(b) Choose $\boldsymbol{v}_2$ and $\boldsymbol{u}_2$ in U and V. Then $A = U\Sigma V^{\mathrm{T}} = \boldsymbol{u}_1\sigma_1\boldsymbol{v}_1^{\mathrm{T}}$ (one term only).

6 Substitute the SVD for A and A^{T} to show that $A^{\mathrm{T}}A$ has its eigenvalues in $\Sigma^{\mathrm{T}}\Sigma$ and AA^{T} has its eigenvalues in $\Sigma\Sigma^{\mathrm{T}}$. Since a diagonal $\Sigma^{\mathrm{T}}\Sigma$ has the same nonzeros as $\Sigma\Sigma^{\mathrm{T}}$, we see again that $A^{\mathrm{T}}A$ and AA^{T} have the same nonzero eigenvalues.

7 If $(A^{\mathrm{T}}A)\boldsymbol{v} = \sigma^2\boldsymbol{v}$, multiply by A. *Move the parentheses to get* $(AA^{\mathrm{T}})A\boldsymbol{v} = \sigma^2(A\boldsymbol{v})$. **If $\boldsymbol{v}$ is an eigenvector of $A^{\mathrm{T}}A$, then ____ is an eigenvector of AA^{T}.**

8 Find the eigenvalues and unit eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2$ of $A^{\mathrm{T}}A$. Then find $\boldsymbol{u}_1 = A\boldsymbol{v}_1/\sigma_1$:

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} \text{ and } A^{\mathrm{T}}A = \begin{bmatrix} 10 & 20 \\ 20 & 40 \end{bmatrix} \text{ and } AA^{\mathrm{T}} = \begin{bmatrix} 5 & 15 \\ 15 & 45 \end{bmatrix}.$$

Verify that $\boldsymbol{u}_1$ is a unit eigenvector of AA^{T}. Complete the matrices U, Σ, V.

SVD
$$\begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 \end{bmatrix} \begin{bmatrix} \sigma_1 & \\ & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{bmatrix}^{\mathrm{T}}.$$

9 Write down orthonormal bases for the four fundamental subspaces of this A.

10 (a) Why is the trace of $A^{\mathrm{T}}A$ equal to the sum of all a_{ij}^2 ? In Example 3 it is 50.

(b) For every rank-one matrix, why is $\sigma_1^2 =$ sum of all a_{ij}^2?

11 Find the eigenvalues and unit eigenvectors of $A^{\mathrm{T}}A$ and AA^{T}. Keep each $A\boldsymbol{v} = \sigma\boldsymbol{u}$. Then construct the singular value decomposition and verify that A equals $U\Sigma V^{\mathrm{T}}$.

Fibonacci matrix
$$A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

12 Use the **svd** part of the MATLAB demo **eigshow** to find those $\boldsymbol{v}$'s graphically.

13 If $A = U\Sigma V^{\mathrm{T}}$ is a square invertible matrix then $A^{-1} =$ ____ ____ ____. Check $A^{-1}A$. This shows that *the singular values of* A^{-1} *are* $1/\sigma_i$.

Note: The largest singular value of A^{-1} is therefore $1/\sigma_{\min}(A)$. The largest eigenvalue $|\lambda(A^{-1})|_{\max}$ is $1/|\lambda(A)|_{\min}$. Then equation (14) says that $\sigma_{\min}(A) \le |\lambda(A)|_{\min}$.

14 Suppose $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ are orthonormal bases for $\mathbf{R}^n$. Construct the matrix $A = U\Sigma V^{\mathrm{T}}$ that transforms each $\boldsymbol{v}_j$ into $\boldsymbol{u}_j$ to give $A\boldsymbol{v}_1 = \boldsymbol{u}_1, \ldots, A\boldsymbol{v}_n = \boldsymbol{u}_n$.

15 Construct the matrix with rank one that has $A\boldsymbol{v} = 12\boldsymbol{u}$ for $\boldsymbol{v} = \frac{1}{2}(1,1,1,1)$ and $\boldsymbol{u} = \frac{1}{3}(2,2,1)$. Its only singular value is $\sigma_1 =$ ____.

16 Suppose A has orthogonal columns $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_n$ of lengths $\sigma_1, \sigma_2, \ldots, \sigma_n$. What are U, Σ, and V in the SVD?

17 Suppose A is a 2 by 2 symmetric matrix with unit eigenvectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$. If its eigenvalues are $\lambda_1 = 3$ and $\lambda_2 = -2$, what are the matrices U, Σ, V^{T} in its SVD?

18 If $A = QR$ with an orthogonal matrix Q, the SVD of A is almost the same as the SVD of R. Which of the three matrices U, Σ, V is changed because of Q ?

19 Suppose A is invertible (with $\sigma_1 > \sigma_2 > 0$). Change A by ***as small a matrix as possible*** to produce a singular matrix A_0. Hint: U and V do not change:

$$\text{From} \quad A = \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 \end{bmatrix} \begin{bmatrix} \sigma_1 & \\ & \sigma_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{bmatrix}^{\mathrm{T}} \quad \text{find the nearest } A_0.$$

20 Find the singular values of A from the command **svd**(A) or by hand.

$$A = \begin{bmatrix} 1 & 0 \\ 100 & 1 \end{bmatrix}. \text{ Why is } \sigma_2 = \frac{1}{\sigma_1} \text{ for this matrix?}$$

21 Why doesn't the SVD for $A + I$ just use $\Sigma + I$?

22 If $A = U\Sigma V^{\mathrm{T}}$ then $Q_1 A Q_2^{\mathrm{T}} = (Q_1 U)\Sigma(Q_2 V)^{\mathrm{T}}$. Why will any orthogonal matrices Q_1 and Q_2 leave $Q_1 U =$ orthogonal matrix and $Q_2 V =$ orthogonal matrix? Then Σ sees **no change in the singular values**: $\boldsymbol{Q_1 A Q_2^{\mathrm{T}}}$ **has the same** $\boldsymbol{\sigma}$**'s as** $\boldsymbol{A}$.

23 If Q is an orthogonal matrix, why do all its singular values equal 1 ?

24 (a) Find the maximum of $\dfrac{\boldsymbol{x}^{\mathrm{T}} S \boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}} \boldsymbol{x}} = \dfrac{3x_1^2 + 2x_1 x_2 + 3x_2^2}{x_1^2 + x_2^2}$. What matrix is S ?

(b) Find the maximum of $\dfrac{(x_1 + 4x_2)^2}{x_1^2 + x_2^2}$. For what matrix A is this $\dfrac{||A\boldsymbol{x}||^2}{||\boldsymbol{x}||^2}$?

25 What are the **minimum values** of the ratios $\dfrac{\boldsymbol{x}^{\mathrm{T}} S \boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}} \boldsymbol{x}}$ and $\dfrac{||A\boldsymbol{x}||^2}{||\boldsymbol{x}||^2}$? We should take $\boldsymbol{x}$ to be which eigenvectors of S ? Should $\boldsymbol{x}$ always be an eigenvector of A ?

26 Every matrix $A = U\Sigma V^{\mathrm{T}}$ takes **circles to ellipses**. $AV = U\Sigma$ says that the radius vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ of the circle go to the semi-axes $\sigma_1 \boldsymbol{u}_1$ and $\sigma_2 \boldsymbol{u}_2$ of the ellipse. Draw the circle and the ellipse for $\theta = 30°$:

$$V = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad U = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}.$$

Section 7.4 will start with an important SVD picture for 2 by 2 matrices:

$A =$ (**rotate**) (**stretch**) (**rotate**). With symmetry $S =$ (**rotate**) (**stretch**) (**rotate back**).

27 This problem looks for all matrices A with a given column space in $\mathbf{R}^m$ and a given row space in $\mathbf{R}^n$. Suppose $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_r$ and $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_r$ are bases for those two spaces. Make them columns of C and B. The goal is to show that A has this form:

$\boldsymbol{A = CMB^{\mathrm{T}}}$ ***for an*** $\boldsymbol{r}$ ***by*** $\boldsymbol{r}$ ***invertible matrix*** $\boldsymbol{M}$. *Hint*: *Start from* $A = U\Sigma V^{\mathrm{T}}$.

The first r columns of U and V must be connected to C and B by invertible matrices, because they contain bases for the same column space (in U) and row space (in V).

7.3 Principal Component Analysis (PCA by the SVD)

1 Data often comes in a matrix : n samples and m measurements per sample.

2 Center each row of the matrix A by subtracting the mean from each measurement.

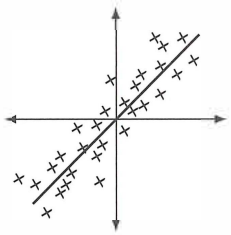
3 The SVD finds combinations of the data that contain the most information.

4 Largest singular value $\boldsymbol{\sigma_1}$ $\leftrightarrow$ greatest variance $\leftrightarrow$ most information in $\boldsymbol{u}_1$.

This section explains a major application of the SVD to statistics and data analysis. Our examples will come from human genetics and face recognition and finance. The problem is to understand a large matrix of data (= measurements). For each of n samples we are measuring m variables. The data matrix A_0 has n columns and m rows.

Graphically, the columns of A_0 are n points in $\mathbf{R}^m$. After we subtract the average of each row to reach A, the n points are often clustered along a line or close to a plane (or other low-dimensional subspace of $\mathbf{R}^m$). What is that line or plane or subspace ?

Let me start with a picture instead of numbers. For $m = 2$ variables like age and height, the n points lie in the plane $\mathbf{R}^2$. Subtract the average age and height to center the data. If the n recentered points cluster along a line, *how will linear algebra find that line* ?



A is $\mathbf{2} \times \boldsymbol{n}$ (large nullspace)

AA^{T} is $\mathbf{2} \times \mathbf{2}$ (small matrix)

$A^{\mathrm{T}}A$ is $\boldsymbol{n} \times \boldsymbol{n}$ (large matrix)

Two singular values $\boldsymbol{\sigma_1 > \sigma_2 > 0}$

Figure 7.2: Data points in A are often close to a line in $\mathbf{R}^2$ or a subspace in $\mathbf{R}^m$.

