \lambda_j \geq \lambda_k \sum_{j=1}^{k} (x^* q_j)^2 = \lambda_k x^* x$$

and so $\lambda(W_k) \geq \lambda_k$. End of Proof.

For our next characterization we return to (12.19) and record higher powers of $B$ onto $x$

$$\begin{aligned} B^k x &= (x^* q_1)\lambda_1^k q_1 + (x^* q_2)\lambda_2^k q_2 + \cdots + (x^* q_n)\lambda_n^k q_n \\ &= (x^* q_1)\lambda_1^k \left( q_1 + \frac{x^* q_2}{x^* q_1} \frac{\lambda_2^k}{\lambda_1^k} q_2 + \cdots + \frac{x^* q_n}{x^* q_1} \frac{\lambda_n^k}{\lambda_1^k} q_n \right). \end{aligned} \tag{12.23}$$

And so, using $\mathrm{sign}(t) \equiv t/|t|$,

$$\frac{B^k x}{\|B^k x\|} = \mathrm{sign}(\lambda_1^k x^* q_1) q_1 + O((\lambda_2/\lambda_1)^k)$$

and note that the latter term goes to zero with increasing $k$ so long as $\lambda_1$ is strictly greater than $\lambda_2$. If, in addition, we assume that $\lambda_1 > 0$, then the first term does not depend on $k$ and we arrive at

**Proposition** 12.10. **The Power Method.** If $B$ is Hermitian and its greatest eigenvalue is simple and positive and the initial guess, $x$, is not orthogonal to $q_1$ then

$$\lim_{k\to\infty} \frac{B^k x}{\|B^k x\|} = \text{sign}(x^* q_1) q_1$$

By way of illustration, we record $B^k x/\|B^k x\|$ for the $B$ of (12.21), a random $x$, and $k = 1$ through 7,

$$\begin{pmatrix} 0.97404 \\ 0.22635 \end{pmatrix}, \begin{pmatrix} 0.95709 \\ -0.28980 \end{pmatrix}, \begin{pmatrix} 0.82030 \\ -0.57194 \end{pmatrix}, \begin{pmatrix} 0.74783 \\ -0.66389 \end{pmatrix}, \begin{pmatrix} 0.72098 \\ -0.69296 \end{pmatrix}, \begin{pmatrix} 0.71176 \\ -0.70242 \end{pmatrix}, \begin{pmatrix} 0.70866 \\ -0.70555 \end{pmatrix}.$$

This is indeed approaching $q_1 = [1, -1]/\sqrt{2}$. As with the Rayleigh quotient it is a simple matter to arrange things as to converge on the eigenvector associated with the smallest eigenvalue. To see this we recall that $1/\lambda_n$ is the largest eigenvalue of $B^{-1}$ and hence $B^{-k}x/\|B^{-k}x\|$ ought to converge to $q_n$. The associated method is called **Inverse Iteration.**

### 12.5. Hückel's Molecular Orbital Theory*

Perhaps the most systematic application of the Minimax Principle has been to the determination of electronic structure of atoms and molecules. We will develop the necessary tools in the context of benzene, $C_6H_6$. It is a planar molecule, Figure 12.1, comprised of 6 carbon atoms and 6 hydrogen happens, coupled by 12 bonds. The single electron of each hydrogen atom lies in its $1s$ orbital. Each carbon has 2 electrons in its $1s$ orbital, 2 electrons in its $2s$ orbital and 2 electrons in its $2p$ orbital. Of these 4 outer carbon electrons it is believed that three are devoted to planar bounds with its three neighbors. This leaves one free electron per carbon to occupy the associated $2p$ out-of-plane orbital, Figure 12.1(B), and to supposedly interact with neighboring molecules.

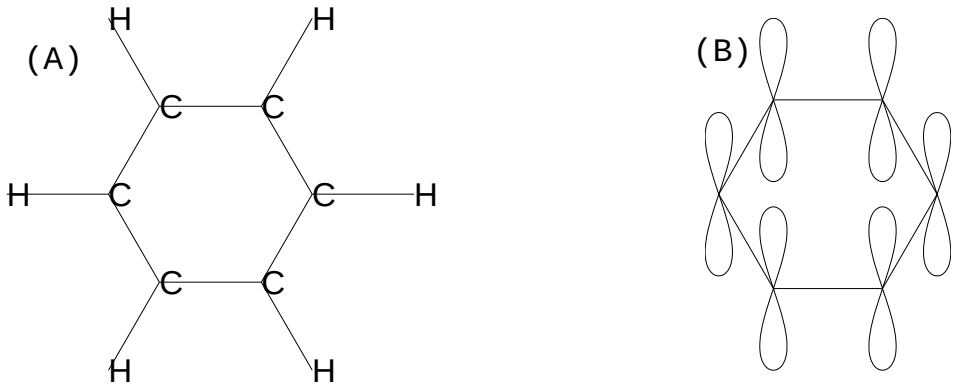



**Figure** 12.1 The structure of benzene. (A) The planar disposition of its six carbons and six hydrogens. (B) The $\pi$-orbitals if its six $\pi$-electrons.

The theory of Hückel provides a means to predict the ground state of these, so-called, $\pi$-orbitals. This ground state is derived from the Schrödinger equation (recall Eq. (9.73))

$$\begin{aligned} -\frac{h^2}{8\pi^2 m}\Delta\Psi(x,y,z,t) + V(x,y,z)\Psi(x,y,z,t) &= \frac{ih}{2\pi}\frac{\partial\Psi(x,y,z,t)}{\partial t} \\ \text{where} \quad \Delta\Psi \equiv \frac{\partial^2\Psi(x,y,z,t)}{\partial x^2} &+ \frac{\partial^2\Psi(x,y,z,t)}{\partial y^2} + \frac{\partial^2\Psi(x,y,z,t)}{\partial z^2}, \end{aligned} \tag{12.24}$$

and $V$ is the potential associated with the six $\pi$-electrons of benzene. We begin our reduction of (12.24) by supposing the wave function, $\Psi$, to have the wave-like time dependence

$$\Psi(x, y, z, t) = \exp(-2\pi i E t/h)\psi(x, y, z). \tag{12.25}$$

On substituting this assumption into (12.24) we find that $\psi$ must obey the eigenvalue problem

$$\mathcal{H}\psi = E\psi \quad \text{where} \quad \mathcal{H} = \frac{-h^2}{8\pi^2 m}\Delta + V. \tag{12.26}$$

Here $E$ denotes energy and we define the **ground state** to be the $\psi$ function associated with the least eigenvalue of (12.26). As $\mathcal{H}$ is indeed a real symmetric linear transformation this least eigenvalue may be characterized by the Rayleigh Principle

$$E_1 = \min_{\psi} \frac{\langle \mathcal{H}\psi, \psi\rangle}{\langle \psi, \psi\rangle} \tag{12.27}$$

where the inner product is the full space integral

$$\langle \psi, \phi\rangle \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi(x, y, z)\phi(x, y, z)\, dxdydz.$$

The beauty of the Hückel method is that it reduces the formidable, infinite dimensional, eigenproblem (12.27) to an eigenvalue problem for the 6-by-6 adjacency matrix associated with benzene's 6 carbons.

To see this we presume that each candidate molecular orbital, $\tilde{\psi}$, is a linear combination of atomic orbitals, $\phi_i$. In particular, we assume that

$$\tilde{\psi} = \sum_{i=1}^{6} c_i\phi_i \tag{12.28}$$

where each $c_i \in \mathbb{R}$. Furthermore, if the atomic orbitals are assumed orthonormal then the denominator in Eq. (12.27) takes the familiar form

$$\langle \tilde{\psi}, \tilde{\psi}\rangle = \sum_{i=1}^{6} c_i^2 = c^T c.$$

Regarding the numerator of Eq. (12.27) it remains to specify $\langle \mathcal{H}\phi_i, \phi_j\rangle$. First off, $\langle \mathcal{H}\phi_i, \phi_i\rangle$ denotes the average energy of the $\pi$–orbital at the $i$th carbon. As all carbons are identical we can expect

$$\langle \mathcal{H}\phi_i, \phi_i\rangle = \alpha$$

for each $i$. (A typical value is $\alpha = -11.2$ eV.) Next, if atoms $i$ and $j$ are not adjacent then we may reasonably expect that their atomic orbitals do not interact and so

$$\langle \mathcal{H}\phi_i, \phi_j\rangle = 0 \quad \text{if atoms } i \text{ and } j \text{ are not adjacent.}$$

Finally, the interaction energy between adjacent orbitals, will be denoted $\beta$. (A typical value for $\beta$ is $-0.7$ eV.) With these approximations, the difficult numerator takes the form

$$\langle \mathcal{H}\tilde{\psi}, \tilde{\psi}\rangle = c^T(\alpha I + \beta A)c$$

where

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{12.29}$$

is the adjacency matrix of the hexagon of carbons. The first row says, e.g., that the first carbon is adjacent to the second and the sixth. Hence, invoking Rayleigh's Principle,

$$E_1 \le \frac{\langle \mathcal{H}\tilde{\psi}, \tilde{\psi}\rangle}{\langle \tilde{\psi}, \tilde{\psi}\rangle} = \frac{c^T(\alpha I + \beta A)c}{c^T c}, \quad \forall c \in \mathbb{R}^6.$$

Hence minimizing over $c$ brings

$$E_1 \le \tilde{E}_1 = \min_{c\in\mathbb{R}^6} \frac{c^T(\alpha I + \beta A)c}{c^T c}. \tag{12.30}$$

It follows that $\tilde{E}_1 = \alpha + \beta\lambda_1$ where $\lambda_1$ is the greatest (recall that $\beta < 0$) eigenvalue of $A$.

It remains only to work out the eigenvalues and vectors of $A$. From its resolvent

$$(sI - A)^{-1} = \frac{1}{s^4 - 5s^2 + 4} \begin{pmatrix} s^3 - 3s & s^2 - 2 & s & 2 & s & s^2 - 2 \\ s^2 - 2 & s^3 - 3s & s^2 - 2 & s & 2 & s \\ s & s^2 - 2 & s^3 - 3s & s^2 - 2 & s & 2 \\ 2 & s & s^2 - 2 & s^3 - 3s & s^2 - 2 & s \\ s & 2 & s & s^2 - 2 & s^3 - 3s & s^2 - 2 \\ s^2 - 2 & s & 2 & s & s^2 - 2 & s^3 - 3s \end{pmatrix}$$

we recognize that $s^4 - 5s^2 + 4 = (s^2 - 4)(s^2 - 1)$ and so the eigenvalues of $A$ are

$$\lambda_1 = 2, \quad \lambda_2 = 1, \quad \lambda_3 = -1 \quad \text{and} \quad \lambda_4 = -2. \tag{12.31}$$

The resulting partial fraction expansion

$$(sI - A)^{-1} = \frac{1}{s+2}P_4 + \frac{1}{s+1}P_3 + \frac{1}{s-1}P_2 + \frac{1}{s-2}P_1$$

where

$$P_4 = \frac{1}{6}\begin{pmatrix} 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \end{pmatrix} \qquad P_3 = \frac{1}{6}\begin{pmatrix} 2 & -1 & -1 & 2 & -1 & -1 \\ -1 & 2 & -1 & -1 & 2 & -1 \\ -1 & -1 & 2 & -1 & -1 & 2 \\ 2 & -1 & -1 & 2 & -1 & -1 \\ -1 & 2 & -1 & -1 & 2 & -1 \\ -1 & -1 & 2 & -1 & -1 & 2 \end{pmatrix}$$

$$P_2 = \frac{1}{6}\begin{pmatrix} 2 & 1 & -1 & -2 & -1 & 1 \\ 1 & 2 & 1 & -1 & -2 & -1 \\ -1 & 1 & 2 & 1 & -1 & -2 \\ -2 & -1 & 1 & 2 & 1 & -1 \\ -1 & -2 & -1 & 1 & 2 & 1 \\ 1 & -1 & -2 & -1 & 1 & 2 \end{pmatrix} \qquad P_1 = \frac{1}{6}\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

reveals the associated eigenprojections. Their traces expose the associated geometric multiplicities, (1,2,2,1), and permit us to pluck the associated eigenvectors from the first one or two columns of each projection. In particular,

$$E_4 = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} \quad E_3 = \begin{pmatrix} 2 & -1 \\ -1 & 2 \\ -1 & -1 \\ 2 & -1 \\ -1 & 2 \\ -1 & -1 \end{pmatrix} \quad E_2 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \\ -1 & 1 \\ -2 & -1 \\ -1 & -2 \\ 1 & -1 \end{pmatrix} \quad E_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

We orthonormalize these to

$$Q_4 = \frac{1}{\sqrt{6}}\begin{pmatrix}1\\-1\\1\\-1\\1\\-1\end{pmatrix} \quad Q_3 = \frac{1}{\sqrt{3}}\begin{pmatrix}1\\-1/2\\-1/2\\1\\-1/2\\-1/2\end{pmatrix}, \begin{pmatrix}0\\1/2\\-1/2\\0\\1/2\\-1/2\end{pmatrix} \quad Q_2 = \frac{1}{\sqrt{3}}\begin{pmatrix}1\\1/2\\-1/2\\-1\\-1/2\\1/2\end{pmatrix}, \begin{pmatrix}0\\1/2\\1/2\\0\\-1/2\\-1/2\end{pmatrix} \quad Q_1 = \frac{1}{\sqrt{6}}\begin{pmatrix}1\\1\\1\\1\\1\\1\end{pmatrix} \tag{12.32}$$

and illustrate the energy levels and eigenmodes in Figure 12.2. Recall that the true energy levels are $\tilde{E}_j = \alpha + \beta\lambda_j$. In order to achieve the least total energy, we place $\pi$-electrons, of alternate spins, at the lowest $\tilde{E}_j$ (including multiplicity). In our case these leads to 2 electrons at $\lambda_1$ and 4 electrons at $\lambda_2$ for a total energy of $6\alpha + 8\beta$.

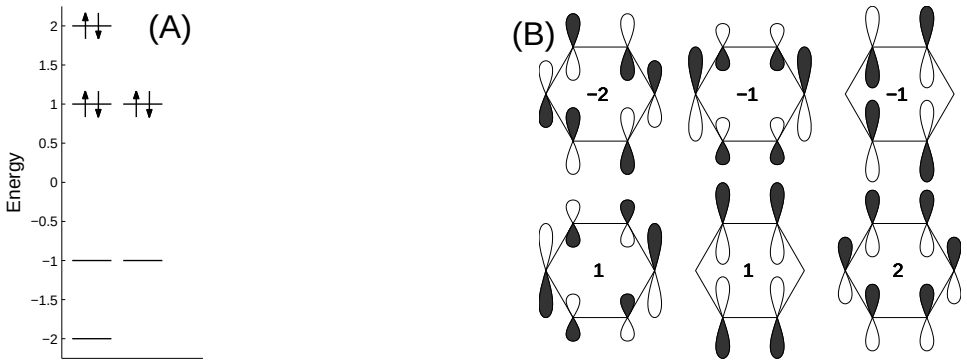



**Figure** 12.2 The $\pi$ energies and orbitals of benzene. (A) We place a pair of electrons at each energy level, starting with the greatest, until there are no electrons remaining. (B) The eigenvectors, (12.32), dictate the relative contribution of the atomic orbital (the $c_i$ coefficients in (12.28)) to the associated molecular orbital. Positive coefficients are black on top. The number at the center of each molecule is the associated eigenvalue of $A$.

The upshot is that the molecular orbitals of $\pi$-electrons of hydrocarbons are completely determined by the eigenvalues and vectors of the molecule's carbon adjacency matrix. As such, quantum chemists have devoted considerable energy to understanding the eigenstructure of adjacency matrices, particularly in molecules with symmetry. We will pursue this briefly in the exercises and then much more fully in Chapter 15 where we use group representation theory to find, by hand, all 60 eigenvalues of the "Buckyball."

### 12.6. Optimal Damping of Mechanical Networks*

We revisit our work, from §8.4, on dynamics of mechanical networks. We close up loose ends, but this is not mere reconciliation, we go considerably beyond. We consider the initial value problem for $u(t) \in \mathbb{R}^n$

$$\begin{gathered} Mu''(t) + Du'(t) + Su(t) = 0, \quad t > 0, \\ u(0) = f,\ u'(0) = g, \end{gathered} \tag{12.33}$$

where $' \equiv d/dt$ and $M,\ D$, and $S$ are each symmetric and positive definite. It will be convenient to consider an equivalent first order system. In particular, with $V(t)$ denoting the $2n - by - 1$ vector

$[u(t),\ u'(t)]$ we find that (12.33) becomes

$$V'(t) = A(D)V(t), \quad V(0) = \begin{pmatrix} f \\ g \end{pmatrix}, \quad A(D) = \begin{pmatrix} 0 & I \\ -M^{-1}S & -M^{-1}D \end{pmatrix} \tag{12.34}$$

We suppose the mass, $M$, and stiffness, $S$, to be prescribed and we venture to produce a damping $D$ for which the energy induced by the initial displacement, $f$, and velocity, $g$, is dissipated as efficiently as possible. The energy we have in mind is simply the sum of the kinetic and potential instantaneous energies, namely

$$E(t) \equiv u'(t)^T M u'(t) + u(t)^T S u(t). \tag{12.35}$$

A more natural measure of damping efficiency is the total energy

$$\int_0^\infty E(t)\, dt.$$

Let us try to make their dependence on $D$ more explicit. As $V(t) = \exp(A(D)t)V(0)$ it follows that

$$E(t) = V(0)^T \exp(A(D)^T t) F \exp(A(D)t) V(0) \quad \text{where} \quad F = \begin{pmatrix} S & 0 \\ 0 & M \end{pmatrix}.$$

As a result, the total energy is

$$\int_0^\infty E(t)\, dt = V(0)^T \mathcal{E}(D) V(0), \quad \text{where} \quad \mathcal{E}(D) = \int_0^\infty \exp(A(D)^T t) F \exp(A(D)t)\, dt. \tag{12.36}$$

One obvious defect with this measure of damping efficiency is its dependence on the initial state, $V(0) = [f,\ g]$. We remedy this by maximizing the above over all initial states of unit energy. That is, we consider the **greatest total energy**

$$\tau_1(D) \equiv \max_{V(0)^T F V(0) = 1} V(0)^T \mathcal{E}(D) V(0). \tag{12.37}$$

One recognizes in (12.37) Rayleigh's Generalized Principle (Exer. 12.9) for the greatest eigenvalue of the pair $(\mathcal{E}(D), F)$. That is, $\tau_1(D)$ is the greatest eigenvalue of the generalized problem

$$\mathcal{E}(D)V = \tau F V. \tag{12.38}$$

With $\mathcal{E}(D)$ as derived in (12.36) this remains a formidable eigenproblem. Over the next few pages we will restrict the damping matrix to natural subspaces in which $\mathcal{E}(D)$ may be expressed as products of the system matrices, $M, K, D$, and their inverses.

The most restricted class is that of so-called **friction damping** where we suppose the damping to be proportional to the mass. The key to this entire section is the following energy equation.

**Proposition** 12.11. If $D = 2aM$ then

$$E(t) = -\frac{d}{dt}\left\{ u(t)^T M u'(t) + a u(t)^T M u(t) + \frac{1}{2a} E(t) \right\}. \tag{12.39}$$

Proof: As $u$ satisfies (12.33) we find

$$E'(t) = 2(Mu''(t) + Su(t))^T u' = -4a u'(t)^T M u'(t). \tag{12.40}$$

Continuing to use (12.33),

$$\begin{aligned} E(t) &= u'(t)^T M u'(t) + u(t)^T S u(t) \\ &= u'(t)^T M u'(t) - u(t)^T M u''(t) - 2a u(t)^T M u'(t) \\ &= u'(t)^T M u'(t) - \{(u(t)^T M u'(t))' - u'(t)^T M u'(t)\} - a(u(t)^T M u(t))' \\ &= -(u(t)^T M u'(t))' - a(u(t)^T M u(t))' - \frac{1}{2a} E'(t), \end{aligned}$$

where in the final line we have used (12.40). End of Proof.

Integrating $E$, this proposition yields

$$\begin{aligned} \int_0^t E(s)\, ds &= -\left\{ u(s)^T M u'(s) + a u(s)^T M u(s) + \frac{1}{2a} E(s) \right\}_{s=0}^{s=t} \\ &= f^T M g + a f^T M f + \frac{1}{2a} E(0) - u(t)^T M u'(t) - a u(t)^T M u(t) - \frac{1}{2a} E(t). \end{aligned}$$

Now if $a > 0$ then $u(t)$, $u'(t)$, and $E(t)$ tend to zero as $t \to \infty$. As a result,

**Corollary** 12.12. If $D = 2aM$ then the total energy is

$$\int_0^\infty E(t)\, dt = \frac{E(0)}{2a} + f^T M(af + g). \tag{12.41}$$

If we now reconcile (12.41) with (12.36) we find that

$$\mathcal{E}(2aM) = \begin{pmatrix} aM + S/(2a) & M/2 \\ M/2 & M/(2a) \end{pmatrix}. \tag{12.42}$$

Hence, if $V = [u, v]$ is an eigenvector of $(\mathcal{E}(2aM), F)$ with eigenvalue $\tau$ then

$$aMu + Su/(2a) + Mv/2 = \tau Su \quad \text{and} \quad Mu/2 + Mv/(2a) = \tau Mv.$$

The latter requires that $u = (2\tau - 1/a)v$ and this in turn reduces the former to

$$(4\tau^2 - 4\tau/a + 1/a^2)Sv = (4a\tau - 1)Mv.$$

This states that $v$ is an eigenvector of $(S, M)$ and

$$\frac{4a\tau - 1}{4\tau^2 - 4\tau/a + 1/a^2} = \lambda_j$$

where $\lambda_j$ is an associated eigenvalue of $(S, M)$. This quadratic in $\tau$ has the two roots

$$\frac{1}{2a} + \frac{a \pm \sqrt{a^2 + \lambda_j}}{2\lambda_j}.$$

$\tau_1(D)$, the largest of these roots, is attained at $j = n$ for the least of the eigenvalues of $(S, M)$. On combining this with the previous corollary we find

**Corollary 12.13.** If $D = 2aM$ then the associated greatest total energy is

$$\tau_1(2aM) = \frac{1}{2a} + \frac{a + \sqrt{a^2 + \lambda_n}}{2\lambda_n}. \tag{12.43}$$

We have graphed this function in Figure 12.3 for the simple choice $\lambda_n = 1$. We note that $\tau_1$ approaches infinity for both small *and* large $a$. For later purposes we record the magnitude, $a$, for which the greatest total energy is least,

$$\hat{a} = \sqrt{\lambda}_n \sqrt{\frac{\sqrt{5} - 1}{2}}, \tag{12.44}$$

as well as the associated leading eigenvector of $(\mathcal{E}(2\hat{a}M), F)$,

$$\hat{V}_1 = q_n(2\tau_1(2\hat{a}M) - 1/\hat{a}, 1) = q_n((\hat{a} + \sqrt{\hat{a}^2 + \lambda_n})/\lambda_n, 1). \tag{12.45}$$

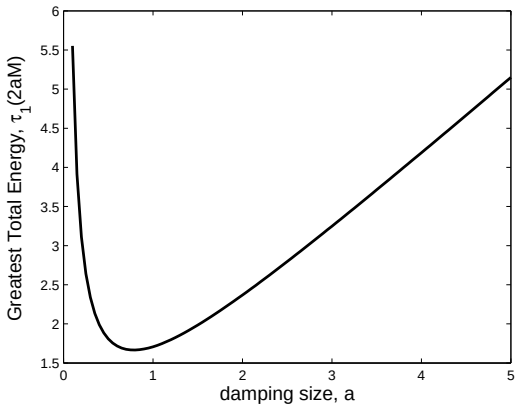



**Figure** 12.3. The greatest total energy for a system with friction damping $D = 2aM$.

This graph enjoys the nice property that it always lies below any chord connecting two of its points. We call functions with such graphs **convex**. You may recall, see (3.36), that the compliance of a mechanical network is a convex function of the fiber stiffnesses. We will need a slight generalization of that result.

**Proposition** 12.13. If the function $v \mapsto H(u, v)$ is convex for each $u$ in some set $\mathcal{U}$ and

$$h(v) \equiv \max_{u \in \mathcal{U}} H(u, v),$$

then $h$ is convex.

Proof: Given $v_1$ and $v_2$ and $t \in [0, 1]$ we note that

$$H(u, tv_1 + (1-t)v_2) \leq tH(u, v_1) + (1-t)H(u, v_2), \quad \forall u \in \mathcal{U}.$$

From the fact that the maximum of a sum is less than the sum of the maximums it follows that

$$\max_{u \in \mathcal{U}} H(u, tv_1 + (1-t)v_2) \leq t \max_{u \in \mathcal{U}} H(u, v_1) + (1-t) \max_{u \in \mathcal{U}} H(u, v_2),$$

i.e., $h(tv_1 + (1-t)v_2) \leq th(v_1) + (1-t)h(v_2)$. End of Proof.

Our goal in the remainder of the section is to show that $\tau_1$ is in fact convex over a larger class of damping matrices and that (12.44) remains the best damping. We look first for a more convenient characterization of $\mathcal{E}(D)$. Recalling the energy equation, Prop. 12.11, we search for an $X$ for which

$$E(t) = -\frac{d}{dt} V(t)^T X V(t). \tag{12.46}$$

Representing $E$ directly in terms of $V$ and computing the derivative on the right side requires of $X$ that

$$\begin{aligned} V(t)^T F V(t) &= -V'(t)^T X V(t) - V(t)^T X V'(t) \\ &= -V(t)^T (XA(D) + A^T(D)X) V(t). \end{aligned}$$

Evaluating this expression at $t = 0$ and using the fact that the initial state, $V(0)$, is perfectly arbitrary we find that $X$ must satisfy the **Liapunov equation**

$$\boxed{A^T(D)X + XA(D) = -F.} \tag{12.47}$$

We will argue in the exercises that this equation possesses a unique symmetric positive definite solution, that we naturally refer to as $\mathcal{E}(D)$. We note however that (12.47) may be solved explicitly when $D$ is drawn from the **Caughy class**, $\mathcal{C}$, of symmetric positive definite matrices for which

$$S^{-1} D M^{-1} = M^{-1} D S^{-1}. \tag{12.48}$$

In particular, if $D \in \mathcal{C}$ then

$$\mathcal{E}(D) = \begin{pmatrix} \frac{1}{2}D + SD^{-1}M & \frac{1}{2}M \\ \frac{1}{2}M & MD^{-1}M \end{pmatrix} \tag{12.49}$$

is the desired solution of (12.47). We note that $2\hat{a}M \in \mathcal{C}$ and now proceed to study the greatest eigenvalue of generalized perturbation problem $(\mathcal{E}(2\hat{a}M + \varepsilon C), F)$ for $C \in \mathcal{C}$. From

$$\begin{aligned} (2\hat{a}M + \varepsilon C)^{-1} &= ((I + \varepsilon C(2\hat{a}M)^{-1})2\hat{a}M)^{-1} \\ &= (2\hat{a}M)^{-1}(I + \varepsilon C(2\hat{a}M)^{-1})^{-1} \\ &= (2\hat{a}M)^{-1} \sum_{k=0}^{\infty} (-\varepsilon)^k (C(2\hat{a}M)^{-1})^k \\ &= (2\hat{a}M)^{-1} - (2\hat{a}M)^{-1} \varepsilon C (2\hat{a}M)^{-1} + O(\varepsilon^2) \end{aligned}$$

we develop

$$\begin{aligned} \mathcal{E}(2\hat{a}M + \varepsilon C) &= \mathcal{E}_0 + \varepsilon \mathcal{E}_1 + O(\varepsilon)^2 \\ &= \mathcal{E}(2\hat{a}M) + \varepsilon \begin{pmatrix} C/2 - SM^{-1}C/(4\hat{a}^2) & 0 \\ 0 & -C/(4\hat{a}^2) \end{pmatrix} + O(\varepsilon^2). \end{aligned} \tag{12.50}$$

We now have all the pieces required to establish that $2\hat{a}M$ is a critical point of $\tau_1$. That is,

**Proposition** 12.14. If $\lambda_n$, the least eigenvalue of the undamped system, $(S, M)$, is simple then

$$\tau_1(2\hat{a}M + \varepsilon C) = \tau_1(2\hat{a}M) + O(\varepsilon^2) \tag{12.51}$$

for each $C \in \mathcal{C}$.

Proof: If $\lambda_n$ is a simple eigenvalue of $(S, M)$ then $\tau_1(2\hat{a}M)$ is a simple eigenvalue of $(\mathcal{E}(2\hat{a}M), F)$ and we may deduce from Prop. 12.7 that

$$\tau_1(2\hat{a}M + \varepsilon C) = \tau_1(2\hat{a}M) + \varepsilon \hat{V}_1^T \mathcal{E}_1 \hat{V}_1 + O(\varepsilon^2)$$

where $\hat{V}_1$ was derived in (12.45) and $\mathcal{E}_1$ in (12.50). We first compute

$$\begin{aligned}\mathcal{E}_1 \hat{V}_1 &= \begin{pmatrix} C/2 - SM^{-1}C/(4\hat{a}^2) & 0 \\ 0 & -C/(4\hat{a}^2) \end{pmatrix} \begin{pmatrix} \frac{\hat{a}+\sqrt{\hat{a}^2+\lambda_n}}{\lambda_n} q_n \\ q_n \end{pmatrix} \\ &= \begin{pmatrix} \frac{\hat{a}+\sqrt{\hat{a}^2+\lambda_n}}{\lambda_n} \left(\frac{1}{2}C - \frac{1}{4\hat{a}^2} SM^{-1}C\right) q_n \\ -\frac{1}{4\hat{a}^2} C q_n \end{pmatrix}.\end{aligned}$$

and then

$$\hat{V}_1^T \mathcal{E}_1 \hat{V}_1 = \left\{ \left( \frac{\hat{a} + \sqrt{\hat{a}^2 + \lambda_n}}{\lambda_n} \right)^2 \left( \frac{1}{2} - \frac{\lambda_n}{4\hat{a}^2} \right) - \frac{1}{4\hat{a}^2} \right\} q_n^T C q_n = 0,$$

because the braced term vanishes thanks to (12.44). End of Proof.

The convexity of $D \mapsto \tau_1(D)$ will follow from Prop. 12.13 and our various maximum principles.

**Proposition** 12.15. The greatest total energy, $D \mapsto \tau_1(D)$, is convex for $D \in \mathcal{C}$.

Proof: In light of Prop. 12.13 and the variational characterization, (12.37), of $\tau_1(D)$, it will suffice to show that $D \mapsto V^T \mathcal{E}(D) V$ is convex for each $V \in \mathbb{R}^{2n}$. We split $V$ into its two $n$–dimensional components, $V = [f \; g]$, and find

$$V^T \mathcal{E}(D) V = \tfrac{1}{2} f^T D f + f^T S D^{-1} M f + f^T M g + g^T M D^{-1} M g. \tag{12.52}$$

As a finite sum of convex functions is convex it suffices to show that each of the summands in (12.52) is convex. The first term is linear in $D$ while the third is independent of $D$ and so both are convex in $D$. With respect to the second term we may deduce (Exer. 12.13) from $D \in \mathcal{C}$ that $SD^{-1}M$ is symmetric and positive definite. As such, it follows from our earlier energy considerations, namely Prop. 3.4, that

$$f^T S D^{-1} M f = \max_{x \in \mathbb{R}^n} f^T x - x^T M^{-1} D S^{-1} x. \tag{12.53}$$

As $2f^T x - x^T M^{-1} D S^{-1} x$ is convex in $D$, it follows from Prop. 12.13 and (12.53) that so too is $D \mapsto f^T S D^{-1} M f$. In a similar fashion, from

$$g^T M D^{-1} M g = \max_{x \in \mathbb{R}^n} 2x^T M g - x^T D x,$$

one deduces the convexity of $D \mapsto g^T M D^{-1} M g$. End of Proof.

It remains only to confirm that a critical point of a convex function is indeed a global minimizer.

**Proposition** 12.16. The greatest total energy, $D \mapsto \tau_1(D)$, attains its minimum over $\mathcal{C}$ at $\hat{D} = 2\hat{a}M$.

Proof: If $D_2 \in \mathcal{C}$ and $\tau_1(D_2) < \tau_1(\hat{D})$ then, as $\tau_1$ is convex,

$$\tau_1(\hat{D} + \varepsilon(D_2 - \hat{D})) = \tau_1(\varepsilon D_2 + (1-\varepsilon)\hat{D}) \leq \varepsilon \tau_1(D_2) + (1-\varepsilon)\tau_1(\hat{D})$$

for arbitrary $\varepsilon$. Hence, for $\varepsilon > 0$,

$$\frac{\tau_1(\hat{D} + \varepsilon(D_2 - \hat{D})) - \tau_1(\hat{D})}{\varepsilon} \leq \tau_1(D_2) - \tau_1(\hat{D}).$$

As $\varepsilon \to 0$ we have argued in (12.51) that the left side goes to zero, in contradiction to $\tau_1(D_2) < \tau_1(\hat{D})$. End of Proof.

Let us apply this finding to the optimal damping of the chain in Figure 3.1. We assume uniform masses, $m$, and stiffnesses, $k$, and seek that damping matrix that produces the least greatest total energy. The respective mass, damping, and stiffness matrices of the form

$$M = m\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad D = \begin{pmatrix} d_{11} & d_{12} & d_{13} \\ d_{12} & d_{22} & d_{23} \\ d_{13} & d_{23} & d_{33} \end{pmatrix}, \quad S = k\begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

This $D$ has six degrees of freedom and one can associate these with 6 dashpots at and between the 3 masses in Figure 3.1. We first restrict our choices to the Caughy Class by enforcing $DS^{-1} = S^{-1}D$. This restricts us to the three dimensional class of damping matrices of the form

$$D = \begin{pmatrix} a_3 & a_2 & a_1 - a_3 \\ a_2 & a_1 & a_2 \\ a_1 - a_3 & a_2 & a_3 \end{pmatrix}. \tag{12.54}$$

The eigenvalues of this $D$ are $2a_3 - a_1$ and $a_1 \pm a_2\sqrt{2}$ hence $D \in \mathcal{C}$ so long as its parameters, $a_1,\ a_2$ and $a_3$ obey

$$2a_3 > a_1 > \sqrt{2}|a_2|.$$

We note that the least eigenvalue of $(S, M)$ is simple and

$$\lambda_3 = \frac{k}{m}(2 - \sqrt{2}).$$

Invoking Prop. 12.16 we find that $D \mapsto \tau_1(D)$ attains its minimum over $\mathcal{C}$ at $D = 2\hat{a}M$. In terms of the parametrization (12.54) this requires

$$a_1 = a_3 = \sqrt{2km(2 - \sqrt{2})(\sqrt{5} - 1)} \quad \text{and} \quad a_2 = 0. \tag{12.55}$$

### 12.7. Notes and Exercises

We have followed Kato (1980). For more on Molecular Orbital Theory see Streitwieser (1961).

1. The stiffness matrix associated with the unstable frame of Figure 3.3 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) Show that $\mathcal{R}(P_3) = \mathcal{N}(S)$.

(iii) Assemble

$$S^+ = \frac{1}{\lambda_1} P_1 + \frac{1}{\lambda_2} P_2$$

and check your result against `pinv(S)` in MATLAB.

(iv) 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.

(v) 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. Lets consider the vibration of the equilateral triangle. Given the incidence matrix, $A$, of Exer. 3.9 assemble

$$S = A^T A = \frac{1}{4} \begin{pmatrix} 5 & \sqrt{3} & -4 & 0 & -1 & -\sqrt{3} \\ \sqrt{3} & 3 & 0 & 0 & -\sqrt{3} & -3 \\ -4 & 0 & 5 & -\sqrt{3} & -1 & \sqrt{3} \\ 0 & 0 & -\sqrt{3} & 3 & \sqrt{3} & -3 \\ -1 & -\sqrt{3} & -1 & \sqrt{3} & 2 & 0 \\ -\sqrt{3} & -3 & \sqrt{3} & -3 & 0 & 6 \end{pmatrix}$$

and use `poly` and `roots` to conclude that its characteristic polynomial is

$$\chi_S(\lambda) = \lambda^3(\lambda - 3)(\lambda - 3/2)^2.$$

We identified the eigenvectors corresponding to $\lambda = 0$ in Exer. 3.9. Compute those corresponding to $\lambda_3 = 3$ and $\lambda_2 = 3/2$ and associate them with the figure below. They are $v_3 = (\sqrt{3}, 1, -\sqrt{3}, 1, 0 - 2)$ and $v_{2,1} = (\sqrt{3}, -1, -\sqrt{3}, -1, 0, 2)$ and $v_{2,2} = (-\sqrt{3}, -1, 0, 2, \sqrt{3}, -1)$.

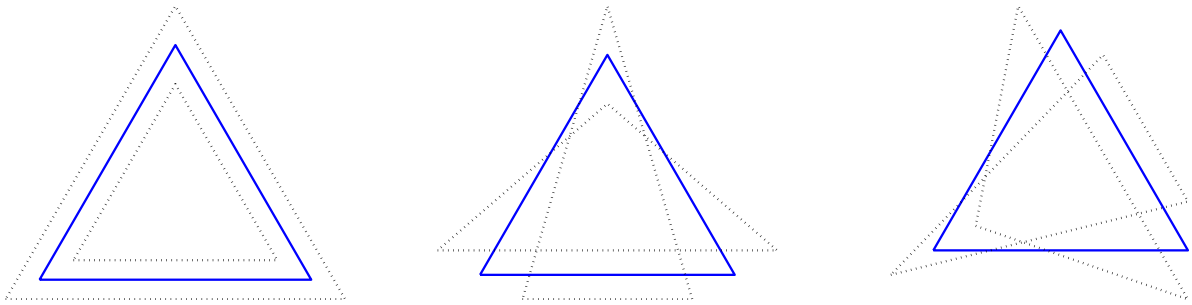


**Figure** 12.4 The 3 modes of vibration of the triangle.

3. Show that if $W$ is Hermitian and positive definite then $\langle x, y \rangle_W \equiv x^* W y$ is an inner product on $\mathbb{C}^n$. Show that if $\langle Ax, Ay \rangle_W = \langle x, y \rangle_W$ then $B \equiv W^{1/2} A W^{-1/2}$ is unitary.

4. We can now prove that the density in Eq. (6.53) obeys

$$\int_{x \in \mathbb{R}^n} \exp(-x^T C^{-1} x)\, dx = \sqrt{(2\pi)^n \det(C)}$$

if $C$ is symmetric and positive definite. Hint: If $C = Q^T \Lambda Q$ then

$$\int_{x\in\mathbb{R}^n} \exp(-x^T C^{-1} x/2)\, dx = \int_{x\in\mathbb{R}^n} \exp(-(Qx)^T \Lambda^{-1} Qx/2)\, dx$$

Now introduce the change of variables $y = Qx$ and the $n$-dimensional analog of the volume distortion equation, (3.33), to argue that $dy = |\det(Q)|dx = dx$ and hence

$$\begin{aligned}\int_{x\in\mathbb{R}^n} \exp(-x^T C^{-1} x/2)\, dx &= \int_{y\in\mathbb{R}^n} \exp(-y_1^2/(2\lambda_1) - \cdots - y_n^2/(2\lambda_n))\, dy \\ &= \prod_{j=1}^n \int_{-\infty}^{\infty} \exp(-y_j^2/(2\lambda_j))\, dy_j \\ &= \prod_{j=1}^n \sqrt{2\lambda_j} \int_{-\infty}^{\infty} \exp(-s^2)\, ds\end{aligned}$$

now use Eq. (6.52).

5. By similar reasoning

$$\begin{aligned}-\int_{\mathbb{R}^n} g_{0,C}(x) \log g_{0,C}(x)\, dx &= \int_{x\in\mathbb{R}^n} \frac{\exp(-x^T C^{-1} x/2)}{2\sqrt{(2\pi)^n \det(C)}} (x^T C^{-1} x + n\log(2\pi) + \log(\det(C))\, dx \\ &= \frac{n\log(2\pi) + \log(\det(C))}{2} + \int_{y\in\mathbb{R}^n} \frac{(y_1^2/(2\lambda_1) + \cdots + y_n^2/(2\lambda_n)) \exp(-y_1^2/(2\lambda_1) - \cdots - y_n^2/(2\lambda_n))}{\sqrt{(2\pi)^n \det(C)}}\, dy \\ &= \frac{n\log(2\pi) + \log(\det(C))}{2} + \frac{n}{\pi^{n/2}} \int_{-\infty}^{\infty} s^2 \exp(-s^2)\, ds \left(\int_{-\infty}^{\infty} \exp(-s^2)\, ds\right)^{n-1}\end{aligned}$$

6. Show that if $A$ is Hermitian and $x^* Ax = 0$ for all $x$ then $A = 0$.

7. Develop perturbation theory for multiple eigenvalues.

8. Use Prop. 12.9 to prove that if $B = B^T \in \mathbb{R}^{n\times n}$ and $Q \in \mathbb{R}^{n\times n-1}$ and $Q^T Q = I$ and $A = Q^T BQ$ then the eigenvalues of $A$ **interlace** those of $B$. That is,

$$\lambda_1(B) \geq \lambda_1(A) \geq \lambda_2(B) \geq \lambda_2(A) \geq \cdots \geq \lambda_{n-1}(A) \geq \lambda_n(B). \tag{12.56}$$

9. Generalize Prop. 12.9 to the case that $S$ and $M$ are symmetric and positive definite:

$$\lambda_k = \max_{\dim(W)=k} \min_{0\neq x\in W} \frac{x^T Sx}{x^T Mx}, \tag{12.57}$$

10. The Power Method. Submit a MATLAB diary of your application of the Power Method to the $S$ matrix in Exercise 1.

11. We claimed in §12.5 that the Schrödinger operator is symmetric. By that we mean that

$$\langle \mathcal{H}\psi, \phi\rangle = \langle \psi, \mathcal{H}\phi\rangle \tag{12.58}$$

for all $\psi$ and $\phi$ for which $\langle \psi, \psi\rangle$ and $\langle \phi, \phi\rangle$ are finite. In order to demonstrate (12.58) please

(a) Show that $\Delta\psi = \nabla \cdot \nabla\psi$ where $\nabla$ is the gradient, $\nabla\psi = (\partial\psi/\partial x, \partial\psi/\partial y, \partial\psi/\partial z)$ and $\nabla\cdot$ is the divergence, $\nabla \cdot (f_1, f_2, f_3) = \partial f_1/\partial x + \partial f_2/\partial y + \partial f_3/\partial z$.

(b) Confirm the product rule

$$\nabla \cdot (\phi\nabla\psi) = \nabla\phi \cdot \nabla\psi + \phi\Delta\psi. \tag{12.59}$$

(c) Use (b) to show that

$$\langle \mathcal{H}\psi, \phi\rangle = \int_{\mathbb{R}^3} (\nabla\psi \cdot \nabla\phi + V\psi\phi - \nabla \cdot (\phi\nabla\psi))\; dxdydz \tag{12.60}$$

(d) The first two terms on the right in (12.60) are indeed symmetric in $\psi$ and $\phi$. Use the Divergence Theorem

$$\int_\Omega \nabla \cdot f\, dxdydz = \int_{\partial\Omega} f \cdot n\, ds \tag{12.61}$$

where $\Omega$ is a subset of $\mathbb{R}^3$ with boundary $\partial\Omega$ and outer unit normal $n$ parametrized by arclength, $s$, to show that

$$\int_{\mathbb{R}^3} \nabla \cdot (\phi\nabla\psi)\, dxdydz = 0.$$

Hint: Argue that since $\langle\phi, \phi\rangle < \infty$ then $\phi$ must essentially vanish outside some big set $\Omega$.

(e) Conclude from (c) and (d) that $\mathcal{H}$ is indeed symmetric.

12. With regard to the hexagonal arrangement of carbons in the benzene molecule of §12.5 note that

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

simply permutes, or shifts, or rotates each carbon into its neighbor.