Let me go more carefully in constructing the data matrix. Start with the measurements in A_0 : the sample data. Find the average (the *mean*) $\mu_1, \mu_2, \ldots, \mu_m$ of each row. *Subtract each mean μ_i from row i to center the data.* The average along each row is now zero, for the centered matrix A. So the point $(0, 0)$ in Figure 7.2 is now the true center of the n points.

The "sample covariance matrix" is defined by $\boldsymbol{S = \dfrac{AA^{\mathrm{T}}}{n-1}}$.

A shows the distance $a_{ij} - \mu_i$ from each measurement to the row average μ_i.

$(AA^{\mathrm{T}})_{11}$ and $(AA^{\mathrm{T}})_{22}$ show the **sum of squared distances (sample variances $\boldsymbol{s_1^2}$, $\boldsymbol{s_2^2}$)**.

$(AA^{\mathrm{T}})_{12}$ shows the **sample covariance** $\boldsymbol{s_{12}}$ = (row 1 of A)·(row 2 of A).

The variance is a key number throughout statistics. An average exam score $\mu = 85$ tells you it was a decent exam. A variance of $s^2 = 25$ (standard deviation $s = 5$) means that most grades were in the 80's: closely packed. A sample variance $s^2 = 225$ ($s = 15$) means that grades were widely scattered. Chapter 12 explains variances.

The *covariance* of a math exam and a history exam is a dot product of those rows of A, with average grades subtracted out. Covariance below zero means: One subject strong when the other is weak. High covariance means: Both strong or both weak.

We divide by $n-1$ instead of n for reasons known best to statisticians. They tell me that one degree of freedom was used by the mean, leaving $n-1$. (I think the best plan is to agree with them.) In any case n should be a big number to count on reliable statistics. Since the rows of A have n entries, the numbers in AA^{T} have size growing like n and the division by $n-1$ keeps them steady.

Example 1 Six math and history scores (notice the zero mean in each row)

$$\boldsymbol{A} = \begin{bmatrix} 3 & -4 & 7 & 1 & -4 & -3 \\ 7 & -6 & 8 & -1 & -1 & -7 \end{bmatrix} \text{ has sample covariance } \boldsymbol{S} = \frac{AA^{\mathrm{T}}}{5} = \begin{bmatrix} 20 & 25 \\ 25 & 40 \end{bmatrix}.$$

The two rows of A are highly correlated: $s_{12} = 25$. Above average math went with above average history. Changing all the signs in row 2 would produce *negative covariance* $s_{12} = -25$. Notice that S has positive trace and determinant; AA^{T} is positive definite.

The eigenvalues of S are near 57 and 3. So the first rank one piece $\sqrt{57}\,\boldsymbol{u}_1\boldsymbol{v}_1^{\mathrm{T}}$ is much larger than the second piece $\sqrt{3}\,\boldsymbol{u}_2\boldsymbol{v}_2^{\mathrm{T}}$. **The leading eigenvector $\boldsymbol{u_1}$ shows the direction that you see in the scatter graph of Figure 7.2.** That eigenvector is close to $\boldsymbol{u}_1 = (.6, .8)$ and the direction in the graph nearly gives a $6-8-10$ or $3-4-5$ right triangle.

The SVD of A (centered data) shows the dominant direction in the scatter plot.

The second singular vector $\boldsymbol{u}_2$ is perpendicular to $\boldsymbol{u}_1$. The second singular value $\sigma_2 \approx \sqrt{3}$ measures the spread across the dominant line. If the data points in A fell exactly on a line ($\boldsymbol{u}_1$ direction), then σ_2 would be zero. Actually there would only be σ_1.

The Essentials of Principal Component Analysis (PCA)

PCA gives a way to understand a data plot in dimension m = the number of measured variables (here age and height). Subtract average age and height ($m = 2$ for n samples) to center the m by n data matrix A. *The crucial connection to linear algebra* is in the singular values and singular vectors of A. Those come from the eigenvalues $\lambda = \sigma^2$ and the eigenvectors $\boldsymbol{u}$ of the sample covariance matrix $S = AA^{\mathrm{T}}/(n-1)$.

- The total variance in the data is the sum of all eigenvalues and of sample variances s^2:
  Total variance $\boldsymbol{T = \sigma_1^2 + \cdots + \sigma_m^2 = s_1^2 + \cdots + s_m^2}$ = **trace** (*diagonal sum*).
- The first eigenvector $\boldsymbol{u}_1$ of S points in the most significant direction of the data. That direction accounts for (or *explains*) a fraction σ_1^2/T of the total variance.
- The next eigenvector $\boldsymbol{u}_2$ (orthogonal to $\boldsymbol{u}_1$) accounts for a smaller fraction σ_2^2/T.
- Stop when those fractions are small. You have the R directions that explain most of the data. The n data points are very near an R-dimensional subspace with basis $\boldsymbol{u}_1$ to $\boldsymbol{u}_R$. These $\boldsymbol{u}$'s are the **principal components** in m-dimensional space.
- R is the "effective rank" of A. The true rank r is probably m or n: full rank matrix.

Perpendicular Least Squares

It may not be widely recognized that the best line in Figure 7.2 (the line in the $\boldsymbol{u}_1$ direction) also solves a problem of *perpendicular least squares* (= orthogonal regression):

The sum of squared distances from the points to the line is a minimum.

Proof. Separate each column $\boldsymbol{a}_j$ into its components along the $\boldsymbol{u}_1$ line and $\boldsymbol{u}_2$ line:

$$\textbf{Right triangles} \qquad \sum_{j=1}^{n} ||\boldsymbol{a}_j||^2 = \sum_{j=1}^{n} |\boldsymbol{a}_j^{\mathrm{T}}\boldsymbol{u}_1|^2 + \sum_{j=1}^{n} |\boldsymbol{a}_j^{\mathrm{T}}\boldsymbol{u}_2|^2 \qquad (1)$$

The sum on the left is fixed by the data points $\boldsymbol{a}_j$ (columns of A). The first sum on the right is $\boldsymbol{u}_1^{\mathrm{T}} A A^{\mathrm{T}} \boldsymbol{u}_1$. So when we maximize that sum in PCA by choosing the eigenvector $\boldsymbol{u}_1$, we minimize the second sum. That second sum (squared distances from the data points to the best line) is a minimum for perpendicular least squares.

Ordinary least squares in Chapter 4 reached a linear equation $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$ by using *vertical distances* to the best line. PCA produces an eigenvalue problem for $\boldsymbol{u}_1$ by using *perpendicular distances*. "Total least squares" will allow for errors in A as well as $\boldsymbol{b}$.

The Sample Correlation Matrix

Data analysis works mostly with A (centered data). But the measurements in A might have different units like inches and pounds and years and dollars. Changing one set of units (inches to meters or years to seconds) would have a big effect on that row of A and S. If scaling is a problem, **we change from covariance matrix S to correlation matrix C**:

A diagonal matrix D rescales A. Each row of DA has length $\sqrt{n-1}$.
The sample correlation matrix $C = DAA^{\mathrm{T}}D/(n-1)$ has 1's on its diagonal.

Chapter 12 on Probability and Statistics will introduce the *expected* covariance matrix V and the *expected* correlation matrix (with diagonal 1's). Those use probabilities instead of actual measurements. The covariance matrix *predicts* the spread of future measurements around their mean, while A and the sample covariances S and the scaled correlation matrix $C = DSD$ use real data. All are highly important—a big connection between statistics and the linear algebra of positive definite matrices and the SVD.

Genetic Variation in Europe

We can follow changes in human populations by looking at genomes. To manage the huge amount of data, one good way to see genetic variation is from SNP's. The uncommon alleles (bases A/C/T/G in a pair from father and mother) are counted by the SNP:

SNP = 0	No change from the common base in that population: normal genotype
SNP = 1	The base pair shows one change from the usual pair
SNP = 2	Both bases are the less common allele

The uncentered matrix A_0 has a column for every person and a row for every base pair. The entries are mostly 0, quite a few 1, not so many 2. We don't test all 3 billion pairs. After subtracting row averages from A_0, the eigenvectors of AA^{T} are extremely revealing. **In Figure 7.4 the first singular vectors of A almost reproduce a map of Europe.**

This means: The SNP's from France and Germany and Italy are quite different. Even from the French and German and Italian parts of Switzerland those "snips" are different! Only Spain and Portugal are surprisingly confounded and harder to separate. More often than not, the DNA of an individual reveals his birthplace within 300 kilometers or 200 miles. A mixture of grandparents usually places the grandchild between their origins.

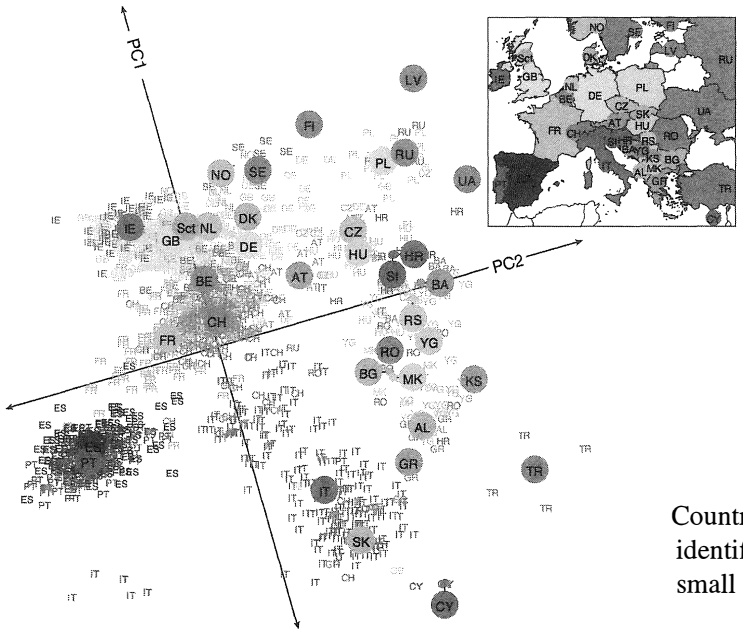


Figure 7.3: *Nature* (2008) Novembre et al: vol. 456 pp.98-101/doc:10.1038/nature07331.

What is the significant message? If we test genomes to understand how they correlate with diseases, we must not forget their spatial variation. Without correcting for geography, what looks medically significant can be very misleading. *Confounding* is a serious problem in medical genetics that PCA and population genetics can help to solve—to remove effects due to geography that don't have medical importance.

In fact "spatial statistics" is a tricky world. *Example*: Every matrix with three diagonals of 1, C, 1 shows a not surprising influence of next door neighbors (from the 1's). But its singular vectors have sine and cosine oscillations going across the map, independent of C. You might think those are true wave-like variations but they can be meaningless.

Maybe statistics produces more arguments than mathematics does? Reducing big data to a single small "P-*value*" can be instructive or it can be extremely deceptive. The expression P-*value* appears in many articles. P stands for the probability that an observation is consistent with the *null hypothesis* (= pure chance). If you see 5 heads in a row, the probability is $P = 1/32$ that this came by chance from a fair coin (or $P = 2/32$ if your observation is taken to be 5 heads or 5 tails in a row). Often a P-value below 0.05 makes the null hypothesis doubtful—maybe a crook is flipping the coin. As here, P-values are not the most reliable guides in statistics—but they are extremely convenient.

Eigenfaces

Recognizing faces would not seem to depend—at first glance—on linear algebra. But an early and well publicized application of the SVD was to **face recognition**. We are not compressing an image, we are identifying it.

The plan is to start with a "training set" A_0 of n images of a wide variety of faces. Each image becomes a very long vector by stacking all pixel grayscales into a column. Then A_0 must be centered : subtract the average of every *column* of A_0 to reach A.

The singular vector $\boldsymbol{v}_1$ of this A tells us the combination of known faces that best identifies a new face. Then $\boldsymbol{v}_2$ tells us the next best combination.

Probably we will use the R best vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_R$ with largest singular values $\sigma_1 \geq \cdots \geq \sigma_R$ of A. Those identify new faces more accurately than any other R vectors. Perhaps $R = 100$ of those **eigenfaces** $A\boldsymbol{v}$ will capture nearly all the variance in the training set. Those R eigenfaces span "face space".

This plan of attack was suggested by Matthew Turk and Alex Pentland. It developed the suggestion by Sirovich and Kirby to use PCA in compressing images of faces. I learned a lot from Jeff Jauregui's description on the Web. His summary is this : **PCA provides a mechanism to recognize geometric/photometric similarity through algebraic means**. He assembled the first principal component (first singular vector) into the first eigenface. Of course the average of each column was added back or you wouldn't see a face!

Note PCA is compared to NMF in a fascinating letter to *Nature* (Lee and Seung, vol. 401, 21 Oct. 1999). Nonnegative Matrix Factorization does not allow the negative entries that always appear in the singular vectors $\boldsymbol{v}$. So everything *adds*—which needs more vectors but they are often more meaningful.

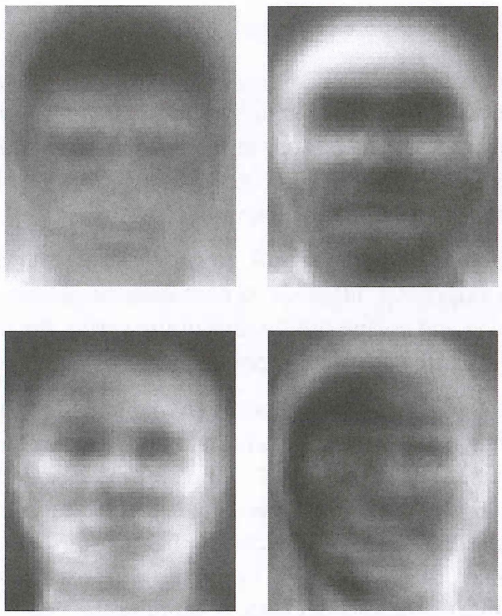


Figure 7.4: Eigenfaces pick out hairline and mouth and eyes and shape.

Applications of Eigenfaces

The first commercial use of PCA face recognition was for law enforcement and security. An early test at Super Bowl 35 in Tampa produced a very negative reaction from the crowd! The test was without the knowledge of the fans. Newspapers began calling it the "Snooper Bowl". I don't think the original eigenface idea is still used commercially (even in secret).

New applications of the SVD approach have come for other identification problems: Eigenvoices, Eigengaits, Eigeneyes, Eigenexpressions. I learned this from Matthew Turk (now in Santa Barbara, originally an MIT grad student. He told me he was in my class). The original eigenfaces in his thesis had problems accounting for rotation and scaling and lighting in the facial images. But the key ideas live on.

In the end, face space is nonlinear. So eventually we want nonlinear PCA.

Model Order Reduction

For a large-scale dynamic problem, the computational cost can become unmanageable. "Dynamic" means that the solution $\boldsymbol{u}(t)$ evolves as time goes forward. Fluid flow, chemical reactions, wave propagation, biological growth, electronic systems, these problems are everywhere. **A reduced model tries to identify important states of the system.** From a reduced problem we compute the needed information at much lower cost.

Model reduction is a truly important computational approach. Many good ideas have been proposed to reduce the original large problem. One simple and often useful idea is to take "snapshots" of the flow, put them in a matrix A, find the principal components (the left singular vectors of A), and work in their much smaller subspace:

A **snapshot** is a column vector that describes the state of the system
It can be an approximation to a typical true state $\boldsymbol{u}(t^*)$
From n snapshots, build a matrix A whose columns span a useful range of states

Now find the first R left singular vectors $\boldsymbol{u}_1$ to $\boldsymbol{u}_R$ of A. They are a basis for a Proper Orthogonal Decomposition (**POD** basis). In practice we choose R so that

Variance ≈ Energy $\quad \sigma_1^2 + \cdots + \sigma_R^2$ is 99% or 99.9% of $\sigma_1^2 + \cdots + \sigma_n^2$.

These vectors are an optimal basis for reconstructing the snapshots in A. If those snapshots are well chosen, then combinations of $\boldsymbol{u}_1$ to $\boldsymbol{u}_R$ will be close to the exact solution $\boldsymbol{u}(t)$ for desired times t and parameters p.

So much depends on the snapshots! *SIAM Review* 2015 includes an excellent survey by Beiner, Gugercin, and Willcox. The SVD compresses data as well as images.

Searching the Web

We believe that Google creates rankings by a walk that follows web links. When this walk goes often to a site, the ranking is high. The frequency of visits gives the leading eigenvector ($\lambda = 1$) of the "Web matrix"—the largest eigenvalue problem ever solved.

That Markov matrix has more than 3 billion rows and columns, from 3 billion web sites.

Many of the important techniques are well-kept secrets of Google. Probably they start with an earlier eigenvector as a first approximation, and they run the random walk very fast. To get a high ranking, you want a lot of links from important sites.

Here is an application of the SVD to web search engines. When you google a word, you get a list of web sites in order of importance. You could try typing "four subspaces".

The HITS algorithm was an early proposal to produce that ranked list. It begins with about 200 sites found from an index of key words. After that we look *only at links between pages*. Search engines are link-based more than content-based.

Start with the 200 sites and all sites that link to them and all sites they link to. That is our list, to be put in order. Importance can be measured by links out and links in.

1. The site may be an ***authority***: *Links come in* from many sites. Especially from hubs.
2. The site may be a ***hub***: *Links go out* to many sites in the list. Especially to authorities.

We want numbers $\boldsymbol{x_1}, \ldots, \boldsymbol{x_N}$ to rank the authorities and $\boldsymbol{y_1}, \ldots, \boldsymbol{y_N}$ to rank the hubs. Start with a simple count: $\boldsymbol{x_i^0}$ and $\boldsymbol{y_i^0}$ count the links into and out of site i.

Here is the point: *A good authority has links from important sites* (like hubs). Links from universities count more heavily than links from friends. *A good hub is linked to important sites* (like authorities). A link to **amazon.com** unfortunately means more than a link to **wellesleycambridge.com**. The raw counts $\boldsymbol{x^0}$ and $\boldsymbol{y^0}$ are updated to $\boldsymbol{x^1}$ and $\boldsymbol{y^1}$ by taking account of *good* links (measuring their quality by $\boldsymbol{x^0}$ and $\boldsymbol{y^0}$):

$$\textbf{Authority / Hub} \quad \boldsymbol{x_i^1} \text{ / } \boldsymbol{y_i^1} = \text{Add up } \boldsymbol{y_j^0} \text{ / } \boldsymbol{x_j^0} \text{ for all links } \textbf{into } \boldsymbol{i} \textbf{ / out from } \boldsymbol{i} \qquad (2)$$

In matrix language those are $\boldsymbol{x^1} = A^{\mathrm{T}}\boldsymbol{y^0}$ and $\boldsymbol{y^1} = A\boldsymbol{x^0}$. The matrix A contains 1's and 0's, with $a_{ij} = 1$ when i links to j. In the language of graphs, A is an "adjacency matrix" for the Web (an enormous matrix). The new $\boldsymbol{x^1}$ and $\boldsymbol{y^1}$ give better rankings, but not the best. Take another step like (2), to reach $\boldsymbol{x^2}$ and $\boldsymbol{y^2}$ from $A^{\mathrm{T}}A\boldsymbol{x^0}$ and $AA^{\mathrm{T}}\boldsymbol{y^0}$:

$$\textbf{Authority} \quad \boldsymbol{x^2} = A^{\mathrm{T}}\boldsymbol{y^1} = A^{\mathrm{T}}A\boldsymbol{x^0} \qquad \textbf{Hub} \quad \boldsymbol{y^2} = A\boldsymbol{x^1} = AA^{\mathrm{T}}\boldsymbol{y^0}. \qquad (3)$$

In two steps we are multiplying by $A^{\mathrm{T}}A$ and AA^{T}. Twenty steps will multiply by $(A^{\mathrm{T}}A)^{10}$ and $(AA^{\mathrm{T}})^{10}$. **When we take powers, the largest eigenvalue $\boldsymbol{\sigma_1^2}$ begins to dominate**. The vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ line up with the leading eigenvectors $\boldsymbol{v_1}$ and $\boldsymbol{u_1}$ of $A^{\mathrm{T}}A$ and AA^{T}. We are computing the top terms in the SVD, by the **power method** that is discussed in Section 11.3. It is wonderful that linear algebra helps to understand the Web.