(a) Show that $AP = PA$ for the $A$ in (12.29).

(b) Deduce from (a) that every eigenspace of $A$ must be invariant under $P$. That is, show that if $Ax = \lambda x$ then $A(Px) = \lambda(Px)$ as well.

13. Use (12.48) to show that (12.49) is symmetric and positive definite and a solution to (12.47).

14. Prove that the Caughy class, $\mathcal{C}$, is convex.

15. One general approach to the Lyapunov Equation,

$$AX^* + XA^* = -S \tag{12.62),}$$

is to begin with the Schur decomposition, $A = U^*TU$. For then $U^*T^*UX + XU^*TU = -S$ is simplified via left multiplication by $U$ and right multiplication by $U^*$, to

$$T^*B + BT = -C, \quad \text{where} \quad B = UXU^* \quad \text{and} \quad C = USU^*. \tag{12.63}$$

This triangular system can be solved by forward substitution. To wit, examine the (1,1) element and find

$$\overline{t}_{1,1}b_{1,1} + b_{1,1}t_{1,1} = -c_{1,1}$$

and next the (1,2) element and find

$$\overline{t}_{1,1}b_{1,2} + b_{1,1}t_{1,2} + b_{1,2}t_{2,2} = -c_{1,2}.$$

Generalize and deduce that so long as no two eigenvalues of $A$ are the negative conjugates of one another then (12.62) possesses a unique, Hermitian, positive definite solution, $X$.

# 13. The Singular Value Decomposition

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 chapter 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 indeed tell us 'everything' about $A$ can be gleaned from

$$\begin{aligned} &\mathcal{N}(A^T A) = \mathcal{N}(A), \quad \mathcal{N}(AA^T) = \mathcal{N}(A^T), \\ &\mathcal{R}(A^T A) = \mathcal{R}(A^T), \quad \text{and} \quad \mathcal{R}(AA^T) = \mathcal{R}(A). \end{aligned} \tag{13.1}$$

You have proven the first of these in Exer. 4.10. The proof of the second is identical. The row and column space results follow from the first two via orthogonality. In light of (13.1) we will see that the singular value decomposition of $A$ delivers orthonormal bases for the four fundamental subspaces of $A$.

Beyond the satisfaction of reinforcing the Fundamental Theorem of Linear Algebra the singular value decomposition will also shed new light on both least squares and the pseudoinverse. These results together indicate that the SVD may be a useful means for constructing accurate low rank approximations of large matrices. Our experimental investigation of this hunch in §13.2 will be followed by theoretical confirmation in §13.3. In addition, it is a fundamental tool in information and data sciences.

### 13.1. The Decomposition

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 3.3)

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{13.2}$$

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 12.1 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** 13.1. 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 $A^TAx = \lambda x$ then $x^TA^TAx = \lambda x^Tx$, 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

$$A^TAx_{j,k} = \lambda_j x_{j,k} \tag{13.3}$$

we find, on multiplying through (from the left) by $A$ that

$$AA^TAx_{j,k} = \lambda_j Ax_{j,k},$$

i.e., $\lambda_j$ 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,

$$y_{j,k} \equiv \frac{Ax_{j,k}}{\sqrt{\lambda}_j}, \quad 1 \leq k \leq n_j \tag{13.4}$$

is a collection of unit eigenvectors of $AA^T$ associated with $\lambda_j$. Let us now show that these vectors are orthonormal for fixed $j$.

$$y_{j,i}^Ty_{j,k} = \frac{1}{\lambda_j}x_{j,i}^TA^TAx_{j,k} = x_{j,i}^Tx_{j,k} = 0.$$

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 (13.3) 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

$$A^TA = X\Lambda_nX^T \tag{13.5}$$

where

$$X = [X_1 \cdots X_h], \quad \text{where} \quad X_j = [x_{j,1} \cdots x_{j,n_j}]$$

and $\Lambda_n$ is the $n$-by-$n$ diagonal matrix with $\lambda_1$ in the first $n_1$ slots, $\lambda_2$ in the next $n_2$ slots, *etc.* Similarly

$$AA^T = Y\Lambda_mY^T \tag{13.6}$$

where

$$Y = [Y_1 \cdots Y_h], \quad \text{where} \quad Y_j = [y_{j,1} \cdots y_{j,n_j}].$$

and $\Lambda_m$ is the $m$-by-$m$ diagonal matrix with $\lambda_1$ in the first $n_1$ slots, $\lambda_2$ in the next $n_2$ slots, *etc.* The $y_{j,k}$ were defined in (13.4) 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 $\Sigma$ denote the $m$-by-$n$ matrix diagonal matrix with $\sigma_1$ in the first $n_1$ slots and $\sigma_2$ in the next $n_2$ slots, *etc.* Notice that

$$\Sigma^T\Sigma = \Lambda_n \quad \text{and} \quad \Sigma\Sigma^T = \Lambda_m. \tag{13.7}$$

Now recognize that (13.4) 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 $A$,

$$\boxed{A = Y\Sigma X^T.} \tag{13.8}$$

Let us confirm this on the $A$ matrix in (13.2). 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,

$$\Sigma = \begin{pmatrix} \sqrt{2} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{13.9}$$

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** 13.2. 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 $\Sigma$.
(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 $\Sigma^+$ 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 $\Sigma^+$ to zero. Now, one defines the **pseudo-inverse** of $A$ to be

$$A^+ \equiv X\Sigma^+ Y^T.$$

Taking the $A$ of (13.2) we find

$$\Sigma^+ = \begin{pmatrix} 1/\sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

and so

$$A^+ = \begin{pmatrix} -1/\sqrt{2} & 0 & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 & 0 \\ 1/\sqrt{2} & 0 & 0 & 1/\sqrt{2} \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & -1/2 & 0 \\ 1 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

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

$$\begin{aligned} (A^TA)A^+b &= X\Lambda_n X^T X\Sigma^+ Y^T b \qquad \text{by (13.5)} \\ &= X\Lambda_n \Sigma^+ Y^T b \qquad \text{because } X^TX = I \\ &= X\Sigma^T \Sigma\Sigma^+ Y^T b \qquad \text{by (13.7)} \\ &= X\Sigma^T Y^T b \qquad \text{because } \Sigma^T\Sigma\Sigma^+ = \Sigma^T \\ &= A^T b \qquad \text{by (13.8)}, \end{aligned}$$

that is, $A^+b$ satisfies the **least-squares problem** $A^TAx = A^Tb$.

### 13.2. 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 13.2 is a 262-by-165 matrix. Such a matrix-image is read, displayed and 'decomposed' by

```
M = imread('JohnBrown.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 13.1.

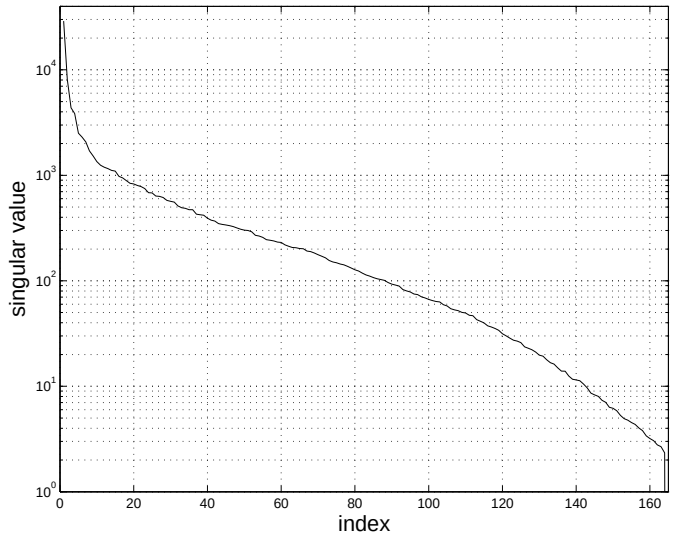



**Figure** 13.1. The singular values of John Brown.

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 $k$.

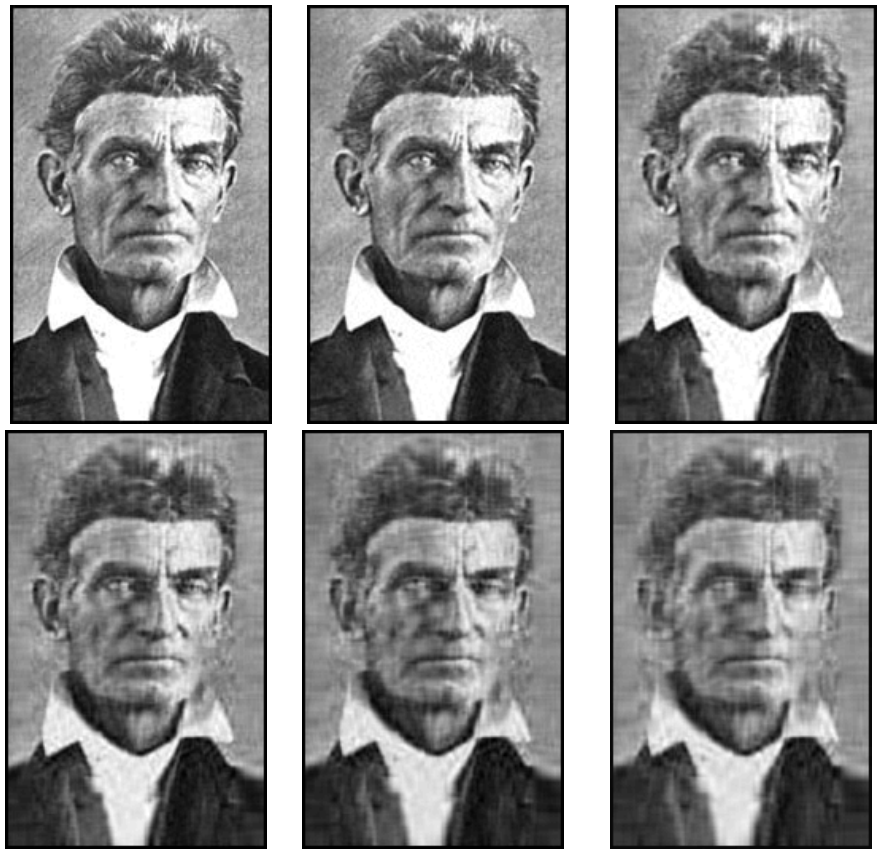


**Figure** 13.2. The results of `imagesc(Ak)` for, starting at the top left, `k=165, 64, 32` and moving right, and then starting at the bottom left and moving right, `k=24,20,16`.

### 13.3. Low Rank Approximation*

In this section we establish a precise version of last section's folk theorem. To prepare the way we note that if $B = B^T \in \mathbb{R}^{m\times m}$ and we list its eigenvalues, including multiplicities, $\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ then, from Cor. 11.8 it follows that

$$\mathrm{tr}(B) = \sum_{i=1}^{m} \lambda_j(B). \tag{13.10}$$

This together with our trace formulation of the Frobenius norm of $A \in \mathbb{R}^{m\times n}$, recall (1.18), yields

$$\|A\|_F^2 = \mathrm{tr}(AA^T) = \sum_{i=1}^{m} \lambda_i(AA^T) = \sum_{i=1}^{m} \sigma_i^2, \tag{13.11}$$

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

$$\|QA\|_F^2 = \mathrm{tr}(QAA^TQ^T) = \mathrm{tr}(AA^TQ^TQ) = \mathrm{tr}(AA^T) = \|A\|_F^2. \tag{13.12}$$

The same argument reveals that $\|AQ\|_F = \|A\|_F$. We may now establish

**Proposition** 13.3. Given an $m$-by-$n$ matrix $A$ (with SVD $A = Y\Sigma X^T$) and a whole number $k \leq \min\{m, n\}$ 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 $\Sigma$ 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.

The expression, (13.11), of the Frobenius norm in terms of singular values leads naturally to the question of like bounds on the associated inner product. Upon associating to each unitary matrix, $U$, a doubly stochastic matrix $D$, via $D_{i,j} = |U_{i,j}|^2$, we note that our Mixing Theorem, Prop. 5.8, is well suited to this task. In particular,

**Proposition** 13.4. If $A$ and $B$ lie in $\mathbb{C}^{m\times n}$ then

$$|\mathrm{tr}(A^*B)| \le \sigma(A)^T\sigma(B). \tag{13.13}$$

Proof: Given the two singular value decompositions $A = USV^*$ and $B = WTX^*$ we express

$$\mathrm{tr}(A^*B) = \mathrm{tr}(VSU^*WTX^*) = \mathrm{tr}(X^*VSU^*WT) = \mathrm{tr}(Q^*SPT)$$

where $P \equiv U^*W$ and $Q = V^*X$ are unitary and so the matrices with elements $|p_{ij}|^2$ and $|q_{ij}|^2$ are doubly stochastic and so

$$\begin{aligned}
|\mathrm{tr}(A^*B)| &= |\mathrm{tr}((SQ)^*PT)| \\
&\le \sum_{i=1}^n\sum_{j=1}^n s_it_j|\overline{q}_{ij}p_{ij}| \\
&\le \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n s_it_j|q_{ij}|^2 + \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n s_it_j|p_{ij}|^2 \\
&\le s^Tt.
\end{aligned}$$

The first inequality is the triangle inequality. The second is $2|\overline{z}_1z_2| \le (|z_1|^2 + |z_2|^2)$ and the third is our Mixture conclusion, (5.9), on realizing that the matrix with elements $(|p_{ij}|^2 + |q_{ij}|^2)/2$ is also doubly stochastic. End of Proof.

### 13.4. Principal Component Analysis*

### 13.5. Independent Component Analysis*

The data is $n$ samples of $m$ signals, $X \in \mathbb{R}^{m\times n}$. To compute the first component we maximize the normalized kurtosis of $w^TX$.

$$\mathcal{K}(w) = \frac{E[(w^TX)^4]}{E^2[(w^TX)^2]} = \frac{n\sum_{j=1}^n\left(\sum_{i=1}^m w_iX_{i,j}\right)^4}{\left(\sum_{j=1}^n\left(\sum_{i=1}^m w_iX_{i,j}\right)^2\right)^2}$$

We first record its gradient

$$\frac{\partial\mathcal{K}(w)}{\partial w_k} = 4n\frac{\sum_{j=1}^n\left(\sum_{i=1}^m w_iX_{i,j}\right)^2\sum_{j=1}^n\left(\sum_{i=1}^m w_iX_{i,j}\right)^3X_{k,j} - \sum_{j=1}^n\left(\sum_{i=1}^m w_iX_{i,j}\right)^4\sum_{j=1}^n\sum_{i=1}w_iX_{i,j}X_{k,j}}{\left(\sum_{j=1}^n\left(\sum_{i=1}^m w_iX_{i,j}\right)^2\right)^3}$$

Hence,

$$\nabla\mathcal{K}(w) = \frac{4n}{(y^Ty)^2}X(y^3) - \frac{4n(y^2)^T(y^2)}{(y^Ty)^3}Xy, \quad \text{where} \quad y = X^Tw$$

and $y^m$ denotes the element wise product `y.^m`. we then make the step $w = w + t\nabla\mathcal{K}(w)$ where $t$ is chosen to maximize

$$f(t) \equiv \mathcal{K}(w + t\nabla\mathcal{K}(w)).$$

To reveal its dependence on $t$ we set

$$z = X^T\nabla\mathcal{K}(w) \quad \text{and express} \quad f(t) = n\frac{p(t)}{q^2(t)} - 3$$

where $p$ and $q$ are the polynomials in $t$

$$p(t) = \sum_{k=0}^{4} p_k t^k \quad \text{and} \quad q(t) = \sum_{k=0}^{2} q_k t^k$$

where

$$p_4 = (z^2)^T z^2, \quad p_3 = 4(z^3)^T y, \quad p_2 = 6(y^2)^T z^2, \quad p_1 = 4(y^3)^T z, \quad p_0 = (y^2)^T y^2$$
$$q_2 = z^T z, \quad q_1 = 2y^T z, \quad q_0 = y^T y.$$

It follows that the critical points of $f$ are the roots of

$$r(t) = q(t)p'(t) - 2p(t)q'(t) = \sum_{k=0}^{4} r_k t^k$$

where

$$r_4 = 2p_4 q_1 - p_3 q_2, \quad r_3 = p_3 q_1 - 2p_2 q_2 + 4p_4 q_0, \quad r_2 = 3p_3 q_0 - 3p_1 q_2,$$
$$r_1 = 2p_2 q_0 - p_1 q_1 - 4p_0 q_2, \quad r_0 = p_1 q_0 - 2p_0 q_1.$$

### 13.6. Notes and Exercises

1. Suppose that $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Set $x^+ = A^+ b$ and suppose $x$ satisfies $A^T Ax = A^T b$. Prove that $\|x^+\| \leq \|x\|$. (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. We use the singular value decomposition to improve our ability to search for terms across documents. The SVD will make associations that are latent or implicit. We denote by $A$ our term by document matrix and its full and reduced SVD by

$$A = Y\Sigma X^T \quad \text{and} \quad A_k = Y_k \Sigma_k X_k^T.$$

We view a document vector as a column of $A$ or $A_k$ and record

$$A_k(:, j) = Y_k \Sigma_k X_k^T(:, j)$$

and so, given a query vector $q$ we consider

$$\hat{q} \equiv \Sigma_k^{-1} Y_k^T q$$

and compute the cosines

$$\cos(\theta_j) = \frac{\hat{q}^T X_k^T(:, j)}{\|\hat{q}\| \|X_k^T(:, j)\|}$$

4. (a) Show that

$$\mathcal{D} \equiv \{A : A = A^T, A \geq 0, \mathrm{tr} A = 1\}$$

is convex.

(b) Show that the extreme points of $\mathcal{D}$ are the orthogonal rank one projections. (pure states)

# 14. Matrix Groups*

Matrix Groups, and their representations, are powerful tools for identifying and exploiting symmetries in both the physical and mathematical worlds. In particular, they allow for a significant reduction in complexity of highly symmetric structures. For example, we will see that the determination of the 60 electronic energy levels (eigenvalues) of the $C_{60}$ molecule (Buckminsterfullerene) may be may be reduced to the determination of the roots of a few cubic polynomials.

To get there we begin with the group of orthogonal matrices, paying special attention to its finite subgroups that preserve the symmetries of regular polyhedra in $\mathbb{R}^2$ and $\mathbb{R}^3$. We can often view such symmetries as invariance under a permutation of vertices. This leads us to consider the important group of permutation matrices. Our final class of finite groups will arise from restricting matrix elements to a finite sequence of integers and using modular arithmetic during matrix multiplication. Throughout the chapter we will encounter and master these matrix groups by building their multiplication tables, sketching their Cayley graphs, and decomposing them into disjoint unions of conjugacy classes. We close with an investigation of group action and its application to difficult counting problems.

## 14.1. Orthogonal Groups

Recall that $Q \in \mathbb{R}^{n\times n}$ is said to be orthogonal when $Q^TQ = I$. It follows that $Q^T = Q^{-1}$. If $S \in \mathbb{R}^{n\times n}$ is also orthogonal then $(QS)^TQS = (SQ)^T(SQ) = I$ and so both $QS$ and $SQ$ are also orthogonal matrices. A property that survives matrix multiplication is important enough to merit a

**Definition** 14.1. A set of matrices, $G$, is a **matrix group** when
(1) The product of any two matrices in $G$ is also in $G$, and
(2) Every matrix in $G$ is invertible and its inverse lies in $G$.

It follows from the consideration that motivated this definition that

$$\boxed{\mathrm{O}_n(\mathbb{R}) \equiv \{Q \in \mathbb{R}^{n\times n} : Q^TQ = I\}}$$

is a matrix group. We call it the **orthogonal group** of transformations of $\mathbb{R}^n$. It follows from the fundamental properties of the determinant, e.g., from Eq. (3.21) and Exer. 11.4, that if $Q^TQ = I$ then

$$1 = \det(I) = \det(Q^TQ) = \det(Q^T)\det(Q) = \det(Q)^2$$

and so $\det(Q) = \pm 1$ for each $Q \in \mathrm{O}_n(\mathbb{R})$. By the same argument, if $\det(Q) = 1$ and $\det(S) = 1$ then $\det(QS) = 1$. As a result,

$$\boxed{\mathrm{SO}_n(\mathbb{R}) \equiv \{Q \in \mathrm{O}_n(\mathbb{R}) : \det(Q) = 1\}}$$

is also a matrix group. We refer to it as the **special orthogonal group**.

We note that groups are a gateway from linear to "nonlinear" algebra. For example, $\mathrm{O}_n$ is nothing like a subspace, for if $Q \in \mathrm{O}_n$ then $2Q \notin \mathrm{O}_n$!. To see what these groups do look like we specialize to the planar, $n = 2$, and spatial, $n = 3$, cases.

Let us show that each $Q \in \mathrm{O}_2$ is either a rotation or an improper rotation, i.e., a reflection followed by a rotation. If $Q = (q_1,\ q_2) \in \mathrm{O}_2(\mathbb{R})$, as its first column, $q_1$, is a unit vector it must

be $q_1 = (\cos\theta,\ \sin\theta)$ for some $\theta$. As $\|q_2\| = 1$ and $q_1^T q_2 = 0$ it follows that $q_2 = \pm(-\sin\theta,\ \cos\theta)$. Hence $Q$ is either

$$Q_+ = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \quad \text{or} \quad Q_- = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}$$

With $e_1 = (1,\ 0)^T$ and $e_2 = (0,\ 1)^T$ we note that $Q_+ e_1 = (\cos\theta,\ \sin\theta)$ and $Q_+ e_2 = (-\sin\theta,\ \cos\theta)$ and so $Q_+ = R_\theta$ is counterclockwise rotation by $\theta$.

Next, as $Q_- e_1 = (\cos\theta,\ \sin\theta)$ and $Q_- e_2 = (\sin\theta,\ -\cos\theta)$ it follows that $Q_- = R_\theta H_{e_2^\perp}$ is rotation by $\theta$ after reflection across the line $e_2^\perp$. We encountered such **reflection matrices**

$$H_{e_2^\perp} \equiv I - 2e_2 e_2^T = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

back in (1.42). As $\det(Q_\pm) = \pm 1$ we established that each $Q \in \mathrm{SO}_2(\mathbb{R})$ is a rotation while each $Q$ in the complement of $\mathrm{SO}_2$, i.e., $Q \in \mathrm{O}_2(\mathbb{R}) \setminus \mathrm{SO}_2(\mathbb{R})$, is a reflection followed by a rotation. We next confirm that this remains true for $n = 3$.

To begin, we need to clarify precisely what we mean by a rotation in $\mathbb{R}^3$. The basic ingredient is an axis $a \in \mathbb{R}^3$ with $\|a\| = 1$. We note that $aa^T$ and $I - aa^T$ are projections onto $a$ and onto its orthogonal complement, respectively. The third direction is achieved by the cross product matrix of $a$

$$\mathtt{X}(a) = \begin{pmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{pmatrix} \tag{14.1}$$

that we built and analyzed in Exer. 1.21. With these we find that

$$\boxed{R_{a,\theta} = aa^T + \sin(\theta)\mathtt{X}(a) + \cos(\theta)(I - aa^T)} \tag{14.2}$$

achieves counterclockwise rotation by $\theta$ about $a$. In Exer. 14.2 we step through the proof that $R_{a,\theta}$ indeed resides in $\mathrm{SO}_3(\mathbb{R})$. We now show that every matrix in $\mathrm{SO}_3(\mathbb{R})$ looks like $R_{a,\theta}$. To begin, we need

**Proposition** 14.2. If $Q \in \mathrm{SO}_3(\mathbb{R})$ then 1 is an eigenvalue of $Q$.

Proof: Each eigenvalue of $Q$ has magnitude 1 and 1 is their product, so one of them must be 1. End of Proof.

For $Q \in \mathrm{SO}_3$ it follows that its spectrum comes in one of three flavors. In each case, as $Q$ is normal, we will use the fact, Exer. 11.12, that $Q$ has an orthonormal basis of eigenvectors.

1. If 1 is a triple eigenvalue of $Q$ then its spectral representation takes the form $Q = q_1 q_1^T + q_2 q_2^T + q_3 q_3^T$ where the $q_j$ are real and orthonormal. In which case $Q^T = Q$ and $I = Q^T Q = q_1 q_1^T + q_2 q_2^T + q_3 q_3^T = Q$, so $Q$ is the trivial rotation.

2. If 1 is a simple eigenvalue and $-1$ is a double eigenvalue, then its spectral representation takes the form $Q = q_1 q_1^T - q_2 q_2^T - q_3 q_3^T$ where the $q_j$ are real and orthonormal. Again $Q^T = Q$ and $I = Q^T Q = q_1 q_1^T + q_2 q_2^T + q_3 q_3^T$ which in this case tells us that

$$q_2 q_2^T + q_3 q_3^T = I - q_1 q_1^T$$

and so

$$Q = q_1 q_1^T - (I - q_1 q_1^T)$$

which, when compared to Eq. (14.2) tells us that $Q$ is rotation by $\pi$ around $q_1$.

3. If 1 is a simple eigenvalue and $-1$ is not an eigenvalue then the remaining eigenvalues occur in a complex conjugate pair, $\exp(\pm i\theta)$, with associated eigenvectors $q_2$ and $\bar{q}_2$. In this case $Q$ enjoys the spectral representation

$$Q = q_1 q_1^T + \exp(i\theta) q_2 \bar{q}_2^T + \exp(-i\theta) \bar{q}_2 q_2^T = q_1 q_1^T + 2\Re\{\exp(i\theta) q_2 \bar{q}_2^T\}.$$

Writing $q_2 = x + iy$ where both $x$ and $y$ lie in $\mathbb{R}^3$ we deduce from $q_1^T q_2 = 0$ that $q_1^T x = q_1^T y = 0$. Next $\bar{q}_2^T \bar{q}_2 = 0$ means $(x - iy)^T (x - iy) = 0$ means $x^T x = y^T y = 1/2$ and $x^T y = 0$ so

$$q_2 \bar{q}_2^T = (x + iy)(x - iy)^T = xx^T + yy^T + i(yx^T - xy^T)$$

and so

$$\Re\{\exp(i\theta) q_2 \bar{q}_2^T\} = \cos(\theta)(xx^T + yy^T) - \sin(\theta)(yx^T - xy^T).$$

As a result we find

$$Q = q_1 q_1^T - 2\sin(\theta)(yx^T - xy^T) + 2\cos(\theta)(xx^T + yy^T).$$

Next, as $2(xx^T + yy^T)$ is a rank 2 orthogonal projection that annihilates $q_1$ we find

$$2(xx^T + yy^T) = I - q_1 q_1^T.$$

Similarly, in building a cross vector from $-2(yx^T - xy^T)$ we find something of unit norm that is orthogonal to both $x$ and $y$, hence can be only $\pm q_1$.

So $\mathrm{SO}_3(\mathbb{R})$ is the group of rotations.

### 14.2. Symmetry Groups

Given a geometric figure $F$ (e.g., triangle, square, tetrahedron,...) the **rotation group** of $F$ is the group of all $Q \in \mathrm{SO}_3(\mathbb{R})$ for which $QF = F$. The **full symmetry group** of $F$ is the group of all $Q \in \mathrm{O}_3(\mathbb{R})$ for which $QF = F$.

We begin with planar figures – denoting the full symmetry group of the regular (equal sides, equal angles) $n$–sided polygon by $\mathrm{Dih}_n$ (for Dihedral) and its associated group of rotations by $\mathrm{SDih}_n$.

The equilateral triangle, Figure 14.1, is invariant under rotations by 0, $2\pi/3$ and $4\pi/3$ and so its rotation group is

$$\mathrm{SDih}_3 = \{I, R_{2\pi/3}, R_{4\pi/3}\} \tag{14.3}$$

where

$$R_{2\pi/3} = \begin{pmatrix} \cos(2\pi/3) & -\sin(2\pi/3) \\ \sin(2\pi/3) & \cos(2\pi/3) \end{pmatrix} = \frac{1}{2}\begin{pmatrix} -1 & -\sqrt{3} \\ \sqrt{3} & -1 \end{pmatrix},$$
$$R_{4\pi/3} = \begin{pmatrix} \cos(4\pi/3) & -\sin(4\pi/3) \\ \sin(4\pi/3) & \cos(4\pi/3) \end{pmatrix} = \frac{1}{2}\begin{pmatrix} -1 & \sqrt{3} \\ -\sqrt{3} & -1 \end{pmatrix}.$$

To confirm that this is a group compute $R_{2\pi/3} R_{4\pi/3}$. You might also check that $R_{2\pi/3}$ generates $\mathrm{SDih}_3$ in the sense that all members are powers of $R_{2\pi/3}$.

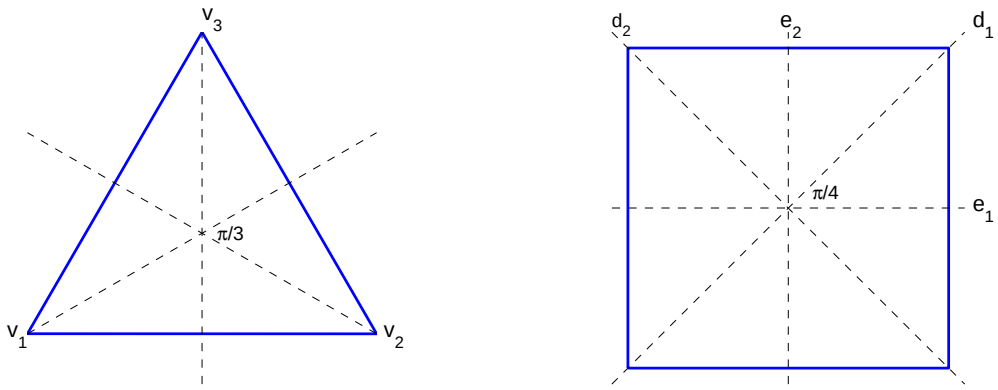



**Figure** 14.1. The equilateral triangle and square.

To arrive at the full symmetry group, $\mathrm{Dih}_3$, we append the three reflections in lines through the vertices $v_1$, $v_2$ and $v_3$ indicated in Figure 14.1. With $v_3 = (0,1)$, $v_2 = (\sqrt{3},-1)/2$ and $v_1 = (-\sqrt{3},-1)/2$ we arrive at the associated reflections

$$\begin{aligned} H_{v_3} &= I - 2v_3^{\perp}(v_3^{\perp})^T = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, \\ H_{v_2} &= I - 2v_2^{\perp}(v_2^{\perp})^T = \frac{1}{2}\begin{pmatrix} 1 & -\sqrt{3} \\ -\sqrt{3} & -1 \end{pmatrix} = R_{2\pi/3}H_{v_3}, \\ H_{v_1} &= I - 2v_1^{\perp}(v_1^{\perp})^T = \frac{1}{2}\begin{pmatrix} 1 & \sqrt{3} \\ \sqrt{3} & -1 \end{pmatrix} = H_{v_3}R_{2\pi/3}. \end{aligned} \tag{14.4}$$

It follows that $\mathrm{Dih}_3$ is generated by products of $R_{2\pi/3}$ and $H_{v_3}$. We capture the full story in the multiplication table below.

| $\mathrm{Dih}_3$ | $I$ | $R_{2\pi/3}$ | $R_{4\pi/3}$ | $H_{v_3}$ | $H_{v_2}$ | $H_{v_1}$ |
|---|---|---|---|---|---|---|
| $I$ | $I$ | $R_{2\pi/3}$ | $R_{4\pi/3}$ | $H_{v_3}$ | $H_{v_2}$ | $H_{v_1}$ |
| $R_{2\pi/3}$ | $R_{2\pi/3}$ | $R_{4\pi/3}$ | $I$ | $H_{v_2}$ | $H_{v_1}$ | $H_{v_3}$ |
| $R_{4\pi/3}$ | $R_{4\pi/3}$ | $I$ | $R_{2\pi/3}$ | $H_{v_1}$ | $H_{v_3}$ | $H_{v_2}$ |
| $H_{v_3}$ | $H_{v_3}$ | $H_{v_1}$ | $H_{v_2}$ | $I$ | $R_{4\pi/3}$ | $R_{2\pi/3}$ |
| $H_{v_2}$ | $H_{v_2}$ | $H_{v_3}$ | $H_{v_1}$ | $R_{2\pi/3}$ | $I$ | $R_{4\pi/3}$ |
| $H_{v_1}$ | $H_{v_1}$ | $H_{v_2}$ | $H_{v_3}$ | $R_{4\pi/3}$ | $R_{2\pi/3}$ | $I$ |

**Table 14.1.** The multiplication table for $\mathrm{Dih}_3$. The products are ordered by left multiplication of the row matrix onto the column matrix.

We encode the symmetries of the square, see Figure 14.1, via four rotation matrices and four reflection matrices. $\mathrm{SDih}_4 = \{I, R_{\pi/2}, R_\pi, R_{3\pi/2}\}$ where

$$R_{\pi/2} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad R_\pi = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \text{and} \quad R_{3\pi/2} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Note that $R_{\pi/2}$ generates $\mathrm{SDih}_4$. The square however enjoys 4 additional symmetries, namely the reflections about lines through $e_1 = (1,0)$, $e_2 = (0,1)$, $d_1 = (1,1)/\sqrt{2}$ and $d_2 = (-1,1)/\sqrt{2}$. Namely, from Eq. (1.42),

$$H_{e_1} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad H_{d_1} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad H_{e_2} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} \quad H_{d_2} = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}.$$

Being reflections, each of these is its own inverse. Products of reflections however, though not reflections, are rotations and products of rotations and reflections are reflections. Its group structure is fully revealed in its multiplication table

| $\mathrm{Dih}_4$ | $I$ | $R_{\pi/2}$ | $R_\pi$ | $R_{3\pi/2}$ | $H_{e_1}$ | $H_{d_1}$ | $H_{e_2}$ | $H_{d_2}$ |
|---|---|---|---|---|---|---|---|---|
| $I$ | $I$ | $R_{\pi/2}$ | $R_\pi$ | $R_{3\pi/2}$ | $H_{e_1}$ | $H_{d_1}$ | $H_{e_2}$ | $H_{d_2}$ |
| $R_{\pi/2}$ | $R_{\pi/2}$ | $R_\pi$ | $R_{3\pi/2}$ | $I$ | $H_{d_2}$ | $H_{e_1}$ | $H_{d_1}$ | $H_{e_2}$ |
| $R_\pi$ | $R_\pi$ | $R_{3\pi/2}$ | $I$ | $R_{\pi/2}$ | $H_{e_2}$ | $H_{d_2}$ | $H_{e_1}$ | $H_{d_1}$ |
| $R_{3\pi/2}$ | $R_{3\pi/2}$ | $I$ | $R_{\pi/2}$ | $R_\pi$ | $H_{d_1}$ | $H_{e_2}$ | $H_{d_2}$ | $H_{e_1}$ |
| $H_{e_1}$ | $H_{e_1}$ | $H_{d_1}$ | $H_{e_2}$ | $H_{d_2}$ | $I$ | $R_{\pi/2}$ | $R_\pi$ | $R_{3\pi/2}$ |
| $H_{d_1}$ | $H_{d_1}$ | $H_{e_2}$ | $H_{d_2}$ | $H_{e_1}$ | $R_{3\pi/2}$ | $I$ | $R_{\pi/2}$ | $R_\pi$ |
| $H_{e_2}$ | $H_{e_2}$ | $H_{d_2}$ | $H_{e_1}$ | $H_{d_1}$ | $R_\pi$ | $R_{3\pi/2}$ | $I$ | $R_{\pi/2}$ |
| $H_{d_2}$ | $H_{d_2}$ | $H_{e_1}$ | $H_{d_1}$ | $H_{e_2}$ | $R_{\pi/2}$ | $R_\pi$ | $R_{3\pi/2}$ | $I$ |

**Table 14.2.** The multiplication table for $\mathrm{Dih}_4$. The products are ordered by left multiplication of the row matrix onto the column matrix.

We discern from the table that $R_{\pi/2}$ and $H_{e_1}$ together generate $\mathrm{Dih}_4$ in the sense that every matrix in $\mathrm{Dih}_4$ is a product of powers of $R_{\pi/2}$ and $H_{e_1}$. This in turn provides yet another means by which to visualize $\mathrm{Dih}_4$.

**Definition** 14.3. Given a group $G$ we suppose that $S$ is a **symmetric** subset of $G$ in the sense that for each $s \in S$ its inverse, $s^{-1}$ also lies in $S$. The **Cayley graph** of $G$ with respect to $S$, written $\mathrm{Cay}(G, S)$, has as its vertex set the elements of $G$, with two vertices, $g$ and $h$, joined by an edge whenever $h = sg$ for some $s \in S$.

For example, in Figure 14.2, we have drawn the **Cayley Graphs** $\mathrm{Cay}(\mathrm{Dih}_3, \{R_{2\pi/3}, R_{4\pi/3}, H_{v_3}\})$ and $\mathrm{Cay}(\mathrm{Dih}_4, \{R_{\pi/2}, R_{3\pi/2}, H_{e_1}\})$.

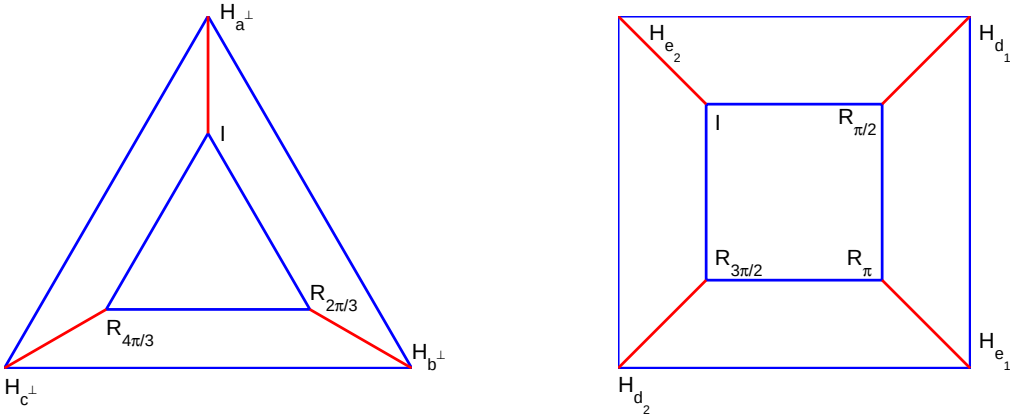



**Figure** 14.2. (A) The Cayley Graph of $\mathrm{Dih}_3$ with respect to $\{R_{2\pi/3}, R_{4\pi/3}, H_{v_3}\}$. The vertices of the graph are elements of $\mathrm{Dih}_3$ and edges are colored blue for multiplication by $R_{2\pi/3}$ or $R_{4\pi/3}$ and red for multiplication by $H_{v_3}$. (B) The Cayley Graph of $\mathrm{Dih}_4$ with respect to $\{R_{\pi/2}, R_{3\pi/2}, H_{e_1}\}$ with edges colored blue for multiplication by $R_{\pi/2}$ or $R_{3\pi/2}$ and red for multiplication by $H_{e_1}$.

We have observed that one reflection suffices to fill out the full group.

**Proposition** 14.4. Let $F \subset \mathbb{R}^3$ be a geometric figure with full symmetry group $G$ and rotation group $SG = G \cap \mathrm{SO}_3(\mathbb{R})$. If $F$ admits a reflection symmetry $H$ then $G = SG \cup SGH$.

Proof: Suppose $A \in G\backslash R$. Then $\det(A) = -1$ and $\det(AJ) = 1$, so $AJ \in R$. Thus $A = (AJ)J \in RJ$. End of Proof.

We next consider symmetries of the regular tetrahedron, Figure 14.3.

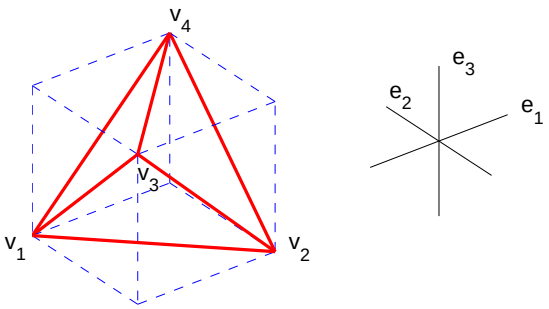



**Figure** 14.3. A tetrahedron embedded in a cube - and its associated coordinate system.

To build concrete symmtries we express the vertices in Figure 14.3 in terms of the coordinates

$$e_1 = (1\ 0\ 0)^T, \quad e_2 = (0\ 1\ 0)^T, \quad e_3 = (0\ 0\ 1)^T$$

as

$$v_1 = (-1\ \ -1\ \ -1)^T, \quad v_2 = (1\ \ -1\ \ -1)^T, \quad v_3 = (-1\ \ -1\ 1)^T, \quad v_4 = (1\ 1\ 1)^T.$$

We first note that lines from the origin through a vertex pass through the centroid of the opposite face of the tetrahedron. As such, to each vertex we may assign rotation by $2\pi/3$ and $4\pi/3$. Recalling our parametric form, Eq. (14.2), we arrive at the 8 distinct rotation matrices

$$R_{v_4,2\pi/3} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \quad R_{v_4,4\pi/3} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad R_{v_3,2\pi/3} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & -1 \\ -1 & 0 & 0 \end{pmatrix}, \quad R_{v_3,4\pi/3} = \begin{pmatrix} 0 & 0 & -1 \\ 1 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix},$$
$$R_{v_1,2\pi/3} = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 1 & 0 & 0 \end{pmatrix}, \quad R_{v_1,4\pi/3} = \begin{pmatrix} 0 & 0 & 1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \end{pmatrix}, \quad R_{v_2,2\pi/3} = \begin{pmatrix} 0 & -1 & 0 \\ 0 & 0 & 1 \\ -1 & 0 & 0 \end{pmatrix}, \quad R_{v_2,4\pi/3} = \begin{pmatrix} 0 & 0 & -1 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}. \tag{14.5}$$

Next there are rotations by $\pi$ on the three axes through centers of opposite faces of cube:

$$R_{e_1,\pi} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \quad R_{e_2,\pi} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}, \quad \text{and} \quad R_{e_3,\pi} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{14.6}$$

These, together with the identity, comprise the 12 elements of STet, the rotation group of the tetrahedron. On confirmation of its group status via computation of its multiplication table we find that $R_{v_4,2\pi/3}$ and $R_{e_1,\pi}$ generate STet. We illustrate it via the Cayley Graph of Figure 14.4.

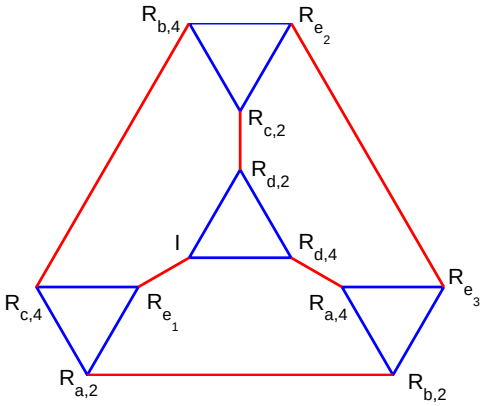



**Figure** 14.4. The Cayley Graph of STet with respect to $\{R_{v_4,2\pi/3}, R_{v_4,4\pi/3}, R_{e_1,\pi}\}$, with edges colored blue for action by $R_{v_4,2\pi/3}$ or $R_{v_4,4\pi/3}$ and red for action by $R_{e_1,\pi}$. We note that $(R_{e_1,\pi}R_{v_4,2\pi/3})^3 = I$ and that the Cayley graph is a truncated tetrahedron.

We will investigate the cube in the exercises and so move onto the icosahedron, illustrated in Figure 14.5. It has 12 vertices, 20 faces and 30 edges.

Each vertex of the icosahedron lies on an axis of 5–fold symmetry. We illustrate one such axis, through vertices 1 and 4, in Figure 14.5(A). There are 6 such axes, and for each axis there are 4 distinct (non–identity) rotations, for a total of 24 rotations.

Each centroid of a face lies on an axis of 3–fold symmetry. We illustrate one such axis, through faces 1-5-6 and 4-7-8, in Figure 14.5(B). There are 10 such axes, and for each axis there are 2 distinct (non–identity) rotations, for a total of 20 rotations.

Each midpoint of an edge lies on an axis of 2–fold symmetry. We illustrate one such axis, through edges 9-2 and 3-12, in Figure 14.5(C). There are 15 such axes, and for each axis there is one distinct (non–identity) rotation, for a total of 15 rotations.

With the identity element, SIco, the group of rotations of the icosahedron, has 1+24+20+15=60 elements.

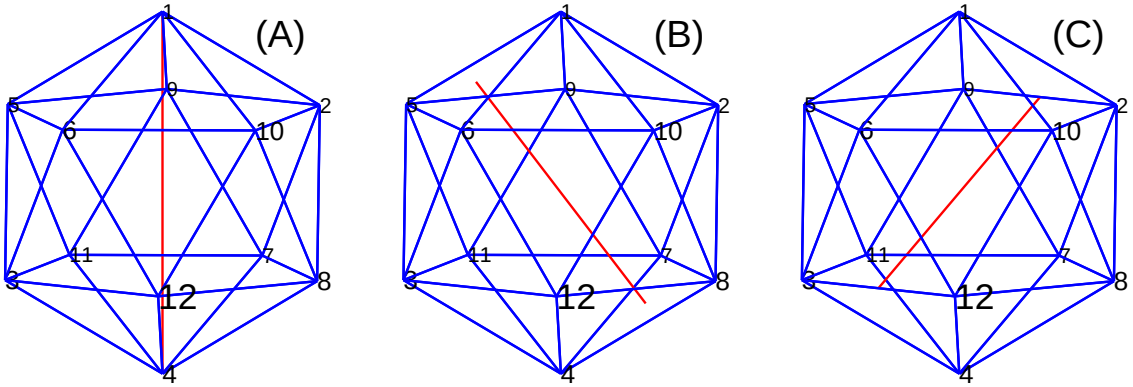



**Figure** 14.5. Axes of symmetry of the icosahedron. We have numbered the 12 vertices, with fontsize decreasing with distance from vertex 12 to indicate depth. (A) The icosahedron is invariant to rotation about the red axis by multiples of $2\pi/5$. (B) The icosahedron is invariant to rotation about the red axis by multiples of $2\pi/3$. (C) The icosahedron is invariant to rotation about the red axis by multiples of $2\pi/2$. `icosasym.m`

We denote by $R_A$, $R_B$ and $R_C$ the rotations depicted in Figure 14.5(A-C) respectively. We will

argue in Exer. 14.3 that

$$R_B = R_C R_A^4 R_C R_A. \tag{14.7}$$

By similar reasoning we can prove that $R_A$ and $R_C$ generate SIco. The associated Cayley graph, $\mathrm{Cay}(\mathrm{SIco}, \{R_A, R_A^{-1}, R_C\})$, is the truncated icosahedron depicted in Figure 14.6. This object is better known as a football, or soccer ball in the U.S., or "Buckyball" following the discovery of the Buckminsterfullerenes.

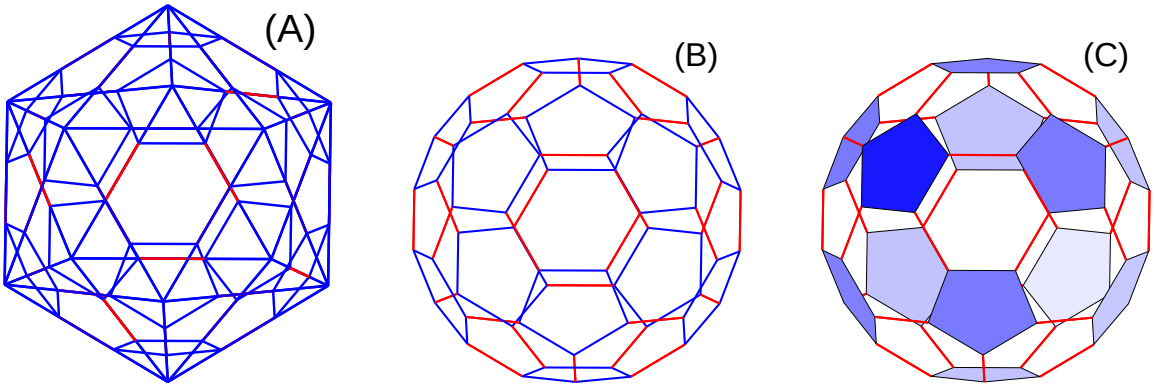



**Figure** 14.6. The Cayley Graph $\mathrm{Cay}(\mathrm{SIco}, \{R_A, R_A^{-1}, R_C\})$ is a Buckyball. (A) To illustrate "truncation" we have superimposed the icosahedron and Buckyball. Truncation of the 12 vertices produces 12 pentagons. The resulting 60 vertices are linked by the edge fragments remaining from the icosahedron to form 20 hexagons. (B) In the Cayley Graph each vertex is a unique element of SIco. The two blue edges incident at a vertex correspond to multiplication by $R_A$ and $R_A^{-1}$ while the red edge corresponds to multiplication by $R_C$. (C) This is the same view as (B) but with pentagonal shading (dark in foreground) to aid in the determination of depth. `bucky.m`

### 14.3. Permutation Groups

In the previous section we equated the symmetries of the triangle with the permutation of its vertices, and associated displacements. The associated permutation matrices of a given order comprise the most well studied of all of the groups. The simplest permutation matrix is the Elementary Permutation Matrix introduced in Eq. (3.16) as a tool for row swapping in Gaussian Elimination. For example,

$$P_{(12)} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

is obtained by interchanging rows 1 and 2 of the 3-by-3 identity matrix. There are two other Elementary Permutations, $P_{(13)}$ and $P_{(23)}$, as well as two interesting compositions

$$P_{(123)} \equiv P_{(13)}P_{(12)} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \quad \text{and} \quad P_{(132)} \equiv P_{(12)}P_{(13)} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}. \tag{14.8}$$

The subscript of each permutation is called a **cycle** and is customary shorthand for the more cumbersome notation of the previous section. For example, rather than writing $(1, 2, 3) \to (2, 1, 3)$ we write (12) and read that as "1 goes to 2" and "2 goes into the starting slot, in this case, 1." In addition, regarding Eq. (14.8), rather than writing $(1, 2, 3) \to (2, 3, 1)$ we write (123) and say "1

goes to 2, 2 goes to 3, and 3 cycles round to the start, 1." With this convention it should be clear that $P_{(312)} = P_{(123)}$.

Together with $I = P_{()}$ these five matrices comprise a group that we will designate $\mathrm{Per}_3$. To show that $\mathrm{Per}_3$ is indeed a group we compile its multiplication table.

| $\mathrm{Per}_3$ | $I$ | $P_{(12)}$ | $P_{(13)}$ | $P_{(23)}$ | $P_{(123)}$ | $P_{(132)}$ |
|---|---|---|---|---|---|---|
| $I$ | $I$ | $P_{(12)}$ | $P_{(13)}$ | $P_{(23)}$ | $P_{(123)}$ | $P_{(132)}$ |
| $P_{(12)}$ | $P_{(12)}$ | $I$ | $P_{(132)}$ | $P_{(123)}$ | $P_{(23)}$ | $P_{(13)}$ |
| $P_{(13)}$ | $P_{(13)}$ | $P_{(123)}$ | $I$ | $P_{(132)}$ | $P_{(12)}$ | $P_{(23)}$ |
| $P_{(23)}$ | $P_{(23)}$ | $P_{(132)}$ | $P_{(123)}$ | $I$ | $P_{(13)}$ | $P_{(12)}$ |
| $P_{(123)}$ | $P_{(123)}$ | $P_{(13)}$ | $P_{(23)}$ | $P_{(12)}$ | $P_{(132)}$ | $I$ |
| $P_{(132)}$ | $P_{(132)}$ | $P_{(23)}$ | $P_{(12)}$ | $P_{(13)}$ | $I$ | $P_{(123)}$ |

**Table 14.3.** The multiplication table for $\mathrm{Per}_3$. The products are ordered by left multiplication of the row matrix onto the column matrix.

We discern from this table that $P_{(12)}$ and $P_{(123)}$ generate $\mathrm{Per}_3$ and proceed to illustrate its Cayley Graph in Figure 14.7.

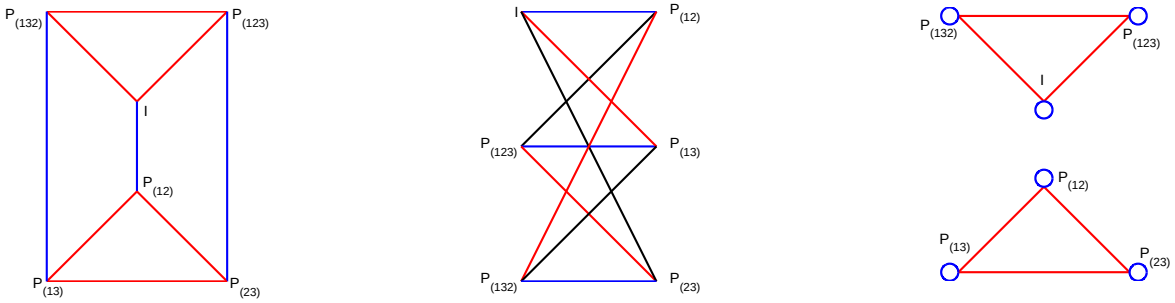



**Figure** 14.7. Three Cayley graphs of $\mathrm{Per}_3$. (A) $\mathrm{Cay}(\mathrm{Per}_3, \{P_{(12)}, P_{(123)}\})$. Red edges correspond to multiplication by $P_{(123)}$, blue edges to multiplication by $P_{(12)}$.

(B) $\mathrm{Cay}(\mathrm{Per}_3, \{P_{(12)}, P_{(13)}, P_{(23)}\})$. Red edges correspond to multiplication by $P_{(13)}$, blue edges to multiplication by $P_{(12)}$ and black edges to multiplication by $P_{(23)}$.

(C) $\mathrm{Cay}(\mathrm{Per}_3, \mathrm{Alt}_3)$. Red edges correspond to multiplication by $P_{(123)}$ and $P_{(132)}$, blue edges to multiplication by $I$.

We learn from Tab. 14.3 and Figure 14.7 that

$$\mathrm{Alt}_3 \equiv \{I, P_{(123)}, P_{(132)}\} \tag{14.9}$$

is a subgroup of $\mathrm{Per}_3$. It is referred to as the **alternating group** and looking back over our small list of groups we recognize that $\mathrm{Alt}_3$ and $\mathrm{Per}_3$ closely resemble the triangle groups $\mathrm{SDih}_3$ and $\mathrm{Dih}_3$. This notion is made precise through

**Definition** 14.5 Two groups, $G$ and $M$, are said to be **isomorphic** if there exists a function $\phi$, called an **isomorphism**, that takes $G$ to $M$ and satisfies
(a) $\phi$ is one–to–one, i.e., if $g_1 \neq g_2$ then $\phi(g_1) \neq \phi(g_2)$.
(b) $\phi$ is onto, i.e., for each $m \in M$ there exists a $g \in G$ for which $m = \phi(g)$.
(c) $\phi$ respects composition, i.e., $\phi(g_1 g_2) = \phi(g_1)\phi(g_2)$.
When $G$ and $M$ are isomorphic we write $G \sim M$.

A function that satisfies (a) and (b) is called a **bijection**. A function that satisfies (c) is called a **homomorphism**. For example, we construct an isomorphism, $\phi$, between $\mathrm{Alt}_3$ and $\mathrm{SDih}_3$ by taking generator to generator, i.e.,

$$\phi(P_{(123)}) \equiv R_{2\pi/3} \tag{14.10}$$

and then define its action on the remainder so as to satisfy property (c) and the respective multiplication tables. Namely,

$$\phi(P_{(132)}) = \phi(P_{(123)}P_{(123)}) = \phi(P_{(123)})\phi(P_{(123)}) = R_{2\pi/3}R_{2\pi/3} = R_{4\pi/3} \tag{14.11}$$

and

$$\phi(I) = \phi(P_{(123)}P_{(132)}) = \phi(P_{(123)})\phi(P_{(132)}) = R_{2\pi/3}R_{4\pi/3} = I. \tag{14.12}$$

These groups are both **cyclic** in the sense they are generated by powers of a single element. The most elementary cyclic groups are

**Proposition** 14.6. For integer $n > 0$ the set $\{0, 1, \ldots, n-1\}$ together with operation of addition modulo $n$ is a cyclic group, denoted $\mathbb{Z}_n$.

Proof: The identity element is 0. If $z \in \mathbb{Z}_n$ is nonzero then $n - z \in \mathbb{Z}_n$ and $z + (n - z) = n = 0 \bmod n$ and hence each member has an inverse. $\mathbb{Z}_n$ is cyclic because each element is a "power" of 1. End of Proof.

It follows that $\mathrm{Alt}_3 \sim \mathrm{SDih}_3 \sim \mathbb{Z}_3$. More generally, cyclic groups of the same size must be isomorphic, and so $\mathrm{SDih}_n \sim \mathbb{Z}_n$ for all $n$.

Returning to $\mathrm{Alt}_3 \sim \mathrm{SDih}_3$ we may extend it to an isomorphism of $\mathrm{Per}_3$ and $\mathrm{Dih}_3$ by connecting the additional generators via

$$\phi(P_{(12)}) \equiv H_{v_3} \tag{14.13}$$

and then, as above simply follow their lead. In particular

$$\begin{aligned} \phi(P_{(13)}) &= \phi(P_{(12)}P_{(123)}) = \phi(P_{(12)})\phi(P_{(123)}) = H_{v_3}R_{4\pi/3} = H_{v_2} \\ \phi(P_{(23)}) &= \phi(P_{(13)}P_{(132)}) = \phi(P_{(13)})\phi(P_{(132)}) = H_{v_2}R_{4\pi/3} = H_{v_1}. \end{aligned} \tag{14.14}$$

As we consider $\mathrm{Per}_n$ for large $n$ there will be an increasing variety of possible cycle shapes and combinations. In navigating this terrain it will helpful to know that

**Proposition** 14.7 Each $P_\sigma \in \mathrm{Per}_n$ can be written uniquely (up to order) as a product $P_{\sigma^1} \cdots P_{\sigma^k}$ of disjoint permutations.

Proof: Argue by induction, using the fact that disjoint permutations commute with one another. See Exer. 14.5. End of Proof.

This permits us to define $\mathrm{Alt}_n$ to be the subgroup of those matrices in $\mathrm{Per}_n$ that can be expressed as the product of an even number of elementary permutations (2–cycles). For example the $12 = 4!/2$ matrices of $\mathrm{Alt}_4$ are the identity, three order-2 matrices

$$P_{(12)(34)},\ P_{(13)(24)},\ P_{(14)(23)}, \tag{14.15}$$

and eight order-3 matrices

$$P_{(132)},\ P_{(123)} = P_{(132)}^2,\ P_{(142)},\ P_{(241)} = P_{(142)}^2,\ P_{(143)},\ P_{(341)} = P_{(143)}^2,\ P_{(243)},\ P_{(342)} = P_{(243)}^2. \tag{14.16}$$

We note the strong resemblance to STet, and construct an isomorphism that maps the eight rotations of order 3 to the eight permutations of order 3 and the 3 rotations of order 2 to the 3 permutations of order 2. In fact, like STet, one matrix of each order suffices to generate $\mathrm{Alt}_4$. Hence, on completing

$$\phi(R_{e_1,\pi}) \equiv P_{(12)(34)} \quad \text{and} \quad \phi(R_{v_4,2\pi/3}) \equiv P_{(132)}$$

to respect multiplication establishes the isomorphism

$$\boxed{\mathrm{STet} \sim \mathrm{Alt}_4.} \tag{14.17}$$

In addition to multiplication tables and Cayley graphs we will also find it convenient to view groups as unions of conjugacy classes. The **conjugacy class** of $g \in G$ is

$$\mathrm{Conj}_g(G) \equiv \{h^{-1}gh : h \in G\}.$$

Note that $\mathrm{Conj}_I(G)$ is always the singleton $\{I\}$ and that if $G$ is abelian then $\mathrm{Conj}_g(G) = \{g\}$ for every $g \in G$. For the nonabelian group $\mathrm{Per}_3$ we glean from its multiplication table, Tab. 14.3,

$$\begin{aligned}
\mathrm{Conj}_{(12)}(\mathrm{Per}_3) &= \{B^{-1}P_{(12)}B : B \in \mathrm{Per}_3\} \\
&= \{I^{-1}P_{(12)}I, P_{(12)}^{-1}P_{(12)}P_{(12)}, P_{(13)}^{-1}P_{(12)}P_{(13)}, P_{(23)}^{-1}P_{(12)}P_{(23)}, P_{(123)}^{-1}P_{(12)}P_{(123)}, P_{(132)}^{-1}P_{(12)}P_{(132)}\} \\
&= \{P_{(12)}, P_{(12)}, P_{(13)}P_{(12)}P_{(13)}, P_{(23)}P_{(12)}P_{(23)}, P_{(132)}P_{(12)}P_{(123)}, P_{(123)}P_{(12)}P_{(132)}\} \\
&= \{P_{(12)}, P_{(12)}, P_{(13)}P_{(132)}, P_{(23)}P_{(123)}, P_{(132)}P_{(23)}, P_{(123)}P_{(13)}\} \\
&= \{P_{(12)}, P_{(12)}, P_{(23)}, P_{(13)}, P_{(13)}, P_{(23)}\} \\
&= \{P_{(12)}, P_{(13)}, P_{(23)}\}.
\end{aligned}$$

This tedious calculation is easily automated, but our chief interest is in discerning patterns. As we proceed to compute conjugacy classes of the remaining elements of $\mathrm{Per}_3$ we find only one additional class. Hence

$$I, \quad \{P_{(12)}, P_{(13)}, P_{(23)}\} \quad \text{and} \quad \{P_{(123)}, P_{(132)}\}. \tag{14.18}$$