This HITS algorithm is described in the 1999 *Scientific American* (June 16). But I don't think the SVD is mentioned there. . . The excellent book by Langville and Meyer, *Google's PageRank and Beyond*, explains in detail the science of search engines.

PCA in Finance: The Dynamics of Interest Rates

The mathematics of finance constantly applies linear algebra and PCA. We choose one application: the **yield curve for Treasury securities**. The "yield" is the interest rate paid on the bonds or notes or bills. That rate depends on time to maturity. For longer bonds (3 years to 20 years) the rate increases with length. The Federal Reserve adjusts short term yields to slow or stimulate the economy. This is the *yield curve*, used by risk managers and traders and investors.

Here is data for the first 6 business days of 2001—each column is a yield curve for investments on a particular day. The time to maturity is the "tenor". The six columns at the left are the interest rates, changing from day to day. The five columns at the right are interest rate *differences between days*, with the mean difference subtracted from each row. **This is the centered matrix A with its rows adding to zero**. A real world application might start with 252 business days instead of 5 or 6 (a year instead of a week).

Table 1. U.S. Treasury Yields : 6 Days and 5 Centered Daily Differences

	US Treasury Yields in 2001						Matrix A in Basis Points (0.01 %)				
Tenor	Jan 3	Jan 4	Jan 5	Jan 6	Jan 7	Jan 10	Jan 4	Jan 5	Jan 6	Jan 7	Jan 10
3 MO	5.87	5.69	5.37	5.12	5.19	5.24	−5.4	−19.4	−12.4	19.6	17.6
6 MO	5.58	5.44	5.20	4.98	5.03	5.11	−4.6	−14.6	−12.6	14.4	17.4
1 YR	5.11	5.04	4.82	4.60	4.61	4.71	1.0	−14.0	−14.0	9.0	18.0
2 YR	4.87	4.92	4.77	4.56	4.54	4.64	9.6	−10.4	−16.4	2.6	14.0
3 YR	4.82	4.92	4.78	4.57	4.55	4.65	13.4	−10.6	−17.6	1.4	13.4
5 YR	4.76	4.94	4.82	4.66	4.65	4.73	18.6	−11.4	−15.4	−0.4	8.6
7 YR	4.97	5.18	5.07	4.93	4.94	4.98	20.8	−11.2	−14.2	0.8	3.8
10 YR	4.92	5.14	5.03	4.93	4.94	4.98	20.8	−12.2	−11.2	−0.2	2.8
20 YR	5.46	5.62	5.56	5.50	5.52	5.53	14.6	−7.4	−7.4	0.6	−0.4

With five columns we might expect five singular values. But the five column vectors add to the zero vector (since every row of A adds to zero after centering). So $S = AA^{\mathrm{T}}/(5-1)$ has four nonzero eigenvalues $\sigma_1^2 > \sigma_2^2 > \sigma_3^2 > \sigma_4^2$. Here are the singular values σ_i and their squares σ_i^2 and the fractions of the total variance $T = \sigma_1^2 + \cdots + \sigma_4^2 =$ trace of S that are "explained" by each principal component (each eigenvector $\boldsymbol{u}_i$ of S).

	$\boldsymbol{\sigma_i}$	$\boldsymbol{\sigma_i^2}$	$\boldsymbol{\sigma_i^2/T}$
Principal component $\boldsymbol{u}_1$	36.39	1323.9	.7536
Principal component $\boldsymbol{u}_2$	19.93	397.2	.2261
Principal component $\boldsymbol{u}_3$	5.85	34.2	.0195
Principal component $\boldsymbol{u}_4$	1.19	1.4	.0008
Principal component $\boldsymbol{u}_5$	0.00	0.0	.0000
		$T = 1756.7$	1.0000

A "scree plot" graphs those fractions σ_i^2/T dropping quickly to zero. In a larger problem you often see fast dropoff followed by a flatter part at the bottom (near $\sigma^2 = 0$). Locating the elbow between those two parts (significant and insignificant PC's) is important.

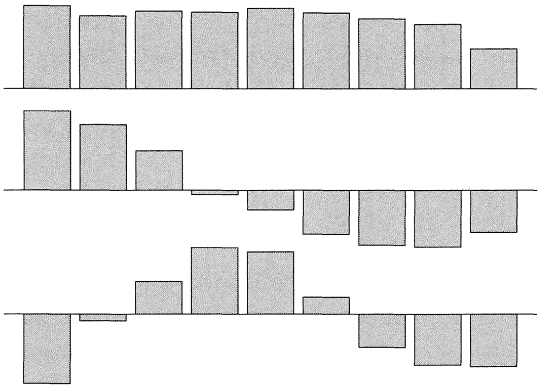
We also aim to understand each principal component. Those singular vectors $\boldsymbol{u}_i$ of A are eigenvectors of S. The entries in those vectors are the "*loadings*". Here are $\boldsymbol{u}_1$ to $\boldsymbol{u}_5$ for this yield curve example (with $S\boldsymbol{u}_5 = \boldsymbol{0}$).

	$\boldsymbol{u}_1$	$\boldsymbol{u}_2$	$\boldsymbol{u}_3$	$\boldsymbol{u}_4$	$\boldsymbol{u}_5$
3 MO	0.383	0.529	−0.478	0.060	0.084
6 MO	0.336	0.436	−0.046	0.210	−0.263
1 YR	0.358	0.263	0.225	−0.491	0.237
2 YR	0.352	−0.028	0.460	0.096	0.242
3 YR	0.371	−0.131	0.430	0.258	−0.555
5 YR	0.349	−0.293	0.117	−0.188	0.446
7 YR	0.323	−0.365	−0.228	0.459	0.081
10 YR	0.297	−0.378	−0.351	−0.579	−0.470
20 YR	0.184	−0.280	−0.361	0.227	0.268

Those five $\boldsymbol{u}$'s are orthonormal. They give bases for the four-dimensional column space of A and the one-dimensional nullspace of A^{T}. What financial meaning do they have?

- $\boldsymbol{u}_1$ measures a weighted average of the daily changes in the 9 yields
- $\boldsymbol{u}_2$ gauges the daily change in the yield spread between long and short bonds
- $\boldsymbol{u}_3$ shows daily changes in the curvature (short and long bonds versus medium)

These graphs show the nine loadings on $\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3$ above from 3 months to 20 years.



The output from a typical code (written in R) will include two more tables—which are going on the book's website. One will show the *right* singular vectors $\boldsymbol{v}_i$ of A. These are eigenvectors of $A^{\mathrm{T}}A$. They are proportional to the vectors $A^{\mathrm{T}}\boldsymbol{u}$. They have 5 components and they show the movement of yields and short-long spreads during the week.

The total variance $T = 1756.7$ (the trace $\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2$ of S) is also the sum of the diagonal entries of S. Those are the sample variances of the rows of A. Here they are:
$\boldsymbol{s_1^2}+\cdots+\boldsymbol{s_9^2} = 313.3+225.8+199.5+172.3+195.8+196.8+193.7+178.7+80.8 = \mathbf{1756.7}$.
Every s^2 is below σ_1^2. And 1756.7 is also the trace of $A^{\mathrm{T}}A/(n-1)$: column variances.

Note that this PCA section 7.3 is working with centered *rows* in A. In some applications (like finance), the matrix is usually transposed and the *columns* are centered. Then the sample covariance matrix S uses $A^{\mathrm{T}}A$, and the $\boldsymbol{v}$'s are the more important principal components. Linear algebra with practical interpretations tells us so much.

Problem Set 7.3

1 Suppose A_0 holds these 2 measurements of 5 samples:

$$A_0 = \begin{bmatrix} 5 & 4 & 3 & 2 & 1 \\ -1 & 1 & 0 & 1 & -1 \end{bmatrix}$$

Find the average of each row and subtract it to produce the centered matrix A. Compute the sample covariance matrix $S = AA^{\mathrm{T}}/(n-1)$ and find its eigenvalues λ_1 and λ_2. What line through the origin is closest to the 5 samples in columns of A?

2 Take the steps of Problem 1 for this 2 by 6 matrix A_0 :

$$A_0 = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 2 & 3 & 3 & 2 & 1 \end{bmatrix}$$

3 The sample variances s_1^2, s_2^2 and the sample covariance s_{12} are the entries of S. What is S (after subtracting means) when $A_0 = \begin{bmatrix} 1 & 2 & 3 \\ 5 & 2 & 2 \end{bmatrix}$? What is σ_1 ?

4 From the eigenvectors of $S = AA^{\mathrm{T}}$, find the line (the $\boldsymbol{u}_1$ direction through the center point) and then the plane ($\boldsymbol{u}_1, \boldsymbol{u}_2$ directions) closest to these four points in three-dimensional space :

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

5 From this sample covariance matrix S, find the correlation matrix DSD with 1's down its main diagonal. D is a positive diagonal matrix that produces those 1's.

$$S = \begin{bmatrix} 4 & 2 & 0 \\ 2 & 4 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

6 Choose the diagonal matrix D that produces DSD and find the correlations c_{ij}:

$$S = \begin{bmatrix} s_1^2 & s_{12} & s_{13} \\ s_{12} & s_2^2 & s_{23} \\ s_{13} & s_{23} & s_3^2 \end{bmatrix} \qquad DSD = \begin{bmatrix} 1 & c_{12} & c_{13} \\ c_{12} & 1 & c_{23} \\ c_{13} & c_{23} & 1 \end{bmatrix}.$$

7 Suppose A_0 is a 5 by 10 matrix with average grades for 5 courses over 10 years. How would you create the centered matrix A and the sample covariance matrix S ? When you find the leading eigenvector of S, what does it tell you ?

7.4 The Geometry of the SVD

1 A typical square matrix $A = U\Sigma V^{\mathrm{T}}$ factors into (rotation) (stretching) (rotation).

2 The geometry shows how A transforms vectors $\boldsymbol{x}$ on a circle to vectors $A\boldsymbol{x}$ on an ellipse.

3 The **norm** of A is $||A|| = \sigma_1$. This singular value is its maximum growth factor $||A\boldsymbol{x}|| \,/\, ||\boldsymbol{x}||$.