comprise the three conjugacy classes of $\mathrm{Per}_3$ – and we quickly observe that their members indeed exhaust $\mathrm{Per}_3$ and, more interestingly, each class contains only cycles of the same "type." We now prove that these observations hold for all $\mathrm{Per}_n$.

**Definition** 14.8. If $\sigma \in \mathrm{Per}_n$ and $\sigma = \prod \sigma_i$ is its decomposition into disjoint cycles then the $j$th element of the **cycle type** of $\sigma$ is $\mathrm{type}_j(\sigma) \equiv \#\{\sigma_i : |\sigma_i| = j\}$, $j = 1, 2, \ldots, n$.

For example, the types of cycles appearing in (14.18) are

$$\mathrm{type}((1)(2)(3)) = (3,0,0), \quad \mathrm{type}((12)(3)) = (1,1,0) \quad \text{and} \quad \mathrm{type}((123)) = (0,0,1).$$

More generally, suppose $P_\sigma$ and $P_\pi$ belong to $\mathrm{Per}_n$. If $\sigma = (\sigma_1, \sigma_2, \cdots, \sigma_k)$ we say $\sigma$ takes $\sigma_m$ to $\sigma_{m+1}$ and write $\sigma(\sigma_m) = \sigma_{m+1}$. We next define $\pi \circ \sigma = (\pi(\sigma_1), \pi(\sigma_2), \cdots, \pi(\sigma_k))$ and prove

$$P_\pi P_\sigma = P_{\pi\circ\sigma} P_\pi, \tag{14.19}$$

by applying each to $e_j$, the $j$th coordinate vector. Starting with the left side of (14.19) we find

$$P_\pi P_\sigma e_j = P_\pi e_{\sigma(j)} = e_{\pi(\sigma(j))},$$

while on the right

$$P_{\pi\circ\sigma}P_\pi e_j = P_{\pi\circ\sigma}e_{\pi(j)} = e_{(\pi\circ\sigma)(\pi(j))},$$

hence it remains to confirm that

$$(\pi\circ\sigma)(\pi(j)) = (\pi(\sigma_1),\pi(\sigma_2),\ldots,\pi(\sigma_k))(\pi(j)) = \begin{cases}\pi(j) & \text{if} \quad j\notin\sigma\\ \pi(\sigma_{i+1}) & \text{if} \quad j=\sigma_i,\end{cases}$$

is indeed precisely $\pi(\sigma(j))$. On rearranging Eq. (14.19) we find $P_\pi P_\sigma P_\pi^{-1} = P_{\pi\circ\sigma}$ and so every member of the conjugacy class of a $k$-cycle is a $k$-cycle. Conversely, if $P_\sigma$ and $P_\phi$ are $k$-cycles then $P_\pi$ will do where $\pi(\sigma_j)\equiv\phi_j$.

Next suppose that $P_\sigma = P_{\sigma^1}P_{\sigma^2}$ is the product of disjoint cycles and that $P_\pi P_\sigma P_\pi^{-1} = P_\phi$. To see that $P_\phi$ has the same cycle type as $P_\sigma$ note that, using (14.19),

$$P_\pi P_\sigma = P_\pi P_{\sigma^1}P_{\sigma^2} = P_{\pi\circ\sigma^1}P_\pi P_{\sigma^2} = P_{\pi\circ\sigma^1}P_{\pi\circ\sigma^2}P_\pi$$

and hence

$$P_\phi = P_\pi P_\sigma P_\pi^{-1} = P_{\pi\circ\sigma^1}P_{\pi\circ\sigma^2}. \tag{14.20}$$

As the cycle lengths of $\pi\circ\sigma^1$ and $\pi\circ\sigma^2$ are the same as $\sigma_1$ and $\sigma_2$ respectively, and the cycle decomposition of $P_\phi$ is unique (Prop. 14.2) it follows from (14.20) that $P_\sigma$ and $P_\phi$ have the same cycle type. Conversely, if $P_\sigma = P_{\sigma^1}P_{\sigma^2}$ and $P_\phi = P_{\phi^1}P_{\phi^2}$ are two permutations with identical cycle types then we may define $P_\pi = P_{\pi^1}P_{\pi^2}$ where $\pi^k(\sigma_j^k) = \phi^k(j)$ and confirm that $P_\pi P_\sigma P_\pi^{-1} = P_\phi$. As this argument generalizes immediately to arbitrary cycle types, we have proven

**Proposition** 14.9 The conjugacy class of $P\in\mathrm{Per}_n$ is the set of permutation matrices with the same cycle type as $P$. That is

$$\mathrm{Conj}_P(\mathrm{Per}_n) = \{Q\in\mathrm{Per}_n : \mathrm{type}(Q) = \mathrm{type}(P)\}.$$

The conjugacy classes of $\mathrm{Per}_n$ are therefore disjoint and their union is all of $\mathrm{Per}_n$.

Regarding the sizes of our conjugacy classes, in building a 2-cycle in $\mathrm{Per}_n$ there are $n$ choices for the first element and $(n-1)$ choices for the second. As $P_{(ij)} = P_{(ji)}$ we have over counted by a factor of 2 hence

$$\left|\mathrm{Conj}_{(12)}(\mathrm{Per}_n)\right| = n(n-1)/2.$$

More generally, for $k$-cycles

$$\left|\mathrm{Conj}_{(12\cdots k)}(\mathrm{Per}_n)\right| = \frac{n!}{k(n-k)!}.$$

For products of disjoint cycles

$$\left|\mathrm{Conj}_{(12)(34)}(\mathrm{Per}_n)\right| = \frac{n(n-1)}{2}\frac{(n-2)(n-3)}{2}\frac{1}{2} = \frac{n!}{8(n-4)!}$$

and

$$\left|\mathrm{Conj}_{(12)(345)}(\mathrm{Per}_n)\right| = \frac{n(n-1)}{2}\frac{(n-2)(n-3)(n-4)}{3} = \frac{n!}{6(n-5)!}.$$

With this we can present the conjugacy classes of $\mathrm{Per}_n$ in tabular form. For example,

| Rep | $I$ | $P_{(12)}$ | $P_{(123)}$ | $P_{(12)(34)}$ | $P_{(1234)}$ |
|---|---|---|---|---|---|
| Size | 1 | 6 | 8 | 3 | 6 |

**Table 14.4**. The Conjugacy Classes of $\mathrm{Per}_4$. Here Rep stands for representative – note that we may choose any member with the same cycle type.

| Rep | $I$ | $P_{(12)}$ | $P_{(123)}$ | $P_{(12)(34)}$ | $P_{(1234)}$ | $P_{(12)(345)}$ | $P_{(12345)}$ |
|---|---|---|---|---|---|---|---|
| Size | 1 | 10 | 20 | 15 | 30 | 20 | 24 |

**Table 14.5**. The Conjugacy Classes of $\mathrm{Per}_5$.

The conjugacy classes of $\mathrm{Alt}_n$ are closely related to those of $\mathrm{Per}_n$, however the explicit calculation

$$\mathrm{Conj}_{(123)}(\mathrm{Alt}_4) \neq \mathrm{Conj}_{(132)}(\mathrm{Alt}_4)$$

already indicates that they are not defined simply by their type. The answer will depend on the nature of $\mathrm{Alt}_n$ as a subgroup of $\mathrm{Per}_n$. Cosets and quotient groups are excellent tools for studying subgroups.

### 14.4. Linear, Free, and Quotient Groups

In this section we begin with the infinite groups of invertible matrices and then select finite subgroups by restricting our matrix elements to finite sets of integers.

We begin with the **General Linear Group**,

$$\mathrm{GL}_n(\mathbb{R}) = \{A \in \mathbb{R}^{n\times n} : \det(A) \neq 0\},$$

of invertible $n$-by-$n$ real matrices and its subgroup, called the **Special Linear Group**,

$$\mathrm{SL}_n(\mathbb{R}) = \{A \in \mathbb{R}^{n\times n} : \det(A) = 1\},$$

of $n$-by-$n$ real matrices with unit determinant, and its subgroup

$$\mathrm{SL}_n(\mathbb{Z}) = \{A \in \mathbb{Z}^{n\times n} : \det(A) = 1\}$$

of integer matrices with determinant 1. We note that both

$$A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix} \tag{14.21}$$

lie in $\mathrm{SL}_2(\mathbb{Z})$ and proceed to show that they generate a free subgroup. This will give us an indication of the vastness of $\mathrm{SL}_2(\mathbb{Z})$ and a start on the construction (to be finished in Chapter 16) of regular graphs with large girth.

A group $G$ is called **free** if there is a subset $S$ of $G$ such that every element of $G$ can be written in one and only one way as a product of finitely many elements of $S$ and their inverses (disregarding trivial variations such as $AB^{-1} = AC^{-1}CB^{-1}$). An important example is

**Proposition** 14.10. If $G$ is the collection of all products of $\{A, A^{-1}, B, B^{-1}\}$ for the $A$ and $B$ of Eq. (14.21) then $G$ is free.

Proof: We note that elements of $G$ belong to one of four types

(a) starting and finishing with a power of $A$:

$$A^{k_1}B^{\ell_1}A^{k_2}B^{\ell_2}\cdots A^{k_r}B^{\ell_r}A^{k_{r+1}},$$

(b) starting and finishing with a power of $B$:

$$B^{k_1}A^{\ell_1}B^{k_2}A^{\ell_2}\cdots B^{k_r}A^{\ell_r}B^{k_{r+1}},$$

(c) starting with a power of $A$ and finishing with a power of $B$:

$$A^{k_1}B^{\ell_1}A^{k_2}B^{\ell_2}\cdots A^{k_r}B^{\ell_r},$$

(d) starting with a power of $B$ and finishing with a power of $A$:

$$B^{k_1}A^{\ell_1}B^{k_2}A^{\ell_2}\cdots B^{k_r}A^{\ell_r}.$$

To show that $G$ is free we must show that none of these products reduce to the identity. The simplest way to see this is by arguing that they move vectors from place to place. The places we have in mind are

$$E=\{x\in\mathbb{R}^2:|x_2|>|x_1|\}\quad\text{and}\quad F=\{x\in\mathbb{R}^2:|x_1|>|x_2|\}.$$

Recall that we showed in Exer. 1.3 that $A^n$ takes $E$ to $F$ while $B^n$ takes $F$ to $E$ for all positive integer $n$. Please confirm that in fact these same mappings hold for all negative integer powers as well.

Applying these results a product of type (a) we find, that if $x\in E$ then $A^{k_{r+1}}x\in F$ and so $B^{\ell_r}A^{k_{r+1}}x\in E$ and that continuing in this fashion produces

$$A^{k_1}B^{\ell_1}A^{k_2}B^{\ell_2}\cdots A^{k_r}B^{\ell_r}A^{k_{r+1}}x\in F.$$

That is, products of type (a) take vectors in $E$ to vectors in $F$. As $E$ and $F$ are disjoint it follows that no such product acts like $I$.

By identical reasoning it follows that products of type (b) take $F$ to $E$ and so can not be $I$.

This reasoning however does not apply to products of type (c) or (d). We therefore transform them to products of type (a) or (b). In particular given, $W_3=A^{k_1}B^{\ell_1}\cdots A^{k_r}B^{\ell_r}$, of type (c) we choose a power $k\neq k_1$ and note that $A^{-k}W_3A^k$ is a product of type (a) and so $A^{-k}W_3A^k\neq I$. It follows $W_3\neq A^kIA^{-k}=I$. By identical reasoning products of type (d) may be transformed to products of type (b). End of Proof.

This result gives us a sense of the infinite scope of $\mathrm{SL}_2(\mathbb{Z})$. Things become finite and more concrete upon reducing all integers modulo $q$. Recall that $\mathrm{mod}_q(x)$ is the remainder of $x$ after division by $q$. More precisely

$$\boxed{\mathrm{mod}_q(x)\equiv x-q\cdot\mathrm{floor}(x/q).}$$

For prime $q$ we may then define the **Field** $\mathbb{F}_q$ comprised of the integers $\{0,1,\ldots,q-1\}$ where addition and multiplication are reduced modulo $q$. For example, with $q=3$ we find

$$\begin{aligned}&0+0=0,\quad 0+1=1,\quad 0+2=2,\quad 1+1=2,\quad 2+1=0,\quad 2+2=1\\&0\cdot 0=0,\quad 0\cdot 1=0,\quad 0\cdot 2=0,\quad 1\cdot 1=1,\quad 2\cdot 1=2,\quad 2\cdot 2=1.\end{aligned}$$

With this we may define the unit determinant $n$-by-$n$ matrices with elements in $\mathbb{F}_q$,

$$\mathrm{SL}_n(q)=\{A\in\mathbb{F}_q^{n\times}:\det(A)=1\}.$$

For example $\mathrm{SL}_2(2)$ is comprised of the identity and the five matrices

$$A_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad A_4 = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \quad A_5 = A_4^2 = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}. \tag{14.22}$$

The first three are order 2 and the second two are order 3 and we note that $\mathrm{SL}_2(2)$ is generated by $A_1$ and $A_4$ and that

$$\phi(P_{(12)}) \equiv A_1 \quad \text{and} \quad \phi(P_{(123)}) \equiv A_4 \tag{14.23}$$

establishes an isomorphism between $\mathrm{SL}_2(2)$ and $\mathrm{Per}_3$.

We note that the $A$ and $B$ of Eq. (14.21) lie in $\mathrm{SL}_2(3)$ and that they generate $\mathrm{SL}_2(3)$ in the sense that every matrix in $\mathrm{SL}_2(3)$ is a finite product of $A$ and $B$. We illustrate this in Figure 14.8 via the Cayley Graph of $\mathrm{SL}_2(3)$ with respect to $S = \{A, B, A^{-1}, B^{-1}\}$ where

$$A^{-1} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad B^{-1} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \tag{14.24}$$

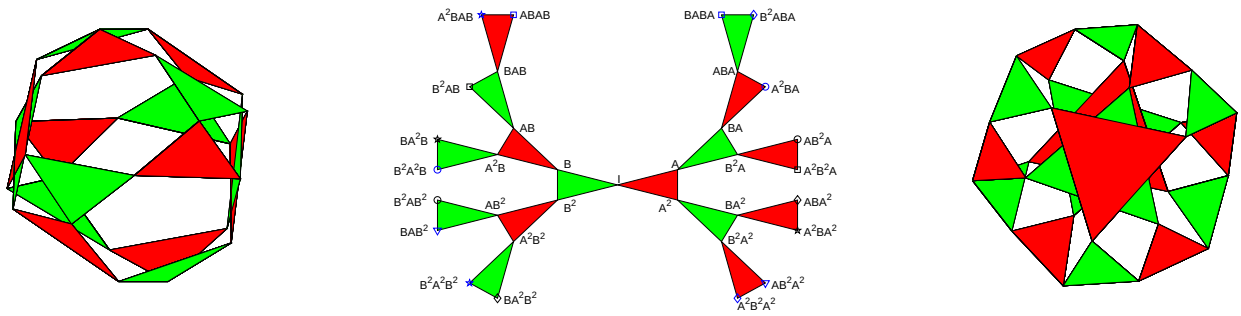



**Figure** 14.8. The Cayley graph $\mathrm{Cay}(\mathrm{SL}_2(3), S)$ flat and rolled.

Regarding the size of $\mathrm{SL}_2(q)$ we note that there are $q^2 - 1$ ways of choosing the first (nonzero) column in

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Once that is done we must choose $b$ and $d$ to assure that $\mathrm{mod}_q(ad - bc) = 1$ and so

$$|\mathrm{SL}_2(q)| = q(q^2 - 1). \tag{14.25}$$

There is a natural map that associates groups of linear transformations of $\mathbb{R}^{n+1}$ to groups of Möbius transformations of $\mathbb{R}^n_\infty \equiv \mathbb{R}^n \cup \{\infty\}$. To fix ideas we develop this for $n = 1$. Given a matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

in $\mathrm{GL}_2(\mathbb{R})$ we associate the **Möbius Transformation** of $\mathbb{R}_\infty$,

$$\phi_A(x) = \frac{ax + b}{cx + d}, \tag{14.26}$$

where, if $c = 0$ we define $\phi_A(\infty) = \infty$ while if $c \neq 0$ we set $\phi_A(\infty) = a/c$ and $\phi_A(-d/c) = \infty$.

Given $A_1$ and $A_2$ in $\mathrm{GL}_2(\mathbb{R})$ it is not difficult to show that $\phi_{A_2}(\phi_{A_1}(x)) = \phi_{A_2A_1}(x)$ and so the class of Möbius Transformations, $\mathbb{M}(\mathbb{R}_\infty)$, form a group under composition. Of course several matrices get mapped to the same Möbius transformation. The root redundancy is captured by the set of matrices in $\mathrm{GL}_2(\mathbb{R})$ that map onto the identity in $\mathbb{M}(\mathbb{R}_\infty)$,

$$N_G \equiv \{A \in \mathrm{GL}_2(\mathbb{R}) : \phi_A(x) = x\} = \{rI : r \in \mathbb{R} \setminus \{0\}\}. \tag{14.27}$$

We now proceed to show that $\phi$ induces an isomorphism, $\Phi$, between the cosets of $N_G$ and the elements of $\mathbb{M}(\mathbb{R}_\infty)$. A right **coset** (left coset) of $N_G$ is the collection of all right (left) multiples of some $B \in \mathrm{GL}_2(\mathbb{R})$ by elements of $N_G$. Namely

$$N_GB = \{rB : r \in \mathbb{R} \setminus \{0\}\} = BN_G = \{Br : r \in \mathbb{R} \setminus \{0\}\}.$$

We so define the map from cosets of $N_G$ to Möbius transformations

$$\Phi(N_GA) = \phi_A \tag{14.28}$$

and confirm that it is one-to-one and onto. If $\Phi(N_GA) = \Phi(N_GB)$ then $\phi_A = \phi_B$ and so $(\phi_{B^{-1}} \circ \phi_A)(x) = x$ and so $B^{-1}A \in N_G$ and so $A$ is a nonzero multiple of $B$ and so in the same coset. We have therefore proven that $\Phi$ is one-to-one.

We next define multiplication of cosets as

$$N_GA * N_GB \equiv N_GAB = \{rAB : r \in \mathbb{R} \setminus \{0\}\},$$

and note that under this operation the cosets of $N_G$ constitute a group and that

$$\Phi(N_GA * N_GB) = \Phi(N_GAB) = \phi_{AB} = \phi_A \circ \phi_B = \Phi(N_GA) \circ \Phi(N_GB).$$

It is time that we named this set of cosets of $N_G$. A subgroup is said to be **normal** when its left and right cosets coincide. The set of cosets of a normal subgroup defines the **quotient** group, written $\mathrm{GL}_2(\mathbb{R})/N_G$. We have proven that this quotient is isomorphic to the Möbius transformations of $\mathbb{R}_\infty$. As the latter can also be shown to be isomorphic to the symmetry group of the projective line, we write

$$\mathrm{PGL}_2(\mathbb{R}) \equiv \mathrm{GL}_2(\mathbb{R})/N_G,$$

and speak of it as the **Projective General Linear group**. This procedure is considerably more general than our one example. For example, if we restrict $\phi_A$ to those $A \in \mathrm{SL}_2(\mathbb{R})$ then the associated kernel,

$$N_S \equiv \{A \in \mathrm{SL}_2(\mathbb{R}) : \phi_A(x) = x\} = \{I, -I\}, \tag{14.29}$$

is a restriction of $N_G$. The cosets of $N_S$ again comprise a group under

$$N_SA * N_SB \equiv N_SAB = \{AB, -AB\},$$

that defines

$$\mathrm{PSL}_2(\mathbb{R}) \equiv \mathrm{SL}_2(\mathbb{R})/N_S,$$

the **Projective Special Linear group**. Via $\Phi(N_SA) = \phi_A$ we find $\mathrm{PSL}_2(\mathbb{R})$ isomorphic to the subgroup of Möbius transformations for which $ad - bc = 1$.

We may also vary the field, for example $\mathrm{PSL}_2(\mathbb{Z})$ is the modular group and $\mathrm{PSL}_2(q) = \mathrm{SL}_2(q)/N_S$ where $N_S = \{I, (q-1)I\}$. As such $\mathrm{PSL}_2(3)$ is isomorphic to the group of 12 Möbius transformations: the identity

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \to x$$

three order 2 elements

$$\begin{pmatrix}0&2\\1&0\end{pmatrix},\begin{pmatrix}0&1\\2&0\end{pmatrix}\to\frac{2}{x},\quad \begin{pmatrix}1&1\\1&2\end{pmatrix},\begin{pmatrix}2&2\\2&1\end{pmatrix}\to\frac{x+1}{x+2},\quad \begin{pmatrix}2&1\\1&1\end{pmatrix},\begin{pmatrix}1&2\\2&2\end{pmatrix}\to\frac{2x+1}{x+1}$$

and 8 elements of order 3

$$\begin{pmatrix}1&1\\0&1\end{pmatrix},\begin{pmatrix}2&2\\0&2\end{pmatrix}\to x+1,\quad \begin{pmatrix}1&2\\0&1\end{pmatrix},\begin{pmatrix}2&1\\0&2\end{pmatrix}\to x+2=(x+1)\circ(x+1),$$
$$\begin{pmatrix}0&2\\1&1\end{pmatrix},\begin{pmatrix}0&1\\2&2\end{pmatrix}\to\frac{2}{x+1},\quad \begin{pmatrix}2&2\\1&0\end{pmatrix},\begin{pmatrix}1&1\\2&0\end{pmatrix}\to\frac{2x+2}{x}=\frac{2}{x+1}\circ\frac{2}{x+1}$$
$$\begin{pmatrix}1&0\\1&1\end{pmatrix},\begin{pmatrix}2&0\\2&2\end{pmatrix}\to\frac{x}{x+1},\quad \begin{pmatrix}1&0\\2&1\end{pmatrix},\begin{pmatrix}2&0\\1&2\end{pmatrix}\to\frac{x}{2x+1}=\frac{x}{x+1}\circ\frac{x}{x+1}$$
$$\begin{pmatrix}0&1\\2&1\end{pmatrix},\begin{pmatrix}0&2\\1&2\end{pmatrix}\to\frac{1}{2x+1},\quad \begin{pmatrix}1&2\\1&0\end{pmatrix},\begin{pmatrix}2&1\\2&0\end{pmatrix}\to\frac{x+2}{x}=\frac{1}{2x+1}\circ\frac{1}{2x+1}$$

**Proposition** 14.11. $\mathrm{PSL}_2(3)$ is isomorphic to $\mathrm{Alt}_4(\mathbb{R})$.

We close with an investigation of quotient groups, and more generally cosets, as a means to better approach conjugacy classes.

**Proposition** 14.12. If $H$ is a subgroup of $G$ then $H$ is normal if and only if it is a union of conjugacy classes.

Proof: Suppose $H$ is a union of conjugacy classes. Then, for each $h\in H$ and $g\in G$ it follows that $g^{-1}hg\in H$. That is $hg\in gH$ and $g^{-1}h\in Hg^{-1}$ for each $h\in H$, and so $Hg\subset gH$ and $g^{-1}H\subset Hg^{-1}$ for each $g\in G$. It follows that the left and right cosets of $H$ coincide, i.e., that $H$ is normal.

Conversely, suppose that $H$ is normal. Then, for each $g\in G$ it follows $Hg=gH$. That is, $g^{-1}Hg=H$, and so $g^{-1}hg\in H$ for each $h\in H$. Hence $\mathrm{Conj}_h(G)\subset H$ for each $h\in H$. And so

$$H=\bigcup_{h\in H}h\subset\bigcup_{h\in H}\mathrm{Conj}_h(G)\subset H,$$

and so equality must hold throughout. End of Proof.

If $H$ is a subgroup of $G$ then the number of left cosets of $H$ in $G$ is called the **index** of $H$ in $G$, and written $[G:H]$. The main result is.

**Proposition** 14.13. **Lagrange's Theorem**. If $H$ is a subgroup of the group $G$ then $[G:H]=|G|/|H|$.

Proof: Let's first show that distinct cosets are actually disjoint. If $aH\cap bH\neq\emptyset$ then there exist $h_1,h_2\in H$ such that $ah_1=bh_2$. Hence $a=bh_2h_1^{-1}$ and so $ah=bh_2h_1^{-1}h\in bH$ for each $h\in H$. This shows that $aH\subset bH$. By like reasoning we may write $b=ah_1h_2^{-1}$ and so conclude that $bH\subset aH$.

Although disjoint, we note that $(ba^{-1})ah=bh$ and so $(ba^{-1})$ acts as an isomorphism of $aH$ and $bH$. It follows that $|aH|=|bH|=|H|$.

Finally, for each $a\in G$ we note that $a\in aH$ and so $G$ is the union of its distinct cosets. As all cosets have the same size it follows that that size, $|H|$, divides the size of $|G|$. In other words, the number of distinct cosets of $H$ is $|G|/|H|$. End of Proof.

The smallest index is two and in that case $H$ and its complement, $G \setminus H$, are the two left cosets of $H$. As they are also the two right cosets we have established: If $[G : H] = 2$ then $H$ is normal. Index 2 subgroups also have the nice property that

$$ab \in H \quad \text{iff} \quad a, b \in H \quad \text{or} \quad a, b \in G \setminus H. \tag{14.30}$$

To see this, suppose $a, b \in G \setminus H$. It follows that $aH = bH = G \setminus H$ (for if $aH = H$ then $a \in H$.) Now if $ab \in G \setminus H$ then as above $abH = aH = bH = G \setminus H$. But then $a^{-1}abH = a^{-1}aH = H$ which contradicts $b \notin H$.

**Proposition** 14.14. $\mathrm{Alt}_n$ is the only subgroup of $\mathrm{Per}_n$ of index 2.

Proof: If $H$ is a subgroup of $\mathrm{Per}_n$ of index 2 then $H$ is normal. Hence if it contains one element of a conjugacy class then it contains the whole class. It follows from Prop. 14.9 that if $H$ contains one element of a certain cycle structure then it contains all elements of that cycle structure. Now if $H$ contains the 2–cycles, then being a group it contains all products of 2–cycles, but this gives all of $\mathrm{Per}_n$. Hence $H$ does not contain a 2–cycle.

We next deduce from Eq. (14.30) that $H$ contains all even products of two cycles, i.e., $H = \mathrm{Alt}_n$. End of Proof.

With this result we can now establish the isomorphism

$$\boxed{\mathrm{Alt}_5 \sim \mathrm{SIco.}} \tag{14.31}$$

This is accomplished by identifying 5 objects faithfully permuted by SIco. This will in turn translate into an injective homomorphism of SIco into $\mathrm{Per}_5$. As its image is a subgroup of $\mathrm{Per}_5$ with 60 elements, we will conclude, by above, that this subgroup is $\mathrm{Alt}_5$. One way to "see" these 5 distinguished objects is to return to the 15 two–fold symmetries of Figure 14.5(C). With regard to Figure 14.9 we reconstruct the red axis through the midpoints of segments 9:2 and 3:12. The plane through these two segments bisects segments 5:11 and 10:8 and begets a second red axis orthogonal to the first. The plane through these latter segments bisects segments 6:1 and 4:7 and begets a third red axis, orthogonal to the first two.

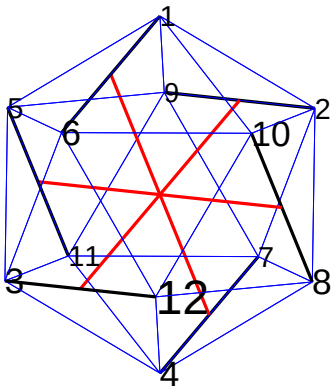



**Figure** 14.9. A red coordinate frame and a family of 6 black edges permuted by SIco.

Let us denote the red coordinate frame in Figure 14.9 by $C_{3:12}$. By rotation of $2\pi/5$ about the 1:4 axis (recall Figure 14.5(A)) we discern 4 more distinct coordinate frames, $C_{12:8} = R_A C_{3:12}$,

$C_{8:7} = R_A^2 C_{3:12}$, $C_{7:11} = R_A^3 C_{3:12}$ and $C_{11:3} = R_A^4 C_{3:12}$. Each $R \in \mathrm{SIco}$ acts to permute these 5 frames. To see that this mapping is injective note that if an element of SIco leaves each frame invariant then it leaves each set of 6 black edges invariant - and hence leaves all 30 edges of the icosahedron invariant, and hence can be only the identity element. It follows that SIco is isomorphic to a 60 element subgroup of $\mathrm{Per}_5$. Our claim, Eq. (14.31), now follows from Prop. 14.14.

This isomorphism and the Cayley graph illustrated in Figure 14.6 will permit us to use the considerable body of knowledge regarding $\mathrm{Alt}_5$ to understand the electronic structure (in the sense of §12.5) of the Buckyball. This body of knowledge begins with conjugacy classes.

We first observe that the size of the conjugacy class of $g \in G$ can be expressed in terms of the size of its **centralizer**

$$\mathrm{Cent}_g(G) = \{h \in G : hg = gh\}.$$

We establish in Exer. 14.15 that each $\mathrm{Cent}_g(G)$ is a subgroup of $G$ and that

$$\boxed{|\mathrm{Conj}_g(G)| = |G : \mathrm{Cent}_g(G)| = |G|/|\mathrm{Cent}_g(G)|.} \tag{14.32}$$

From here we can now classify conjugacy classes in $\mathrm{Alt}_n$.

**Proposition** 14.15. Suppose $P \in \mathrm{Alt}_n$. If $P$ commutes with an odd element of $\mathrm{Per}_n$ then

$$\mathrm{Conj}_P(\mathrm{Per}_n) = \mathrm{Conj}_P(\mathrm{Alt}_n),$$

else $\mathrm{Conj}_P(\mathrm{Alt}_n)$ and $\mathrm{Conj}_{P_{(12)}PP_{(12)}}(\mathrm{Alt}_n)$ are disjoint and the same size and

$$\mathrm{Conj}_P(\mathrm{Per}_n) = \mathrm{Conj}_P(\mathrm{Alt}_n) \cup \mathrm{Conj}_{P_{(12)}PP_{(12)}}(\mathrm{Alt}_n).$$

Proof: Suppose that $PR = RP$ for odd $R$ and that $Y \in \mathrm{Conj}_P(\mathrm{Per}_n)$. The latter implies that $Y = H^{-1}PH$ for some $H \in \mathrm{Per}_n$. If $H$ is even then $Y$ is even and $Y \in \mathrm{Conj}_P(\mathrm{Alt}_n)$; while if $H$ is odd then $RH \in \mathrm{Alt}_n$ and

$$Y = H^{-1}PH = H^{-1}R^{-1}RPH = H^{-1}R^{-1}PRH = (RH)^{-1}P(RH),$$

so again $Y \in \mathrm{Conj}_P(\mathrm{Alt}_n)$. Thus $\mathrm{Conj}_P(\mathrm{Per}_n) \subset \mathrm{Conj}_P(\mathrm{Alt}_n)$ and so $\mathrm{Conj}_P(\mathrm{Per}_n) = \mathrm{Conj}_P(\mathrm{Alt}_n)$.

Conversely, if $P$ does not commute with any odd permutations then $\mathrm{Cent}_P(\mathrm{Per}_n) = \mathrm{Cent}_P(\mathrm{Alt}_n)$ and so Eq. (14.32) implies that

$$\begin{aligned} |\mathrm{Conj}_P(\mathrm{Alt}_n)| &= |\mathrm{Alt}_n : \mathrm{Cent}_P(\mathrm{Alt}_n)| = |\mathrm{Per}_n : \mathrm{Cent}_P(\mathrm{Alt}_n)|/2 \\ &= |\mathrm{Per}_n : \mathrm{Cent}_P(\mathrm{Per}_n)|/2 = |\mathrm{Conj}_P(\mathrm{Per}_n)|/2. \end{aligned}$$

Next, we observe that

$$\{HPH^{-1} : H \text{ is odd}\} = \mathrm{Conj}_{P_{(12)}^{-1}PP_{(12)}}(\mathrm{Alt}_n)$$

since every odd permutation has the form $P_{(12)}A$ for some $A \in \mathrm{Alt}_n$. Now

$$\begin{aligned} \mathrm{Conj}_P(\mathrm{Per}_n) &= \{HPH^{-1} : H \text{ is even}\} \cup \{HPH^{-1} : H \text{ is odd}\} \\ &= \mathrm{Conj}_P(\mathrm{Alt}_n) \cup \mathrm{Conj}_{P_{(12)}PP_{(12)}}(\mathrm{Alt}_n). \end{aligned}$$

Since $|\mathrm{Conj}_P(\mathrm{Alt}_n)| = |\mathrm{Conj}_P(\mathrm{Per}_n)|/2$ it follows that $\mathrm{Conj}_P(\mathrm{Alt}_n)$ and $\mathrm{Conj}_{P_{(12)}PP_{(12)}}(\mathrm{Alt}_n)$ are disjoint and of the same size. End of Proof.

The nonidentity elements of $\mathrm{Alt}_4$ have cycle types $(0,2,0,0)$ and $(1,0,1,0)$. As $(12)(34)$ commutes with the odd permutation $(12)$ Prop. 14.15 implies that

$$\mathrm{Conj}_{(12)(34)}(\mathrm{Alt}_4) = \mathrm{Conj}_{(12)(34)}(\mathrm{Per}_4) = \{(12)(34), (13)(24), (14)(23)\}.$$

The 3–cycle $(123)$ does not commute with any odd permutations (Exer. 1416) and so Prop. 14.15 reveals (for $(12)^{-1}(123)(12) = (132)$)

$$\mathrm{Conj}_{(123)}(\mathrm{Per}_4) = \mathrm{Conj}_{(123)}(\mathrm{Alt}_4) \cup \mathrm{Conj}_{(132)}(\mathrm{Alt}_4)$$

and $|\mathrm{Conj}_{(123)}(\mathrm{Alt}_4)| = |\mathrm{Conj}_{(132)}(\mathrm{Alt}_4)| = 4$. We record this information in Tab. 14.6.

| Rep | $I$ | $P_{(123)}$ | $P_{(132)}$ | $P_{(12)(34)}$ |
|---|---|---|---|---|
| Size | 1 | 4 | 4 | 3 |

**Table 14.6.** The Conjugacy Classes of $\mathrm{Alt}_4$.

The nonidentity elements of $\mathrm{Alt}_5$ have cycle types $(2,0,1,0,0)$, $(1,2,0,0,0)$ and $(0,0,0,0,1)$. The elements $(123)$ and $(23)(45)$ commute with the odd $(45)$, but $(12345)$ commutes with no odd (Exer. 1416) permutation hence, by Prop. 14.15, the nonidentity conjugacy classes are

$$\begin{aligned} &\mathrm{Conj}_{(12)(34)}(\mathrm{Alt}_5) = \mathrm{Conj}_{(12)(34)}(\mathrm{Per}_5), \quad \mathrm{Conj}_{(123)}(\mathrm{Alt}_5) = \mathrm{Conj}_{(123)}(\mathrm{Per}_5), \\ &\mathrm{Conj}_{(12345)}(\mathrm{Alt}_5) \quad \text{and} \quad \mathrm{Conj}_{(13452)}(\mathrm{Alt}_5), \end{aligned} \tag{14.33}$$

with sizes

$$\begin{aligned} |\mathrm{Conj}_{(12)(34)}(\mathrm{Alt}_5)| = 15, \quad |\mathrm{Conj}_{(123)}(\mathrm{Alt}_5)| = 20, \\ \text{and} \quad |\mathrm{Conj}_{(12345)}(\mathrm{Alt}_5)| = |\mathrm{Conj}_{(13452)}(\mathrm{Alt}_5)| = 12. \end{aligned} \tag{14.34}$$

We record this information in Tab. 14.7.

| Rep | $I$ | $P_{(123)}$ | $P_{(12)(34)}$ | $P_{(12345)}$ | $P_{(13452)}$ |
|---|---|---|---|---|---|
| Size | 1 | 20 | 15 | 12 | 12 |

**Table 14.7.** The Conjugacy Classes of $\mathrm{Alt}_5$.

### 14.5. Group Action and Counting Theory

In the previous section we have seen how Lagrange's Theorem, Prop. 14.13, permitted us through (14.32) to count the number of elements in conjugacy classes in $\mathrm{Alt}_4$ and $\mathrm{Alt}_5$. Via the notion of Group Action we will develop significant generalizations of this line of thought. This will permit us to count the number of permissible isomers in Chapter 16. As an added benefit, the notion of Group Action is a natural precursor of the representation theory of the next chapter.

An **action** of a group $G$ on a set $X$ is a homomorphism $\phi$ from $G$ into $\mathrm{Per}_X$, the set of permutations of elements of $X$. Given $x \in X$ we study its **stabilizer**

$$\mathrm{Stab}_x(G) \equiv \{g \in G : gx = x\}$$

and its **orbit**

$$\mathrm{Orb}_x(G) \equiv \{gx : g \in G\}. \tag{14.35}$$

For example, the dihedral group $\mathrm{Dih}_4$ acts on the solid square $X = \{(x_1, x_2) : |x_1| \leq 1,\ |x_2| \leq 1\}$. Every element stabilizes the origin, i.e., $\mathrm{Stab}_{(0,0)}(\mathrm{Dih}_4) = \mathrm{Dih}_4$. The stabilizer of any point (other than the origin) on any of the diagonals (see Figure 14.1) is the two element group of the identity and the proper reflection. For example, if $0 < a \leq 1$,

$$\mathrm{Stab}_{(a,a)}(\mathrm{Dih}_4) = \{I, H_{d_1}\} \quad \text{and} \quad \mathrm{Stab}_{(0,a)}(\mathrm{Dih}_4) = \{I, H_{e_2}\}.$$

The identity is the only group element that stabilizes the remaining points. The orbit of the origin is just the origin, $\mathrm{Orb}_{((0,0))}(\mathrm{Dih}_4) = (0, 0)$. The orbit of any point (other than the origin) on a diagonal hits 4 places, e.g.,

$$\mathrm{Orb}_{(a,a)}(\mathrm{Dih}_4) = \{(a, a), (-a, a), (-a, -a), (a, -a)\}$$
$$\mathrm{Orb}_{(0,a)}(\mathrm{Dih}_4) = \{(0, a), (-a, 0), (0, -a), (a, 0)\}$$

while orbits of remaining points hit 8 places, e.g., if $0 < a < 1$,

$$\mathrm{Orb}_{(1,a)}(\mathrm{Dih}_4) = \{(1, a), (-1, a), (-1, -a), (1, -a), (-a, 1), (-a, -1), (a, -1), (a, 1)\}.$$

In this case we note that each $\mathrm{Stab}_x(\mathrm{Dih}_4)$ is a subgroup of $\mathrm{Dih}_4$ and that the product of the sizes, $|\mathrm{Stab}_x(\mathrm{Dih}_4)||\mathrm{Orb}_x(\mathrm{Dih}_4)|$, simply the size, $|\mathrm{Dih}_4|$. Both of these observations are true in general.

**Proposition** 14.16. If the group $G$ acts on the set $X$ then $\mathrm{Stab}_x(G)$ is a subgroup of $G$ for each $x \in X$.

Proof: Exercise 14.. End of Proof.

**Proposition** 14.17. If the group $G$ acts on the set $X$ then $\mathrm{Orb}_x(G)$ is isomorphic to the set of left cosets of $\mathrm{Stab}_x(G)$, for each $x \in X$.

Proof: Consider the mapping $\psi(g\mathrm{Stab}_x(G)) = gx$. To see that it is injective note that $gx = hx$ iff $g^{-1}hx = x$, i.e., iff $g^{-1}h \in \mathrm{Stab}_x(G)$. The latter is equivalent to $h \in g\mathrm{Stab}_x(G)$. As $h$ trivially lies in $h\mathrm{Stab}_x(G)$ and distinct cosets are disjoint we have shown that $gx = hx$ iff $g\mathrm{Stab}_x(G) = h\mathrm{Stab}_x(G)$. To see that $\psi$ is surjective note that if $y \in \mathrm{Orb}_x(G)$ then $y = gx$ for some $g \in G$ and so $\psi(g\mathrm{Stab}_x(G)) = gx = y$. End of Proof.

**Proposition** 14.18. If the finite group $G$ acts on the set $X$ then

$$|\mathrm{Stab}_x(G)||\mathrm{Orb}_x(G)| = |G| \tag{14.36}$$

for each $x \in X$.

Proof: The previous result states that $|\mathrm{Orb}_x(G)|$ is the number of left cosets of $\mathrm{Stab}_x(G)$. The claim (14.36) then follows from Lagrange's Theorem, Prop. 14.13. End of Proof.

Already this has important applications. For example, we can deduce that the number of ways to choose $k$ objects from $n$ objects is

$$\frac{n!}{k!(n-k)!}. \tag{14.37}$$

To see this, let $X$ be the set of $k$ element subsets of $\{1, 2, \ldots, n\}$. Our goal is compute its size, $|X|$. The group $\mathrm{Per}_n$ acts on $X$ via

$$g\{a_1, a_2, \ldots, a_k\} = \{ga_1, ga_2, \ldots, ga_k\}.$$

It follows that the stabilizer of $x = \{1, 2, \ldots, k\}$ is the group of permutations that don't mix $\{1, 2, \ldots, k\}$ and $\{k+1, k+2, \ldots, n\}$. That is, $\mathrm{Stab}_x(\mathrm{Per}_n) = \mathrm{Per}_k \times \mathrm{Per}_{n-k}$. It follows from (14.36) that

$$|\mathrm{Orb}_x(\mathrm{Per}_n)| = \frac{|\mathrm{Per}_n|}{|\mathrm{Stab}_x(\mathrm{Per}_n)|} = \frac{n!}{k!(n-k)!}.$$

It remains only to note that this action has but one orbit, i.e, that $X = \mathrm{Orb}_x(\mathrm{Per}_n)$. We call such actions **transitive**. This generalizes nicely (see Exer. 14.18) to the statement that the number of sequences of $r_1$ 1's, $r_2$ 2's, ..., and $r_k$ $k$'s is

$$\frac{(r_1 + r_2 + \cdots + r_k)!}{r_1! r_2! \cdots r_k!}. \tag{14.38}$$

For example the number of sequences of 1 C, 2 G, 3 A and 4 T nucleotides is $10!/(1!2!3!4!) = 12600$. If this 10-nucleotide being has a circular genome then sequences that are mere rotations or reflections of one another would be equivalent. We ask then for the number of distinct genomes. To set the ideas lets start a bit smaller, say, how many genomes are there with 2 C's and 2 A's? There are 6 sequences and we picture each nucleotide to occupy a vertex of the square. As $\mathrm{Dih}_4$ acts on these vertices we may consider the associated orbits. In particular the two orbits

$$\mathrm{Orb}_{CAAC}(\mathrm{Dih}_4) = \{CAAC, AACC, ACCA, CCAA\} \quad \text{and} \quad \mathrm{Orb}_{CACA}(\mathrm{Dih}_4) = \{CACA, ACAC\}$$

exhaust the 6 sequences - and we recognize that this process reveals that there are two distinct genomes. We next ask, can we count without listing?

**Proposition** 14.19. If the finite group $G$ acts on the finite set $X$ then there exist $x_i \in X$, $i = 1, 2, \ldots, m$ such that

$$\mathrm{Orb}_{x_i}(G) \cap \mathrm{Orb}_{x_j}(G) = \emptyset \quad \text{if} \quad i \neq j \quad \text{and} \quad X = \bigcup_{i=1}^{m} \mathrm{Orb}_{x_i}(G).$$

By analogy to the group index we will denote the number of orbits of the action $G \curvearrowright X$ by $[G : X]$. This number is determined by the size of $G$ and the size of the sets fixed by individual group elements

$$\mathrm{Fix}_g(X) \equiv \{x \in X : gx = x\}.$$

We note that if $g$ is a nonzero rotation then no points are fixed while if $g$ is reflection across $e_1$ or $e_2$ then it fixes 2 midpoints while if $g$ is reflection across $d_1$ or $d_2$ then it fixes 2 vertices. In general

**Proposition** 14.20. If the finite group $G$ acts on the finite set $X$ then the number of orbits is the average number of fixed points, i.e.,

$$[G : X] = \frac{1}{|G|} \sum_{g \in G} |\mathrm{Fix}_g(X)| \tag{14.39}$$

Proof: Define $F \equiv \{(g,x) \in G \times X : gx = x\}$ and note that

$$|F| = \sum_{g \in G} |\{x : gx = x\}|$$

We start with - suppose $\{x_i : i = 1, \ldots, [G:X]\}$ to be representatives of the orbits of $G \curvearrowright X$. Then

$$\frac{1}{|G|} \sum_{x \in X} |\mathrm{Stab}_x(G)| = \sum_{x \in X} \frac{1}{|\mathrm{Orb}_x(G)|} = \sum_{i=1}^{[G:X]} \sum_{x \in \mathrm{Orb}_{x_i}(G)} \frac{1}{|\mathrm{Orb}_x(G)|} = \sum_{i=1}^{[G:X]} 1 = [G:X].$$

End of Proof.

Lets first check our small genome, with $X = C^2A^2$ please confirm that

$$\begin{aligned} &|\mathrm{Fix}_X(I)| = 6, \quad |\mathrm{Fix}_X(R_{\pi/2})| = |\mathrm{Fix}_X(R_{3\pi/2})| = 0 \quad \text{and} \\ &|\mathrm{Fix}_X(R_\pi)| = |\mathrm{Fix}_X(H_{e_2})| = |\mathrm{Fix}_X(H_{e_1})| = |\mathrm{Fix}_X(H_{d_2})| = |\mathrm{Fix}_X(H_{d_1})| = 2 \end{aligned} \tag{14.40}$$

and so $[\mathrm{Dih}_4 : C^2A^2] = 16/8 = 2$ as above. Now this should help in larger problems when Fix is easier to compute than Orb.

Lets try it out on $[\mathrm{Dih}_{10} : CG^2A^3T^4]$. As there is only 1 C no nontrivial rotation can fix a sequence. For a reflection to fix a sequence, it must be reflection across a diagonal connecting a C vertex to an A vertex and the sequence must be symmetric across this axis. There are 5 such diagonals, and for each axis there are

$$2\frac{(1+1+2)!}{1!1!2!} = 24$$

sequences (for there are 2 choices of AC/CA and then we place 1 A, 1 G and 2 T's on a side and then reflect). Hence there are

$$[\mathrm{Dih}_{10} : CG^2A^3T^4] = \frac{12600 + 5 \cdot 24}{20} = 636$$

distinct genomes with 1 C, 2 G's, 3 A's and 4 T's.

**Patterns:** The action of $G$ on $X$ induces an action of $G$ on $Y^X$, the set of functions from $X$ to $Y$,

$$\mathrm{Orb}_f(G) = \{f \circ g : g \in G\}$$

The distinct orbits are called **patterns**.

**Weights:** If each $y \in Y$ has weight $w(y)$ it lifts to $f \in Y^X$ via

$$w(f) \equiv \prod_{x \in X} w(f(x)).$$

For $h \in \mathrm{Orb}_f(G)$ we know $h = f \circ g$ for some $g \in G$ and so

$$w(h) = \prod_{x \in X} w(f(gx)) = w(f),$$

(because $\{gx : x \in X\} = X$). Hence each element of an orbit/pattern has the same weight - so we just call this the weight of the pattern - $w(F)$.

**Proposition** 14.21. Suppose that $X$ and $Y$ are finite sets and that the finite group $G$ acts on $X$. This induces an action $G \curvearrowright Y^X$. If $w$ is a weight function on $Y$ then $w$ extends to $Y^X$ and is constant on orbits of $G \curvearrowright Y^X$. If $f_i$ lie in the $[G : Y^X]$ distinct orbits then

$$\sum_{i=1}^{[G:Y^X]} w(f_i) = \frac{1}{|G|} \sum_{g \in G} \prod_{k=1}^{n} \left( \sum_{y \in Y} w(y)^k \right)^{c_k(g)} . \tag{14.41}$$

Proof: Suppose $w_0$ is a legitimate weight and gather

$$S(w_0) \equiv \{ f \in Y^X : w(f) = w_0 \}.$$

As $G$ acts on $S(w_0)$ it follows from Prop. 14.20 that the number of patterns, relative to $S$, is

$$[G : S(w_0)] = \frac{1}{|G|} \sum_{g \in G} |\mathrm{Fix}_g(S(w_0))| \tag{14.42}$$

Hence, on multiplication of both sides by $w_0$ and summing over $w_0$ we find

$$\sum_{i=1}^{[G:Y^X]} w(f_i) = \frac{1}{|G|} \sum_{w_0} \sum_{g \in G} |\mathrm{Fix}_g(S(w_0))| w_0 \tag{14.43}$$

Exchanging sums we work out the right hand side

$$\sum_{w_0} |\mathrm{Fix}_g(S(w_0))| w_0 = \sum_{fg=f} w(f)$$

and now invoke the disjoint cycle decomposition $g = g_1 g_2 \cdots g_n$ and the associated disjoint union $X = X_1 \cup X_2 \cup \cdots \cup X_n$. We now show that $fg = f$ iff $f$ is constant on each $X_i$. Hence

$$\begin{aligned} \sum_{fg=f} w(f) &= \sum_{fg=f} \prod_{x \in X} w(f(x)) = \sum_{f(X_i)=y_i \in Y} \prod_{i=1}^{n} \prod_{x \in X_i} w(f(x)) \\ &= \sum_{y_i \in Y} \prod_{i=1}^{n} w(y_i)^{|X_i|} = \prod_{i=1}^{n} \sum_{y \in Y} w(y)^{|X_i|} = \prod_{k=1}^{n} \left( \sum_{y \in Y} w(y)^k \right)^{c_k(g)} , \end{aligned}$$

because $c_k(g) = |\{i : |X_i| = k\}|$. Inserting this into (14.43) we arrive at (14.41) as claimed. End of Proof.

Returning to our small genome, let $X$ be the 4 vertices of the square, $G = \mathrm{Dih}_4$, $Y$ the labels $C$ and $A$, with weights $w(A) = x$ and $w(C) = y$. We need the decompositions, with type

$$\begin{aligned} I &= (1)(2)(3)(4), \quad \{4,0,0,0\} \\ R_{\pi/2} &= (1234), \quad \{0,0,0,1\} \\ R_{\pi} &= (13)(24), \quad \{0,2,0,0\} \\ R_{3\pi/2} &= (1432), \quad \{0,0,0,1\} \\ R_{e_1} &= (12)(34), \quad \{0,2,0,0\} \\ R_{e_2} &= (14)(32), \quad \{0,2,0,0\} \\ R_{d_1} &= (1)(3)(24), \quad \{2,1,0,0\} \\ R_{d_2} &= (2)(4)(13), \quad \{2,1,0,0\} \end{aligned}$$

The full inventory is

$$\frac{1}{8}\left((x+y)^4+2(x^4+y^4)+3(x^2+y^2)^2+2(x+y)^2(x^2+y^2)\right)=x^4+x^3y+2x^2y^2+xy^3+y^4.$$

The coefficient of $x^2y^2$ is indeed 2, the number of $A^2C^2$ genomes.

Onto the big genome, the identity is type $c_1=10$. The five reflections through vertices are of type $(c_1=2,c_2=4)$. The five reflections through edge midpoints are of type $c_2=5$. $R_{\pi/5}$ and $R_{6\pi/5}$ is of type $c_{10}=1$. $R_{2\pi/5}$ and $R_{7\pi/5}$ is of type $c_5=2$. $R_{3\pi/5}$ and $R_{8\pi/5}$ is of type $c_{10}=1$. $R_{4\pi/5}$ and $R_{9\pi/5}$ is of type $c_5=2$. $R_\pi$ is of type $c_2=5$. So the full inventory is

$$\frac{1}{20}\big((A+G+C+T)^{10}+5(A+G+C+T)^2(A^2+G^2+C^2+T^2)^4+6(A^2+G^2+C^2+T^2)^5 \\ +4(A^{10}+G^{10}+C^{10}+T^{10})+4(A^5+G^5+C^5+T^5)^2\big).$$

The coefficient of $CG^2A^3T^4$ is indeed 636.

We will need the lovely

$$\sum_{n=0}^{\infty}\mathcal{P}_{\mathrm{Per}_n}(f(x),\ldots,f(x^n))=\exp\left(\sum_{k=1}^{\infty}f(x^k)/k\right). \tag{14.44}$$

Also, if $A=\mathrm{Per}_n$ and $C$ is constructed by restriction to bijective $f$ then

$$C(x)=\mathcal{P}_{\mathrm{Alt}_n}(c(x),\ldots,c(x^n))-\mathcal{P}_{\mathrm{Per}_n}(c(x),\ldots,c(x^n)). \tag{14.45}$$

### 14.6. Notes and Exercises

For a thorough introduction to abstract algebra see Goodman (1997). For more on Cayley Graphs see Krebs and Shaheen (2011). Our work on conjugacy classes was drawn from James and Liebeck (2001). Pólya's Theory of counting is drawn from Polya and Read (2011).

1. Please justify each of the three claims made in the proof of Prop. 14.2.

2. (a) Confirm that the rotation matrix, $R_{a,\theta}$, of Eq. (14.2) is an orthogonal matrix. Hint: Use the properties of $\mathtt{X}(a)$ established in Exer. 1.21.

   (b) Show that $\det(R_{a,\theta})=1$. Hint: $R_{a,0}=I$ is easy. At the other extreme, $R_{a,\pi}=2aa^T-I$ has a two dimensional eigenspace, $v_3$, associated with the double eigenvalue, $-1$. Now assume $0<\theta<\pi$. Confirm that $R_{a,\theta}a=a$ so that one eigenvalue $\lambda_1=1$. Confirm that the other two eigenvalues are nonreal complex conjugates, $\lambda_2$ and $\overline{\lambda}_2$. Now confirm that $\mathrm{tr}R_{a,\theta}=1+2\cos\theta$ and deduce that $\Re\lambda_2=\cos\theta$. Now confirm that $|\lambda_2|=1$ and deduce that $\lambda_2=\cos\theta+i\sin\theta$ and finally that $\det(R_{a,\theta})=\lambda_1\lambda_2\overline{\lambda}_2=1$.

3. Use `icosasym.m` to rotate Figure 14.5(C) and confirm that $R_C$ takes vertices

   $$(1,2,3,4,5,6,7,8,9,10,11,12)\quad\text{to}\quad(7,9,12,6,8,4,1,5,2,11,10,3).$$

   $R_B$ clearly takes vertex 6 to vertex 1. Confirm that $R_CR_A^4R_CR_A$ also takes 6 to 1. Confirm that $R_CR_A^4R_CR_A=R_B$ on vertices 1 and 5 as well.

4. The cube has four 3–fold axes through pairs of opposite vertices, giving eight rotations. There are also three 4–fold axes through the centroids of opposite faces, adding nine rotations, Finally, the cube has six 2–fold axes through centers of opposite edges, adding 6 rotations. With the identity, we have 24 rotations. Construct each rotatation matrix. Show that its dual, the octahedron, has the same rotation group.

5. Prove Prop. 14.7.

6. Regarding the argument that resulted in Prop. 14.9 please show that if $P_\sigma$ and $P_\phi$ are $k-$cycles in $\mathrm{Per}_n$ and $\pi(\sigma_j) \equiv \phi_j$ then
$$P_\pi P_\sigma P_\pi^{-1} = P_\phi.$$

7. Given $S \in \mathbb{R}^{n\times n}$ show that
$$\mathcal{S} = \{A \in \mathrm{GL}_n(\mathbb{R}) : A^T S A = S\}$$
is a group. This is a generalization of the orthogonal group in the sense that $\mathcal{S} = \mathrm{O}_n$ if $S = I$. In the case that
$$S = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$
show that $\mathcal{S} = \{A \in \mathrm{GL}_2(\mathbb{R}) : \det(A) = 1\}$. In higher dimensions, with
$$S = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix}$$
where $I_n$ is the $n$-by-$n$ identity matrix, $\mathcal{S}$ is known as the **symplectic group**.

8. (a) Show that
$$U_1 \equiv \left\{ \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} : x \in \mathbb{R} \right\} \quad \text{and} \quad L_1 \equiv \left\{ \begin{pmatrix} 1 & 0 \\ y & 1 \end{pmatrix} : y \in \mathbb{R} \right\}$$
are subgroups of $\mathrm{SL}_2(\mathbb{R})$.
(b) Show that every $A \in \mathrm{SL}_2(\mathbb{R})$ can be expressed as a finite product of matrices in $U_1$ and $L_1$.

9. Define and study the derived subgroup following the lovely examples in Grove, Algebra.

10. Show that each element of the Modular group takes the upper half plane to itself. Hint: Show
$$\Im(f(z)) = \frac{\Im z}{|cz+d|^2}.$$

11. Confirm that $\mathrm{PSL}_2(3)$ is isomorphic to $\mathrm{Alt}_4(\mathbb{R})$.

12. Use Prop. 14.12 and Table 14.6 to show that the **Klein 4-group**
$$V_4 \equiv \{I, P_{(12)(34)}, P_{(13)(24)}, P_{(14)(23)}\} \tag{14.46}$$
is a normal subgroup of $\mathrm{Alt}_4$. Compute the three cosets of the quotient group $\mathrm{Alt}_4/V_4$ and show that this quotient is isomorphic to $\mathrm{Alt}_3$.

13. Show that $V_4$ is abelian and isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2$.

14. Show that $\mathrm{SDih}_4$ is a normal subgroup of $\mathrm{Dih}_4$ and that $\mathrm{Dih}_4/\mathrm{SDih}_4 \sim V_4$.

15. Prove that $\mathrm{Cent}_g(G)$ is a subgroup of $G$ and that the mapping $f(g^{-1}xg) \equiv g\mathrm{Cent}_x(G)$ is a bijection of $\mathrm{Conj}_x(G)$ and the set of left cosets of $\mathrm{Cent}_x(G)$. Explain how Eq. (14.32) follows from your work. Hint: Show that $g^{-1}xg = h^{-1}xh$ iff $g^{-1}\mathrm{Cent}_x(G) = h^{-1}\mathrm{Cent}_x(G)$.

16. Show that (123) can not commute with an odd element of $\mathrm{Per}_4$. Show that (12345) can not commute with an odd element of $\mathrm{Per}_5$.

17. By Lagrange's Theorem there are $5 = |\mathrm{Alt}_5|/|\mathrm{Alt}_4|$ left cosets of $\mathrm{Alt}_4$ in $\mathrm{Alt}_5$. Explain why $\{(), (125), (152), (345), (354)\}$ is a complete set of representatives of these 5 cosets.

18. Let $X$ be the set of sequences of $r_1$ 1's, $r_2$ 2's, ..., and $r_k$ $k$'s. Show that

$$|X| = \frac{(r_1 + r_2 + \cdots + r_k)!}{r_1! r_2! \cdots r_k!}.$$

Hint: Set $n = r_1 + r_2 + \cdots + r_k$ and show that $\mathrm{Per}_n$ acts on $X$ transitively. Show that the stabilizer of

$$(1, \ldots, 1, 2, \ldots, 2, \ldots, k, \ldots, k)$$

with $r_1$ ones, $r_2$ 2 *etc.*, is $\mathrm{Per}_{r_1} \times \mathrm{Per}_{r_2} \times \cdots \times \mathrm{Per}_{r_k}$.

19. Confirm the 8 equations in (14.40).

20. Find the number of distinct genomes with 2 C, 2 G, 2 A and 2 T nucleotides.

# 15. Group Representation Theory*

Representation Theory is essentially a conduit by which the results of linear algebra can be brought to bear on the study of groups. In some cases this conduit has established a bridge between two groups of very different origin. For example, Wiles use of representation theory to equate certain modular groups with groups associated with elliptic curves is at the heart of his resolution of Fermat's Last Theorem. In other cases Representation Theory serves to systematically reduce the complexity of large structured problems by expressing them in terms of irreducible, and often tractable, objects.

After building the requisite representation and character theory in the first two sections we will show how it may be used to explicitly determine the electronic structure of the Buckyball – that is all 60 eigenvalues of the adjacency matrix of the graph in Figure 14.6. The theory is even simpler, and in fact takes on a familiar character, when applied to abelian groups. We close the chapter with both the theory, and application, of Fourier analysis on abelian Groups.

### 15.1. Representations

A **representation** of a group $G$ is a pair $(V, \pi)$ where $V$ is a complex vector space and $\pi$ is a homomorphism from $G$ to $\mathrm{GL}(V)$, the group of invertible linear transformations of $V$. We call $d = \dim_{\mathbb{C}}(V)$ the **degree** of $(V, \pi)$. We note that as $\pi$ is a homomorphism it follows that $\pi(I) = \pi(gg^{-1}) = \pi(g)\pi(g^{-1}) = \pi(g)\pi(g)^{-1} = I$. That is, $\pi$ takes the identity element of $G$ to the identity transformation in $GL(V)$.

For example, if $(\mathbb{C}, \pi)$ is a (degree 1) representation of $\mathrm{Alt}_3 = \{(123), (123)^2, (123)^3\}$ then

$$1 = \pi(I) = \pi((123))^3,$$

hence $\pi((123))$ is a cube root of unity. The three cube roots give rise to three distinct representations

$$\begin{aligned} \pi_1(()) &= 1, \quad \pi_1((123)) = 1, \quad \pi_1(132) = 1 \\ \pi^{(2)}(()) &= 1, \quad \pi^{(2)}((123)) = \exp(2\pi i/3), \quad \pi^{(2)}(132) = \exp(-2\pi i/3) \\ \pi_3(()) &= 1, \quad \pi_3((123)) = \exp(-2\pi i/3), \quad \pi^{(2)}(132) = \exp(2\pi i/3). \end{aligned} \tag{15.1}$$

We will soon see that this is a complete set of representations of $\mathrm{Alt}_3$.

We next construct a few concrete representations of $\mathrm{Per}_3$. Every group sports the constant representation, $(\mathbb{C}, \pi_1)$, here

$$\pi_1(\sigma) = 1, \quad \forall \sigma \in \mathrm{Per}_3, \tag{15.2}$$

and every permutation group features the sign representation, $(\mathbb{C}, \pi^{(2)})$, here

$$\pi^{(2)}(\sigma) = \begin{cases} 1 & \text{if } \sigma \text{ is even} \\ -1 & \text{if } \sigma \text{ is odd.} \end{cases} \tag{15.3}$$

These are both of degree 1. Note the isomorphism between $\mathrm{Per}_3$ and $\mathrm{Dih}_3$ constructed in Eq. (14.10)–(14.14) is a representation that we will denote $(\mathbb{C}^2, \pi_3)$. For convenience we reproduce it here

$$\begin{aligned} \pi_3(I) = I, \quad \pi_3((12)) = H_{a^\perp}, \quad \pi_3((13)) = H_{b^\perp}, \\ \pi_3((23)) = H_{c^\perp}, \quad \pi_3((123)) = R_{2\pi/3}, \quad \pi_3((132)) = R_{4\pi/3}. \end{aligned} \tag{15.4}$$

There is also a natural degree three representation, namely $(\mathbb{C}^3, \sigma \mapsto P_\sigma)$.

In addition to the constant representation the next most common example is $(\mathbb{C}[G], R_G)$ where $\mathbb{C}[G]$ is the vector space of complex valued functions on $G$ and $R_G$ is defined via

$$(R_G(h)f)(g) \equiv f(gh). \tag{15.5}$$

To confirm that it is a homomorphism we compute

$$(R_G(h_1h_2)f)(g) = f(gh_1h_2) \quad \text{and} \quad (R_G(h_1)(R_G(h_2)f))(g) = (R_G(h_2)f)(gh_1) = f(gh_1h_2),$$

and so indeed $R_G(h_1h_2) = R_G(h_1)R_G(h_2)$. We call $(\mathbb{C}[G], R_G)$ the **right regular representation** of $G$. The role of the right regular representation in exploring the adjacency matrix, $A$, of a Cayley graph $\mathrm{Cay}(G, S)$ is exposed on refraining from coordinates and instead interpreting $A$ as a linear transformation of $\mathbb{C}[G]$. In particular, for $f \in \mathbb{C}[G]$ we recognize $Af \in \mathbb{C}[G]$ as

$$(Af)(g) = \sum_{s\in S} f(gs) = \sum_{s\in S} (R_G(s)f)(g), \quad \text{for each } g \in G. \tag{15.6}$$

On writing Eq. (15.6) more succinctly as

$$\boxed{A_{\mathrm{Cay}(G,S)} = \sum_{s\in S} R_G(s)} \tag{15.7}$$

we can better expose the role of $R_G$ in the eigendecomposition of $A$. For if $S$ is a conjugacy class of $G$ then, for each $g \in G$,

$$AR_G(g) = \sum_{s\in S} R_G(s)R_G(g) = \sum_{s\in S} R_G(gsg^{-1})R_G(g) = \sum_{s\in S} R_G(gsg^{-1}g) = \sum_{s\in S} R_G(gs) = R_G(g)A.$$

Now if $Ax = \lambda x$ then $R_G(g)Ax = AR_G(g)x = \lambda R_G(g)x$ and so each eigenspace of $A$ is a $G$–**invariant subspace** of $R_G$. The true power of (15.7) is revealed on showing that the right regular representation is similar to a direct sum of simpler objects, the so–called irreducible representations.

A representation, $(V, \pi)$, of a group $G$ is called **irreducible** when 0 and $V$ are its **only** $G$–invariant subspaces and **reducible** when $(V, \pi)$ can be expressed as the direct sum, $(V_1 \oplus V_2, \pi_1 \oplus \pi^{(2)})$ where $V_1$ and $V_2$ are both proper $G$–invariant subspaces.

Regarding $\mathrm{Per}_3$ we note that the constant and sign representations, (15.2) and (15.3), are both degree 1 and so are irreducible. As the $(\mathbb{C}^2, \pi_3)$ defined in (15.4) has degree 2, any proper nonzero $\mathrm{Per}_3$–invariant subspace, $W$, must be one dimensional. Now $\pi_3(P)W = W$ for every $P \in \mathrm{Per}_3$ means that $BW = W$ for every $B \in \mathrm{Dih}_3$. In other words, there exists a $w \in \mathbb{C}^2$ that is an eigenvector *for every* $B \in \mathrm{Dih}_3$. As $R_{2\pi/3}$ and $H_{a^\perp}$ share no eigenvectors we conclude that $\pi_3$ is irreducible. We will soon see that $\pi_1$, $\pi^{(2)}$ and $\pi_3$ are a complete set of irreducible representations in the sense that every representation of $\mathrm{Per}_3$ is similar to a unique (up to equivalence) direct sum of $\pi_1$, $\pi^{(2)}$ and $\pi_3$. In particular, we will show that

$$R_{\mathrm{Per}_3} \sim \pi_1 \oplus \pi^{(2)} \oplus 2\pi_3. \tag{15.8}$$

This will follow from a general theory that establishes the existence of such a similarity transform. We will offer an explicit construction in §15.5. The general theory begins with the study of unitary representations.

**Proposition** 15.1. If $(V, \pi)$ is a unitary representation of the group $G$, i.e., $\pi(g)^* = \pi(g^{-1})\ \forall\, g \in G$, then $\pi$ is either irreducible or reducible.

Proof: If $\pi$ is not irreducible then there exists a proper $G$-invariant subspace $W \subset V$. We denote the orthogonal complement of $W$ by $W^\perp$ and note that $V = W \oplus W^\perp$. We now show that $W^\perp$ is $G$-invariant, i.e., that $\pi(g)W^\perp \perp W$ for every $g \in G$. So, with $w \in W$ and $v \in W^\perp$ we find, as $\pi$ is unitary, that

$$\langle \pi(g)v, w\rangle = \langle \pi(g^{-1})\pi(g)v, \pi(g^{-1})w\rangle = \langle v, \pi(g^{-1})w\rangle = 0,$$

The final equality follows from $\pi(g^{-1})w \in W$ as $W$ is invariant under $\pi(g^{-1})$. End of Proof.

You should confirm that right regular representation, $(\mathbb{C}[G], R_G)$, is unitary. It will suffice that every representation is similar to a unitary representation.

**Proposition** 15.2. If $(V, \pi)$ is a representation of the group $G$ then there exists a $T \in \mathrm{GL}(V)$ such that $\phi(g) \equiv T\pi(g)T^{-1}$ is a unitary representation of $G$.

Proof: Note that

$$\Pi \equiv \sum_{g\in G} \pi(g)^*\pi(g)$$

is self-adjoint and positive definite and so defines a weighted inner product, $\langle v, w\rangle_\Pi = \langle \Pi v, w\rangle$, on $V$. Now

$$\langle \pi(h)v, \pi(h)w\rangle_\Pi = \sum_{g\in G} \langle \pi(g)\pi(h)v, \pi(g)\pi(h)w\rangle = \sum_{g\in G} \langle \pi(gh)v, \pi(gh)w\rangle = \langle v, w\rangle_\Pi$$

where the final equality stems from the realization that groups satisfy $\{gh : g \in G\} = G$ for each $h \in G$. Finally, it follows from Exer. 12.3 that $\phi(h) \equiv \Pi^{1/2}\pi(h)\Pi^{-1/2}$ is unitary. End of Proof.

We have now shown that every representation is similar to a representation that is either reducible or irreducible. This is enough to establish our first expansion.

**Proposition** 15.3. (**Maschke's Theorem.**) Each representation, $(V, \pi)$, of a group $G$ may be **completely reduced** in the sense that there exist irreducible representations $\{(V_j, \pi_j)\}_{j=1}^s$ such that

$$V = V_1 \oplus V_2 \oplus \cdots \oplus V_s \quad \text{and} \quad \pi \sim \pi_1 \oplus \pi^{(2)} \oplus \cdots \oplus \pi_s.$$

Proof: If $\dim_{\mathbb{C}}(V) = 1$ then $V$ has no proper subspaces and so $(V, \pi)$ is irreducible and so completely reduced. We now argue by induction. Suppose that representations of degree $\le n$ may be completely reduced and consider the unitary representation $(V, \pi)$ of degree $n + 1$. If $(V, \pi)$ is irreducible then it is completely reduced. If $(V, \pi)$ is not irreducible then $V = V_1 \oplus V_2$ where each $V_j$ is a proper subspace that is $G$-invariant under $\pi$. It follows that each $\dim V_j \le n$ and so, by the inductive hypothesis, each $(V_j, \pi|_{V_j})$ may be completely reduced. As such, each $V_j$ is the direct sum

$$V_j = V_{j,1} \oplus V_{j,2} \oplus \cdots \oplus V_{j,m_j}$$

of proper $G$-invariant subspaces and $(V_{j,k}, \pi|_{V_{j,k}})$ is irreducible. On gathering we find

$$V = V_{1,1} \oplus V_{1,2} \oplus \cdots \oplus V_{1,m_1} \oplus V_{2,1} \oplus V_{2,2} \oplus \cdots \oplus V_{2,m_2}$$

and so $(V, \pi)$ is completely reduced. End of Proof.

This proof does not reveal the number $s$, of irreducible representations nor does is tell us how often a given representation appears. The first step toward a more quantitative expansion is the following Lemma of Schur.

**Proposition** 15.4. **Schur's Lemma**. Suppose that $(V_1, \pi_1)$ and $(V_2, \pi^{(2)})$ are two irreducible representations of the group $G$. If they are **intertwined** by the operator $T$, i.e.,

$$T\pi_1(g) = \pi^{(2)}(g)T, \quad \forall\, g \in G,$$

then
(a) If $T \neq 0$ then $T$ is invertible.
(b) If $\pi_1 \not\sim \pi^{(2)}$ then $T = 0$.
(c) If $\pi_1 = \pi^{(2)}$ then $T = \lambda I$ for some $\lambda \in \mathbb{C}$.

Proof: If $v \in \mathcal{N}(T)$ then $T\pi_1(g)v = \pi^{(2)}(g)Tv = 0$ and so $\mathcal{N}(T)$ is $G$-invariant under $\pi_1$. As $(V_1, \pi_1)$ is irreducible it follows that $\mathcal{N}(T) = \{0\}$ or $\mathcal{N}(T) = V$. It follows that if $T \neq 0$ then $T$ is injective.

Next, we interpret the right side of $T\pi_1(g)y = \pi^{(2)}(g)Ty$ as $\pi^{(2)}(g)$ acting on $\mathcal{R}(T)$ and note that the left side always lies in $\mathcal{R}(T)$. It follows that $\mathcal{R}(T)$ is $G$-invariant under $\pi^{(2)}$. Again, by irreducibility we know that $\mathcal{R}(T) = \{0\}$ or $\mathcal{R}(T) = V$, from which it follows that if $T \neq 0$ then $T$ is surjective.

Claim (a) follows: if $T \neq 0$ then $T$ is invertible. Regarding claim (b), if $\pi_1 \not\sim \pi^{(2)}$ then $T$ is not invertible and so by (a) $T = 0$.

Regarding claim (c), as $T$ and $I$ both commute with $\pi_1$ it follows that so too does $\lambda I - T$ for any $\lambda \in \mathbb{C}$. It then follows from (a) that $\lambda I - T$ is either zero or invertible. We can exclude the latter by choosing $\lambda$ to be an eigenvalue of $T$ and so force the former, i.e., $T = \lambda I$ as claimed. End of Proof.

This lemma reveals that individual elements of irreducible representations are orthogonal to one another in the inner product of $\mathbb{C}[G]$. As such the next result is often referred to as the **Grand Orthogonality Theorem** or **Wonderful Orthogonality Theorem** in the applied literature.

**Proposition** 15.5. If $(V, \phi)$ is a unitary irreducible representation of $G$ of degree $n$ then

$$\langle \phi_{ij}, \phi_{rs} \rangle = \frac{|G|}{n} \delta_{i,r} \delta_{j,s}. \tag{15.9}$$

If $(W, \pi)$ is a unitary irreducible representation of $G$ that is not similar to $(V, \phi)$ then

$$\langle \phi_{ij}, \pi_{rs} \rangle = 0. \tag{15.10}$$

Proof: We start with the latter by constructing the matrices

$$T \equiv \sum_{g \in G} \pi(g) E_{j,s} \phi(g^{-1}), \tag{15.11}$$

where $E_{j,s}$ is the $m$-by-$n$ matrix with the value one in row $j$ and column $s$ and zeros elsewhere. For $h \in G$ we find

$$\pi(h)T = \sum_{g \in G} \pi(hg) E_{j,s} \phi(g^{-1}) = \sum_{x \in G} \pi(x) E_{j,s} \phi(x^{-1}h) = T\phi(h),$$

and so $T$ intertwines $\pi$ and $\phi$. As these two are presumed not similar it follows from Schur's Lemma that $T = 0$. The $(i, r)$ element of Eq. (15.11) becomes

$$0 = \sum_{g\in G} \pi_{ij}(g) E_{j,s} \phi_{sr}(g^{-1}) = \sum_{g\in G} \pi_{ij}(g) E_{j,s} \overline{\phi_{rs}(g)} = \langle \pi_{ij}, \phi_{rs} \rangle$$

and we obtain Eq. (15.10).

If we substitute $\pi$ for $\phi$ in Eq. (15.11) we arrive at a matrix

$$T \equiv \sum_{g\in G} \pi(g) E_{j,s} \pi(g^{-1}), \tag{15.12}$$

that commutes with $\pi$. Again by Schur's Lemma it follows that $T = \lambda I$ and so the nondiagonal elements of Eq. (15.12) must vanish. That is, for $i \neq r$,

$$0 = \sum_{g\in G} \pi_{ij}(g) E_{j,s} \pi_{sr}(g^{-1}) = \sum_{g\in G} \pi_{ij}(g) E_{j,s} \overline{\pi_{rs}(g)} = \langle \pi_{ij}, \pi_{rs} \rangle \tag{15.13}$$

When $i = r$ we find

$$\lambda = \sum_{g\in G} \pi_{ij}(g) E_{j,s} \pi_{s,i}(g^{-1}) = \sum_{g\in G} \pi_{ij}(g) E_{j,s} \overline{\pi_{is}(g)} = \langle \pi_{ij}, \pi_{is} \rangle \tag{15.14}$$

It remains to compute $\lambda$. The trace of Eq. (15.12) reveals

$$\lambda n = \sum_{g\in G} \mathrm{tr}(\pi(g) E_{j,s} \pi(g^{-1})) = |G| \mathrm{tr}(E_{j,s}) = |G| \delta_{j,s}. \tag{15.15}$$

Combining Eqs. (15.13)–(15.15) establishes Eq. (15.9). End of Proof.

**Proposition** 15.6. The number of equivalence classes of irreducible representations of $G$ does not exceed $|G|$. If $\{(V^{(k)}, \pi^{(k)}) : 1 \leq k \leq s\}$ is a set of representatives from each equivalence class of irreducible representations of $G$, then

$$Q \equiv \{\sqrt{d_k}\, \pi^{(k)}_{ij} : 1 \leq i, j \leq d_k,\ 1 \leq k \leq s\} \quad \text{where} \quad d_k = \dim(V^{(k)}), \tag{15.16}$$

is an orthonormal set in $\mathbb{C}[G]$ and hence $s \leq d_1^2 + \cdots + d_s^2 \leq |G|$.

Proof: Each equivalence class contains a unitary member. As $\dim \mathbb{C}[G] = |G|$ no linearly independent set in $\mathbb{C}[G]$ can have more than $|G|$ members. But Prop. 15.5 states that the entries of inequivalent unitary representations form an orthogonal set of nonzero members of $\mathbb{C}[G]$. Hence the number of such classes is bounded by $|G|$. End of Proof.

### 15.2. Characters

The trace was the final ingredient in the proof of Prop. 15.5. We now use it to define the **character** of a representation, $\pi$,

$$\boxed{\chi_\pi(g) \equiv \mathrm{tr}(\pi(g)).} \tag{15.17}$$

The character of an irreducible representation is called an **irreducible character**.

Representations of degree one are irreducible and coincide with their characters and so, recalling (15.1), we note that we have already found 3 irreducible characters of $\mathrm{Alt}_3$. As $|\mathrm{Alt}_3| = 3$ it follows from the previous proposition, Prop. 15.6, that there are no other distinct characters. We find it best to display them through the character table below.

| $\mathrm{Alt}_3$ | () | (123) | (132) |
|---|---|---|---|
| $\chi_1$ | 1 | 1 | 1 |
| $\chi_2$ | 1 | $\omega_3$ | $\omega_3^2$ |
| $\chi_3$ | 1 | $\omega_3^2$ | $\omega_3$ |

**Table 15.1.** The character table for $\mathrm{Alt}_3$.

A short calculation reveals that its columns in Tab. 15.1 are orthogonal in the natural inner product on $\mathbb{C}[\mathrm{Alt}_3]$. That is,

$$\langle \chi_j, \chi_k \rangle \equiv \sum_{m=0}^{2} \chi_j((123)^m)\overline{\chi_k((123)^m)} = 3\delta_{j,k}.$$

Hence, the characters of $\mathrm{Alt}_3$ provide an orthogonal basis for $\mathbb{C}[\mathrm{Alt}_3]$. We will see that such is the case for all abelian groups. In the general setting we work over a natural subspace of $\mathbb{C}[G]$ dictated by the fact that characters can not distinguish conjugate elements. To be precise, as characters are traces of homomorphisms, it follows that

$$\begin{aligned}\chi_\pi(h^{-1}gh) &= \mathrm{tr}(\pi(h^{-1}gh)) = \mathrm{tr}(\pi(h^{-1})\pi(g)\pi(h)) = \mathrm{tr}(\pi(h)^{-1}\pi(g)\pi(h)) \\ &= \mathrm{tr}(\pi(h)^{-1}\pi(h)\pi(g)) = \mathrm{tr}(\pi(g)) = \chi_\pi(g).\end{aligned}$$

In other words, characters are constant on conjugacy classes. This feeds the general definition. The **class functions** of a group $G$ is the subspace

$$\mathrm{Class}[G] \equiv \{f \in \mathbb{C}[G] : f(g) = f(h^{-1}gh)\ \forall g \text{ and } h \in G\}. \tag{15.18}$$

We proceed to construct three characters for $\mathrm{Per}_3$ and show that they comprise an orthogonal basis for $\mathrm{Class}[\mathrm{Per}_3]$. Defining $\chi_j = \mathrm{tr}(\pi_j)$ for the three $\pi_j$ in (15.2)–(15.4) we arrive at Table 15.2.

| $\mathrm{Per}_3$ | $I$ | (12) | (123) |
|---|---|---|---|
| $\chi_1$ | 1 | 1 | 1 |
| $\chi_2$ | 1 | $-1$ | 1 |
| $\chi_3$ | 2 | 0 | $-1$ |

**Table 15.2.** The character table for $\mathrm{Per}_3$. As characters are constant on conjugacy classes we may index each column with a representative from each conjugacy class. Recall that conjugacy classes on $\mathrm{Per}_n$ are determined solely by cycle type.

We note that the characters of $\mathrm{Per}_3$ satisfy

$$\langle \chi_j, \chi_k \rangle = 6\delta_{j,k} = |\mathrm{Per}_3|\delta_{j,k}$$

and so comprise an orthogonal basis for the class functions of $\mathrm{Per}_3$. Returning to the general case we find,

**Proposition** 15.7. If $\pi$ and $\phi$ are both irreducible representations of $G$ then

$$\langle \chi_\pi, \chi_\phi \rangle = \begin{cases} |G| & \text{if } \pi \sim \phi, \\ 0, & \text{otherwise.} \end{cases}$$

Conversely, if $\langle \chi_\pi, \chi_\pi \rangle = |G|$ then $\pi$ is irreducible.

Proof: We set $m \equiv \deg(\pi)$ and $n \equiv \deg(\phi)$ and develop

$$\begin{aligned} \langle \chi_\pi, \chi_\phi \rangle &= \sum_{g \in G} \mathrm{tr}(\pi(g)) \mathrm{tr}(\overline{\phi(g)}) = \sum_{g \in G} \sum_{i=1}^{m} \pi_{ii}(g) \sum_{j=1}^{n} \overline{\phi_{jj}(g)} \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{g \in G} \pi_{ii}(g) \overline{\phi_{jj}(g)} = \sum_{i=1}^{m} \sum_{j=1}^{n} \langle \pi_{ii}, \phi_{jj} \rangle. \end{aligned} \tag{15.19}$$

If $\pi \not\sim \phi$ then each inner product vanishes. If $\pi \sim \phi$ then $\chi_\pi = \chi_\phi$ and so we may substitute $\phi = \pi$ in Eq. (15.19), invoke $\langle \pi_{ii}, \pi_{jj} \rangle = |G| \delta_{ij}/n$ and conclude that $\langle \chi_\pi, \chi_\pi \rangle = |G|$.

Regarding the converse, from Maschke's Theorem,

$$\pi \sim \pi^{(1)} \oplus \pi^{(2)} \oplus \cdots \oplus \pi^{(s)} \tag{15.20}$$

where each $\pi^{(j)}$ is irreducible. We note that

$$\chi_\pi = \sum_{i=1}^{s} \chi_{\pi^{(i)}} \quad \text{and} \quad \langle \chi_\pi, \chi_\pi \rangle = \sum_{i=1}^{s} \langle \chi_{\pi^{(i)}}, \chi_{\pi^{(i)}} \rangle = \sum_{i=1}^{s} |G|.$$

Hence if $\langle \chi_\pi, \chi_\pi \rangle = |G|$ then $n = s$ in Eq. (15.20) and so $\pi$ is irreducible. End of Proof.

If we consider the $(V^{(k)}, \pi^{(k)})$ to act as basis elements it makes sense to grant them weights when considering general expansions. More precisely, if $(V, \pi)$ is a representation then $(mV, m\pi)$ denotes the representation where $mV$ is the direct sum of $m$ copies of $V$ and $m\pi$ is the direct sum of $m$ copies of $\pi$.

**Proposition** 15.8. Let $\{(V^{(k)}, \pi^{(k)} : 1 \leq k \leq s\}$ denote a complete set of representatives of the equivalence classes of irreducible representations of $G$. If $(V, \pi)$ is a representation of $G$ then

$$\pi \sim m_1 \pi^{(1)} \oplus m_2 \pi^{(2)} \oplus \cdots \oplus m_s \pi^{(s)} \quad \text{where} \quad m_k = \langle \chi_\pi, \chi_{\pi^{(k)}} \rangle / |G|.$$

Consequently, the decomposition of $\pi$ into irreducible components is unique and is determined up to equivalence by its character.

Proof: We note that

$$\chi_\pi = \sum_{k=1}^{s} m_k \chi_{\pi^{(k)}} \Rightarrow \langle \chi_\pi, \chi_{\pi^{(i)}} \rangle = \sum_{k=1}^{s} m_k \langle \chi_{\pi^{(k)}}, \chi_{\pi^{(i)}} \rangle = m_i |G|.$$

End of Proof.

These weights take on familiar values when we expand the right regular representation. We first compute its character. If $\{e_g\}_{g\in G}$, where $e_g(i) = \delta_{g,i}$, is the Euclidian orthonormal basis for $\mathbb{C}[G]$ then

$$\chi_{R_G}(h) = \mathrm{tr}(R_G(h)) = \sum_{g\in G} e_g^T R_G(h) e_g = \sum_{g\in G}\sum_{i\in G} e_g(i)e_g(ih) = \sum_{g\in G} e_g(gh) = |G|\delta_{h,I} \tag{15.21}$$

Hence, if

$$R_G \sim m_1\pi_1 \oplus m_2\pi^{(2)} \oplus \cdots \oplus m_s\pi_s$$

then

$$m_i|G| = \langle \chi_{R_G}, \chi_{\pi^{(i)}}\rangle = \sum_{g\in G} \chi_{R_G}(g)\overline{\chi_{\pi^{(i)}}(g)} = |G|\overline{\chi_{\pi^{(i)}}(I)} = d_i|G|.$$

Now evaluating $\chi_{R_G}$ at the identity we find

$$\boxed{|G| = \chi_{R_G}(I) = \sum_{k=1} d_k\chi_{\pi^{(k)}}(I) = \sum_{k=1}^{s} d_k^2.} \tag{15.22}$$

From which it follows that the $Q$ of Eq. (15.16) is an orthonormal basis for $\mathbb{C}[G]$.

**Proposition** 15.9. The irreducible characters of $G$ constitute an orthonormal basis for the class functions, Class$[G]$.

Proof: We show that the characters span Class$[G]$. If $f \in$ Class$[G]$) then $f \in \mathbb{C}[G]$ and as $Q$ forms an orthonormal basis of $\mathbb{C}[G]$ we may develop

$$f(g) = \sum_{i,j,k} c_{i,j}^{(k)} \pi_{i,j}^{(k)}(g)$$

Being a class function brings

$$\begin{aligned}
f(x)|G| &= \sum_{g\in G} f(gxg^{-1}) = \sum_{g\in G}\sum_{i,j,k} c_{i,j}^{(k)}\pi_{i,j}^{(k)}(gxg^{-1}) = \sum_{i,j,k} c_{i,j}^{(k)} \sum_{g\in G} \pi_{i,j}^{(k)}(gxg^{-1}) \\
&= \sum_{i,j,k} c_{i,j}^{(k)} \left[\sum_{g\in G} \pi^{(k)}(g)\pi^{(k)}(x)\pi^{(k)}(g^{-1})\right]_{i,j} \\
&= \sum_{i,j,k} c_{i,j}^{(k)} \left[\sum_{g\in G} \pi^{(k)}(g)\left(\sum_{r,s} \pi_{r,s}^{(k)}(x)E_{r,s}\right)\pi^{(k)}(g^{-1})\right]_{i,j} \\
&= \sum_{i,j,k} c_{i,j}^{(k)} \left[\sum_{r,s} \pi_{r,s}^{(k)}(x)\sum_{g\in G} \pi^{(k)}(g)E_{r,s}\pi^{(k)}(g^{-1})\right]_{i,j} \\
&= \sum_{i,j,k} c_{i,j}^{(k)} \sum_{r,s} \pi_{r,s}^{(k)}(x)\sum_{g\in G} \pi_{i,r}^{(k)}(g)\overline{\pi_{j,s}^{(k)}(g)} \\
&= \sum_{i,j,k} c_{i,j}^{(k)} \sum_{r,s} \pi_{r,s}^{(k)}(x)\langle \pi_{i,r}^{(k)}, \pi_{j,s}^{(k)}\rangle \\
&= \sum_{i,j,k} c_{i,j}^{(k)} \sum_{r,s} \pi_{r,s}^{(k)}(x)\delta_{i,j}\delta_{r,s}|G|/d_k = |G|\sum_{i,k} c_{i,i}^{(k)} \sum_{r} \pi_{r,r}^{(k)}(x)/d_k = |G|\sum_{i,k} \frac{c_{i,i}^{(k)}}{d_k}\chi_{\pi^{(k)}}(x)
\end{aligned}$$

which lies in the span of the $\chi_{\pi^{(k)}}$. End of Proof.

From here we can finally show that character tables have orthogonal columns.

**Proposition** 15.10. Let $\{\chi_j\}_{j=1}^s$ be the irreducible characters of $G$. If $C$ and $C'$ are conjugacy classes in $G$ with $g \in C$ and $g' \in C'$ then

$$\sum_{j=1}^{s} \chi_j(g)\overline{\chi}_j(g') = \begin{cases} |G|/|C| & \text{if} \quad C = C' \\ 0 & \text{if} \quad C \neq C'. \end{cases}$$

Proof: The characteristic function of $C'$ is a class function and so by the previous proposition we may write

$$\begin{aligned} \mathbb{1}_{C'}(g) &= \sum_{j=1}^{s} \langle \mathbb{1}_{C'}, \chi_j \rangle \chi_j(g) \\ &= \sum_{j=1}^{s} \frac{1}{|G|} \sum_{x \in G} \mathbb{1}_{C'}(x)\overline{\chi}_j(x)\chi_j(g) \\ &= \sum_{j=1}^{s} \frac{1}{|G|} \sum_{x \in C'} \overline{\chi}_j(x)\chi_j(g) \\ &= \frac{|C'|}{|G|} \sum_{j=1}^{s} \chi_j(g)\overline{\chi}_j(g'). \end{aligned}$$

As the left hand side is one when $g \in \mathcal{C}'$ and zero otherwise we arrive at the claim. End of Proof.

### 15.3. New Representations from Old

In order to apply such results we require concrete means of constructing representations and characters. We present here two means for extending known representations of subgroups. This will permit us to bootstrap from characters of $\mathrm{Alt}_3$ to characters of $\mathrm{Alt}_4$ and from characters of $\mathrm{Alt}_4$ to representations of $\mathrm{Alt}_5$.

**Lifting:** If $N$ is a normal subgroup of $G$ and $(V, \pi_{G/N})$ is a representation of the quotient group $G/N$ then we define, for each $g \in G$ via

$$\pi_G(g) \equiv \pi_{G/N}(Ng). \tag{15.23}$$

To see that it is also a representation

$$\pi_G(gh) = \pi_{G/N}(N(gh)) = \pi_{G/N}(Ng \circ Nh) = \pi_{G/N}(Ng)\pi_{G/N}(Nh) = \pi_G(g)\pi_G(h).$$

On taking the trace of each side of Eq. (15.23) It follows that each character, $\chi_{G/N}$, of $G/N$ may be **lifted** to a character

$$\chi_G(g) \equiv \chi_{G/N}(Ng) \tag{15.24}$$

of $G$. Lets use these to lift the characters of $\mathrm{Alt}_3$ to $\mathrm{Alt}_4$ via the subgroup

$$V_4 \equiv \{I, (12)(34), (13)(24), (14)(23)\}.$$

We argued in Exer. 14.12 that $V_4$ was a normal subgroup of $\mathrm{Alt}_4$ and that $\mathrm{Alt}_4/V_4 \sim \mathrm{Alt}_3$. In particular

$$\mathrm{Alt}_4/V_4 = \{V_4 I, V_4(123), V_4(132)\} \sim \{I, (123), (132)\}.$$

Hence

$$\begin{aligned}
\chi_{\mathrm{Alt}_4,j}(I) &= \chi_{\mathrm{Alt}_4/V_4,j}(V_4 I) = \chi_{\mathrm{Alt}_3,j}(I) \\
\chi_{\mathrm{Alt}_4,j}((123)) &= \chi_{\mathrm{Alt}_4/V_4,j}(V_4(123)) = \chi_{\mathrm{Alt}_3,j}((123)) \\
\chi_{\mathrm{Alt}_4,j}((132)) &= \chi_{\mathrm{Alt}_4/V_4,j}(V_4(132)) = \chi_{\mathrm{Alt}_3,j}((132)) \\
\chi_{\mathrm{Alt}_4,j}((12)(34)) &= \chi_{\mathrm{Alt}_4/V_4,j}(V_4(12)(34)) = \chi_{\mathrm{Alt}_4/V_4,j}(V_4 I) = \chi_{\mathrm{Alt}_3,j}(I)
\end{aligned}$$

and we so arrive, on recalling the Character Table 15.1 of $\mathrm{Alt}_3$, at the first three rows of Table 15.3. Regarding the fourth character, from the sum formula Eq. (15.22) we deduce that $\chi_{\mathrm{Alt}_4,4}(I) = 3$. The rest follows by orthogonality

| $\mathrm{Alt}_4$ | $I$ | $(12)(34)$ | $(123)$ | $(132)$ |
|---|---|---|---|---|
| $\chi_1$ | 1 | 1 | 1 | 1 |
| $\chi_2$ | 1 | 1 | $\omega_3$ | $\omega_3^2$ |
| $\chi_3$ | 1 | 1 | $\omega_3^2$ | $\omega_3$ |
| $\chi_4$ | 3 | $-1$ | 0 | 0 |

**Table 15.3.** The character table for $\mathrm{Alt}_4$. $\omega_3 = \exp(2\pi i/3)$

**Induction:** If $(V, \pi)$ is a representation for $H$ and $H$ is a subgroup $G$ it is natural to attempt to induce from $(V, \pi)$ a representation of $G$. We begin with the simple extension by zero,

$$\pi^0(g) \equiv \begin{cases} \pi(g) & \text{if} \quad g \in H \\ 0 & \text{otherwise,} \end{cases}$$

and establish

**Proposition** 15.11. Suppose that $(\mathbb{C}^d, \pi)$ is a representation for $H$, a subgroup of $G$, and that $\{t_1, t_2, \ldots, t_m\}$ is a complete set of representatives of left cosets of $H$. Then $(\mathbb{C}^{dm}, \mathrm{Ind}_H^G \pi)$, where

$$(\mathrm{Ind}_H^G \pi(g))_{i,j} \equiv \pi^0(t_i^{-1} g t_j) \tag{15.25}$$