4 **Polar decomposition** factors A into QS: rotation $Q = UV^{\mathrm{T}}$ times stretching $S = V\Sigma V^{\mathrm{T}}$.

5 The **pseudoinverse** $A^{+} = V\Sigma^{+}U^{\mathrm{T}}$ brings $A\boldsymbol{x}$ in the column space back to $\boldsymbol{x}$ in the row space.

The SVD separates a matrix into three steps: (**orthogonal**)×(**diagonal**)×(**orthogonal**). Ordinary words can express the geometry behind it: (**rotation**)×(**stretching**)×(**rotation**). $U\Sigma V^{\mathrm{T}}\boldsymbol{x}$ starts with the rotation to $V^{\mathrm{T}}\boldsymbol{x}$. Then Σ stretches that vector to $\Sigma V^{\mathrm{T}}\boldsymbol{x}$, and U rotates to $A\boldsymbol{x} = U\Sigma V^{\mathrm{T}}\boldsymbol{x}$. Here is the picture.

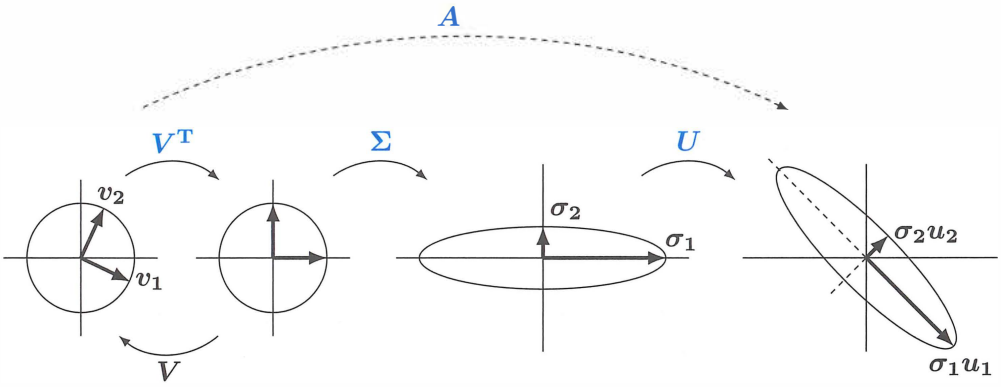


Figure 7.5: U and V are rotations and possible reflections. Σ stretches circle to ellipse.

Admittedly, this picture applies to a 2 by 2 matrix. And not every 2 by 2 matrix, because U and V didn't allow for a reflection—all three matrices have determinant > 0. This A would have to be invertible because the three steps are shown as invertible:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \sigma_1 & \\ & \sigma_2 \end{bmatrix} \begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix} = U\Sigma V^{\mathrm{T}}. \quad (1)$$

The four numbers a, b, c, d in the matrix A led to four numbers $\theta, \sigma_1, \sigma_2, \phi$ in its SVD.

This picture will guide us to three neat ideas in the algebra of matrices:

1 **The norm $||A||$ of a matrix**—its maximum growth factor.

2 **The polar decomposition $A = QS$**—orthogonal Q times positive definite S.

3 **The pseudoinverse A^{+}**—the best inverse when the matrix A is not invertible.

The Norm of a Matrix

If I choose one crucial number in the picture it is $\boldsymbol{\sigma_1}$. That number is the *largest growth factor of any vector* $\boldsymbol{x}$. If you follow the vector $\boldsymbol{v}_1$ on the left, you see it rotate to $(1,0)$ and stretch to $(\sigma_1, 0)$ and finally rotate to $\sigma_1 \boldsymbol{u}_1$. The statement $A\boldsymbol{v}_1 = \sigma_1 \boldsymbol{u}_1$ is exactly the SVD equation. This largest singular value σ_1 is the "*norm*" of the matrix A.

The norm $||A||$ is the largest ratio $\dfrac{||A\boldsymbol{x}||}{||\boldsymbol{x}||}$

$$||A|| = \max_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{||A\boldsymbol{x}||}{||\boldsymbol{x}||} = \boldsymbol{\sigma_1} \tag{2}$$

MATLAB uses norm $(\boldsymbol{x})$ for vector lengths and the same word norm (A) for matrix norms. The math symbols have double bars: $||\boldsymbol{x}||$ and $||A||$. Here $||\boldsymbol{x}||$ means the standard length of a vector with $||\boldsymbol{x}||^2 = |x_1|^2 + \cdots + |x_n|^2$. The matrix norm comes from this vector norm when $\boldsymbol{x} = \boldsymbol{v}_1$ and $A\boldsymbol{x} = \sigma_1 \boldsymbol{u}_1$ and $||A\boldsymbol{x}|| \,/\, ||\boldsymbol{x}|| = \sigma_1 =$ largest ratio $= ||A||$.

Two valuable properties of that number norm (A) come directly from its definition:

Triangle inequality	$\|\|A + B\|\| \leq \|\|A\|\| + \|\|B\|\|$	**Product inequality**	$\|\|AB\|\| \leq \|\|A\|\| \, \|\|B\|\|$	(3)

The definition (2) says that $||\boldsymbol{A}\boldsymbol{x}|| \leq ||\boldsymbol{A}|| \, ||\boldsymbol{x}||$ for every vector $\boldsymbol{x}$. That is what we know! Then the triangle inequality for vectors leads to the triangle inequality for matrices:

For vectors $||(A+B)\boldsymbol{x}|| \leq ||A\boldsymbol{x}|| + ||B\boldsymbol{x}|| \leq ||A|| \, ||\boldsymbol{x}|| + ||B|| \, ||\boldsymbol{x}||$.

Divide this by $||\boldsymbol{x}||$. Take the maximum over all $\boldsymbol{x}$. Then $||\boldsymbol{A} + \boldsymbol{B}|| \leq ||\boldsymbol{A}|| + ||\boldsymbol{B}||$.

The product inequality comes quickly from $||AB\boldsymbol{x}|| \leq ||A|| \, ||B\boldsymbol{x}|| \leq ||A|| \, ||B|| \, ||\boldsymbol{x}||$. Again divide by $||\boldsymbol{x}||$. Take the maximum over all $\boldsymbol{x}$. The result is $||\boldsymbol{AB}|| \leq ||\boldsymbol{A}|| \, ||\boldsymbol{B}||$.

Example 1 A rank-one matrix $A = \boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}$ is as basic as we can get. It has one nonzero eigenvalue λ_1 and one nonzero singular value σ_1. Neatly, its eigenvector is $\boldsymbol{u}$ and its singular vectors (left and right) are $\boldsymbol{u}$ and $\boldsymbol{v}$.

Eigenvector $A\boldsymbol{u} = (\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}})\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}) = \lambda_1 \boldsymbol{u}$ So $\lambda_1 = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}$

Singular vector $A^{\mathrm{T}}A\boldsymbol{v} = (\boldsymbol{v}\boldsymbol{u}^{\mathrm{T}})(\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}})\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u})(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}) = \sigma_1^2 \boldsymbol{v}$ So $\sigma_1 = ||\boldsymbol{u}|| \, ||\boldsymbol{v}||$.

It makes you feel good that $|\lambda_1| \leq \sigma_1$ is exactly the Schwarz inequality $|\boldsymbol{v}^{\mathrm{T}}\boldsymbol{u}| \leq ||\boldsymbol{u}|| \, ||\boldsymbol{v}||$.

How do we know that $|\lambda_1| \leq \sigma_1$? The eigenvector for $A\boldsymbol{x} = \lambda_1 \boldsymbol{x}$ will give the ratio $||A\boldsymbol{x}|| \,/\, ||\boldsymbol{x}|| = ||\lambda_1 \boldsymbol{x}|| \,/\, ||\boldsymbol{x}||$ which is $|\lambda_1|$. The maximum ratio σ_1 can't be less than $|\lambda_1|$.

Is it also true that $|\lambda_2| \leq \sigma_2$? **No**. That is completely wrong. In fact a 2 by 2 matrix will have $|\det A| = |\lambda_1 \lambda_2| = \sigma_1 \sigma_2$. In this case $|\lambda_1| \leq \sigma_1$ will force $|\lambda_2| \geq \sigma_2$.

The closest rank k matrix to A is $A_k = \sigma_1 u_1 v_1^T + \cdots + \sigma_k u_k v_k^T$

This is the key fact in matrix approximation: The Eckart-Young-Mirsky Theorem says that

$$||A - B|| \geq ||A - A_k|| = \sigma_{k+1} \text{ for all matrices } B \text{ of rank } k.$$

To me this completes the Fundamental Theorem of Linear Algebra. The v's and u's give orthonormal bases for the four fundamental subspaces, and the *first k v's and u's and σ's* give the best matrix approximation to A.

Polar Decomposition $A = QS$

Every complex number $x + iy$ has the polar form $re^{i\theta}$. A number $r \geq 0$ multiplies a number $e^{i\theta}$ on the unit circle. We have $x + iy = r\cos\theta + ir\sin\theta = r(\cos\theta + i\sin\theta) = re^{i\theta}$. Think of these numbers as 1 by 1 matrices. Then $e^{i\theta}$ is an *orthogonal matrix* Q and $r \geq 0$ is a *positive semidefinite matrix* (call it S). The ***polar decomposition*** extends the same idea to n by n matrices: orthogonal times positive semidefinite, $A = QS$.

Every real square matrix can be factored into $A = QS$, where Q is ***orthogonal*** and S is ***symmetric positive semidefinite***. If A is invertible, S is positive definite.

For the proof we just insert $V^T V = I$ into the middle of the SVD:

Polar decomposition
$$A = U\Sigma V^T = (UV^T)(V\Sigma V^T) = (Q)(S). \quad (4)$$

The first factor UV^T is Q. The product of orthogonal matrices is orthogonal. The second factor $V\Sigma V^T$ is S. It is positive semidefinite because its eigenvalues are in Σ.

If A is invertible then Σ and S are also invertible. ***S is the symmetric positive definite square root of $A^T A$***, because $S^2 = V\Sigma^2 V^T = A^T A$. So the eigenvalues of S are the singular values of A. The eigenvectors of S are the singular vectors v of A.