is a representation of $G$.

Proof: To show that $\mathrm{Ind}_H^G \pi$ respects multiplication we take $g$ and $h$ in $G$ and compute

$$\begin{aligned}
(\mathrm{Ind}_H^G \pi(g) \mathrm{Ind}_H^G \pi(h))_{i,j} &= \sum_{k=1}^{m} (\mathrm{Ind}_H^G \pi(g))_{i,k} (\mathrm{Ind}_H^G \pi(h))_{k,j} \\
&= \sum_{k=1}^{m} \pi^0(t_i^{-1} g t_k) \pi^0(t_k^{-1} h t_j)
\end{aligned} \tag{15.26}$$

For $(\mathrm{Ind}_H^G \pi(h))_{k,j}$ to contribute to the sum requires that $t_k^{-1} h t_j \in H$ or, equivalently, $t_k H = h t_j H$. If $t_r$ is the coset representative of $h t_j H$ then Eq. (15.26) takes the form

$$(\mathrm{Ind}_H^G \pi(g) \mathrm{Ind}_H^G \pi(h))_{i,j} = \pi^0(t_i^{-1} g t_r) \pi(t_r^{-1} h t_j). \tag{15.27}$$

Similarly, for the $t_i^{-1}gt_r$ term to contribute we require that $t_iH = gt_rH$. Now as $t_rH = ht_jH$ from above it follows that $t_iH = ght_rH$ and so $t_i^{-1}ght_r \in H$ and Eq. (15.27) becomes

$$\begin{aligned}(\mathrm{Ind}_H^G\pi(g)\mathrm{Ind}_H^G\pi(h))_{i,j} &= \pi(t_i^{-1}gt_r)\pi(t_r^{-1}ht_j)\\ &= \pi(t_i^{-1}gt_rt_r^{-1}ht_j) = \pi(t_i^{-1}ght_j) = (\mathrm{Ind}_H^G\pi(gh))_{i,j},\end{aligned}$$

and so $\mathrm{Ind}_H^G\pi$ is a representation. End of Proof.

We put this to immediate use by inducing the degree 5 representation of $\mathrm{Alt}_5$

$$\Psi_3 \equiv \mathrm{Ind}_{\mathrm{Alt}_4}^{\mathrm{Alt}_5}\chi_3$$

from the degree one representation, $\chi_3$, of $\mathrm{Alt}_4$. Recalling Eq. (15.25) we must compute

$$(\Psi_3(g))_{i,j} = \chi_3^0(t_i^{-1}gt_j) \tag{15.28}$$

where the $t_i$ are a complete set of representatives of the 5 cosets of $\mathrm{Alt}_4$. Following Exer. 14.17 we select

$$t_1 = I,\ t_2 = (125),\ t_3 = (152),\ t_4 = (345),\ t_5 = (354).$$

Now we compute the next hundred permutations, $t_i^{-1}gt_j$, in `induce.m`

$$\Psi_3(123) = \chi_3^0\begin{pmatrix}(123) & (13)(25) & (153) & (12345) & (12354)\\ (235) & (135) & (12)(35) & (234) & (23)(45)\\ (15)(23) & (132) & (253) & (15234) & (15423)\\ (12543) & (15243) & (143) & (125) & (12534)\\ (12453) & (14523) & (13)(45) & (12435) & (124)\end{pmatrix} = \begin{pmatrix}\omega_3 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & \omega_3^2 & 0\\ 0 & \omega_3^2 & 0 & 0 & 0\\ 0 & 0 & \omega_3^2 & 0 & 0\\ 0 & 0 & 0 & 0 & \omega_3^2\end{pmatrix}$$

$$\Psi_3((12)(34)) = \chi_3^0\begin{pmatrix}(12)(34) & (25)(34) & (15)(34) & (12)(45) & (12)(35)\\ (25)(34) & (15)(34) & (12)(34) & (254) & (253)\\ (15)(34) & (12)(34) & (25)(34) & (154) & (153)\\ (12)(45) & (245) & (145) & (12)(35) & (12)(34)\\ (12)(35) & (235) & (135) & (12)(34) & (12)(45)\end{pmatrix} = \begin{pmatrix}1 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0\\ 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 1 & 0\end{pmatrix}$$

$$\Psi_3((12345)) = \chi_3^0\begin{pmatrix}(12345) & (13452) & (345) & (12354) & (123)\\ (234) & (13425) & (15342) & (23)(45) & (235)\\ (15234) & (134) & (12534) & (15423) & (15)(23)\\ (125) & (152) & () & (12534) & (12543)\\ (12435) & (14352) & (354) & (124) & (12453)\end{pmatrix} = \begin{pmatrix}0 & 0 & 0 & 0 & \omega_3\\ \omega_3^2 & 0 & 0 & 0 & 0\\ 0 & \omega_3 & 0 & 0 & 0\\ 0 & 0 & 1 & 0 & 0\\ 0 & 0 & 0 & \omega_3^2 & 0\end{pmatrix}$$

and

$$\Psi_3((13452)) = \chi_3^0\begin{pmatrix}(13452) & (345) & (12345) & (13542) & (132)\\ (13425) & (15342) & (234) & (13254) & (13)(25)\\ (134) & (12534) & (15234) & (13)(45) & (135)\\ (152) & () & (125) & (15342) & (15432)\\ (14352) & (354) & (12435) & (142) & (14532)\end{pmatrix} = \begin{pmatrix}0 & 0 & 0 & 0 & \omega_3^2\\ 0 & 0 & \omega_3^2 & 0 & 0\\ \omega_3 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0 & \omega_3 & 0\end{pmatrix}.$$

Regarding the reducibilty of $\Psi_3$ we evaluate its character, $\psi_3 \equiv \mathrm{tr}\Psi_3$, finding

$$\psi_3(I) = 5, \quad \psi_3((123)) = -1, \quad \psi_3((12)(34)) = 1, \quad \psi_3((12345)) = \psi_3((13452)) = 0$$

from which we arrive at

$$\begin{aligned}\langle\psi_3,\psi_3\rangle = \sum_{g\in\mathrm{Alt}_5}\psi_3(g)\overline{\psi_3(g)} &= 25|\mathrm{Conj}_{()}(\mathrm{Alt}_5)| + |\mathrm{Conj}_{(123)}(\mathrm{Alt}_5)| + |\mathrm{Conj}_{(12)(34)}(\mathrm{Alt}_5)|\\ &= 25 + 20 + 15 = 60 = |\mathrm{Alt}_5|.\end{aligned}$$

It follows from Prop. 15.7 that $\Psi_3$ is indeed irreducible.

### 15.4. The Electronic Structure of the Buckyball

The Buckyball is comprised of 60 carbon atoms bound to their three nearest neighbors, via overlapping atomic orbitals, along the pentagonal (blue) and hexagonal (red) edges of Figure 14.6(B). The hexagonal edges are typically a bit shorter and so may indicate double bonds. Our interest here is in demonstrating how representation theory may be used to ascertain the electronic structure of this large symmetric molecule. The electronic structure is used to gauge the stability and reactivity of the molecule and corresponds to determining the energy levels, and distribution, of the 60 $\pi$–electrons in the ground state.

We follow the Molecular Orbital Theory of Hückel outlined in §12.5. Namely, we build a putative molecular orbital from 60 atomic orbitals

$$\tilde{\psi} = \sum_{i=1}^{60} c_i \phi_i$$

and deduce from Rayleigh's Principle, for the Schrödinger operator $\mathcal{H}$, that

$$E_1 \leq \frac{\langle \mathcal{H}\tilde{\psi}, \tilde{\psi} \rangle}{\langle \tilde{\psi}, \tilde{\psi} \rangle}$$

for each $c$. Proceeding exactly as in §12.5 we find

$$\langle \tilde{\psi}, \tilde{\psi} \rangle = \sum_{i=1}^{60} c_i^2 = c^T c \quad \text{and} \quad \langle \mathcal{H}\tilde{\psi}, \tilde{\psi} \rangle = c^T(\alpha I + \beta A(t))c$$

where $A(t)$ is the weighted adjacency matrix of the bucky ball. In particular, it is 60-by-60 and it is 1 (or $t$) in row $i$ column $j$ if atom $i$ is adjacent to atom $j$ via a pentagonal (or hexagonal) edge. Hence, the ground state of the Buckyball may be approximated by solving

$$\tilde{E}_1 = \min_{c \in \mathbb{R}^{60}} \frac{c^T(\alpha I + \beta A(t))c}{c^T c}. \tag{15.29}$$

It follows that $\tilde{E}_1 = \alpha + \beta\lambda_1(t)$ where $\lambda_1(t)$ is the largest eigenvalue of $A(t)$. Remarkably, we will now calculate, by hand, the 60 eigenvalues of $A(t)$ without ever constructing $A(t)$ itself. This calculation will take some time and will generate ideas and results with applicability well beyond the Buckyball. So however as to to not lose sight of this application we have plotted its (to be computed) eigenvalues in Figure 15.1(A) as $t$ varies from 0 to 3. With these eigenvalues we may follow the conventions of HMO and assign the 60 $\pi$–electrons in pairs, of opposite spin, to the 30 lowest energy levels. As $E = \alpha + \lambda(t)\beta$, and $\beta$ is negative, these levels correspond to the 30 largest eigenvalues of $A(t)$. We depict these energy levels, in Figure 15.1(B), with horizontal bars crossed by arrows of opposite orientation (spin), for the case the bond ratio $t = 1$.

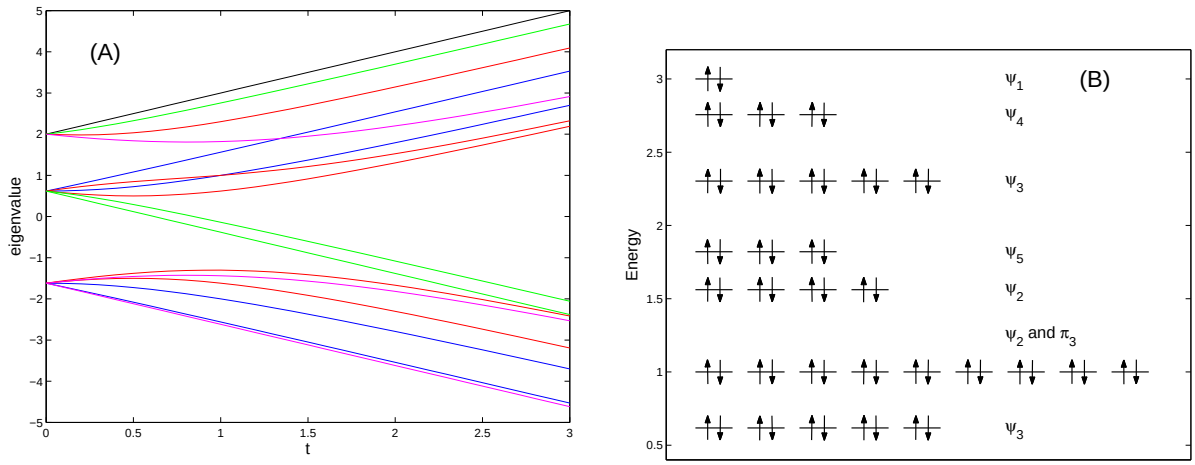



**Figure** 15.1 Electronic structure of $C_{60}$. (A) The eigenvalues of $A(t)$. Each color corresponds to a particular irreducible representation, $\Psi_j$, of $\mathrm{Alt}_5$. The degree of the representation coincides with the multiplicity of the eigenvalue. Black corresponds to $\Psi_1$ and has degree 1, blue to $\Psi_2$ of degree 4, red to $\Psi_3$ of degree 5, green to $\Psi_4$ of degree 3 and magenta to $\Psi_5$ also of degree 3. (B) Hückel energy levels at $t = 1$.

The explicit calculation of the eigenvalues of the adjacency matrix of the Buckyball hinges on five remarkable results.

1. The Buckyball is the Cayley graph of SIco, the symmetry group of the icosahedron, with respect to the set $S = \{R_A, R_A^{-1}, R_C\}$. The elements $R_A$ and $R_C$ are defined in Figure 14.5 as rotations about 5–fold and 2–fold axes of the icosahedron.

2. SIco is isomorphic to $\mathrm{Alt}_5$, the group of alternating permutations on $\mathbb{R}^5$ – and hence

$$\boxed{\mathrm{Cay}(\mathrm{SIco}, \{R_A, R_A^{-1}, R_C\}) = \mathrm{Cay}(\mathrm{Alt}_5, \{(12345), (15432), (12)(34)\}).}$$

3. The adjacency matrix of a Cayley graph can be expressed via the right regular representation

$$\boxed{A(t) = R_{\mathrm{Alt}_5}((12345)) + R_{\mathrm{Alt}_5}((15432)) + tR_{\mathrm{Alt}_5}((12)(34))}.$$

4. The right regular representation is similar to the direct sum of irreducible representations

$$\boxed{R_{\mathrm{Alt}_5} \sim \Psi_1 \oplus 4\Psi_2 \oplus 5\Psi_3 \oplus 3\Psi_4 \oplus 3\Psi_5}$$

   of $\mathrm{Alt}_5$.

5. It follows that the eigenvalues of $A(t)$ are the eigenvalues of

$$\boxed{A_j(t) = \Psi_j((12345)) + \Psi_j((15432)) + t\Psi_j((12)(34))} \tag{15.30}$$

   for $j = 1, \ldots, 5$. The biggest of these matrices is 5-by-5, and in that case its characteristic polynomial factors naturally into a cubic and a quadratic.

It remains only to compute the reduced adjacency matrices of (15.30). We begin by first building character tables for $\mathrm{Per}_4$, $\mathrm{Per}_5$ and $\mathrm{Alt}_5$.

**Irreducible Characters of** $\mathrm{Per}_4$**.** As always we assign $\chi_1(P) = 1$ for each $P \in \mathrm{Per}_4$, and $\chi_2(P)$ to be the sign character, Eq. (15.3). Recalling the conjugacy classes of $\mathrm{Per}_4$, Table 14.4, we have built the first two rows of Table 15.4. Regarding the remaining three characters we next observe that

$$\chi_t(P) = \mathrm{tr}\, P$$

defines a character on $\mathrm{Per}_4$ (in fact on every $\mathrm{Per}_n$). On evaluation of $\chi_t$ on the conjugacy classes of $\mathrm{Per}_4$, recall Table 14.4, we find

$$\chi_t(I) = 4, \quad \chi_t((12)) = 2, \quad \chi_t((123)) = 1 \quad \text{and} \quad \chi_t((12)(34)) = \chi_t((1234)) = 0.$$

Although $\chi_t$ is not irreducible we note that

$$\chi_3 \equiv \chi_t - \chi_1$$

obeys $\langle \chi_3, \chi_3 \rangle = 24 = |\mathrm{Per}_4|$ and so is irreducible by Prop. 15.7.

Next we note that

$$\chi_4(P) \equiv \chi_2(P)\chi_3(P)$$

is a distinct character for which $\langle \chi_4, \chi_4 \rangle = 24$. We get the degree of the fifth and final irreducible character by Eq. (15.22),

$$1 + 1 + 3^2 + 3^2 + d_5^2 = |\mathrm{Per}_4| = 24,$$

so $d_5 = \chi_5(I) = 2$. The remaining elements of the fifth row of Table 15.4 then follow by orthogonality with column one.

| $\mathrm{Per}_4$ | $I$ | (12) | (123) | (12)(34) | (1234) |
|---|---|---|---|---|---|
| $\chi_1$ | 1 | 1 | 1 | 1 | 1 |
| $\chi_2$ | 1 | −1 | 1 | 1 | −1 |
| $\chi_3$ | 3 | 1 | 0 | −1 | −1 |
| $\chi_4$ | 3 | −1 | 0 | −1 | 1 |
| $\chi_5$ | 2 | 0 | −1 | 2 | 0 |

**Table 15.4.** The character table for $\mathrm{Per}_4$.

**Irreducible Characters of** $\mathrm{Per}_5$**.** The first 4 characters of $\mathrm{Per}_4$ also work for $\mathrm{Per}_5$ giving the first 4 rows of Table 15.5. For the remainder we construct the pair

$$\chi_{3\pm}(P) \equiv (\chi_3^2(P) \pm \chi_3(P^2))/2.$$

and find

$$\begin{aligned}
\chi_{3\pm}(I) &= (\chi_3^2(I) \pm \chi_3(I))/2 = (16 \pm 4)/2, \\
\chi_{3\pm}((12)) &= (\chi_3^2((12)) \pm \chi_3(I))/2 = (4 \pm 4)/2, \\
\chi_{3\pm}((123)) &= (\chi_3^2((123)) \pm \chi_3((132)))/2 = (1 \pm 1)/2, \\
\chi_{3\pm}((12)(34)) &= (\chi_3^2((12)(34)) \pm \chi_3(I))/2 = (0 \pm 4)/2, \\
\chi_{3\pm}((1234)) &= (\chi_3^2((1234)) \pm \chi_3((13)(24))/2 = (0 \pm 0)/2, \\
\chi_{3\pm}((123)(45)) &= (\chi_3^2((123)(45)) \pm \chi_3((132))/2 = (1 \pm 1)/2, \\
\chi_{3\pm}((12345)) &= (\chi_3^2((12345)) \pm \chi_3((13524))/2 = (1 \mp 1)/2.
\end{aligned}$$

We check that

$$\langle \chi_{3-}, \chi_{3-} \rangle = 36 + 4|\mathrm{Conj}_{(12)(34)}(\mathrm{Per}_5)| + |\mathrm{Conj}_{(12345)}(\mathrm{Per}_5)| = 36 + 4 \cdot 15 + 24 = 120 = |\mathrm{Per}_5|$$

and so find $\chi_{3-}$ irreducible, and so call it $\chi_5$. Although $\chi_{3+}$ is not irreducible, we find, on computing $\langle\chi_{3+},\chi_1\rangle$ and $\langle\chi_{3+},\chi_3\rangle$ that

$$\chi_6 \equiv \chi_{3+} - \chi_1 - \chi_3$$

is a new irreducible character. The final character can then be constructed as in $\mathrm{Per}_4$ with the sum of squares formula Eq. (15.22) and column orthogonality.

| $\mathrm{Per}_5$ | $I$ | (12) | (123) | (12)(34) | (1234) | (123)(45) | (12345) |
|---|---|---|---|---|---|---|---|
| $\chi_1$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $\chi_2$ | 1 | −1 | 1 | 1 | −1 | −1 | 1 |
| $\chi_3$ | 4 | 2 | 1 | 0 | 0 | −1 | −1 |
| $\chi_4$ | 4 | −2 | 1 | 0 | 0 | 1 | −1 |
| $\chi_5$ | 6 | 0 | 0 | −2 | 0 | 0 | 1 |
| $\chi_6$ | 5 | 1 | −1 | 1 | −1 | 1 | 0 |
| $\chi_7$ | 5 | −1 | −1 | 1 | 1 | −1 | 0 |

**Table 15.5.** The character table for $\mathrm{Per}_5$.

**Irreducible Characters of** $\mathrm{Alt}_5$**.** We note that the restriction of $\mathrm{Per}_5$ characters 1, 3 and 6 are irreducible characters of $\mathrm{Alt}_5$, and denote these $\psi_1$, $\psi_2$ and $\psi_3$. For our next character, $\psi_4$, we use the trace of the isomorphism with SIco established in (14.31). Noting that $\operatorname{tr} R_{a,\theta} = 2\cos(\theta)+1$ we find

$$\psi_4((123)) = 2\cos(4\pi/3)+1 = 0, \quad \psi_4((12)(34))) = 2\cos(\pi)+1 = -1,$$

and

$$\psi_4((12345)) = 2\cos(2\pi/5)+1 = (1+\sqrt{5})/2, \quad \psi_4((123)) = 2\cos(6\pi/5)+1 = (1-\sqrt{5})/2,$$

The table is then completed with the sum of squares formula Eq. (15.22) and column orthogonality.

| $\mathrm{Alt}_5$ | $I$ | (123) | (12)(34) | (12345) | (13452) |
|---|---|---|---|---|---|
| $\psi_1$ | 1 | 1 | 1 | 1 | 1 |
| $\psi_2$ | 4 | 1 | 0 | −1 | −1 |
| $\psi_3$ | 5 | −1 | 1 | 0 | 0 |
| $\psi_4$ | 3 | 0 | −1 | $\alpha$ | $1-\alpha$ |
| $\psi_5$ | 3 | 0 | −1 | $1-\alpha$ | $\alpha$ |

**Table 15.6.** The character table for $\mathrm{Alt}_5$, $\alpha = (1+\sqrt{5})/2$.

We can now embark on the final step, the construction of the

**Irreducible Representations of** $\mathrm{Alt}_5$**.**

The first irreducible is simply the constant representation, $\Psi_1 = 1$. It follows that the associated reduced adjacency matrix (recall (15.30)), is $A_1(t) = 2+t$ and hence the eigenvalue of $A_1(t)$ is $2+t$. This is the black curve in Figure 15.1.

For $\Psi_2$ we begin with the representation $(\mathbb{C}^5, \Psi(\sigma) = P_\sigma)$ and note that $\operatorname{tr}\Psi = \psi_1+\psi_2$. It follows that $\Psi$ is not irreducible, and at the same time suggests, on recalling Prop. 15.1, that $\mathbb{C}^5 = W\oplus W^\perp$ and $\Psi \sim \Psi_1\oplus\Psi_2$ where $W = \mathrm{span}\{(1,1,1,1,1)^T\}$ is the eigenspace of $\Psi$ with eigenvalue 1. As

$1 = \Psi_1(\sigma)$ we may reveal $\Psi_2$ by assembling a similarity transformation composed of basis vectors of $W$ and $W^\perp$. In particular

$$X = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 1 & 0 \\ 1 & 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & 0 & -1 \end{pmatrix} \quad \text{yields} \quad X^{-1}\Psi(\sigma)X = \begin{pmatrix} 1 & 0 \\ 0 & \Psi_2(\sigma) \end{pmatrix}.$$

As $\sigma$ runs over the representatives of the conjugacy classes of $\mathrm{Alt}_5$ we find

$$\Psi_2(()) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \Psi_2((123)) = \begin{pmatrix} 0 & -1 & 1 & 0 \\ 1 & -1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \Psi_2((12)(34)) = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\Psi_2((12345)) = \begin{pmatrix} 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \quad \Psi_2((13452)) = \begin{pmatrix} -1 & 1 & 0 & 0 \\ -1 & 1 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

As its character indeed jibes with the $\psi_2$ of Tab. 15.6 we may conclude that $\Psi_2$ is irreducible.

Its associated reduced adjacency matrix (recall (15.30)) is

$$A_2(t) = \Psi_2((12345)) + (\Psi_2((12345)))^{-1} + t\Psi_2((12)(34)) = \begin{pmatrix} -1-t & 1+t & 0 & -1 \\ 0 & t & 1 & -1 \\ -1 & 1+t & -t & t \\ -1 & 0 & 1 & -1+t \end{pmatrix}.$$

Now `poly` and `simple` reveal the characteristic polynomial

$$\boxed{\det(\lambda I - A_2(t)) = (\lambda^2 + \lambda - 1 - t^2)(\lambda^2 + \lambda - (t+1)^2).} \tag{15.31}$$

Its roots give the blue eigen–curves in Figure 15.1.

We constructed the third irreducible, $\Psi_3$, via induction in the previous section. You may wish to confirm that its character indeed coincides with the $\psi_3$ of Tab. 15.6. The associated reduced adjacency matrix is

$$A_3(t) = \Psi_3((12345)) + (\Psi_3((12345)))^{-1} + t\Psi_3((12)(34)) = \begin{pmatrix} t & \omega_3 & 0 & 0 & \omega_3 \\ \omega_3^2 & 0 & \omega_3^2 + t & 0 & 0 \\ 0 & \omega_3 + t & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & \omega_3 + t \\ \omega_3^2 & 0 & 0 & \omega_3^2 + t & 0 \end{pmatrix}$$

and `poly` and `simple` reveal the characteristic polynomial

$$\boxed{\det(\lambda I - A_3(t)) = (\lambda^2 + \lambda - 1 + t - t^2)(\lambda^3 - (1+t)\lambda^2 - (3 - 2t + t^2)\lambda + t^3 - t^2 + t + 2).}$$

Its roots are the red eigen–curves of Figure 15.1.

Regarding $\Psi_4$ we recall that $\psi_4$ was constructed from the isomorphism $\mathrm{Alt}_5 \sim \mathrm{SIco}$. We note that (12345) corresponds to the five-fold rotation

$$\Psi_4((12345)) = R_A = \begin{pmatrix} \cos(2\pi/5) & \sin(2\pi/5) & 0 \\ -\sin(2\pi/5) & \cos(2\pi/5) & 0 \\ 0 & 0 & 1 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} \alpha - 1 & \sqrt{\alpha+2} & 0 \\ -\sqrt{\alpha+2} & \alpha - 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

where $\alpha = (1+\sqrt{5})/2$ as in Tab. 15.6. Regarding (12)(34), as we are leaving frame 5 invariant this can only be an order-2 rotation about an axis in frame 5. In fact rotation of $\pi$ through the midpoint of edge $4:8$ does the job

$$\Psi_4((12)(34)) = R_{m(4:8),\pi}.$$

We express the midpoint as $m(4:8) = (\alpha, -\sqrt{3-\alpha}, -2\alpha)/(2\sqrt{\alpha+2})$ and so find

$$R_{m(4:8),\pi} = \frac{1}{2\sqrt{5}} \begin{pmatrix} \alpha & -\sqrt{3-\alpha} & -2\alpha \\ -\sqrt{3-\alpha} & (3-\alpha)/\alpha & 2\sqrt{3-\alpha} \\ -2\alpha & 2\sqrt{3-\alpha} & 4\alpha \end{pmatrix} - I.$$

All together this yields the reduced adjacency matrix

$$\begin{aligned} A_4(t) &= R_A + R_A^T + tR_{m(4:8),\pi} \\ &= \begin{pmatrix} \alpha-1 & 0 & 0 \\ 0 & \alpha-1 & 0 \\ 0 & 0 & 2 \end{pmatrix} + \frac{t}{2\sqrt{5}} \begin{pmatrix} \alpha-2\sqrt{5} & -\sqrt{3-\alpha} & -2\alpha \\ -\sqrt{3-\alpha} & -((\alpha-3)/\alpha+2\sqrt{5}) & 2\sqrt{3-\alpha} \\ -2\alpha & 2\sqrt{3-\alpha} & 2(\alpha-\sqrt{5}) \end{pmatrix} \end{aligned}$$

and characteristic polynomial

$$\boxed{\det(\lambda I - A_4(t)) = (t+\lambda-\alpha+1)(\alpha t^2 + t - \alpha\lambda^2 + \alpha(1+\alpha)\lambda - 2).}$$

Its roots are the green eigen–curves of Figure 15.1.

Finally, with reference to Tab. 15.6, we note that $\psi_5$ may be obtained from $\psi_4$ by interchanging the last two conjugacy classes. As these two classes differ (recall Prop. 14.15) only by conjugation with (12) its seems natural to inspect $\Psi_5(\sigma) \equiv \Psi_4((12)\sigma(12))$. In this way

$$\Psi_5((12)(34)) = \Psi_4((12)(12)(34)(12)) = \Psi_4((12)(34)) = R_{m(4:8),\pi}$$

while

$$\Psi_5(12345) = \Psi_4((12)(12345)(12)) = \Psi_4((13542))$$

delivers a new matrix. We recognize (13452) to be rotation by $6\pi/5$ through vertices 12 and 9,

$$\Psi_4((13452)) = R_{v_{12:9},6\pi/5} \quad \text{where} \quad v_{12:9} = (-1/\alpha, -\sqrt{4\alpha+3}/\alpha, -1)/\sqrt{5}.$$

The associated reduced adjacency matrix is therefore

$$\begin{aligned} A_5(t) &= R_{v_{12:9},6\pi/5} + R_{v_{12:9},6\pi/5}^T + tR_{m(4:8),\pi} \\ &= \frac{1}{5} \begin{pmatrix} \sqrt{5}/\alpha - 5\alpha & \sqrt{20\alpha+15}/\alpha & \sqrt{5} \\ \sqrt{20\alpha+15}/\alpha & 5 & \sqrt{20\alpha+15} \\ \sqrt{5} & \sqrt{20\alpha+15} & -2\sqrt{5} \end{pmatrix} + \frac{t}{2\sqrt{5}} \begin{pmatrix} \alpha-2\sqrt{5} & -\sqrt{3-\alpha} & -2\alpha \\ -\sqrt{3-\alpha} & -((\alpha-3)/\alpha+2\sqrt{5}) & 2\sqrt{3-\alpha} \\ -2\alpha & 2\sqrt{3-\alpha} & 2(\alpha-\sqrt{5}) \end{pmatrix} \end{aligned}$$

with characteristic polynomial

$$\boxed{\det(\lambda I - A_5(t)) = \lambda^3 + \lambda^2(t+\sqrt{5}-1) + \lambda((\sqrt{5}-1)t - t^2 - \alpha - \sqrt{5}) - t^3 - (\alpha-1)t - \sqrt{5} - 3.}$$

Its roots are the magenta eigen–curves of Figure 15.1. This completes our explicit expression of the reduced characteristic polynomials of the buckyball via the explicit expression of each of the

irreducicle representations of $\mathrm{Alt}_5$. We will develop a more subtle approach in the exercises that discerns the same results from the mere character table.

### 15.5. Block Diagonalization of Symmetric Structures

We now move from eigenvalues of the adjacency matrix of a network to eigenvalues of the stiffness matrix, $S$, of the associated mechanical network. We suppose that the network has $\nu$ nodes and that each node has $n$ degrees of freedom. If the undeformed network has a symmetry group, $G \subset \mathrm{O}_n$, then we define the **stiffness representation**, $\sigma : G \to \mathrm{GL}_{\nu n}$, by asking $\sigma(g)$ to transform the vector displacements in accordance with the nodes transformed by $g$. It will then follow that the similarity transformation that expresses the stiffness representation as a direct sum of irreducible representations of $G$ also serves to block diagonalize the associated stiffness matrix. To fix ideas, we consider the equilateral triangle in Figure 15.2(A).

The equilateral triangle has $\nu = 3$ nodes each with $n = 2$ degrees of freedom and so a displacement vector $x \in \mathbb{R}^6$. We recall the stiffness matrix studied in Exer. 12.2,

$$S = A^T A = \frac{1}{4} \begin{pmatrix} 5 & \sqrt{3} & -4 & 0 & -1 & -\sqrt{3} \\ \sqrt{3} & 3 & 0 & 0 & -\sqrt{3} & -3 \\ -4 & 0 & 5 & -\sqrt{3} & -1 & \sqrt{3} \\ 0 & 0 & -\sqrt{3} & 3 & \sqrt{3} & -3 \\ -1 & -\sqrt{3} & -1 & \sqrt{3} & 2 & 0 \\ -\sqrt{3} & -3 & \sqrt{3} & -3 & 0 & 6 \end{pmatrix}. \tag{15.32}$$

The symmetry group of the equilateral triangle is $\mathrm{Dih}_3$ and we understand that it maps vertices to vertices. Our first task is to construct the stiffness representation, $\sigma$, that associates with each $g \in \mathrm{Dih}_3$ a mapping of planar vectors at vertices to planar vectors at vertices.

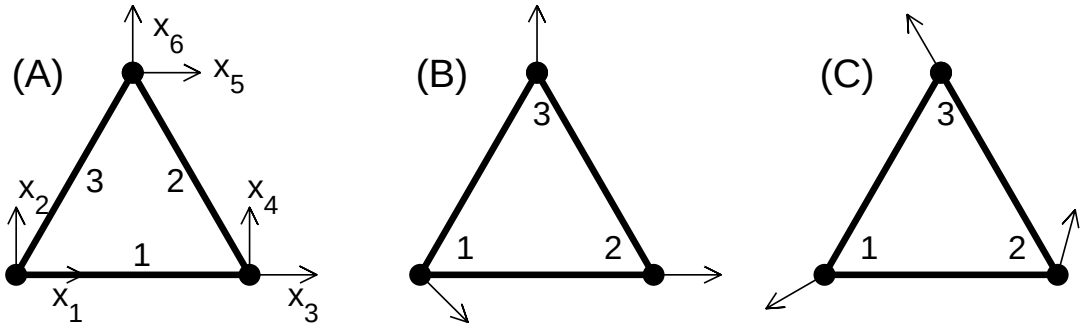



**Figure** 15.2. (A) An equilateral triangle with labeled edges and degrees of freedom. (B) Nodal displacement vectors. (C) The transformation of the vectors in (B) by $\sigma(R_{2\pi/3})$.

To begin, note that rotation by $2\pi/3$ permutes the vertices, $(1, 2, 3)$ to $(2, 3, 1)$ and so the associated displacements, $(x_1, x_2, x_3, x_4, x_5, x_6)$, to $(x_3, x_4, x_5, x_6, x_1, x_2)$. We can encode this permutation in the 6-by-6 matrix

$$P_{2\pi/3} = \begin{pmatrix} 0 & 0 & I \\ I & 0 & 0 \\ 0 & I & 0 \end{pmatrix}$$

of 2-by-2 blocks. The desired transformation is then to rotate each vector about its vertex and then

to rotate the whole frame about the centroid of the triangle. This is accomplished by

$$\sigma(R_{2\pi/3}) = P_{2\pi/3}\begin{pmatrix} R_{2\pi/3} & 0 & 0 \\ 0 & R_{2\pi/3} & 0 \\ 0 & 0 & R_{2\pi/3}\end{pmatrix} = \begin{pmatrix} 0 & 0 & R_{2\pi/3} \\ R_{2\pi/3} & 0 & 0 \\ 0 & R_{2\pi/3} & 0\end{pmatrix} \tag{15.33}$$

and illustrated in Figure 15.2(B–C). We next transform the other generator of $\mathrm{Dih}_3$. Note that the reflection $H_{v_3}$ built in (14.4) exchanges vertices 1 and 2 and so permutes the displacement vector $(x_1, x_2, x_3, x_4, x_5, x_6)$ to $(x_3, x_4, x_1, x_2, x_5, x_6)$. We encode this in the permutation matrix

$$P_{v_3} = \begin{pmatrix} 0 & I & 0 \\ I & 0 & 0 \\ 0 & 0 & I\end{pmatrix}$$

of 2-by-2 blocks and proceed to represent

$$\sigma(H_{v_3}) = P_{v_3}\begin{pmatrix} H_{v_3} & 0 & 0 \\ 0 & H_{v_3} & 0 \\ 0 & 0 & H_{v_3}\end{pmatrix} = \begin{pmatrix} 0 & H_{v_3} & 0 \\ H_{v_3} & 0 & 0 \\ 0 & 0 & H_{v_3}\end{pmatrix}. \tag{15.34}$$

We extend $\sigma$ to the remaining elements of $\mathrm{Dih}_3$ by respecting the multiplication of Table 14.1. For example,

$$\sigma(R_{4\pi/3}) = \sigma(R_{2\pi/3}R_{2\pi/3}) = \sigma(R_{2\pi/3})\sigma(R_{2\pi/3}).$$

The central result is that the stiffness representation commutes with the stiffness matrix.

**Proposition** 15.12. If the mechanical network has a symmetry group $G$ then its mass and stiffness matrices commute with its stiffness representation. That is,

$$\sigma(g)M = M\sigma(g) \quad \text{and} \quad \sigma(g)S = S\sigma(g) \qquad \forall\, g \in G. \tag{15.35}$$

Proof: To see this we return to the network's instantaneous energy, (12.35),

$$E(u(t), u'(t)) = u'(t)^T M u'(t) + u(t)^T S u(t)$$

associated with displacement $u(t)$ and velocity $u'(t)$. As each $\sigma(g)$ merely relabels the network's displacements and velocities it does not effect the energy. That is

$$E(\sigma(g)u(t), \sigma(g)u'(t)) = E(u(t), u'(t)) \quad \forall g \in G \quad \text{and} \quad u(t), u'(t). \tag{15.36}$$

It follows that $u^T\sigma(g)^T M\sigma(g)u = u^T M u$ and $v^T\sigma(g)^T S\sigma(g)v = v^T S v$ for all $u$ and $v$ in $\mathbb{R}^{n\nu}$. From here we may infer from Exer.12.6 that that $\sigma(g)^T M\sigma(g) = M$ and $\sigma(g)^T K\sigma(g) = K$. As each $\sigma(g) \in \mathrm{O}_{n\nu}$ we then arrive at $M\sigma(g) = \sigma(g)M$ and $S\sigma(g) = \sigma(g)S$. End of Proof.

The next step is to transform $\sigma$ into a direct sum of irreducibles of $\mathrm{Per}_3$. We begin by noting that its character is simply

$$\mathrm{tr}(\sigma(g)) = 6\delta_{g,I}.$$

On recalling our work in (15.21) it follows that $\sigma$ is similar to $R_{\mathrm{Per}_3}$ and hence

$$\sigma \sim \pi^{(1)} \oplus \pi^{(2)} \oplus 2\pi^{(3)}. \tag{15.37}$$

To begin we define the family of matrices

$$P_i^{(j)} \equiv \frac{d_j}{|G|} \sum_{g \in G} \pi_{i,i}^{(j)}(g)\sigma(g) \quad i = 1, \ldots, d_j, \quad j = 1, \ldots, s \tag{15.38}$$

and establish

**Proposition** 15.13. Each $P_i^{(j)}$ is symmetric and

$$P_i^{(j)} P_{i'}^{(j')} = \delta_{j,j'}\delta_{i,i'} P_i^{(j)}, \qquad \dim \mathcal{R}(P_i^{(j)}) = m_j \quad \text{and} \quad \sum_{j=1}^{s} \sum_{i=1}^{d_j} P_i^{(j)} = I_{n\nu} \tag{15.39}$$

and $S$ commutes with each $P_i^{(j)}$ and each $\mathcal{R}(P_i^{(j)})$ is an invariant subspace of $S$.

Proof: We first establish symmetry

$$(P_i^{(j)})^T \equiv \frac{d_j}{|G|} \sum_{g \in G} \pi_{i,i}^{(j)}(g)\sigma^T(g)$$

but $\sigma^T(g) = \sigma^{-1}(g) = \sigma(g^{-1}) = \sigma(g^T)$ and $\pi^{(j)}(g^T) = (\pi^{(j)}(g))^T$ so $\pi_{i,i}^{(j)}(g^T) = \pi_{i,i}^{(j)}(g)$ and hence

$$(P_i^{(j)})^T \equiv \frac{d_j}{|G|} \sum_{g \in G} \pi_{i,i}^{(j)}(g)\sigma^T(g) = \frac{d_j}{|G|} \sum_{g \in G} \pi_{i,i}^{(j)}(g^T)\sigma(g^T) = P_i^{(j)}.$$

To establish (15.39) we express the stiffness representation in terms of the irreducibles of $G$,

$$\sigma(g) = X \bigoplus_{k=1}^{s} m_k \pi^{(k)}(g) X^T$$

and invoke the Grand Orthogonality Theorem to deduce

$$\begin{aligned} P_i^{(j)} &= \frac{d_j}{|G|} \sum_{g \in G} \pi_{i,i}^{(j)}(g)\sigma(g) \\ &= \frac{d_j}{|G|} X \sum_{g \in G} \pi_{i,i}^{(j)}(g) \bigoplus_{k=1}^{s} m_k \pi^{(k)}(g) X^T \\ &= X \mathrm{diag}(0, \ldots, 0, e_i e_i^T, \ldots, e_i e_i^T, 0, \ldots, 0) X^T \end{aligned} \tag{15.40}$$

where $e_i \in \mathbb{R}^{d_j}$ and $e_i(j) = \delta_{i,j}$ and there are $m_j$ copies of $e_i e_i^T$. It follows that

$$P_i^{(j)} P_{i'}^{(j)} = X \mathrm{diag}(0, \ldots, 0, e_i e_i^T e_{i'} e_{i'}^T, \ldots, e_i e_i^T e_{i'} e_{i'}^T, 0, \ldots, 0) X^T = \delta_{i,i'} P_i^{(j)}.$$

As $P_i^{(j)}$ is a projection it follows that its rank is its trace. Its trace, $m_j$, follows directly from (15.40). The orthogonality in $j$ also follows directly from (15.40) for the $m_j$ copies of $e_i e_i^T$ share no common slots with the $m_{j'}$ copies of $e_{i'} e_{i'}^T$. On summing over $i$ and $j$ in (15.40) we find

$$\sum_{j=1}^{s} \sum_{i=1}^{d_j} P_i^{(j)} = X \mathrm{diag}(I_{d_1}, \ldots, I_{d_1}, I_{d_2}, \ldots, I_{d_2}, \ldots, I_{d_s}, \ldots, I_{d_s}) X^T = X I X^T = I_{n\nu}.$$

Finally, $SP_i^{(j)} = P_i^{(j)}S$ follows from Prop. 15.12 and the definition, (15.38). As $P_i^{(j)}$ is a projection it follows that if $v \in \mathcal{R}(P_i^{(j)})$ then $P_i^{(j)}v = v$ and so

$$Sv = SP_i^{(j)}v = P_i^{(j)}Sv,$$

which in turn implies that $Sv \in \mathcal{R}(P_i^{(j)})$. This proves that $\mathcal{R}(P_i^{(j)})$ is an invariant subspace of $S$. End of Proof.

This proposition guarantees that if we assemble the matrix

$$V = [V_1^{(1)} \cdots V_{d_1}^{(1)}\ V_1^{(2)} \cdots V_{d_2}^{(2)}\ \cdots\ \cdots\ V_1^{(s)} \cdots V_{d_s}^{(s)}] \tag{15.41}$$

where $V_i^{(j)}$ is an orthonormal basis for $\mathcal{R}(P_i^{(j)})$ then $V^TV = I$ and $V^TSV$ will be block diagonal. We illustrate this first on the equilateral triangle and then on the methane molecule.

The stiffness representation of the equilateral triangle is similar to

$$\pi^1 \oplus \pi^{(2)} \oplus 2\pi^{(3)}$$

where the $\pi^{(j)}$ are the irreducible representations of $\mathrm{Per}_3$ established in (15.2)–(15.4). With these we construct the two rank 1 projections

$$\begin{aligned}
P_1^{(1)} &= \frac{1}{6}\sum_{g\in \mathrm{Per}_3} \pi^{(1)}(g)\sigma(g) = \frac{1}{6}\sum_{g\in \mathrm{Per}_3} \sigma(g) \\
&= \frac{1}{6}\begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix} + \frac{1}{6}\begin{pmatrix} 0 & 0 & R_{2\pi/3} \\ R_{2\pi/3} & 0 & 0 \\ 0 & R_{2\pi/3} & 0 \end{pmatrix} + \frac{1}{6}\begin{pmatrix} 0 & R_{4\pi/3} & 0 \\ 0 & 0 & R_{4\pi/3} \\ R_{4\pi/3} & 0 & 0 \end{pmatrix} \\
&\quad + \frac{1}{6}\begin{pmatrix} 0 & H_{v_3} & 0 \\ H_{v_3} & 0 & 0 \\ 0 & 0 & H_{v_3} \end{pmatrix} + \frac{1}{6}\begin{pmatrix} 0 & 0 & H_{v_2} \\ 0 & H_{v_2} & 0 \\ H_{v_2} & 0 & 0 \end{pmatrix} + \frac{1}{6}\begin{pmatrix} H_{v_1} & 0 & 0 \\ 0 & 0 & H_{v_1} \\ 0 & H_{v_1} & 0 \end{pmatrix} \\
&= \frac{1}{12}\begin{pmatrix} 3 & \sqrt{3} & -3 & \sqrt{3} & 0 & -2\sqrt{3} \\ \sqrt{3} & 1 & -\sqrt{3} & 1 & 0 & -2 \\ -3 & -\sqrt{3} & 3 & -\sqrt{3} & 0 & 2\sqrt{3} \\ \sqrt{3} & 1 & -\sqrt{3} & 1 & 0 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -2\sqrt{3} & -2 & 2\sqrt{3} & -2 & 0 & 4 \end{pmatrix}
\end{aligned}$$

and

$$P_1^{(2)} = \frac{1}{6}\sum_{g\in \mathrm{Per}_3} \pi^{(2)}(g)\sigma(g) = \frac{1}{12}\begin{pmatrix} 1 & -\sqrt{3} & 1 & \sqrt{3} & -2 & 0 \\ -\sqrt{3} & 3 & -\sqrt{3} & -3 & 2\sqrt{3} & 0 \\ 1 & -\sqrt{3} & 1 & \sqrt{3} & -2 & 0 \\ \sqrt{3} & -3 & \sqrt{3} & 3 & -2\sqrt{3} & 0 \\ -2 & 2\sqrt{3} & -2 & -2\sqrt{3} & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and the two rank two projections

$$P_1^{(3)} = \frac{2}{6}\sum_{g\in \mathrm{Per}_3} \pi^{(3)}_{1,1}(g)\sigma(g) = \frac{2}{6}\sum_{g\in \mathrm{Dih}_3} g_{1,1}\sigma(g) = \frac{1}{12}\begin{pmatrix} 5 & \sqrt{3} & 5 & -\sqrt{3} & 2 & 0 \\ \sqrt{3} & 3 & \sqrt{3} & -3 & -2\sqrt{3} & 0 \\ 5 & \sqrt{3} & 5 & -\sqrt{3} & 2 & 0 \\ -\sqrt{3} & -3 & -\sqrt{3} & 3 & 2\sqrt{3} & 0 \\ 2 & -2\sqrt{3} & 2 & 2\sqrt{3} & 8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and

$$P_2^{(3)} = \frac{2}{6}\sum_{g\in \mathrm{Per}_3} \pi^{(3)}_{2,2}(g)\sigma(g) = \frac{2}{6}\sum_{g\in \mathrm{Dih}_3} g_{2,2}\sigma(g) = \frac{1}{12}\begin{pmatrix} 3 & -\sqrt{3} & -3 & -\sqrt{3} & 0 & 2\sqrt{3} \\ -\sqrt{3} & 5 & \sqrt{3} & 5 & 0 & 2 \\ -3 & \sqrt{3} & 3 & \sqrt{3} & 0 & -2\sqrt{3} \\ -\sqrt{3} & 5 & \sqrt{3} & 5 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 2\sqrt{3} & 2 & -2\sqrt{3} & 2 & 0 & 8 \end{pmatrix}.$$

From these we build the block diagonalizer, $V$, of (15.41). The first two projections are rank 1 and so the first two columns of $V$ are determined up to sign. We take

$$V_1^{(1)} = \frac{1}{2\sqrt{3}}\begin{pmatrix}\sqrt{3} & 1 & -\sqrt{3} & 1 & 0 & -2\end{pmatrix}^T \quad \text{and} \quad V_1^{(2)} = \frac{1}{2\sqrt{3}}\begin{pmatrix}1 & -\sqrt{3} & 1 & \sqrt{3} & -2 & 0\end{pmatrix}^T.$$

The next two projections are rank 2 and so there is a wide choice of bases. On letting MATLAB choose, via `orth` as in `triblock.m`, we arrive at

$$\begin{aligned} V_1^{(3)} &= \frac{1}{\sqrt{15}}\begin{pmatrix} -5/2 & -\sqrt{3}/2 & -5/2 & \sqrt{3}/2 & -1 & 0 \\ 0 & -\sqrt{3} & 0 & \sqrt{3} & 3 & 0 \end{pmatrix}^T \\ V_2^{(3)} &= \frac{1}{\sqrt{3}}\begin{pmatrix} -\sqrt{3}/2 & 1/2 & \sqrt{3}/2 & 1/2 & 0 & -1 \\ 0 & -1 & 0 & -1 & 0 & -1 \end{pmatrix}^T \end{aligned} \tag{15.42}$$

and arrive at the block diagonalization

$$V^T S V = \begin{pmatrix} 3 & & & & & \\ & 0 & & & & \\ & & 0.3 & 0.6 & & \\ & & 0.6 & 1.2 & & \\ & & & & 1.5 & 0 \\ & & & & 0 & 0 \end{pmatrix} \tag{15.43}$$

of the stiffness matrix. As the characteristic polynomial of the 2-by-2 block is $\lambda(\lambda - 3/2)$ it follows that the eigenvalues of $S$ are 3, with multiplicity 1,3/2 with multiplicity 2, and 0 with multiplicity 3.

Regarding eigenvectors, it follows from (15.43) that $V_1^{(1)}$ is an eigenvector associated with the eigenvalue 3 (that we recognize as the pure stretch mode depicted in Figure 12.4(A)), while the $V_1^{(2)}$ is an eigenvector associated with eigenvalue 0, that we recognize as a pure rotation. By the same reasoning, the first column of $V_2^{(3)}$ is an eigenvector associated to eigenvalue 3/2 (that we recognize as one of the scissor modes depicted in Figure 12.4(B–C)), while the second column of $V_2^{(3)}$ is an eigenvector associated to eigenvalue 0, that we recognize as vertical translation.

simply normalized versions of the first columns of their respective projections. The first of these, associated with eigenvalue 3, is the The first column of $V_1^{(3)}$ comes from normalization of $\sqrt{3}(P_1^{(3)})_{:,1} - (P_1^{(3)})_{:,2}$. It is associated with eigenvalue 0 and corresponds to horizontal translation. The second column, determined (up to sign) by orthogonality, is associated with eigenvalue 3/2 and correspods to one of the scissor modes depicted in Figure 12.4(B–C). Similarly, The first column of $V_2^{(3)}$ is the normalization of $(P_2^{(3)})_{:,1} + \sqrt{3}(P_2^{(3)})_{:,2}$. It is associated with eigenvalue 0 and corresponds to vertical translation. The second column, determined (up to sign) by orthogonality, is associated with eigenvalue 3/2 and correspods to the other scissor mode of Figure 12.4(B–C).

We next consider vibration of the methane molecule of Figure 3.11. Its geometric incidence matrix, $A$, derived in (3.35) using the 15 degrees of freedom illustrated in Figure 15.3, yields the 15-by-15 stiffness matrix $S = A^T A$.

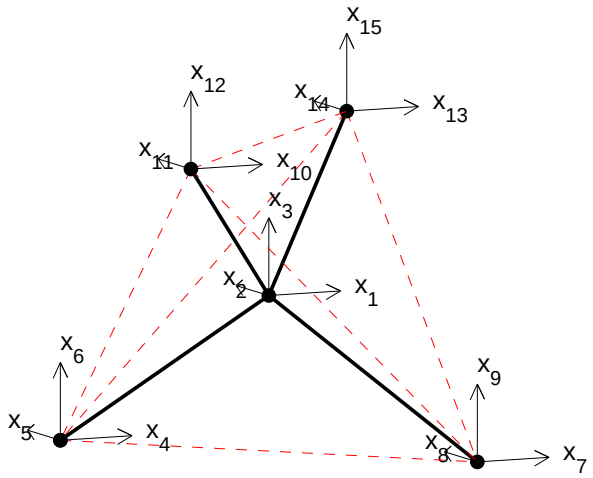



**Figure** 15.3 A re-illustration of the tetrahedron in Figure 3.11.

The molecule has the symmetry of the tetrahedron and as Tet $\sim$ Per$_4$ the associated stiffness representation, $\sigma$, takes Per$_4$ to GL$_{15}$. We define $\sigma$ on the generators (first column is central carbon, subsequent columns are hydrogens at vertices $v_1$, $v_2$, $v_3$ and $v_4$ in Figure 14.3)

$$\sigma(R_{v_4,2\pi/3}) = \begin{pmatrix} R_{v_4,2\pi/3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & R_{v_4,2\pi/3} & 0 \\ 0 & R_{v_4,2\pi/3} & 0 & 0 & 0 \\ 0 & 0 & R_{v_4,2\pi/3} & 0 & 0 \\ 0 & 0 & 0 & 0 & R_{v_4,2\pi/3} \end{pmatrix}$$

$$\sigma(R_{e_1,\pi}) = \begin{pmatrix} R_{e_1,\pi} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & R_{e_1,\pi} & 0 \\ 0 & 0 & 0 & 0 & R_{e_1,\pi} \\ 0 & R_{e_1,\pi} & 0 & 0 & 0 \\ 0 & 0 & R_{e_1,\pi} & 0 & 0 \end{pmatrix}$$

$$\sigma(H_{3,4}) = \begin{pmatrix} H_{3,4} & 0 & 0 & 0 & 0 \\ 0 & 0 & H_{3,4} & 0 & 0 \\ 0 & H_{3,4} & 0 & 0 & 0 \\ 0 & 0 & 0 & H_{3,4} & 0 \\ 0 & 0 & 0 & 0 & H_{3,4} \end{pmatrix} \quad \text{where} \quad H_{3,4} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

is reflection across the plane that contains both the origin and the line through vertices $v_3$ and $v_4$. It simply swaps vertices $v_1$ and $v_2$.

The next step is the decomposition of $\sigma$ into the irreducible representations of Per$_4$,

$$\sigma = m_1\pi^{(1)} \oplus m_2\pi^{(2)} \oplus m_3\pi^{(3)} \oplus m_4\pi^{(4)} \oplus m_5\pi^{(5)} \quad \text{where} \quad m_k = \langle \chi_\sigma, \chi_k \rangle / |\text{Per}_4|.$$

We proceed then to compute the character of $\sigma$ at reprentatives of the five cojugacy classes of Per$_4$. The first, $\chi_\sigma(I) = 15$ is easy.

Regarding the character table Tab. 15.4, the permutation (12) correponds to our reflection $H_{3,4}$ and so $\chi_\sigma((12)) = 3\text{tr}(H_{3,4}) = 3$.

Next (123) corresponds to a three-fold rotation and so $\chi_\sigma(123) = 2\text{tr}(R_{d,2\pi/3}) = 0$.

Next (12)(34) corresponds to a two-fold rotation and so $\chi_\sigma((12)(34)) = \text{tr}(R_{e_1,\pi}) = -1$.

Finally (1234) corresponds to the product of a two-fold rotation and a reflection and so $\chi_\sigma((1234)) = \text{tr}(\sigma(H_{3,4})\sigma(R_{e_1,\pi})) = \text{tr}(H_{3,4}R_{e_1,\pi}) = -1$.

With these we may now evaluate the multiplicities

$$m_1 = \langle \chi_\sigma, \chi_1 \rangle / 24 = (1\cdot 1\cdot 15 + 6\cdot 1\cdot 3 + 8\cdot 1\cdot 0 + 3\cdot 1\cdot -1 + 6\cdot 1\cdot -1)/24 = 1$$
$$m_2 = \langle \chi_\sigma, \chi_2 \rangle / 24 = (1\cdot 1\cdot 15 + 6\cdot -1\cdot 3 + 8\cdot 1\cdot 0 + 3\cdot 1\cdot -1 + 6\cdot -1\cdot -1)/24 = 0$$
$$m_3 = \langle \chi_\sigma, \chi_3 \rangle / 24 = (1\cdot 3\cdot 15 + 6\cdot 1\cdot 3 + 8\cdot 0\cdot 0 + 3\cdot -1\cdot -1 + 6\cdot -1\cdot -1)/24 = 3$$
$$m_4 = \langle \chi_\sigma, \chi_4 \rangle / 24 = (1\cdot 3\cdot 15 + 6\cdot -1\cdot 3 + 8\cdot 0\cdot 0 + 3\cdot -1\cdot -1 + 6\cdot 1\cdot -1)/24 = 1$$
$$m_5 = \langle \chi_\sigma, \chi_5 \rangle / 24 = (1\cdot 2\cdot 15 + 6\cdot 0\cdot 3 + 8\cdot -1\cdot 0 + 3\cdot 2\cdot -1 + 6\cdot 0\cdot -1)/24 = 1$$

and hence

$$\sigma = \pi_1 \oplus 3\pi_3 \oplus \pi_4 \oplus \pi_5.$$

Next we build the projections

$$P^{(j)} \equiv \frac{d_j}{24} \sum_{g \in \mathrm{Per}_4} \chi_j(g)\sigma(g) \quad \text{and} \quad V^{(j)} = \mathrm{orth}(\mathcal{R}(P^{(j)})).$$

When $j = 3$ we subdivide

$$P_i^{(3)} \equiv \frac{1}{8} \sum_{g \in \mathrm{Tet}} \pi_{i,i}^{(3)}(g)\sigma(g) = \frac{1}{8} \sum_{g \in \mathrm{Tet}} g_{i,i}\sigma(g). \quad \text{and} \quad V_i^{(3)} = \mathrm{orth}(\mathcal{R}(P_i^{(3)})).$$

and construct

$$V = [V^{(1)}\ V_1^{(3)}\ V_2^{(3)}\ V_3^{(3)}\ V^{(4)}\ V^{(5)}]$$

and find

$$V^T S V = \begin{pmatrix} 1 & & & & & \\ & S_1 & & & & \\ & & S_1 & & & \\ & & & S_2 & & \\ & & & & 0_5 & \end{pmatrix}$$

where

$$S_1 = \frac{1}{3}\begin{pmatrix} 2 & -2\sqrt{2} & -\sqrt{2} \\ -2\sqrt{2} & 4 & 2 \\ -\sqrt{2} & 2 & 1 \end{pmatrix}, \quad \text{and} \quad S_2 = \frac{1}{3}\begin{pmatrix} 4 & 2\sqrt{2} & 2 \\ 2\sqrt{2} & 2 & \sqrt{2} \\ 2 & \sqrt{2} & 1 \end{pmatrix}.$$

Each of these $S_i$ have easy identical spectra, $(0, 0, 4/3)$.

### 15.6. Fourier Analysis on Abelian Groups

If $G = \{g_1, g_2, \ldots, g_n\}$ is abelian then every element is its own conjugacy class and so there are $n$ irreducible characters, $\hat{G} \equiv \{\chi_1, \chi_2, \ldots, \chi_n\}$. It follows from our work above that $\mathrm{Class}[G] = \mathbb{C}[G]$ and so the elements of $\hat{G}$ constitute an orthogonal basis for $\mathbb{C}[G]$. Hence, if $u \in \mathbb{C}[G]$ then we may express

$$u(g_i) = \sum_{j=1}^{n} u_j \chi_j(g_i) \tag{15.44}$$

where, on account of $\langle \chi_j, \chi_k \rangle = n\delta_{j,k}$ it follows that $u_j = \langle u, \chi_j \rangle / n$. The collection of these coefficients comprises an element $\hat{u} \in \mathbb{C}[\hat{G}]$, deemed the **Fourier Transform** of $u$:

$$\boxed{\hat{u}(\chi_j) \equiv \langle u, \chi_j \rangle = \sum_{i=1}^{n} u(g_i)\overline{\chi_j(g_i)}.} \tag{15.45}$$

With this definition our initial expansion is seen as the Fourier series

$$u(g_i) = \frac{1}{n}\sum_{j=1}^{n}\langle u, \chi_j\rangle \chi_j(g_i) = \frac{1}{n}\sum_{j=1}^{n}\hat{u}(\chi_j)\chi_j(g_i). \tag{15.46}$$

In order to reconcile this with the Fourier Series considered in 9.5 we recall that in building the characters of $\mathrm{Alt}_3$ we were lead (by the nose) to the three cube roots of unity. For cyclic groups of order $n$, e.g., $\mathbb{Z}_n$, this same reasoning leads to $n$ $\mathrm{n}^{th}$ roots of unity and the associated characters

$$\chi_j(p) \equiv \exp(2\pi ijp/n), \quad 0 \le j, p < n. \tag{15.47}$$

In this case, (15.45) takes the form

$$\hat{u}(\chi_j) = \sum_{p=1}^{n} u(p)\exp(-2\pi ijp/n)$$

which is precisely the component form of the Discrete Fourier Transform as expressed in (9.60). The Parseval Theorem for the DFT, recall (9.71), extends easily to the group setting.

**Proposition** 15.14. **Parseval's Theorem**. Suppose that $G$ is abelian and $n = |G|$. For $u$ and $v$ in $\mathbb{C}[G]$ we find

$$\langle u, v\rangle = \frac{1}{n}\langle \hat{u}, \hat{v}\rangle$$

Proof:

$$\begin{aligned}\langle u, v\rangle = \sum_{i=1}^{n} u(g_i)\overline{v(g_i)} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{u}(\chi_j)\chi_j(g_i)\sum_{k=1}^{n}\overline{\hat{v}(\chi_k)\chi_k(g_i)} \\ &= \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\hat{u}(\chi_j)\overline{\hat{v}(\chi_k)}\sum_{i=1}^{n}\chi_j(g_i)\overline{\chi_k(g_i)} \\ &= \frac{1}{n}\sum_{j=1}^{n}\hat{u}(\chi_j)\overline{\hat{v}(\chi_j)} = \frac{1}{n}\langle \hat{u}, \hat{v}\rangle.\end{aligned}$$

End of Proof.

We next extend the notion of discrete convolution first posed in (9.65). Given $u, v \in \mathbb{C}[G]$,

$$(u \star v)(g_i) \equiv \sum_{j=1}^{n} u(g_i g_j^{-1})b(g_j). \tag{15.48}$$

And note that it performs well under the Fourier Transform.

**Proposition** 15.15. If $G$ is abelian then $\widehat{u \star v} = \hat{u}\hat{v}$ for each $u$ and $v$ in $\mathbb{C}[G]$.

Proof: We define $g_{i,j} = g_i g_j^{-1}$ and note that for fixed $j$ the $g_{i,j}$ exhaust $G$ as $i$ runs from 1 to $n$. We

proceed then to compute

$$\begin{aligned}
\widehat{u \star v}(\chi) = \langle u \star v, \chi \rangle &= \sum_{i=1}^{n} (u \star v)(g_i)\overline{\chi(g_i)} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} u(g_i g_j^{-1}) v(g_j)\overline{\chi(g_i)} \\
&= \sum_{j=1}^{n} v(g_j) \sum_{i=1}^{n} u(g_i g_j^{-1})\overline{\chi(g_i)} \\
&= \sum_{j=1}^{n} v(g_j) \sum_{i=1}^{n} u(g_{i,j})\overline{\chi(g_{i,j} g_j)} \\
&= \sum_{j=1}^{n} v(g_j) \sum_{i=1}^{n} u(g_{i,j})\overline{\chi(g_{i,j})}\chi(g_j) \\
&= \sum_{j=1}^{n} v(g_j)\overline{\chi(g_j)} \sum_{i=1}^{n} u(g_{i,j})\overline{\chi(g_{i,j})} = \hat{v}(\chi)\hat{u}(\chi) = \hat{u}(\chi)\hat{v}(\chi).
\end{aligned}$$

End of Proof.

This allows us to express the eigenvalues and vectors of convolution operators. We then show that adjacency matrices of Cayley graphs of Abelian groups are convolutions. For $a \in \mathbb{C}[G]$ we define the convolution operator on $\mathbb{C}[G]$ as simply

$$C_u v \equiv u \star v.$$

**Proposition** 15.16. If $G$ is abelian then $C_u \chi_k = \hat{u}(\chi_k)\chi_k$ for each $\chi_k \in \hat{G}$.

Proof: The previous proposition provides for

$$\widehat{C_u \chi_k}(\chi_j) = \widehat{u \star \chi_k}(\chi_j) = \hat{u}(\chi_k)\hat{\chi}_k(\chi_j) = \hat{u}(\chi_k) n \delta_{j,k}.$$

From here we invoke (15.46) to arrive at

$$(C_u \chi_k)(g) = \frac{1}{n}\sum_{j=1} \widehat{C_u \chi_k}(\chi_j)\chi_j(g) = \sum_{j=1} \hat{u}(\chi_k)\delta_{j,k}\chi_j(g) = \hat{u}(\chi_k)\chi_k(g).$$

End of Proof.

**Proposition** 15.17. Let $G = \{g_1, \ldots, g_n\}$ be an abelian group with irreducible characters $\{\chi_1, \ldots, \chi_n\}$. If $S \subset G$ is symmetric and $A$ is the adjacency matrix of $\mathrm{Cay}(G, S)$ then the eigenvalues of $A$ are

$$\lambda_k = \sum_{s \in S} \chi_k(s), \quad k = 1, \ldots, n,$$

with associated eigenvectors, $v_k = (\chi_k(g_1), \chi_k(g_2), \ldots, \chi_k(g_n))^T$.

Proof: We employ the basis $\{e_g : g \in G\}$ for $\mathbb{C}[G]$ where $e_g(h) = \delta_{g,h}$ to express

$$u \equiv \sum_{s \in S} e_s.$$

By the previous proposition the eigenvalues of the associated convolution operator, $C_u$, are

$$\hat{u}(\chi_k) = \sum_{i=1} u(g_i)\overline{\chi_k(g_i)} = \sum_{s \in S} \overline{\chi_k(s)} = \sum_{s \in S} \chi_k(s),$$

where the last equality is due to the symmetry of $S$.

It remains to show that $C_u$ is precisely the adjacency matrix of $\mathrm{Cay}(G, S)$. To wit

$$\begin{aligned}(C_u e_g)(h) = (u \star e_g)(h) = \sum_{s \in S}(e_s \star e_g)(h) = \sum_{s \in S}\sum_{j=1}^{n} e_s(hg_j^{-1})e_g(g_j) = \sum_{s \in S} e_s(hg^{-1}) \\ = \sum_{s \in S} e_{s^{-1}}(hg^{-1}) = \sum_{s \in S} e_g(sh) = \sum_{s \in S} e_g(hs) = \sum_{s \in S}(R_G(s)e_g)(h) = (Ae_g)(h).\end{aligned}$$

where the first equality of the second line follows from the symmetry of $S$. The next equality is due to $s^{-1} = hg^{-1}$ when $g = sh$, and the next from the abelian hypothesis. The final equalities are drawn from our definition of right regular representation and Eq. (15.7). End of Proof.

For example, we note that $\mathrm{Cay}(\mathrm{Alt}_3, \{(123), (132)\})$ is the equilateral triangle with adjacency matrix

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

and eigenvalues

$$\begin{aligned}\lambda_1 &= \chi_1((123)) + \chi_1((132)) = 2, \\ \lambda_2 &= \chi_2((123)) + \chi_2((132)) = \omega_3 + \omega_3^2 = \exp(2\pi i/3) + \exp(4\pi i/3) = -1 \\ \lambda_3 &= \chi_3((123)) + \chi_3((132)) = \omega_3^2 + \omega_3^4 = \exp(4\pi i/3) + \exp(8\pi i/3) = -1.\end{aligned}$$

More generally, we note that

$$\chi_j(p) \equiv \exp(2\pi ijp/n), \quad 0 \le j, p < n$$

are the irreducible characters of $\mathbb{Z}_n$. The Cayley Graph of $(\mathbb{Z}_n, \{1, n-1\})$, is the regular $n$-gon and the eigenvalues of its adjacency matrix are

$$\lambda_j = \chi_j(1) + \chi_j(n-1) = 2\cos(2\pi j/n), \quad 0 \le j < n. \tag{15.49}$$

Continuing in this vein, the irreducible characters of the product $\mathbb{Z}_n^2 = \mathbb{Z}_n \times \mathbb{Z}_n$ are

$$\chi_{j,k}(p, q) = \exp(2\pi ijp/n)\exp(2\pi ikq/n), \qquad 0 \le j, k, p, q < n$$

and hence the eigenvalues of the discrete torus, $T_n \equiv \mathrm{Cay}(\mathbb{Z}_n^2, \{(1,0), (n-1,0), (0,1), (0,n-1)\})$, are

$$\begin{aligned}\lambda_{j,k} &= \chi_{j,k}(1,0) + \chi_{j,k}(n-1,0) + \chi_{j,k}(0,1) + \chi_{j,k}(0,n-1) \\ &= 2\cos(2\pi j/n) + 2\cos(2\pi k/n), \qquad 0 \le j, k < n.\end{aligned} \tag{15.50}$$

We note that the $T_n$ are a growing sequence of 4-regular graphs whose eigenvalues do not exceed 4 in magnitude. 4 is an eigenvalue for each $n$ and $-4$ is an eigenvalue only for even $n$, and that the spectral gap $2 - 2\cos(2\pi/n)$ approaches zero like $1/n^2$.

### 15.7. Fourier Series and Characters of the Circle

The unit circle

$$\mathbb{T} \equiv \{z \in \mathbb{C} : |z| = 1\}$$

is an infinite abelian group with an infinite family of cyclic subgroups

$$\mathbb{T}_n \equiv \{\exp(2\pi i m/n) : m = 0, \ldots, n-1\}, \quad n = 1, \ldots.$$

With an infinite number of group elements one is naturally compelled to recognize the "distance" between elements and to expect that any character take near-by group elements to near-by numbers in $\mathbb{C}^*$. More precisely we shall restrict characters of the circle to be **continuous** homomorphisms of the circle. Here continuity of $\phi$ means that $\phi(z_n) \to \phi(z_*)$ whenever $z_n \to z_*$ in $\mathbb{T}$. This latter limit is unambiguous in the sense that $\mathbb{T}$ is **closed** in the sense that it contains all of its limit points.

**Proposition** 15.18. If $\phi$ is a character of $\mathbb{T}$ then $\ker\phi \equiv \{z \in \mathbb{T} : \phi(z) = 1\}$ is a closed subgroup of $\mathbb{T}$.

Proof: We proved long ago that kernels of homomorphisms are subgroups. If $\{z_n\}_n \subset \ker\phi$ and $z_n \to z_*$ then $\phi(z_n) \to \phi(z_*)$. As $\phi(z_n) = 1$ for all $n$ it follows that $\phi(z^*) = 1$, i.e., $z_* \in \ker\phi$. End of Proof.

We next show that the $\mathbb{T}_n$ are the only closed subgroups of $\mathbb{T}$. Note that if $z = \exp(2\pi i t)$ does not belong to any $\mathbb{T}_n$ then $t$ is not a ratio of two integers, i.e., $t$ is irrational. We will now show that powers of such a $z$ effectively cover the disk. More precisely, given $w \in \mathbb{T}$ and an $\varepsilon > 0$ we will show that there exists a natural number $N$ such that $|z^N - w| < \varepsilon$. We write $w = \exp(2\pi i s)$ for $s \in [0, 1)$ and $z^N = \exp(2\pi i N t) = \exp(2\pi i\,\mathrm{frac}(Nt))$, where $\mathrm{frac}(x)$ is the fractional part of $x$. We note that it suffices to prove

**Proposition** 15.19. Suppose $t \in [0, 1)$ is irrational. Given $\varepsilon > 0$ there exists a natural number $m$ such that $\mathrm{frac}(mt) \neq 0$ and $|\mathrm{frac}(mt)| < \varepsilon$.

Proof: Choose $N$ such that $1/N < \varepsilon$ and define the $N$ intervals $I_n = [(n-1)/N, n/N)$, $n = 1, \ldots, N$.Each of the $N+1$ numbers $\mathrm{frac}(jt)$, $j = 1, \ldots, N+1$ must lie in one of the intervals and, as $N + 1 > N$ it follows that one interval must contain two of these numbers. In particular, there exist $j \neq k$ and $n$ such $\mathrm{frac}(jt) \in I_n$ and $\mathrm{frac}(kt) \in I_n$. Of course this means that they are close, namely, $|\mathrm{frac}(jt) - \mathrm{frac}(kt)| = |\mathrm{frac}((j-k)t)| < 1/N < \varepsilon$. Finally, as $t$ is irrational it follows that $\mathrm{frac}((j-k)t) \neq 0$. End of Proof.

following the notes of Constantin Teleman

**Proposition** 15.20. The closed proper subgroups of $\mathbb{T}$ are the cyclic subgroups $\mathbb{T}_n$ of $n$th roots of unity, $n > 1$.

Proof: If $q \in \mathbb{T}$ is not a root of unity, then its powers are dense in $\mathbb{T}$ by Prop. 15.19. So, any closed, proper subgroup of $\mathbb{T}$ consists only of roots of unity. Among those, there must be one of smallest

argument (in absolute value) or else there would be a sequence converging to 1; these would again generate a dense subgroup of $\mathbb{T}$. The root of unity of smallest argument is then the generator. End of Proof.

**Proposition** 15.21. A continuous 1–dimensional representation $\mathbb{T} \to \mathbb{C}^*$ has the form $z \to z^n$, for some integer $n$.

Proof: A continuous map Suppose that $\phi : \mathbb{T} \to \mathbb{C}^*$ is a continuous homomorphism. The latter implies that $\phi(z^n) = (\phi(z))^n$. The image must lie on the unit circle, because the integral powers of any other complex number form an unbounded sequence. So $\phi$ is a continuous homomorphism from $\mathbb{T}$ itself. Now, $ker(\phi) \equiv \{z \in \mathbb{T} : \phi(z) = 1\}$ is a closed subgroup of $\mathbb{T}$. If $ker(\phi) = \mathbb{T}$ then $\phi \equiv 1 = z^0$.

If $ker(\phi) = \mu_n$ for $n > 1$, we will now show that $\phi(z) = z^{\pm n}$, with the same choice of sign for all $z$. To see this, define a continuous function

$$\psi : [0, 2\pi/n] \to \mathbb{R}, \psi(0) = 0, \psi(\theta) = arg\phi(\exp(i\theta));$$

in other words, we parametrize $\mathbb{T}$ by the argument $\theta$, start with $\psi(0) = 0$, which is one value of the argument of $\phi(1) = 1$, choose the argument so as to make the function continuous.

Because $ker(\phi) = \mu_n$, $\psi$ must be injective on $[0, 2\pi/n)$. By continuity, it must be monotonically increasing or decreasing (Intermediate Value Theorem), and we must have $\psi(2\pi/n) = \pm 2\pi$: the value zero is ruled out by monotonicity and any other multiple of $2\pi$ would lead to an intermediate value of $\theta$ with $\phi(\exp(i\theta)) = 1$. Henceforth, $\pm$ denotes the sign of $\psi(2\pi/n)$.

Because $\phi$ is a homomorphism and $\phi(\exp(2\pi i/n)) = 1$, $\phi(\exp(2\pi i/mn)$ must be an $m$th root of unity, and so $\psi(\{2\pi k/mn\}) \subset \{\pm 2\pi k/m\},\ k = 0, \ldots, m$. By monotonicity, these $m + 1$ values must be taken exactly once and in the natural order, so $psi(2\pi k/mn) = \pm 2\pi k/m$, for all $m$ and all $k = 0, \ldots, m$. But then, $\psi(\theta) = \pm n \cdot \theta$, by continuity, and $\phi(z) = z^{\pm n}$, as claimed. End of Proof.

### 15.8. Notes and Exercises

We have followed Steinberg (2011), James and Liebeck (2001) and Krebs and Shaheen (2011). Our Buckyball work follows Chung and Sternberg (1993). Our construction of the irreducible representations of $\mathrm{Per}_n$ and $\mathrm{Alt}_n$ for small $n$ follows the bare-handed approach of Maria Wesslen. There is a systematic approach for constructing the irreducible representations of $\mathrm{Per}_n$ for every $n$ via the method of Young Diagrams.

Our exposition on spectra of symmetric structures follows **?**.

For the very big picture, the argument that representation theory was central to the development of math and physics see **?**. This is done more concretely by **?** and **?**.

Exer. 15.8 is taken from Babai (1979).

1. Use the previous exercise to express the adjacency matrix of $\mathrm{Cay}(\mathrm{Per}_3, \{(12), (13), (23)\})$ as

$$A = R_{\mathrm{Per}_3}((12)) + R_{\mathrm{Per}_3}((13)) + R_{\mathrm{Per}_3}((23))$$

and to deduce that $A$ is similar to

$$\Pi((12)) + \Pi((13)) + \Pi((23)) = \mathrm{diag}(3, -3, H_{v_3} + H_{b^\perp} + H_{c^\perp}, H_{v_3} + H_{b^\perp} + H_{c^\perp})$$

and so deduce that $A$ has two simple eigenvalues, $\pm 3$, and a zero eigenvalue of multiplicity 4.

2. It appears that $R_{\mathrm{Per}_3} = \sigma$. Build

$$(\mathbb{C}[\mathrm{Per}_3], R_{\mathrm{Per}_3}) = (V_1, \pi^{(1)}) \oplus (V_2, \pi^{(2)}) \oplus (V_{3,1}, \pi^{(3)}) \oplus (V_{3,2}, \pi^{(3)})$$

where $V_k$ is the column space of

$$\sum_{g \in \mathrm{Per}_3} \pi^{(k)}(g)\sigma(g)$$

for $k = 1, 2$, and $V_{3,i}$ is the column space of

$$\sum_{g \in \mathrm{Per}_3} \pi^{(3)}_{i,1}(g)\sigma(g).$$

3. Show that the right regular representation, $(\mathbb{C}[G], R_G)$, is unitary, i.e., that $(R_G(g))^* = R_G(g^{-1})$ for each $g \in G$.

4. Prove that the set of class functions, Class$[G]$) of Eq. (15.18), is a subspace of $\mathbb{C}[G]$. Prove that the dimension of Class$[G]$) is the number of conjugacy classes of $G$.

5. Use the fact, from Exer. 14.14, that $\mathrm{Dih}_4/\mathrm{SDih}_4 \sim V_4$ and the fact that $V_4$ is abelian to lift the four characters of $V_4$ to $\mathrm{Dih}_4$. Use the sum of squares formula to deduce that the degree of the remaining irreducible representation is 2. Complete the character table with the standard matrix representation.

| $\mathrm{Dih}_4$ | $I$ | $R_{\pi/2},\ R_{3\pi/2}$ | $R_\pi$ | $H_{e_i}$ | $H_{d_i}$ |
|---|---|---|---|---|---|
| $\chi_1$ | 1 | 1 | 1 | 1 | 1 |
| $\chi_2$ | 1 | 1 | 1 | $-1$ | $-1$ |
| $\chi_3$ | 1 | $-1$ | 1 | $-1$ | 1 |
| $\chi_4$ | 1 | $-1$ | 1 | 1 | $-1$ |
| $\chi_5$ | 2 | 0 | $-2$ | 0 | 0 |

**Table 15.4.** The character table for $\mathrm{Dih}_4$.

6. Show that $L_G(h)v(x) = v(h^{-1}x)$ defines a representation $(\mathbb{C}[G], L_G)$. Show that it commutes with $R_G$ in the sense that $R_G(g)L_G(h) = L_G(h)R_G(g)$ for all $g, h \in G$. Use this and Eq. (15.7) to show that $L_G$ commutes with the adjacency matrix of every Cayley graph of $G$. Argue that if $G$ has $p$ irreducibles then each such adjacency matrix has at most $p$ distinct eigenvalues.

7. (a) Show that the graph of benzene carbons is $\mathrm{Cay}(\mathrm{SDih}_6, \{R_{\pi/3}, R_{5\pi/3}\})$. (b) Use (a) and Prop. 15.17 to show that the eigenvalues of the adjacency matrix are

$$\lambda_k = 2\cos((k-1)\pi/3), \quad k = 1, \ldots, 6.$$

(c) Construct the Huckel energy for Benzene. (d) Use Prop. 15.17 to compute the eigenvectors and illustrate their orbitals.

8. Prove that the eigenvalues of $\mathrm{Cay}(G, H)$ obey (need full dress Babai to handle t-bond ratio)

$$\lambda_{i,1}^k + \cdots + \lambda_{i,n_i}^k = \sum_{g_1, \ldots g_k \in H} \chi_i \left( \prod_{s=1}^{k} g_s \right)$$

where the sum is over all $k$-tuples of elements drawn from $H$. Use this to prove that the eigenvalues of $\Pi_4$ (recall Eq. (***) obey

$$\lambda_{4,1} + \lambda_{4,2} + \lambda_{4,3} = \alpha - 1 - t$$
$$\lambda_{4,1}^2 + \lambda_{4,2}^2 + \lambda_{4,3}^2 = 8 - 2\alpha$$
$$\lambda_{4,1}^3 + \lambda_{4,2}^3 + \lambda_{4,3}^3 = 4\alpha + 2 + (6\alpha - 6)t + 6\alpha t^2 - t^3.$$

Now use the Newton Identities, Eq. (11.73), to recover the characteristic polynomial Eq. (***.

# 16. Graph Theory*

We have studied electrical, mechanical and metabolic networks in the preceding chapters. We here study their natural abstraction.

### 16.1. Graphs, Matrices and Groups

A graph, $\Gamma$, is a pair of sets of vertices, $V$, and edges, $E$. We order the vertices $v_1$ through $v_n$ and denote the adjacency matrix of $\Gamma$ by $A_\Gamma$, where $A_\Gamma(i,j)$ is the number of edges between $v_i$ and $v_j$. The **degree** of $v_i$ is denoted $\deg(v_i)$ and is defined to be the number of edges with an end at $v_i$.

Preview...

### 16.2. Trees and Molecules

A tree is a graph without loops.

Given $v \in V$ there exist $\deg(v)$ subtrees, $\Gamma_i(v)$, see Figure 16.1, of $\Gamma$ stemming from $v$. (More precisely, if $v_1, \ldots, v_d$ are the vertices adjacent to $v$ then $\Gamma_i(v)$ is the union of paths whose first edge is $vv_i$.) The number of edges in the $i$th subtree stemming from $v$ is $|E_i(v)|$ and the **Jordan Index** of $v$ is the size of the largest subtree,

$$J(v) \equiv \max_i |E_i(v)|.$$

If $J(v) = |E_k(v)|$ we call $\Gamma_k(v)$ a maximal subtree stemming from $v$.

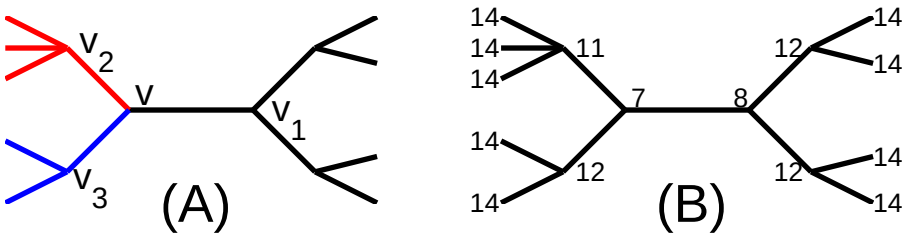



**Figure** 16.1. An illustration of the Jordan Index. (A) As $\deg(v) = 3$ there are three subtrees stemming from $v$. $\Gamma_1(v)$ is colored black, $\Gamma_2(v)$ is red, and $\Gamma_3(v)$ is blue. (B) We have labeled each vertex with its Jordan Index.

Jordan used this index to divide trees into those with a centroid and those with a bicentroid.

**Proposition** 16.1. If $\Gamma = (V, E)$ is a tree then there exists *either* a single vertex (centroid) with Jordan Index less than or equal to $(|V| - 1)/2$ *or* a single pair of adjacent vertices (bicentroid) with Jordan Indices of $|V|/2$. All other vertices have Jordan Indices strictly greater than $|V|/2$.

Proof: Set $n = |V|$ and $m = n - 1$. Let $v_1$ denote a vertex with the smallest Jordan Index, call it $J_1$. We first show that $J_1 \leq n/2$.

Suppose $u \sim v_1$ is a vertex belonging to a maximal subtree stemming from $v_1$. The neighbors of $u$, $\{v_1, v_2, \ldots, v_{\deg(u)}\}$, generate subtrees stemming from $u$ with sizes $|E_i(u)|$. As these subtrees exhaust the full tree we find

$$\sum_{i=1}^{\deg(u)} |E_i(u)| = m. \tag{16.1}$$

Similarly, as there are $m - J_1$ edges outside the maximal tree stemming from $v_1$ then

$$|E_1(u)| = m - J_1 + 1 = n - J_1. \tag{16.2}$$

Also, as $J_1$ is minimal it follows that

$$J_1 \leq J(u) = \max_i |E_i(u)|.$$

As (16.1) implies $|E_i(u)| < J_1$ for $i \geq 2$ it follows that $|E_1(u)| \geq J_1$ and so from (16.2) it follows that $m - J_1 + 1 \geq J_1$, i.e., $J_1 \leq n/2$.

We now show that all other vertices have larger Jordan indices. Take $w \in V$ and note that it lies within one subtree stemming from $v_1$, say $w \in \Gamma_1(v_1)$. It follows that there is subtree stemming from $w$ with at least $m - |E_1(v_1)| + \text{dist}(w, v_1)$ edges. As $|E_1(v_1)| \leq J_1$ and $\text{dist}(w, v_1) \geq 1$ it follows that $J(w) \geq m - |E_1(v_1)| + \text{dist}(w, v_1) \geq m - J_1 + 1$. Hence, if $J_1 < n/2$ than $J(w) > n/2$. If instead $J_1 = n/2$ then there is a unique maximal subtree stemming from $v_1$ (for if two trees had $n/2$ edges together we would have $n > m$ edges). It follows that $J_1(w) > n/2$ unless $w$ is both adjacent to $v_1$ and lies in its maximal subtree - in which case $J(w) \geq n/2$. To establish equality note that the subtree stemming from $w$ that contains $v_1$ has $m - J_1 + 1 = n/2$ edges. End of Proof.

Two vertices are similar if there is an automorphism that takes one to the other. An edge is called a symmetry edge if its two ends are similar. Action/Orbit/Burnside

If $g \in \text{Per}(\Gamma)$ then $J(v) = J(gv)$ for all $v \in V$. Note that $g$ preserves adjacency and so degree, so if $d$ subtrees stem from $v$ then $d$ subtrees stem from $gv$ and their sizes are unchanged. Hence the centroid (bicentroid) is fixed by every $g$. We need: If two adjacent vertices have the same Jordan index then they are the bicentroid. We define

$$p^* \equiv [\text{Per}(\Gamma) : V] \quad \text{and} \quad q^* \equiv [\text{Per}(\Gamma) : E]$$

to deduce $p^* - (q^* - s) = 1$ from Burnside we prove

$$|\text{Fix}_g(V)| = |\text{Fix}_g(E)| + 1$$

for graphs with centroid. To prove this suppose $g$ fixes $e$ but flips its vertices then $e$ is a bicentroid. So if $g$ fixes an edge then it fixes both vertices. Now suppose that $g$ fixes two disjoint edges, consider the unique path that joins them - as $g$ preserves adjacency it must preserve this path. It really follows that $\text{Fix}_g(E)$ is a tree!

Next we suppose $\{\Gamma_i\}$ to be a list of the trees with $p$ vertices, and prove

$$\sum_{i=1}^{t_p} [\text{Per}(\Gamma_i) : V_i] = T_p.$$

(This "feels" like an application of PET itself). To see this, look at small $p$ (4,5) and observe that $[\text{Per}(\Gamma_i) : V_i]$ is the number of roots. Even "harder" is the statement

$$\sum_{i=1}^{t_p} ([\text{Per}(\Gamma_i) : E_i] - s_i) = L_p$$

the number of trees with $p$ points rooted at an edge which is not a symmetry edge. Hence, $t_p = T_p - L_p$. So if $L(x)$ is the counting series for trees rooted at a nonsymmetry line, then

$$t(x) = T(x) - L(x).$$

We now focus on distances within trees as a simple manifestation of the basic question of structure and function. We ask how the arrangement of the carbons in alkanes effects their boiling point. An **alkane** is an acyclic hydrocarbon of the form $C_nH_{2n+2}$. We have illustrated the three isomers of pentane in Figure 16.2

**Figure** 16.2. The three isomers of pentane.

The first question is one of enumeration... (Cayley/Polya)

Let $T_p$ denote the number of rooted trees on $p$ vertices. Its generating function is

$$T(x) = \sum_p T_p x^p,$$

also, if $T^{(n)}(x)$ is the generating function for the rooted trees with root of degree $n$ then

**Proposition** 16.2 The generating function for rooted trees obeys

$$T(x) = x \exp\left(\sum_{k=1}^{\infty} T(x^k)/k\right)$$

Proof: We first find, $T^{(n)}$, the generating function which enumerates rooted trees in which the root has degree $n$. We note that

$$T(x) = \sum_n T^{(n)}(x)$$

and $T^{(0)}(x) = x$ and $T^{(1)}(x) = xT(x)$. Each of the latter trees corresponds in a natural way to a "combination with repetition" of $n$ rooted trees.

More specifically, given a collection of $n$ rooted trees, a new rooted tree is formed by adding one new point and making it adjacent to each of the roots of then given rooted trees. Clearly all trees whose roots have degree $n$ can be formed in this manner. To find out how many there are, we consider the power group $E^{\mathrm{Per}_n}$ with object set $Y^X$ where $E$ is the identity group, $X = \{1, \ldots, n\}$, and $Y$ is the set of all rooted trees. Then each function in $Y^X$ corresponds to an ordered $n$-tuple of rooted trees. We define the weight of each rooted tree in $Y$ to be the number of points in the tree. Then $T(x)$ enumerates the elements of $Y$ by weight and is called the "figure counting series" for $Y$. Thus the weight of each function in $Y^X$, as defined by (***, is the total number of points in the $n$ rooted trees of the $n$-tuple to which the function corresponds.

Since $\mathrm{Per}_n$ consists of all permutations of $X$, the orbits of the power group $E^{\mathrm{Per}_n}$ correspond precisely to rooted trees whose root has degree $n$. Note that the weight of each orbit, which is the weight of any function in it, is just one less than the total number of points in the rooted tree to which the orbit corresponds. Therefore on applying PET with $A = \mathrm{Per}_n$ and $T(x)$ as the figure counting series, we have $Z(\mathrm{Per}_n, T(x))$ as the function counting series, and the coefficient of $x^p$ in $Z(\mathrm{Per}_n, T(x))$ is the number of rooted trees of order $p+1$ whose roots have degree $n$. Multiplication of $Z(\mathrm{Per}_n, T(x))$ by $x$ corrects the weights so that the coefficient of $x^p$ in $xZ(\mathrm{Per}_n, T(x))$ is the number of these trees with $p$ points. Then on summing over all possible values of $n$, $T(x)$ itself is obtained:

$$T(x) = x \sum_{n=0}^{\infty} Z(\mathrm{Per}_n, T(x)). \tag{16.3}$$

The proof is completed by applying the identity (14.44) for sums of cycle indexes to the right side of (16.3). End of Proof.

**Proposition** 16.2 The generating function for rootless trees obeys

$$L(x) = T(x) - (T^2(x) - T(x^2))/2. \tag{16.4}$$

Proof:

End of Proof.

With $\theta(\Gamma)$ the boiling point we denote the **Wiener Index**

$$W(\Gamma) = \frac{1}{2}\sum_{u,v\in V} \text{dist}(u,v)$$

and the Wiener Polarity

$$W_P(\Gamma) = |\{\{u,v\} \in V \times V : \text{dist}(u,v) = 3\}| \tag{16.5}$$

The key result is

$$\theta(\Gamma) - \theta(\Gamma_0) = \frac{98}{n^2}(W(\Gamma) - W(\Gamma_0)) + 5.5(P(\Gamma) - P(\Gamma_0))$$

where $\theta(\Gamma)$ is boiling point..., $\Gamma_0$ denotes the normal (linear) isomer and at this geometry

$$W(\Gamma_0) = n(n+1)(n-1)/6 \quad \text{and} \quad P(\Gamma_0) = n-3$$

and

| Alkane | Methane | Ethane | Propane | Butane | Pentane | Hexane | Heptane | Octane |
|---|---|---|---|---|---|---|---|---|
| Formula | $CH_4$ | $C_2H_6$ | $C_3H_8$ | $C_4H_{10}$ | $C_5H_{12}$ | $C_6H_{14}$ | $C_7H_{16}$ | $C_8H_{18}$ |
| Boiling point | -162 | -89 | -42 | 0 | 36 | 69 | 98 | 126 |

**Table 16.1.** Boiling points of the normal isomers of the first 8 alkanes.

The distance, $d_{ij} \equiv \text{dist}(v_i, v_j)$, from $v_i$ to $v_j$ is the number of edges, without backtracking, required to get from $v_i$ to $v_j$. For example, the adjacency matrix for the tree in Figure 16.3(A) and its associated distance matrix are

$$A = \begin{pmatrix} 0&0&0&0&0&1\\ 0&0&1&0&0&0\\ 0&1&0&1&0&1\\ 0&0&1&0&0&0\\ 0&0&0&0&0&1\\ 1&0&1&0&1&0 \end{pmatrix} \quad \text{and} \quad D = \begin{pmatrix} 0&3&2&3&2&1\\ 3&0&1&2&3&2\\ 2&1&0&1&2&1\\ 3&2&1&0&3&2\\ 2&3&2&3&0&1\\ 1&2&1&2&1&0 \end{pmatrix}$$

Given an orientation on the edges it is of interest to record the **vertex-edge incidence matrix** $B$ where

$$B_{ij} = \pm 1 \quad \text{if vertex } i \text{ is the positive(negative) end of edge } j.$$

Its connection to $D$ is

**Proposition** 16.4. If $\Gamma$ is a tree with $n$ vertices with distance matrix $D$ and vertex-edge incidence matrix $B$ then $B^T D B = -2I_{n-1}$.

Proof: The product $b_{is}d_{ij}b_{it} = 0$ unless $v_i$ is an end of $e_s$ and $v_j$ is an end of $e_i$. Suppose $e_s = \{w, x\}$ and $e_t = \{y, z\}$, where $x$ and $z$ are positive ends. Then

$$\sum_{i,j \in V} b_{is}d_{ij}q_{jt} = d_{wy} - d_{wz} - d_{xy} + d_{xz}.$$

If $s = t$, then $d_{wy} = d_{xz} = 0$ while $d_{wz} = d_{xy} = 1$, so the sum is $-2$. If $s \neq t$, it may still happen that $w = y$ or $x = z$. If $w = y$, then $d_{wy} = 0$, $d_{xz} = 2$, and $d_{wz} = d_{xy} = 1$, so the sum is zero. The case $x = z$ is handled similarly. If $w$, $x$, $y$, and $z$ are four distinct vertices then either $x$ is on the (unique) path from $w$ to $e_t$, or $w$ is on the path from $x$ to $e_t$. These cases are similar. We argue the first, i.e., $d_{wy} = d_{xy} + 1$ and $d_{wz} = d_{xz} + 1$. In this case, the sum is $d_{xy} + 1 - (d_{xz} + 1) - d_{xy} + d_{xz} = 0$. End of Proof.

We next define $K = B^T B$ and note that

$$K = 2I + A(\Gamma^*) \quad \text{and} \quad L = BB^T.$$

and establish

**Proposition** 16.5. If $\Gamma$ is a tree with distance matrix $D$ and vertex-edge incidence matrix $B$ then the eigenvalues of $-2(B^T B)^{-1}$ interlace the eigenvalues of $D$.

Proof: We note that the columns of $B$ are linearly independent and that each contains exactly one 1 and one $-1$ while all other elements are zero. Applying Gram–Schmidt to the columns of $B$ we find $Q = BM$ and we note that $e \in \mathbb{R}^n$, the column of ones, is orthogonal to $\mathcal{R}(B)$ and so the augmented matrix $U \equiv (BM \quad e/\sqrt{n})$ is orthogonal and

$$U^T DU = \begin{pmatrix} M^T B^T DBM & M^T B^T De/\sqrt{n} \\ e^T DBM/\sqrt{n} & e^T De/n \end{pmatrix} = \begin{pmatrix} -2M^T M & M^T B^T De/\sqrt{n} \\ e^T DBM/\sqrt{n} & 2W/n \end{pmatrix} \tag{16.6}$$

Next, from $M^T KM = M^T B^T BM = I$ we can take inverses and find $K^{-1} = MM^T$ and so $K^{-1}$ and $M^T M$ have the same spectrum and so the result follows from Exer. 12.8. End of Proof.

We note that $\mathrm{tr}D = 0$ and take the trace of (16.6) and find

$$0 = \mathrm{tr}D = \mathrm{tr}(U^T DU) = \mathrm{tr}(-2M^T M) + 2W/n$$

which establishes

**Proposition** 16.6. If $\Gamma$ is a tree on $n$ vertices with vertex-edge incidence matrix $B$ then $W(\Gamma) = n\mathrm{tr}((B^T B)^{-1})$

Work out $W(T)$ for the pentane isomers.

### 16.3. Spanning Trees and Electrical Networks

A **spanning tree** of $G$ is a tree that hits every vertex. Even small graphs have many spanning trees, e.g, the graphs of Figure 14.7 have 75 and 81 spanning trees respectively. We have plotted 3 of these in Figure 16.3

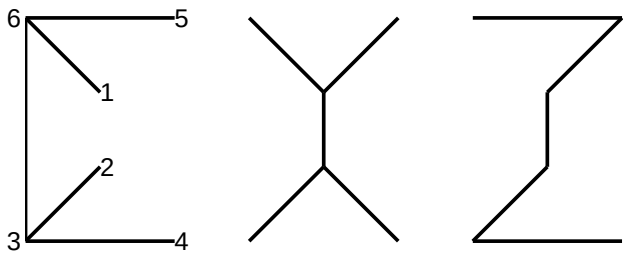



**Figure** 16.3. Three of the 75 spanning trees of the graph in Figure 14.7(A).

Their enumeration would very tedious, linear algebra provides a very simple formula.

We count the number of spanning trees in a graph and proceed to count pathlengths in trees.

The basic construction is deletion and contraction. The subgraph obtained by taking graph G and deleting edge e, but leaving all other edges and vertices as is, is denoted $G - \{e\}$. The contraction of a graph, $G/\{e\}$ is the multigraph (not a subgraph) obtained from $G$ by contracting the edge $e = \{v, w\}$ until the two vertices $v$ and $w$ coincide. Call this new vertex $vw$. We denote by $\kappa(G)$ the number of spanning trees in $G$ and prove the deletion–contraction formula

**Proposition** 16.7. For any edge $e$ in the graph $G$, $\kappa(G) = \kappa(G - \{e\}) + \kappa(G/\{e\})$.

Proof: The edge $e$ divides the spanning trees of $G$ into (i) those that contain $e$ and (ii) those that do not contain $e$.

Regarding (ii), a spanning tree misses $e$ iff it is a spanning tree of $G-\{e\}$. Regarding (i) we show that a spanning tree containing $e$ is equivalent to a spanning tree of $G/\{e\}$. Note that $G/\{e\}$ has one less vertex and one less edge than $G$, but except for $e$, every edge of $G$ corresponds to a unique edge of $G/\{e\}$, and *vice–versa*. If $E(T)$ denotes the edges of $G$ corresponding to a spanning tree then $E(T)-\{e\}$ corresponds to $T'$, a set of edges in $G/\{e\}$. Since the subgraph corresponding to $T'$ is connected and has the right number of edges, $T'$ is a spanning tree of $G/\{e\}$. The correspondence also works in reverse. End of Proof.

This property of $\kappa$ is reminiscent of the multilinearity of the determinant.

**Proposition** 16.8. Elementary row operations do not change the determinant. That is, for $a \in \mathbb{R}$,

$$\det(b_1; \ldots; b_{i-1}; b_i; b_{i+1}; \ldots; b_n) = \det(b_1; \ldots; b_{i-1}; b_i + ab_j; b_{i+1}; \ldots; b_n) \tag{16.7}$$

The determinant is multilinear in the sense that