There is also a polar decomposition $A = KQ$ in the reverse order. Q is the same but now $K = U\Sigma U^T$. Then K is the symmetric positive definite square root of AA^T.

Example 2 The SVD example in Section 7.2 was $A = \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix} = U\Sigma V^T$. Find the factors Q and S (rotation and stretch) in the polar decomposition $A = QS$.

Solution I will just copy the matrices U and Σ and V from Section 7.2:

$$Q = UV^T = \frac{1}{\sqrt{20}}\begin{bmatrix} 1 & -3 \\ 3 & 1 \end{bmatrix}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} = \frac{1}{\sqrt{20}}\begin{bmatrix} 4 & -2 \\ 2 & 4 \end{bmatrix} = \frac{1}{\sqrt{5}}\begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}$$

$$S = V\Sigma V^T = \frac{\sqrt{5}}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 3 & \\ & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} = \sqrt{5}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}. \text{ Then } \boldsymbol{A = QS}.$$

In mechanics, the polar decomposition separates the ***rotation*** (in Q) from the ***stretching*** (in S). The eigenvalues of S give the stretching factors as in Figure 7.5. The eigenvectors of S give the stretching directions (the principal axes of the ellipse). The orthogonal matrix Q includes both rotations U and V^T.

Here is a fact about rotations. $Q = UV^T$ is the **nearest orthogonal matrix** to A. This Q makes the norm $||Q - A||$ as small as possible. That corresponds to the fact that $e^{i\theta}$ is the nearest number on the unit circle to $re^{i\theta}$.

The SVD tells us an even more important fact about nearest singular matrices:

The nearest singular matrix A_0 to A comes by changing the smallest $\sigma_{\min}$ to zero.

So $\sigma_{\min}$ is measuring the distance from A to singularity. For the matrix in Example 2 that distance is $\sigma_{\min} = \sqrt{5}$. If I change $\sigma_{\min}$ to zero, this knocks out the last (smallest) piece in $A = \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^{\mathrm{T}} + \sigma_2 \boldsymbol{u}_2 \boldsymbol{v}_2^{\mathrm{T}}$. Then only the rank-one (singular!) matrix $\sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^{\mathrm{T}}$ will be left: the closest to A. The smallest change had norm $\sigma_2 = \sqrt{5}$ (*smaller than* 3).

In computational practice we often do knock out a very small σ. Working with singular matrices is better than coming too close to zero and not noticing.

The Pseudoinverse A^+

By choosing good bases, A multiplies $\boldsymbol{v}_i$ in the row space to give $\sigma_i \boldsymbol{u}_i$ in the column space. A^{-1} must do the opposite! If $A\boldsymbol{v} = \sigma \boldsymbol{u}$ then $A^{-1}\boldsymbol{u} = \boldsymbol{v}/\sigma$. The singular values of A^{-1} are $1/\sigma$, just as the eigenvalues of A^{-1} are $1/\lambda$. The bases are reversed. The $\boldsymbol{u}$'s are in the row space of A^{-1}, the $\boldsymbol{v}$'s are in the column space.

Until this moment we would have added "*if A^{-1} exists*." Now we don't. A matrix that multiplies $\boldsymbol{u}_i$ to produce $\boldsymbol{v}_i/\sigma_i$ *does* exist. It is the pseudoinverse A^+:

Pseudoinverse of A

$$A^+ = V\Sigma^+ U^{\mathrm{T}} = \underset{n \text{ by } n}{\begin{bmatrix} \boldsymbol{v}_1 \cdots \boldsymbol{v}_r \cdots \boldsymbol{v}_n \end{bmatrix}} \underset{n \text{ by } m}{\begin{bmatrix} \sigma_1^{-1} & & \\ & \ddots & \\ & & \sigma_r^{-1} \end{bmatrix}} \underset{m \text{ by } m}{\begin{bmatrix} \boldsymbol{u}_1 \cdots \boldsymbol{u}_r \cdots \boldsymbol{u}_m \end{bmatrix}^{\mathrm{T}}}$$

The ***pseudoinverse*** A^+ is an n by m matrix. If A^{-1} exists (we said it again), then A^+ is the same as A^{-1}. In that case $m = n = r$ and we are inverting $U\Sigma V^{\mathrm{T}}$ to get $V\Sigma^{-1}U^{\mathrm{T}}$. The new symbol A^+ is needed when $r < m$ or $r < n$. Then A has no two-sided inverse, but it has a *pseudo*inverse A^+ with that same rank r:

$$A^+ \boldsymbol{u}_i = \frac{1}{\sigma_i} \boldsymbol{v}_i \quad \text{for } i \le r \quad \text{and} \quad A^+ \boldsymbol{u}_i = \boldsymbol{0} \quad \text{for } i > r.$$

The vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r$ in the column space of A go back to $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r$ in the row space. The other vectors $\boldsymbol{u}_{r+1}, \ldots, \boldsymbol{u}_m$ are in the left nullspace, and A^+ sends them to zero. When we know what happens to all those basis vectors, we know A^+.

Notice the pseudoinverse of the diagonal matrix Σ. Each σ in Σ is replaced by σ^{-1} in Σ^+. The product $\Sigma^+\Sigma$ is as near to the identity as we can get. It is a projection matrix, $\Sigma^+\Sigma$ is partly I and otherwise zero. We can invert the σ's, but we can't do anything about the zero rows and columns. This example has $\sigma_1 = 2$ and $\sigma_2 = 3$:

$$\Sigma^+\Sigma = \begin{bmatrix} \mathbf{1/2} & 0 & 0 \\ 0 & \mathbf{1/3} & 0 \\ 0 & 0 & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{2} & 0 & 0 \\ 0 & \mathbf{3} & 0 \\ 0 & 0 & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & 0 & 0 \\ 0 & \mathbf{1} & 0 \\ 0 & 0 & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & 0 \\ 0 & \mathbf{0} \end{bmatrix}.$$

The pseudoinverse A^+ is the n by m matrix that makes AA^+ and A^+A into projections.

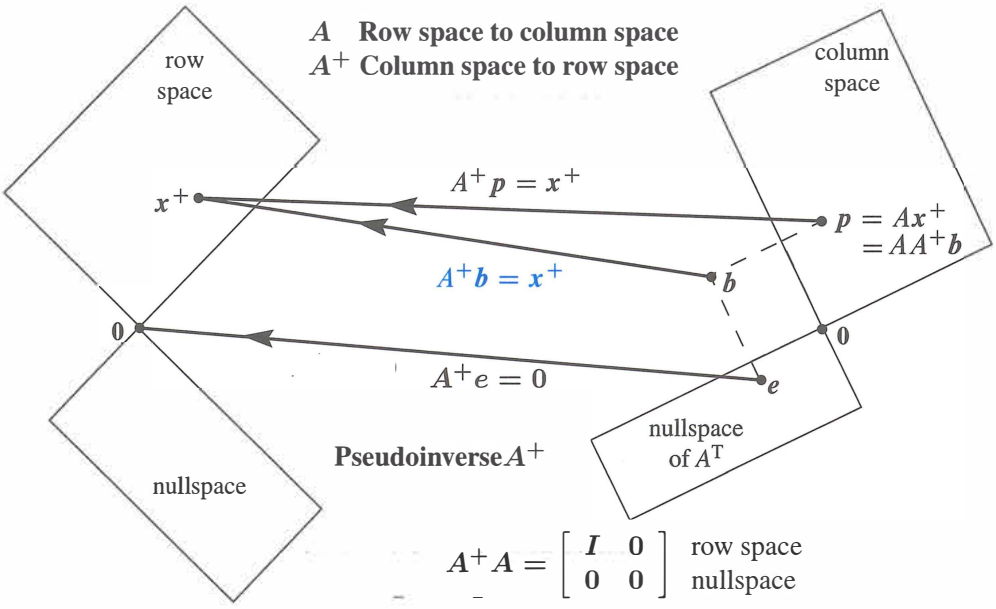


Figure 7.6: $A\boldsymbol{x}^+$ in the column space goes back to $A^+A\boldsymbol{x}^+ = \boldsymbol{x}^+$ in the row space.

Trying for $AA^{-1} = A^{-1}A = I$	AA^+ = projection matrix onto the column space of A A^+A = projection matrix onto the row space of A

Example 3 Every rank one matrix is a column times a row. With unit vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, that is $A = \sigma\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}$. Its pseudoinverse is $A^+ = \boldsymbol{v}\boldsymbol{u}^{\mathrm{T}}/\sigma$. The product AA^+ is $\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$, the projection onto the line through $\boldsymbol{u}$. The product A^+A is $\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}$.

Example 4 Find the pseudoinverse of $A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. This matrix is not invertible. The rank is 1. The only singular value is $\sigma_1 = 2$. That is inverted to $1/2$ in Σ^+ (also rank 1).

$$A^+ = V\Sigma^+U^{\mathrm{T}} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1/2 & 0 \\ 0 & 0 \end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \frac{\mathbf{1}}{\mathbf{4}}\begin{bmatrix} \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} \end{bmatrix}.$$

A^+ also has rank 1. Its column space is always the row space of A.

Least Squares with Dependent Columns

That matrix A with four 1's appeared in Section 4.3 on least squares. It broke the requirement of independent columns. The matrix appeared when we made two measurements, both at time $t = 1$. The closest straight line went halfway between the measurements 3 and 1, but there was no way to decide on the slope of the best line.

In matrix language, $A^{\mathrm{T}}A$ was singular. The equation $A^{\mathrm{T}}A\boldsymbol{x} = \boldsymbol{A^{\mathrm{T}}b}$ **had infinitely many solutions**. The pseudoinverse gives us a way to choose a "best solution" $\boldsymbol{x}^+ = A^+\boldsymbol{b}$.

Let me repeat the unsolvable $A\boldsymbol{x} = \boldsymbol{b}$ and the infinitely solvable $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$:

$$A\boldsymbol{x} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} = \boldsymbol{b} \qquad A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}\begin{bmatrix} \widehat{x}_1 \\ \widehat{x}_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 4 \end{bmatrix} = A^{\mathrm{T}}\boldsymbol{b}$$

Any vector $\widehat{\boldsymbol{x}} = (1 + c,\ 1 - c)$ will solve those normal equations $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$. The purpose of the pseudoinverse is to choose one solution $\widehat{\boldsymbol{x}} = \boldsymbol{x}^+$.

$\boldsymbol{x}^+ = \boldsymbol{A}^+\boldsymbol{b} = (1, 1)$ is the shortest solution to $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$ and $A\widehat{\boldsymbol{x}} = \boldsymbol{p}$.

You can see that $\boldsymbol{x}^+ = (1, 1)$ is shorter than any other solution $\widehat{\boldsymbol{x}} = (1 + c, 1 - c)$. The length squared of $\widehat{\boldsymbol{x}}$ is $(1 + c)^2 + (1 - c)^2 = 2 + 2c^2$. The shortest choice is $c = 0$. That gives the solution $\boldsymbol{x}^+ = (1, 1)$ in the row space of A.

The geometry tells us what A^+ should do: Take the column space of A back to the row space. Both spaces have dimension r. Kill off the error vector $\boldsymbol{e}$ in the left nullspace.

The pseudoinverse A^+ and this best solution $\boldsymbol{x}^+$ are essential in statistics, because experiments often have a matrix with dependent columns as well as dependent rows.

■ REVIEW OF THE KEY IDEAS ■

1. The ellipse of vectors $A\boldsymbol{x}$ has axes along the singular vectors $\boldsymbol{u}_i$.
2. The matrix norm $||A|| = \sigma_1$ comes from the vector length: Maximize $||A\boldsymbol{x}||/||\boldsymbol{x}||$.
3. Invertible matrix = (orthogonal matrix) (positive definite matrix): $A = QS$.
4. Every $A = U\Sigma V^{\mathrm{T}}$ has a pseudoinverse $A^+ = V\Sigma^+U^{\mathrm{T}}$ that sends $\boldsymbol{N}(A^{\mathrm{T}})$ to $\boldsymbol{Z}$.

■ WORKED EXAMPLES ■

7.4 A If A has rank n (full column rank) then it has a **left inverse** $L = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$. This matrix L gives $\boldsymbol{LA} = \boldsymbol{I}$. Explain why the pseudoinverse is $\boldsymbol{A^+ = L}$ in this case.

If A has rank m (full row rank) then it has a **right inverse** $R = A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}$. This matrix R gives $\boldsymbol{AR} = \boldsymbol{I}$. Explain why the pseudoinverse is $\boldsymbol{A^+ = R}$ in this case.

Find L for A_1 and find R for A_2. Find A^+ for all three matrices A_1, A_2, A_3:

$$A_1 = \begin{bmatrix} \mathbf{2} \\ \mathbf{2} \end{bmatrix} \qquad A_2 = \begin{bmatrix} \mathbf{2} & \mathbf{2} \end{bmatrix} \qquad A_3 = \begin{bmatrix} \mathbf{2} & \mathbf{2} \\ \mathbf{1} & \mathbf{1} \end{bmatrix}.$$

Solution If A has independent columns then $A^{\mathrm{T}}A$ is invertible—this is a key point of Section 4.2. Certainly $L = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$ multiplies A to give $LA = I$: a left inverse.

$AL = A(A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}$ is the projection matrix (Section 4.2) on the column space. So L meets the requirements on A^{+}: *LA and AL are projections* on $\boldsymbol{C}(A)$ and $\boldsymbol{C}(A^{\mathrm{T}})$.

If A has rank m (full row rank) then AA^{T} is invertible. Certainly A multiplies $R = A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}$ to give $AR = I$. In the opposite order, $RA = A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}A$ is the projection matrix onto the row space (column space of A^{T}). So R equals the pseudoinverse A^{+}.

The example A_1 has full column rank (for L) and A_2 has full row rank (for R):

$$\boldsymbol{A_1^{+}} = (A_1^{\mathrm{T}}A_1)^{-1}A_1^{\mathrm{T}} = \frac{1}{\sqrt{8}}\begin{bmatrix} 2 & 2 \end{bmatrix} \qquad \boldsymbol{A_2^{+}} = A_2^{\mathrm{T}}(A_2A_2^{\mathrm{T}})^{-1} = \frac{1}{\sqrt{8}}\begin{bmatrix} 2 \\ 2 \end{bmatrix}.$$

Notice $A_1^{+}A_1 = [1]$ and $A_2A_2^{+} = [1]$. But A_3 has no left or right inverse. **Its rank is not full. Its pseudoinverse brings the column space of A_3 to the row space.**

$$A_3^{+} = \begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix}^{+} = \frac{\boldsymbol{v}_1\boldsymbol{u}_1^{\mathrm{T}}}{\sigma_1} = \frac{1}{10}\begin{bmatrix} \mathbf{2} & \mathbf{1} \\ \mathbf{2} & \mathbf{1} \end{bmatrix}.$$

Problem Set 7.4

Problems 1–4 compute and use the SVD of a particular matrix (not invertible).

1 (a) Compute $A^{\mathrm{T}}A$ and its eigenvalues and unit eigenvectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. Find σ_1.

Rank one matrix $A = \begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix}$

(b) Compute AA^{T} and its eigenvalues and unit eigenvectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$.

(c) Verify that $A\boldsymbol{v}_1 = \sigma_1\boldsymbol{u}_1$. Put numbers into $\boldsymbol{A = U\Sigma V^{\mathrm{T}}}$ (this is the SVD).

2 (a) From the $\boldsymbol{u}$'s and $\boldsymbol{v}$'s in Problem 1 write down orthonormal bases for the four fundamental subspaces of this matrix A.

(b) Describe all matrices that have those same four subspaces. Multiples of A?

3 From U, V, and Σ in Problem 1 find the orthogonal matrix $Q = UV^{\mathrm{T}}$ and the symmetric matrix $S = V\Sigma V^{\mathrm{T}}$. Verify the polar decomposition $A = QS$. This S is only semidefinite because _____. Test $S^2 = A$.

4 Compute the pseudoinverse $A^{+} = V\Sigma^{+}U^{\mathrm{T}}$. The diagonal matrix Σ^{+} contains $1/\sigma_1$. Rename the four subspaces (for A) in Figure 7.6 as four subspaces for A^{+}. Compute the projections $A^{+}A$ and AA^{+} on the row and column spaces of A.

Problems 5–9 are about the SVD of an invertible matrix.

5 Compute $A^{\mathrm{T}}A$ and its eigenvalues and unit eigenvectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. What are the singular values σ_1 and σ_2 for this matrix A?

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

6 AA^{T} has the same eigenvalues σ_1^2 and σ_2^2 as $A^{\mathrm{T}}A$. Find unit eigenvectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$. Put numbers into the SVD:

$$A = \begin{bmatrix} 3 & 3 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 \end{bmatrix} \begin{bmatrix} \sigma_1 & \\ & \sigma_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{bmatrix}^{\mathrm{T}}.$$

7 In Problem 6, multiply columns times rows to show that $A = \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^{\mathrm{T}} + \sigma_2 \boldsymbol{u}_2 \boldsymbol{v}_2^{\mathrm{T}}$. Prove from $A = U\Sigma V^{\mathrm{T}}$ that every matrix of rank r is the sum of r matrices of rank one.

8 From U, V, and Σ find the orthogonal matrix $Q = UV^{\mathrm{T}}$ and the symmetric matrix $K = U\Sigma U^{\mathrm{T}}$. Verify the polar decomposition in reverse order $A = KQ$.

9 The pseudoinverse of this A is the same as _____ because _____.

Problems 10–11 compute and use the SVD of a 1 by 3 rectangular matrix.

10 Compute $A^{\mathrm{T}}A$ and AA^{T} and their eigenvalues and unit eigenvectors when the matrix is $A = \begin{bmatrix} 3 & 4 & 0 \end{bmatrix}$. What are the singular values of A?

11 Put numbers into the singular value decomposition of A:

$$A = \begin{bmatrix} 3 & 4 & 0 \end{bmatrix} = [\boldsymbol{u}_1] \begin{bmatrix} \sigma_1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \boldsymbol{v}_3 \end{bmatrix}^{\mathrm{T}}.$$

Put numbers into the pseudoinverse $V\Sigma^{+}U^{\mathrm{T}}$ of A. *Compute* AA^{+} *and* $A^{+}A$:

$$\textbf{Pseudoinverse} \quad A^{+} = \begin{bmatrix} & \\ & \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \boldsymbol{v}_3 \end{bmatrix} \begin{bmatrix} 1/\sigma_1 & \\ 0 & \\ 0 & \end{bmatrix} \begin{bmatrix} \boldsymbol{u}_1 \end{bmatrix}^{\mathrm{T}}.$$

12 What is the only 2 by 3 matrix that has no pivots and no singular values? What is Σ for that matrix? A^{+} is the zero matrix, but what is its shape?

13 If $\det A = 0$ why is $\det A^{+} = 0$? If A has rank r, why does A^{+} have rank r?

14 For vectors in the unit circle $||\boldsymbol{x}|| = 1$, the vectors $\boldsymbol{y} = A\boldsymbol{x}$ in the ellipse will have $||A^{-1}\boldsymbol{y}|| = 1$. This ellipse has axes along the singular vectors with lengths $= \sigma_1, \ldots, \sigma_r$ (as in Figure 7.5). Expand $||A^{-1}\boldsymbol{y}||^2 = 1$ for $A = [2\ \ 1\ ;1\ \ 2]$.

Problems 15–18 bring out the main properties of A^+ and $x^+ = A^+b$.

15 All matrices in this problem have rank one. The vector $\boldsymbol{b}$ is (b_1, b_2).

$$A = \begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix} \quad AA^{\mathrm{T}} = \begin{bmatrix} 8 & 4 \\ 4 & 2 \end{bmatrix} \quad A^{\mathrm{T}}A = \begin{bmatrix} 5 & 5 \\ 5 & 5 \end{bmatrix} \quad A^+ = \begin{bmatrix} .2 & .1 \\ .2 & .1 \end{bmatrix}$$

(a) The equation $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$ has many solutions because $A^{\mathrm{T}}A$ is _____.