$$\det(b_1; \ldots; b_{i-1}; ab_i; b_{i+1}; \ldots; b_n) = a \det(b_1; \ldots; b_{i-1}; b_i; b_{i+1}; \ldots; b_n) \tag{16.8}$$

and

$$\begin{aligned} \det(b_1; \ldots; b_{i-1}; v + w; b_{i+1}; \ldots; b_n) = \det(b_1; \ldots; b_{i-1}; v; b_{i+1}; \ldots; b_n) \\ + \det(b_1; \ldots; b_{i-1}; w; b_{i+1}; \ldots; b_n). \end{aligned} \tag{16.9}$$

Proof: Eq. (16.7) follows from the product formula, Eq. (3.21), $\det(EB) = \det(E)\det(B)$ where $E$ is the elimination matrix comprised of the identity matrix with the addition of $a$ in row $j$ column $i$. As $E$ is unit–triangular we find $\det(E) = 1$ and conclude Eq. (16.7).

Our second claim also follows from the product formula, $\det(AB) = \det(A)\det(B)$ where $A$ is the identity matrix except for $A_{ii} = a$.

Regarding the third claim, if the $n-1$ rows $(b_1; \ldots; b_{i-1}; b_{i+1}; \ldots; b_n)$ are linearly dependent then all three determinants in (16.9) are zero and so equality holds trivially. If they are instead linearly independent that we can complete them to a basis for $\mathbb{R}^n$ with the vector $b_i$. We then expand $v$ and $w$ in this basis as

$$v = \sum_{j=1}^{n} v_j b_j \quad \text{and} \quad w = \sum_{j=1}^{n} w_j b_j.$$

Now it follows from Eq. (16.7) that

$$\det(b_1; \ldots; b_{i-1}; v + w; b_{i+1}; \ldots; b_n) = \det(b_1; \ldots; b_{i-1}; v + w - (v_1 + w_1)b_1; b_{i+1}; \ldots; b_n).$$

Applying this reasoning to the remaining rows brings

$$\begin{aligned} \det(b_1; \ldots; b_{i-1}; v + w; b_{i+1}; \ldots; b_n) &= \det(b_1; \ldots; b_{i-1}; (v_i + w_i)b_i; b_{i+1}; \ldots; b_n) \\ &= (v_i + w_i)\det(b_1; \ldots; b_{i-1}; b_i; b_{i+1}; \ldots; b_n) \end{aligned} \tag{16.10}$$

where the second equality follows from Eq. (16.8). Now rebuild $v$ and $w$ individually using Eq. (16.7) again. More precisely

$$\begin{aligned} v_i \det(b_1; \ldots; b_{i-1}; b_i; b_{i+1}; \ldots; b_n) &= \det(b_1; \ldots; b_{i-1}; v_i b_i; b_{i+1}; \ldots; b_n) \\ &= \det(b_1; \ldots; b_{i-1}; v_i b_i + v_1 b_1; b_{i+1}; \ldots; b_n) \end{aligned}$$

and on continuation we find

$$v_i \det(b_1; \ldots; b_{i-1}; b_i; b_{i+1}; \ldots; b_n) = \det(b_1; \ldots; b_{i-1}; v; b_{i+1}; \ldots; b_n).$$

as the same logic applies to $w$ we deduce Eq. (16.9) from Eq. (16.10). End of Proof.

**Proposition** 16.9. **Kirchhoff Tree Theorem.** The number of spanning trees in a graph is the determinant of the reduced Laplacian. That is $\kappa(G) = \det(L_0(G))$.

We will see that the determinant of the matrix of distances depends solely on $n$. For this we need one more take on the determinant.

We note that if $\tilde{I}$ is an elementary perturbation matrix then $B\tilde{I}$ simply swaps two **columns** of $B$ and $\det(B\tilde{I}) = -\det(B)$. This leads to the nice

**Proposition** 16.10. **Cofactor Expansion.** For $B \in \mathbb{R}^{n\times n}$ we denote its elements by $b_{i,j}$ and denote the matrix achieved by ignoring row $i$ and column $j$ by $B(-i,-j)$. For any row index, $i$, we may expand the $\det(B)$ along this row as

$$\det(B) = \sum_{j=1}^{n} b_{i,j}(-1)^{i+j} \det(B(-i,-j)). \tag{16.11}$$

Proof: To minimize notation let us set (for now) $i = 1$. By multilinearity we can write

$$\det(B) = \sum_{j=1}^{n} b_{1,j} \det(e_j; b_2; \ldots; b_n).$$

The first det can be reduced, by row reduction to

$$\det \begin{pmatrix} 1 & 0_{n-1} \\ 0_{n-1} & B(2:n, 2:n) \end{pmatrix} = \det(B(2:n, 2:n)).$$

while the second det can be reduced, by row reduction to

$$\det \begin{pmatrix} 0 & 1 & 0(1,n-2) \\ B(2:n,1) & 0(n-1,1) & B(2:n,3:n) \end{pmatrix} = -\det(B(-1,-2))$$

after swapping columns. the third det can be reduced, by row reduction to

$$\det \begin{pmatrix} 0(1,2) & 1 & 0(1,n-3) \\ B(2:n,1:2) & 0(n-1,1) & B(2:n,4:n) \end{pmatrix} = \det(B(-1,-3))$$

after swapping columns twice. End of Proof.

We now deduce from these new results on determinants new formulations of the characteristic polynomials for trees. For the tree $\Gamma$ with adjacency matrix $T$ we denote the characteristic polynomial

$$\chi(\Gamma, z) \equiv \det(zI - T).$$

**Proposition** 16.11. Suppose that $\Gamma$ is a tree. If $v_1$ is an isolated vertex then

$$\chi(\Gamma, z) = z\chi(\Gamma_1, z) \tag{16.12}$$

where $\Gamma_1$ is the tree without $v_1$. If the vertex $v_1$ is adjacent only to vertex $v_2$ Then

$$\chi(\Gamma, z) = z\chi(\Gamma_1, z) - \chi(\Gamma_{12}, z) \tag{16.13}$$

where $\Gamma_{12}$ is the graph without $v_1$ and $v_2$.

Proof: We expand by rows

$$\begin{aligned} \chi(\Gamma, z) &= \det(ze_1 - t_1; ze_2 - t_2; \cdots; ze_n - t_n) \\ &= \det(ze_1; ze_2 - t_2; \cdots; ze_n - t_n) - \det(t_1; ze_2 - t_2; \cdots; ze_n - t_n) \\ &= z\chi(\Gamma_1, z) - \chi(\Gamma_{12}, z) \end{aligned}$$

because... End of Proof.

If $\Gamma = P_n$ is simply a path than this permits us to write

$$\chi(P_n, z) = z\chi(P_{n-1}, z) - \chi(P_{n-2}, z)$$

and hence $\chi(P_n, z) = U_n(z/2)$, the $n$th Chebyshev polynomial. As an example lets watch Figure 16.4

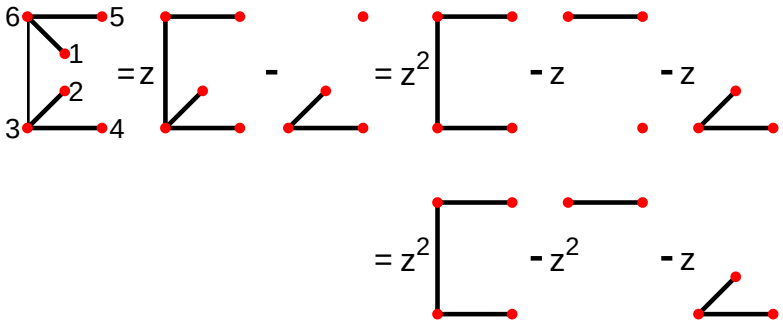



**Figure** 16.4 An illustration of the decomposition of a tree into 3 paths, that facilitate the construction of the full characteristic polynomial.

With reference to Figure 16.4 we find

$$\begin{aligned}\chi(\Gamma, z) &= z^2U_4(z/2) - z^2U_2(z/2) - zU_3(z/2)\\ &= z^2(z^4 - 3z^2 + 1) - z^2(z^2 - 1) - z(z^3 - 2z)\\ &= z^2(z^4 - 5z^2 + 4)\end{aligned}$$

has roots $0, 0, \pm 1, \pm 2$.

### 16.4. Cycles and Girth

A **cycle** of length $r$ starting at vertex $v_1$ is a sequence of $r$ vertices

$$\mathbf{c} = (v_1, v_2, \ldots, v_r)$$

where $v_i$ is adjacent to $v_{i+1}$ for $i = 1, \ldots, r-1$ and $v_r$ is adjacent to $v_1$ and backtracking is not allowed, i.e., $v_{i+1} \neq v_{i-1}$ for $i = 2, \ldots, r-1$ and $v_2 \neq v_r$.

**Trace theorem.** Define $A_r$ via $(A_r)_{i,j}$ is the number of paths of length $r$, without backtracking, from vertex $i$ to vertex $j$. Note that $A_0 = I$ and $A_1 = A$, the adjacency matrix. As an example lets

work with the Cayley graph in Figure 14.7(A). Its adjacency matrix is

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix} \tag{16.14}$$

And so lets record

$$A_2 = \begin{pmatrix} 0 & 0 & 2 & 2 & 1 & 1 \\ 0 & 0 & 1 & 1 & 2 & 2 \\ 2 & 1 & 0 & 1 & 2 & 0 \\ 2 & 1 & 1 & 0 & 0 & 2 \\ 1 & 2 & 2 & 0 & 0 & 1 \\ 1 & 2 & 0 & 2 & 1 & 0 \end{pmatrix} \quad \text{and} \quad A_3 = \begin{pmatrix} 2 & 2 & 3 & 3 & 1 & 1 \\ 2 & 2 & 1 & 1 & 3 & 3 \\ 3 & 1 & 2 & 1 & 3 & 2 \\ 3 & 1 & 1 & 2 & 2 & 3 \\ 1 & 3 & 3 & 2 & 2 & 1 \\ 1 & 3 & 2 & 3 & 1 & 2 \end{pmatrix}$$

**Proposition** 16.12. Suppose the graph is $k$ regular
a) $A_1^2 = A_2 + kI$,
b) for $r \geq 2$, $A_1 A_r = A_r A_1 = A_{r+1} + (k-1)A_{r-1}$.
c) for real $t$

$$\left( \sum_{r=0}^{\infty} t^r A_r \right) (I - tA + (k-1)t^2 I) = (1-t^2)I. \tag{16.15}$$

Proof: Note that

$$(A_1^2)_{i,j} = \sum_{m=1}^{n} (A_1)_{i,m} (A_1)_{m,j}$$

If a product is nonzero then $i$ and $m$ are adjacent and $m$ and $j$ are adjacent. If $i \neq j$ then no backtracking is possible and so $(A_1^2)_{i,j}$ is the number of paths of length 2 between $i$ and $j$. If $i = j$ then a path of length 2 means to step to adjacent vertex and then step back. As there are precisely $m$ vertices we find $(A_1^2)_{i,i} = m$. This proves (a).

Note that

$$(A_r A_1)_{i,j} = \sum_{m=1}^{n} (A_r)_{i,m} (A_1)_{m,j}$$

If a product is nonzero then there is a straight (no bt) path of length $r$ from $i$ to $m$ followed by a path of length one from $m$ and $j$. If the $r-1$st vertex on that path is not $j$ then the composite is a straight path of length $r+1$, and there are $(A_{r+1})_{i,j}$ such paths. Otherwise, backtracking occurred at the last step, and there are $(k-1)(A_{r-1})_{i,j}$ such paths.

For (c) we note that

$$\begin{aligned}
(I+tA_1+t^2A_2+\cdots+t^mA_m)(I-tA+(k-1)t^2I) &= (I+tA_1+t^2A_2+\cdots+t^mA_m)\\
&-(tA_1+t^2A_1^2+t^3A_1A_2+\cdots+t^{m+1}A_1A_m)+(k-1)(t^2I+t^3A_1+t^4A_2+\cdots+t^{m+2}A_m)\\
&= I+t^2(A_2-A_1^2)+t^3(A_3-A_1A_2)+\cdots+t^m(A_m-A_1A_{m-1})-t^{m+1}A_1A_m\\
&\quad+(k-1)(t^2I+t^3A_1+t^4A_2+\cdots+t^{m+2}A_m)\\
&= I-t^2kI-(k-1)(t^3A_1+t^4A_2+\cdots+t^mA_{m-2})-t^{m+1}A_1A_m\\
&\quad+(k-1)(t^2I+t^3A_1+t^4A_2+\cdots+t^{m+2}A_m)\\
&= (1-t^2)I-t^{m+1}A_1A_m+(k-1)t^{m+1}A_{m-1}+(k-1)t^{m+2}A_m\\
&= (1-t^2)I-t^{m+1}A_{m+1}+(k-1)t^{m+2}A_m.
\end{aligned}$$

On taking $m\to\infty$ we arrive at Eq. (16.15). End of Proof.

We can make it a bit cleaner with

$$T_m \equiv \sum_{r=0}^{m/2} A_{m-2r}.$$

For then

$$\left(\sum_{m=0}^{\infty} t^m\mathbf{T}_m\right)(I-tA+(k-1)t^2I) = I. \tag{16.16}$$

To see this

$$\begin{aligned}
\sum_{m=0}^{\infty} t^m\mathbf{T}_m &= \sum_{m=0}^{\infty}\sum_{r=0}^{m/2} A_{m-2r}t^m\\
&= \sum_{r=0}^{\infty}\sum_{m=2r}^{\infty} A_{m-2r}t^m\\
&= \sum_{r=0}^{\infty} t^{2r}\sum_{m=2r}^{\infty} A_{m-2r}t^{m-2r}\\
&= \left(\sum_{r=0}^{\infty} t^{2r}\right)\left(\sum_{m=0}^{\infty} A_m t^m\right)\\
&= \frac{1}{1-t^2}(1-t^2)I = I.
\end{aligned}$$

Lets record,

$$U_0(x)=1,\quad U_1(x)=2x,\quad U_2(x)=4x^2-1,\quad U_3(x)=8x^3-4x,\quad U_4(x)=16x^4-12x^2+1.$$

**Proposition** 16.13. The number, $c_\ell$, of cycles of length $\ell$ through a vertex is $c_0 = k,\ c_1 = 0$

$$2n(c_0 + c_2) = (k-1)\sum_{j=0}^{n-1} U_2\left(\frac{\mu_j}{2\sqrt{k-1}}\right)$$

$$2n(c_1 + c_3 = (k-1)^{3/2}\sum_{j=0}^{n-1} U_3\left(\frac{\mu_j}{2\sqrt{k-1}}\right)$$

$$2n(c_0 + c_2 + c_4) = (k-1)^2\sum_{j=0}^{n-1} U_4\left(\frac{\mu_j}{2\sqrt{k-1}}\right)$$

$$2n(c_1 + c_3 + c_5) = (k-1)^{5/2}\sum_{j=0}^{n-1} U_5\left(\frac{\mu_j}{2\sqrt{k-1}}\right)$$

$$2n(c_4 + c_6) = (k-1)^3\sum_{j=0}^{n-1} U_6\left(\frac{\mu_j}{2\sqrt{k-1}}\right)$$

Now

$$c_1 = 0 = \sum_{j=0}^{n-1} U_1\left(\frac{\mu_j}{2\sqrt{k-1}}\right) = \frac{1}{\sqrt{k-1}}\sum_{j=0}^{n-1}\mu_j$$

while

$$n = n(c_0 + c_2) = (k-1)\sum_{j=0}^{n-1} U_2\left(\frac{\mu_j}{2\sqrt{k-1}}\right) = (k-1)\sum_{j=0}^{n-1}\left(\frac{\mu_j^2}{k-1} - 1\right)$$

and so

$$\boxed{nk = \sum_{j=0}^{n-1}\mu_j^2} \tag{16.17}$$

allows us to read the degree from the eigenvalues. Next

$$nc_3 = n(c_1 + c_3) = \sum_{j=0}^{n-1}\mu_j^3$$

delivers $c_3$. Next

$$\begin{aligned} n(1 + c_4) = n(c_0 + c_2 + c_4) &= (k-1)^2\sum_{j=0}^{n-1} U_4\left(\frac{\mu_j}{2\sqrt{k-1}}\right) \\ &= \sum_{j=0}^{n-1}\mu_j^4 - 3(k-1)\sum_{j=0}^{n-1}\mu_j^2 + n(k-1)^2 \\ &= n(k-1)^2 - 3nk(k-1) + \sum_{j=0}^{n-1}\mu_j^4 \end{aligned}$$

delivers $c_4$.

It now follows that the spectra determines the cycle sequence, and so ask to what degree do the $C_j$ determine the graph. Starting with our example we see that (*** gives $c_3 = 2$ (which as orientation does not count really means $c_3 = 1$) so each vertex is a vertex of one triangle example, its eigenvalues are $3, 1, 0, 0, -2, -2$ and from these we easily deduce $f_3 = 2$ which in fact uniquely determines the graph! Next go to the truncated tetrahedron in Figure 14.4. We find its degree is 3. Then $f_3 = 2$ dictates that every vertex lies in exactly one triangle. So we find 4 triangles. To reach degree each triangle must have at least one edge to another triangle. If two triangles are however joined by two edges then a vertex would have cycle length 4, but $f_4 = 0$ so each triangle has exactly one edge to each triangle, and so we can have only Figure 14.4. With regard to the Buckyball in Figure 14.6 we find $k = 3$ while $f_3 = f_4 = 0$, so no triangles or squares, while $f_5 = 1$ so every vertex is in 1 pentagon so there are $12 = 60/5$ pentagons. The spectrum also reveals that $f_7 = f_8 = 0$.Now if two pentagons were connected by more than one edge we would contradict either $f_4 = f_7 = f_80$ or $f_5 = 1$. And so we get Figure 14.6.

The **girth** of a graph is the length of its shortest cycle.

**Proposition** 16.14. For odd prime $q$ the graph $X_q$ has large girth. In particular

$$\liminf_{q\to\infty} \frac{\text{girth}(X_q)}{\log_3 |X_q|} \geq \frac{1}{3\log_3(1+\sqrt{2})}.$$

Proof: Write $g$ for $\text{girth}(X_q)$. As $X_q$ is vertex transitive it contains a cycle of length $g$ starting and ending at $I \in SL_2(q)$:

$$x_0 = I, x_1, \ldots, x_{g-1}, x_g = I.$$

By the Cayley construction there exist $y_1, y_2, \ldots, y_g \in S_q$ such that $x_i = y_1 y_2 \cdots y_{i-1} y_i$. Let $\tilde{y}_i$ be the unique element in $S$ for which $\tau_q(\tilde{y}_i) = y_i$. Now $\tilde{y}_1\tilde{y}_2\cdots\tilde{y}_g$ is an element of $H$, the group generated by $S$. As we proved that $H$ is free it follows that $\tilde{y}_1\tilde{y}_2\cdots\tilde{y}_g \neq I$. On the other hand, since $\tau_q(\tilde{y}_1\tilde{y}_2\cdots\tilde{y}_g) = y_1y_2\cdots y_g = x_g = I$ it follows that each element of $\tilde{y}_1\tilde{y}_2\cdots\tilde{y}_g - I$ is divisible by $q$ and hence $\|\tilde{y}_1\tilde{y}_2\cdots\tilde{y}_g - I\| \geq q$. End of Proof.

### 16.5. The Isoperimetric Constant and Expanders

An increasing sequence of graphs is said to be an expanding sequence if the number of edges connecting a set to its complement increases in proportion to the size of the smaller set. To make this precise, suppose that $\Gamma = (V, E)$ is a graph. If $U$ is a subset of the vertex set, $V$, we define its boundary

$$\partial U = \{\{v_1, v_2\} \in E : v_1 \in U,\ v_2 \in V \setminus U\}$$

to be those edges that connect $U$ to $V \setminus U$. With this we may define the **expander constant** of $\Gamma$

$$h(\Gamma) \equiv \min_{U\subset V} \frac{|\partial U|}{\min\{|U|, |V\setminus U|\}}. \tag{16.18}$$

A family of graphs, $\{\Gamma_n = (V_n, E_n)\}$, is said to be an **expander family** when $|V_n| \to \infty$ and there exists a $\delta > 0$ such that

$$\liminf_{n\to\infty} h(\Gamma_n) \geq \delta.$$

Examples: The family of torii $T_n$ is not expanding, choose $U$ to be a natural half, so $|U| = n^2/2$ while $|\partial U| = 2n$ so $h(T_n) \leq 4/n$. We will see that a particular scheme for doubling the degree

at each node does indeed produce an expanding family. The proof of expansion will hinge on the connection between the expander constant and the spectral gap of the adjacency matrix.

**Proposition** 16.15. Suppose that $\Gamma$ is a connected regular graph of degree $k$ with adjacency matrix $A_\Gamma$.
(i) $k$ is a simple eigenvalue of $A_\Gamma$.
(ii) Each eigenvalue of $A_\Gamma$ has magnitude less that $k$.

Proof: We number the nodes from 1 to $n$ and use $i \sim j$ to denote that node $i$ is adjacent to node $j$. As $\Gamma$ is regular of degree $k$ it follows that every row sum of $A_\Gamma$ equals $k$. In other words, the vector of ones is an eigenvector of $A_\Gamma$ with eigenvalue $k$. To see that $k$ is simple we suppose $Ax = kx$ and $\|x\| = 1$ and that $|x_j| \geq |x_i|$ for $i = 1 : n$. Now $(Ax)_j = kx_j$ means

$$\sum_{i \sim j} x_i = kx_j$$

which implies that $x_i = x_j$ for all $i \sim j$. By the same argument it follows that all elements adjacent to each of these $i$ is also $x_j$. As $\Gamma$ is connected it follows that all $x_i = x_j$.

Regarding part (ii), suppose $Ax = zx$ and that, as above, $x_j$ is an element of $x$ of maximum magnitude. We suppose, without loss, that $x_j > 0$ and from $(Ax)_j = zx_j$ deduce

$$|z|x_j = |zx_j| = |(Ax)_j| = \left| \sum_{i \sim j} A_{i,j} x_i \right| \leq \sum_{i \sim j} A_{i,j} |x_i| \leq x_j \sum_{i \sim j} A_{i,j} \leq k x_j,$$

and conclude that $|z| \leq k$. End of Proof.

It follows that $k$ is the largest eigenvalue of $\Gamma$ and so its next largest eigenvalue is

$$\lambda(\Gamma) = \max_{x \perp \mathbb{1}} \frac{x^T A_\Gamma x}{x^T x}. \tag{16.19}$$

Its often more convenient to study the associated graph Laplacian

$$\Delta_\Gamma \equiv kI - A_\Gamma.$$

Because

$$k - \lambda(\Gamma) = \mu(\Gamma) = \min_{x \perp \mathbb{1}} \frac{x^T \Delta_\Gamma x}{x^T x} \tag{16.20}$$

and

$$x^T \Delta_\Gamma x = x^T (kI - A_\Gamma) x = \sum_i x_i \left( k x_i - \sum_{j \sim i} x_j \right) = \sum_{j \sim i} (x_j - x_i)^2.$$

The spectral connection is

**Proposition** 16.16. If $\Gamma$ is a connected regular graph of degree $k$ then

$$h(\Gamma) \geq \frac{k - \lambda(\Gamma)}{2}. \tag{16.21}$$

Proof: We partition the vertex set as $V = U \cup (V \setminus U)$ and build the Rayleigh candidate for (16.20)

$$x_j = \begin{cases} \frac{1}{|U|} & \text{if} \quad j \in U \\ \frac{-1}{|V \setminus U|} & \text{if} \quad j \notin U. \end{cases}$$

We note that $x^T \mathbb{1} = 0$ and

$$x^T x = \frac{1}{|U|} + \frac{1}{|V \setminus U|} \quad \text{and} \quad x^T (kI - A_\Gamma) x = |\partial U| \left( \frac{1}{|U|} + \frac{1}{|V \setminus U|} \right)^2.$$

And so

$$\mu(\Gamma) \leq \frac{x^T (kI - A_\Gamma) x}{x^T x} = |\partial U| \left( \frac{1}{|U|} + \frac{1}{|V \setminus U|} \right) \leq \frac{2|\partial U|}{\min\{|U|, |V \setminus U|\}}$$

for each choice of $U$. Taking the minimum over $U \subset V$ we find $\mu(\Gamma) \leq 2h(\Gamma)$. End of Proof.

The **Margulis Construction**. $\Gamma_n = (V, E)$ with vertex set $V = \mathbb{Z}_n^2$. Let

$$T_1 = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, \quad T_2 = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}, \quad e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{16.22}$$

and each vertex $v \in V$ is adjacent to the four vertices $T_1 v$, $T_2 v$, $T_1 v + e_1$ and $T_2 v + e_2$ and their four inverses. We illustrate it in Figure 16.5 when $n = 3$.

We will make use of Fourier Analysis on $\mathbb{Z}_n^2$. So we recall the characters

$$\chi_\zeta(z) = \exp(2\pi i \zeta^T z / n), \quad z = (z_1, z_2) \in \mathbb{Z}_n^2, \quad \zeta = (\zeta_1, \zeta_2) \in \mathbb{Z}_n^2.$$

We will invoke Parseval's Theorem and so must also understand how to transform the Margulis compositions $u(T_i z + e_i)$. In particular we will need

**Proposition** 16.17. Suppose that $A \in \mathrm{GL}_2(n)$, $b \in \mathbb{Z}_n^2$. If $u \in \mathbb{C}[\mathbb{Z}_n^2]$ and $v(z) = u(Az + b)$ then

$$\hat{v}(\zeta) = \exp(2\pi i \zeta^T A^{-1} b / n) \hat{u}(A^{-T} \zeta)$$

Proof: We first record how characters transform

$$\chi_\zeta(A^{-1} z) = \exp(2\pi i \zeta^T A^{-1} z / n) = \exp(2\pi i (A^{-T} \zeta)^T z / n) = \chi_{A^{-T}\zeta}(z) \tag{16.23}$$

and then carry this through to $\hat{v}$,

$$\hat{v}(\zeta) = \langle v, \chi_\zeta \rangle = \sum_{z \in \mathbb{Z}_n^2} v(z) \overline{\chi_\zeta(z)} = \sum_{z \in \mathbb{Z}_n^2} u(Az + b) \overline{\chi_\zeta(z)}.$$

As $z \mapsto Az + b$ is a bijection of $\mathbb{Z}_n^2$ we may change variables, $z' = Az + b$, to arrive, via (16.23), at

$$\begin{aligned} \hat{v}(\zeta) &= \sum_{z' \in \mathbb{Z}_n^2} u(z') \overline{\chi_\zeta(A^{-1}(z' - b))} \\ &= \overline{\chi_\zeta(-A^{-1}b)} \sum_{z' \in \mathbb{Z}_n^2} u(z') \overline{\chi_\zeta(A^{-1} z')} \\ &= \overline{\chi_\zeta(-A^{-1}b)} \sum_{z' \in \mathbb{Z}_n^2} u(z') \overline{\chi_{A^{-T}\zeta}(z')} \\ &= \exp(2\pi i \zeta^T A^{-1} b / n) \hat{u}(A^{-T} \zeta). \end{aligned}$$

End of Proof.

Whereas the girth story hinged on the action of $T_1$ and $T_2$ on the bow-tie, here it will hinge on their action on the diamond.

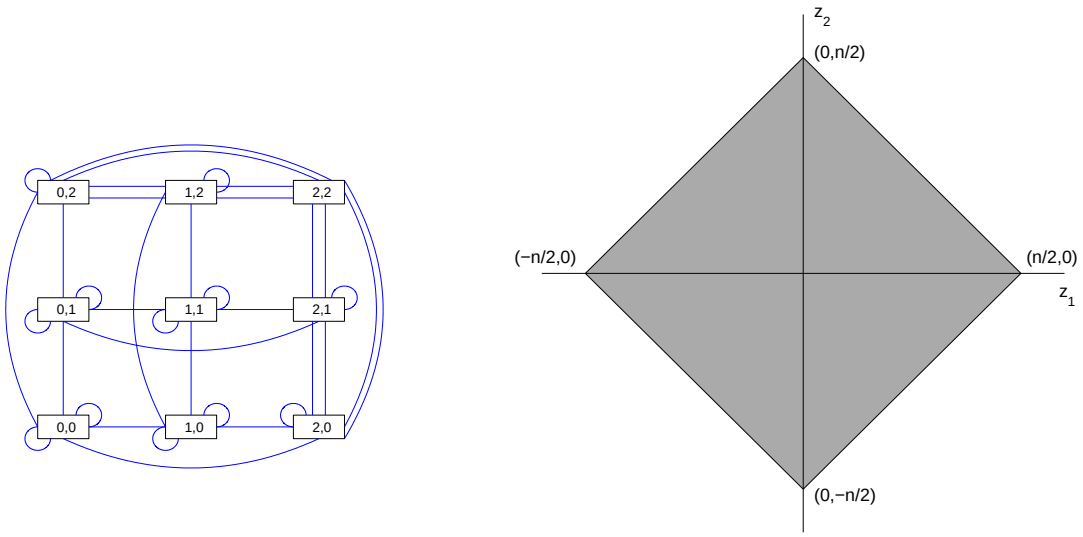



**Figure** 16.5. The Margulis Construction. (Left) $\Gamma_3$. (Right) The diamond, $D_n$.

The proof that $\lambda(\Gamma_n)$ stays clear of $k = 8$ will follow from Rayleigh's Principle and Parseval's formula and a careful study of how $T_1$ and $T_2$ and their inverses deform the diamond in Figure 16.5. We examined this action in Exer. 1.5. In particular, in Figure 1.6 we observed that $T_1^{\pm 1}$ shear right and left while $T_2^{\pm 1}$ shear up and down. To better quantify this action we seek a means for comparing a point $z$ to its image under $T_j^{\pm 1}$. This leads us to seek means for comparing points in $\mathbb{Z}_n^2$. We write $z \in \mathbb{Z}_n^2$ as $z = (z_1, z_2)$ and assume that each $z_j \in [-n/2, n/2)$ and note the natural partial ordering

$$(z_1, z_2) \succ (z_1', z_2') \quad \text{if } |z_1| \geq |z_1'|,\ |z_2| \geq |z_2'| \text{ and } |z_1| + |z_2| > |z_1'| + |z_2'|, \tag{16.24}$$

suffices, in the sense that it permits us to order the action of $T_j^{\pm 1}$ on the diamond

$$D_n \equiv \{z \in \mathbb{Z}_n^2 : 2|z_1| + 2|z_2| < n\},$$

depicted in Figure 16.5(B).

**Proposition** 16.18. For each $z \in D_n \setminus 0$ either:
(a) Three of the four points, $T_j^{\pm 1} z$, are $\succ z$ and one is $\prec z$ or
(b) Two of the four points, $T_j^{\pm 1} z$, are $\succ z$ and two are incomparable with $z$.

Proof: We show that case (b) pertains to the diagonals, $|z_1| = |z_2|$, and the axes, $z_1 = 0$ and $z_2 = 0$. To begin, if $z_1 = 0$ then $T_2^{\pm 1} z = z$ and so $z$ is comparable to neither $T_2 z$ nor $T_2^{-1} z$. As

$$T_1^{\pm 1} \begin{pmatrix} 0 \\ z_2 \end{pmatrix} = \begin{pmatrix} \pm 2 z_2 \mod n \\ z_2 \end{pmatrix},$$

to show that $T_1^{\pm 1} z \succ z$ we need only confirm that $|\pm 2 z_2 \mod n| > 0$. The latter follows from $0 < |z_2| < n/2$. The other axis is handled in just the same way.

Now on the diagonal $z_1 = z_2 = a \neq 0$ we find

$$T_1^{\pm 1}\begin{pmatrix} a \\ a \end{pmatrix} = \begin{pmatrix} (1 \pm 2)a \bmod n \\ a \end{pmatrix} \quad \text{and} \quad T_2^{\pm 1}\begin{pmatrix} a \\ a \end{pmatrix} = \begin{pmatrix} a \\ (1 \pm 2)a \bmod n \end{pmatrix}.$$

From these we see that $z$ is not comparable to $T_i^{-1}z$ while $T_i z \succ z$ if $|3a \bmod n| > |a|$. To confirm $|3a \bmod n| > |a|$ we note that $(a, a) \in D_n$ iff $n \geq 4a+1$. Now, if $3a \leq n/2$ then $3a \bmod n = 3a$ and cleary $|3a| > |a|$. Conversely, if $3a > n/2$ then $3a \bmod n = 3a - n$. Now, from $n \geq 4a + 1$ comes $n - 3a > a$ and so $|3a - n| > |a|$.

Similarly, on the antidiagonal, $z_1 = a$, $z_2 = -a \neq 0$ we find

$$T_1^{\pm 1}\begin{pmatrix} a \\ -a \end{pmatrix} = \begin{pmatrix} (1 \mp 2)a \bmod n \\ -a \end{pmatrix} \quad \text{and} \quad T_2^{\pm 1}\begin{pmatrix} a \\ -a \end{pmatrix} = \begin{pmatrix} a \\ (2 \mp 1)a \bmod n \end{pmatrix}.$$

From these we see that $z$ is not comparable to $T_i z$ while $T_i^{-1}z \succ z$ if $|3a \bmod n| > |a|$. As above, this inequality holds for all $(a, a) \in D_n$.

We now leave the axes and diagonals and focus on the triangle $t_n \equiv \{z \in D_n : z_1 > z_2 > 0\}$. Now $T_1^{-1}z \prec z$ is equivalent to

$$|z_1 - 2z_2 \bmod n| < |z_1|, \quad z \in t_n.$$

We note that $z_1 - z_2 = j > 0$ and find

$$|z_1 - 2z_2| \leq |z_1 - z_2| + |z_2| = j + z_1 - j = z_1 = |z_1|,$$

with equality iff $z_2 = 0$. As $z_2 > 0$ and $|z_1| < n/2$ we have shown that $|z_1 - 2z_2 \bmod n| = |z_1 - 2z_2| < |z_1|$, i.e.,$T_1^{-1}z \prec z$ for $z \in t_n$.

Next, we note that $T_1 z \succ z$ is equivalent to $|z_1 + 2z_2 \bmod n| > |z_1|$. If $z_1 + 2z_2 \leq n/2$ then $(z_1 + 2z_2) \bmod n = z_1 + 2z_2 > z_1$ as $z_2 > 0$. Conversely, if $z_1 + 2z_2 > n/2$ then $(z_1 + 2z_2) \bmod n = z_1 + 2z_2 - n$. As $z \in t_n$ it follows that $n > 2z_1 + 2z_2$ and hence $n - z_1 - 2z_2 > z_1$.

Next, $T_2 z \succ z$ is equivalent to $|2z_1 + z_2 \bmod n| > |z_2|$. If $2z_1 + z_2 \leq n/2$ then $(2z_1 + z_2) \bmod n = 2z_1 + z_2 > z_2$ as $z_1 > 0$. Conversely, if $2z_1 + z_2 > n/2$ then $(2z_1 + z_2) \bmod n = 2z_1 + z_2 - n$. From $n > 2z_1 + 2z_2$ we deduce the required $n - 2z_1 - z_2 > z_2$.

Finally, $T_2^{-1}z \succ z$ is equivalent to $|z_2 - 2z_1 \bmod n| > |z_2|$. If $z_2 - 2z_1 \geq -n/2$ then $(z_2 - 2z_1) \bmod n = z_2 - 2z_1$. Now from $z_1 > z_2$ we deduce $2z_1 > 2z_2$ and $2z_1 - z_2 > z_2$ and finally $|z_2 - 2z_1| > |z_2|$. Conversely, if $z_2 - 2z_1 < -n/2$ then $(z_2 - 2z_1) \bmod n = n + z_2 - 2z_1$. Now $n + z_2 - 2z_1 > z_2$ follows from $n > 2z_1$. End of Proof.

With this we now have all we need to prove that the Margulis Construction is expanding.

**Proposition** 16.19. $\lambda(\Gamma_n) \leq 7.3$ for every positive integer $n$.

Proof: We begin with the numerator of the Rayleigh quotient

$$\langle A_{\Gamma_n} u, u\rangle = 2\sum_{z \in \mathbb{Z}_n^2} u(z)\{u(T_1 z) + u(T_1 z + e_1) + u(T_2 z) + u(T_2 z + e_2)\}. \tag{16.25}$$

for $u \perp \mathbb{1}$. By Parseval's Theorem, Prop. 15.14, and Prop. 16.17 we may write the first half of (16.25) as

$$\sum_{z \in \mathbb{Z}_n^2} u(z)\{u(T_1 z) + u(T_1 z + e_1)\} = \frac{1}{n^2}\sum_{z \in \mathbb{Z}_n^2} \overline{\hat{u}(z)}\hat{u}(T_2^{-1}z)(1 + \omega_n^{z_1}),$$

where $\omega_n = \exp(2\pi i/n)$ and we have used $T_1^{-T} = T_2^{-1}$ and $T_1^{-1}e_1 = e_1$. Applying this same argument to the remaining two terms of (16.25) and noting $|1 + \omega_n^k| = 2|\cos(\pi k/n)|$ we find

$$n^2\langle A_{\Gamma_n}u, u\rangle \le \sum_{z\in\mathbb{Z}_n^2} 4|\hat{u}(z)|\{|\hat{u}(T_2^{-1}z)||\cos(z_1\pi/n)| + |\hat{u}(T_1^{-1}z)||\cos(z_2\pi/n)|\}. \tag{16.26}$$

We also note that $u \perp \mathbb{1}$ translates into

$$\hat{u}(0) = \langle u, \chi_0\rangle = u^T\mathbb{1} = 0. \tag{16.27}$$

Regarding the denominator of the Rayleigh quotient we invoke Parseval's Theorem again to arrive at

$$n^2\langle u, u\rangle = \langle \hat{u}, \hat{u}\rangle = \sum_{z\in\mathbb{Z}_n^2} |\hat{u}(z)|^2. \tag{16.28}$$

In contrasting (16.26) and (16.28) one is led to the hope that we may perhaps factor a term like $|\hat{u}(z)|^2$ out of the former. This hope is realized with the help of the elementary inequality

$$2ab \le a^2c + b^2/c, \tag{16.29}$$

that holds for any nonnegative $a$, $b$, and $c$. This inequality permits great flexibility in the choice of $c$ and, given the complexity of (16.26), we will choose $c$ to in fact vary with $z$ and its action under $T_1^{-1}$ and $T_2^{-1}$ in a manor that exploits the comparisons established in Prop. 16.18. In particular, we will see that

$$c(z, z') = \begin{cases} 5/4 & \text{if } z \succ z', \\ 4/5 & \text{if } z \prec z', \\ 1 & \text{otherwise,} \end{cases}$$

suffices for our purposes. To see this we note that $c(z, z')c(z', z) = 1$ for every $z, z' \in \mathbb{Z}_n^2$ permits us to express (16.29) as

$$2|\hat{u}(z)||\hat{u}(z')| \le c(z, z')|\hat{u}(z)|^2 + c(z', z)|\hat{u}(z')|^2.$$

Using this in (16.26) then brings

$$\begin{aligned} n^2\langle A_{\Gamma_n}u, u\rangle \le & 2\sum_{z\in\mathbb{Z}_n^2} |\cos(\pi z_1/n)|\{c(z, T_2^{-1}z)|\hat{u}(z)|^2 + c(T_2^{-1}z, z)|\hat{u}(T_2^{-1}z)|^2\} \\ & + 2\sum_{z\in\mathbb{Z}_n^2} |\cos(\pi z_2/n)|\{c(z, T_1^{-1}z)|\hat{u}(z)|^2 + c(T_1^{-1}z, z)|\hat{u}(T_1^{-1}z)|^2\}. \end{aligned} \tag{16.30}$$

We can indeed factor out a common $|\hat{u}(z)|^2$ in the first sum on noting that the change of variable $y = T_2^{-1}z$ leaves $y_1 = z_1$. In particular, this allows us to conclude that

$$\sum_{z\in\mathbb{Z}_n^2} |\cos(\pi z_1/n)|c(T_2^{-1}z, z)|\hat{u}(T_2^{-1}z)|^2 = \sum_{z\in\mathbb{Z}_n^2} |\cos(\pi z_1/n)|c(z, T_2z)|\hat{u}(z)|^2.$$

The remaining sum in (16.30) is handled by noting $y = T_1^{-1}z$ leaves $y_2 = z_2$. Hence, from (16.30) we arrive at

$$\begin{aligned} n^2\langle A_{\Gamma_n}u, u\rangle \le & 2\sum_{z\in\mathbb{Z}_n^2} |\hat{u}(z)|^2|\cos(\pi z_1/n)|\{c(z, T_2z) + c(z, T_2^{-1}z)\} \\ & + 2\sum_{z\in\mathbb{Z}_n^2} |\hat{u}(z)|^2|\cos(\pi z_2/n)|\{c(z, T_1z) + c(z, T_1^{-1}z)\}. \end{aligned} \tag{16.31}$$

Contrasting this with (16.28), and using (16.27), our full result will follow from the pointwise bound

$$|\cos(\pi z_1/n)|\{c(z,T_2z)+c(z,T_2^{-1}z)\}+|\cos(\pi z_2/n)|\{c(z,T_1z)+c(z,T_1^{-1}z)\}\le\frac{73}{20} \tag{16.32}$$

for every nonzero $z\in\mathbb{Z}_n^2$.

We split the proof (16.32) into two domains. For $z$ outside $D_n$ we overestimate all the $c$ terms by $5/4$ and verify

$$|\cos(\pi z_1/n)|+|\cos(\pi z_2/n)|\le\sqrt{2}\le\frac{73}{50},$$

which implies the necessary inequality for such $z$. We suppose that $z_1\ge 0$ and $z_2\ge 0$. The other cases follow similarly. Since $z_2\mapsto\cos(\pi z_2/n)$ is decreasing and since we are outside $D_n$, this expression is maximized on the boundary, $z_2=n/2-z_1$, where $\cos(\pi z_2/n)=\sin(\pi z_1/n)$. Hence,

$$\cos(\pi z_1/n)+\cos(\pi z_2/n)=\cos(\pi z_1/n)+\sin(\pi z_1/n)\le\sqrt{2},$$

as needed. Conversely, when $z\in D_n$ we bound eash cosine by 1 and prove

$$c(z,T_2z)+c(z,T_2^{-1}z)+c(z,T_1z)+c(z,T_1^{-1}z)\le\frac{73}{20} \tag{16.33}$$

Ths will follow from the previous proposition. In case (a), the left-hand side of Eq. (16.33) is $3/(5/4)+5/4=73/20$, while in case (b) it is $2/(5/4)+2=72/20$. End of Proof.

### 16.6. Notes and Exercises

For more see Biggs (1994) and Harary (1969). Results on girth and expansion are due to Margulis. For girth we follow the exposition of G. Davidoff and Valette (2003). For expanders we follow the exposition of Hoory and Wigderson (2006).

1. Show that these two trees are isospectral but not isomorphic.

2. Suppose that $T=(V,E)$ is a tree, that $\deg_T(v)$ is the degree of vertex $v$ and that $d_T(u,v)$ is the number of edges traversed on the unique path from $u$ to $v$. Prove that the Wiener Polarity can be written

$$W_P(T)=|\{\{u,v\}:d_T(u,v)=3\}|=\sum_{uv\in E}(\deg_T(u)-1)(\deg_T(v)-1). \tag{16.34}$$

   Hint: $d_T(u,v)=3$ iff $\exists$ $x$ and $y$ in $V$ such that $ux$ and $vy$ lie in $E$. The number of such points is precisely that on the right side of (16.34).

3. (a) Show that

$$\sum_{v\in V}\deg(v)^2=\sum_{uv\in E}\deg(u)+\deg(v).$$

   (b) Use part (a) and the previous exercise to show that

$$W_P(T)=Z_2(T)-Z_1(T)+|E|$$

   where $Z_1$ and $Z_2$ are the first and second Zagreb indices

$$Z_1(T)=\sum_{v\in V}\deg(v)^2\quad\text{and}\quad Z_2(T)=\sum_{uv\in E}\deg(u)\deg(v).$$

(c) For trees with a common degree sequence it follows that to maximize $W_P$ is to maximize $Z_2$. Next show that

$$d^T Ad = 2Z_2(A).$$

so we should choose the $A$ so that $Ad$ is monotone nondecreasing. For example, given $d = (3, 3, 2, 1, 1, 1, 1)$ note that both trees have this sequence.

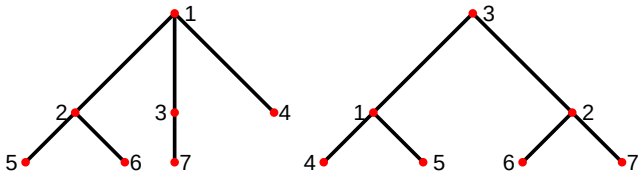



**Figure** 16.6 Two graphs with degree sequence $d = (3, 3, 2, 1, 1, 1, 1)$.

Compute the respective $Z_2$ and the respective adjacency matrices and argue in terms of "similar ordering" why the second graph has a larger Wiener polarity.

(d) Given a degree sequence $d$ argue that there exists an adjacency matrix $\hat{A}$ such that $\hat{A}d$ is nonincreasing and then prove that $Z_2(A) \leq Z_2(\hat{A})$ for any $A$ with the same degree sequence.

4. Show that if $T$ is a tree on $n$ vertices with distance matrix $D_n$ then

$$\det(D_n) = (-1)^{n-1}(n-1)2^{n-2}. \tag{16.35}$$

Proof: If node $n$ is a leaf then

$$D_n = \begin{pmatrix} 0 & d_{1,2} & \cdots & d_{1,n-1} & 1+d_{1,n-1} \\ d_{1,2} & 0 & \cdots & d_{2,n-1} & 1+d_{2,n-1} \\ \vdots & \vdots & & \vdots & \vdots \\ d_{1,n-1} & d_{2,n-1} & \cdots & 0 & 1 \\ 1+d_{1,n-1} & 1+d_{2,n-1} & \cdots & 1 & 0 \end{pmatrix}$$

now subtracting column $n-1$ from column $n$ and subtracting row $n-1$ from row $n$ brings

$$D'_n = \begin{pmatrix} 0 & d_{1,2} & \cdots & d_{1,n-1} & 1 \\ d_{1,2} & 0 & \cdots & d_{2,n-1} & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ d_{1,n-1} & d_{2,n-1} & \cdots & 0 & 1 \\ 1 & 1 & \cdots & 1 & -2 \end{pmatrix}$$

We now imagine pruning the $n$th vertex from $\Gamma_n$ and so arriving at the smaller tree $\Gamma_{n-1}$. The distance matrix for $\Gamma_{n-1}$ is precisely the upper $(n-1)$-by-$(n-1)$ block of $D'_n$. Returning to our previous argument, if node $n-1$ is a leaf of $\Gamma_{n-1}$ then we may reduce $D'_n$ to

$$D''_n = \begin{pmatrix} 0 & d_{1,2} & \cdots & d_{1,n-2} & 1 & 1 \\ d_{1,2} & 0 & \cdots & d_{2,n-2} & 1 & 1 \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ d_{1,n-2} & d_{2,n-2} & \cdots & 0 & 1 & 1 \\ 1 & 1 & \cdots & 1 & -2 & 0 \\ 1 & 1 & \cdots & 1 & 0 & -2 \end{pmatrix}$$

Continuing in this fashion brings

$$D^*_n = \begin{pmatrix} 0 & 1 & \cdots & 1 & 1 \\ 1 & -2 & 0 & \cdots & 0 \\ 1 & 0 & -2 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 & -2 \end{pmatrix}$$

Expanding along the last column then reveals

$$\det(D_n) = (-1)^{n-1}2^{n-2} - 2\det(D_{n-1}), \quad \det(D_1) = 0, \quad \det(D_2) = -1$$

from which we deduce (16.35). End of Proof.

5. Use the Power Method to prove that the random walk converges to the uniform distribution.

# 17. References


H Akaike. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21(1):243–247, 1969.

GS Ammar and WB Gragg. The generalized schur algorithm for the superfast solution of toeplitz systems. In *Rational Approximation and its Applications in Mathematics and Physics*, Lecture Notes in Mathematics Volume 1237, pages 315–330. Springer, 1987.

L Babai. Spectra of cayley graphs. *Journal of Combinatorial Theory*, pages 180–189, 1979.

R Bellman. *Introduction to Matrix Analysis*. SIAM, 1970.

A Ben-Tal and A Nemirovski. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*. SIAM, 2001.

N Biggs. *Algebraic Graph Theory*. Cambridge University Press, 1994.

DR Brillinger. *Time Series: Data Analysis and Theory*. SIAM, 2001.

K Bryan and T Leise. The $25,000,000,000 eigenvector: The linear algebra behind google. *SIAM Review*, 48:569–581, 2006.

F Chung and S Sternberg. Mathematics and the buckyball. *American Scientist*, 81:56–71, 1993.

V Chvatal. *Linear Programming*. WH Freeman, 1983.

David Wright David Mumford, Carlone Series. *Indra's Pearls, The Vision of Felix Klein*. Cambridge University Press, 2002.

NJ Calkin R Girgensohn DR Luke DH Bailey, JM Borwein and V Moll. *Experimental Mathematics in Action*. AK Peters/CRC Press, 2007.

JL Doob. *Stochastic Processes*. Wiley-Interscience, 1990.

Lester R. Ford. *Automorphic Functions*. AMS Chelsea, 1957.

Komei Fukuda and Alain Prodon. Double description method revisited. In *Combinatorics and Computer Science*. Springer, 1996.

P. Sarnak G. Davidoff and A. Valette. *Elementary Number Theory, Group Theory and Ramanujan Graphs*. Cambridge University Press, 2003.

F Gabbiani and SJ Cox. *Mathematics for Neuroscientists*. Elsevier, 2010.

J GN Stephanopoulos, AA Aristidou and Nielsen. *Metabolic Engineering: Principles and Methodologies*. Academic Press, 1998.

Didier Gonze and Wassim Abou-Jaoud. The goodwin model: behind the hill function. *PLoS One*, 8(8):e69573, 2013. doi: 10.1371/journal.pone.0069573. URL `http://dx.doi.org/10.1371/journal.pone.0069573`.

FM Goodman. *Algebra: Abstarct and Concrete*. Prentice Hall, 1997.

F Harary. *Graph Theory*. Addison-Wesley, 1969.

R. Heinrich, S. M. Rapoport, and T. A. Rapoport. Metabolic regulation and mathematical models. *Prog Biophys Mol Biol*, 32(1):1–82, 1977.

DJ Higham and NJ Higham. *Matlab Guide*. SIAM, 2005.

N Hoory, S Linial and Wigderson. Expander graphs and their applications. *Bulletin of the American Mathematical Society*, 43:439–561, 2006.

G James and M Liebeck. *Representations and Characters of Groups*. Cambridge University Press, 2001.

T Kato. *Perturbation Theory for Linear Operators*. Springer, 1980.

M Krebs and A Shaheen. *Expander Families and Cayley Graphs: A Beginner's Guide*. Oxford University Press, 2011.

N Levinson and RM Redheffer. *Complex Variables*. HoldenDay, 1970.

S Lipschutz. *3,000 Solved Problems in Linear Algebra*. McGraw-Hill, 1989.

L. Perko. *Differential Equations and Dynamical Systems*. springer, 1991.

G Polya and RC Read. *Combinatorial Enumeration of Groups, Graphs, and Chemical Compounds*. Springer, 2011.

B Steinberg. *Representation Theory of Finite Groups: An Introductory Approach*. Springer, 2011.

G Strang. *Computational Science and Engineering*. Wellesley-Cambridge Press, 2007.

A Streitwieser. *Molecular Orbital Theory*. Wiley, 1961.

DJ Velleman. *How to Prove It: A Structured Approach*. Cambridge University Press, 2006.

# Index