(b) Verify that $\boldsymbol{x}^+ = A^+\boldsymbol{b} = (.2b_1 + .1b_2, .2b_1 + .1b_2)$ solves $A^{\mathrm{T}}A\boldsymbol{x}^+ = A^{\mathrm{T}}\boldsymbol{b}$.

(c) Add $(1, -1)$ to that $\boldsymbol{x}^+$ to get another solution to $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$. Show that $\|\widehat{\boldsymbol{x}}\|^2 = \|\boldsymbol{x}^+\|^2 + 2$, and $\boldsymbol{x}^+$ is shorter.

16 ***The vector $\boldsymbol{x}^+ = A^+\boldsymbol{b}$ is the shortest possible solution to $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$.*** Reason: The difference $\widehat{\boldsymbol{x}} - \boldsymbol{x}^+$ is in the nullspace of $A^{\mathrm{T}}A$. This is also the nullspace of A, orthogonal to $\boldsymbol{x}^+$. Explain how it follows that $\|\widehat{\boldsymbol{x}}\|^2 = \|\boldsymbol{x}^+\|^2 + \|\widehat{\boldsymbol{x}} - \boldsymbol{x}^+\|^2$.

17 Every $\boldsymbol{b}$ in $\mathbf{R}^m$ is $\boldsymbol{p} + \boldsymbol{e}$. This is the column space part plus the left nullspace part. Every $\boldsymbol{x}$ in $\mathbf{R}^n$ is $\boldsymbol{x}^+ + \boldsymbol{x}_n$. This is the row space part plus the nullspace part. Then

$$AA^+\boldsymbol{p} = \_\_\_\_ \qquad AA^+\boldsymbol{e} = \_\_\_\_ \qquad A^+A\boldsymbol{x}^+ = \_\_\_\_ \qquad A^+A\boldsymbol{x}_n = \_\_\_\_$$

18 Find A^+ and A^+A and AA^+ and $\boldsymbol{x}^+$ for this matrix $A = U\Sigma V^{\mathrm{T}}$ and these $\boldsymbol{b}$:

$$A = \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} .6 & -.8 \\ .8 & .6 \end{bmatrix} \begin{bmatrix} 5 \\ 0 \end{bmatrix} [1] \qquad \boldsymbol{b} = \begin{bmatrix} 3 \\ 4 \end{bmatrix} \text{ and } \boldsymbol{b} = \begin{bmatrix} -4 \\ 3 \end{bmatrix}.$$

19 A general 2 by 2 matrix A is determined by four numbers. If triangular, it is determined by three. If diagonal, by two. If a rotation, by one. If a unit eigenvector, also by one. Check that the total count is four for each factorization of A:

Four numbers in $\boldsymbol{LU}$ $\boldsymbol{LDU}$ $\boldsymbol{QR}$ $\boldsymbol{U\Sigma V^{\mathrm{T}}}$ $\boldsymbol{X\Lambda X^{-1}}$.

20 Following Problem 18, check that LDL^{T} and $Q\Lambda Q^{\mathrm{T}}$ are determined by *three* numbers. This is correct because the matrix is now _____.

21 From A and A^+ show that A^+A is correct and $(A^+A)^2 = A^+A =$ projection.

$$A = \sum_1^r \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathrm{T}} \qquad A^+ = \sum_1^r \frac{\boldsymbol{v}_i \boldsymbol{u}_i^{\mathrm{T}}}{\sigma_i} \qquad A^+A = \sum_1^r \boldsymbol{v}_i \boldsymbol{v}_i^{\mathrm{T}} \qquad AA^+ = \sum_1^r \boldsymbol{u}_i \boldsymbol{u}_i^{\mathrm{T}}$$

22 Each pair of singular vectors $\boldsymbol{v}$ and $\boldsymbol{u}$ has $A\boldsymbol{v} = \sigma\boldsymbol{u}$ and $A^{\mathrm{T}}\boldsymbol{u} = \sigma\boldsymbol{v}$. Show that the double vector $\begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{u} \end{bmatrix}$ is an eigenvector of the symmetric block matrix $M = \begin{bmatrix} 0 & A^{\mathrm{T}} \\ A & 0 \end{bmatrix}$. The SVD of A is equivalent to the diagonalization of that symmetric matrix M.

Chapter 8

Linear Transformations

8.1 The Idea of a Linear Transformation

1 A **linear transformation** T takes vectors $\boldsymbol{v}$ to vectors $T(\boldsymbol{v})$. Linearity requires $\boxed{T(c\,\boldsymbol{v}+d\,\boldsymbol{w}) = cT(\boldsymbol{v}) + dT(w)}$ Note $T(\mathbf{0}) = \mathbf{0}$ so $T(\boldsymbol{v}) = \boldsymbol{v} + \boldsymbol{u}_0$ is not linear.

2 The input vectors $\boldsymbol{v}$ and outputs $T(\boldsymbol{v})$ can be in $\mathbf{R}^{\boldsymbol{n}}$ or matrix space or function space.

3 If A is m by n, $T(\boldsymbol{x}) = A\boldsymbol{x}$ is linear from the input space $\mathbf{R}^{\boldsymbol{n}}$ to the output space $\mathbf{R}^{\boldsymbol{m}}$.

4 The derivative $T(f) = \dfrac{df}{dx}$ is linear. The integral $T^+(f) = \displaystyle\int_0^x f(t)\,dt$ is its pseudoinverse.

5 The product ST of two linear transformations is still linear: $\boxed{(ST)(\boldsymbol{v}) = S(T(\boldsymbol{v}))}$.

When a matrix A multiplies a vector $\boldsymbol{v}$, it "transforms" $\boldsymbol{v}$ into another vector $A\boldsymbol{v}$. ***In goes*** $\boldsymbol{v}$, ***out comes*** $T(\boldsymbol{v}) = A\boldsymbol{v}$. A transformation T follows the same idea as a function. In goes a number x, out comes $f(x)$. For one vector $\boldsymbol{v}$ or one number x, we multiply by the matrix or we evaluate the function. The deeper goal is to see all vectors $\boldsymbol{v}$ at once. We are transforming the whole space $\mathbf{V}$ when we multiply every $\boldsymbol{v}$ by A.

Start again with a matrix A. It transforms $\boldsymbol{v}$ to $A\boldsymbol{v}$. It transforms $\boldsymbol{w}$ to $A\boldsymbol{w}$. Then we *know* what happens to $\boldsymbol{u} = \boldsymbol{v} + \boldsymbol{w}$. There is no doubt about $A\boldsymbol{u}$, it has to equal $A\boldsymbol{v} + A\boldsymbol{w}$. Matrix multiplication $T(\boldsymbol{v}) = A\boldsymbol{v}$ gives a ***linear transformation*** :

A ***transformation*** T assigns an output $T(\boldsymbol{v})$ to each input vector $\boldsymbol{v}$ in $\mathbf{V}$.
The transformation is ***linear*** if it meets these requirements for all $\boldsymbol{v}$ and $\boldsymbol{w}$:

(a) $T(\boldsymbol{v} + \boldsymbol{w}) = T(\boldsymbol{v}) + T(\boldsymbol{w})$ (b) $T(c\boldsymbol{v}) = cT(\boldsymbol{v})$ for all c.

If the input is $\boldsymbol{v} = \mathbf{0}$, the output must be $T(\boldsymbol{v}) = \mathbf{0}$. We combine rules (a) and (b) into one:

Linear transformation $T(c\boldsymbol{v} + d\boldsymbol{w})$ ***must equal*** $cT(\boldsymbol{v}) + dT(\boldsymbol{w})$.

Again I can test matrix multiplication for linearity: $A(c\boldsymbol{v} + d\boldsymbol{w}) = cA\boldsymbol{v} + dA\boldsymbol{w}$ is *true*.

A linear transformation is highly restricted. Suppose T adds $\boldsymbol{u}_0$ to every vector. Then $T(\boldsymbol{v}) = \boldsymbol{v} + \boldsymbol{u}_0$ and $T(\boldsymbol{w}) = \boldsymbol{w} + \boldsymbol{u}_0$. This isn't good, or at least *it isn't linear*. Applying T to $\boldsymbol{v} + \boldsymbol{w}$ produces $\boldsymbol{v} + \boldsymbol{w} + \boldsymbol{u}_0$. That is not the same as $T(\boldsymbol{v}) + T(\boldsymbol{w})$:

Shift is not linear $\boldsymbol{v} + \boldsymbol{w} + \boldsymbol{u}_0$ is not $T(\boldsymbol{v}) + T(\boldsymbol{w}) = (\boldsymbol{v} + \boldsymbol{u}_0) + (\boldsymbol{w} + \boldsymbol{u}_0)$.

The exception is when $\boldsymbol{u}_0 = \mathbf{0}$. The transformation reduces to $T(\boldsymbol{v}) = \boldsymbol{v}$. This is the ***identity transformation*** (nothing moves, as in multiplication by the identity matrix). That is certainly linear. In this case the input space $\mathbf{V}$ is the same as the output space $\mathbf{W}$.

The linear-plus-shift transformation $T(\boldsymbol{v}) = A\boldsymbol{v} + \boldsymbol{u}_0$ is called *"affine"*. Straight lines stay straight although T is not linear. Computer graphics works with affine transformations in Section 10.6, because we must be able to move images.

Example 1 Choose a fixed vector $\boldsymbol{a} = (1, 3, 4)$, and let $T(\boldsymbol{v})$ be the dot product $\boldsymbol{a} \cdot \boldsymbol{v}$:

The input is $\boldsymbol{v} = (v_1, v_2, v_3)$. The output is $T(\boldsymbol{v}) = \boldsymbol{a} \cdot \boldsymbol{v} = v_1 + 3v_2 + 4v_3$.

Dot products are linear. The inputs $\boldsymbol{v}$ come from three-dimensional space, so $\mathbf{V} = \mathbf{R}^3$. The outputs are just numbers, so the output space is $\mathbf{W} = \mathbf{R}^1$. We are multiplying by the row matrix $A = [1 \quad 3 \quad 4]$. Then $T(\boldsymbol{v}) = A\boldsymbol{v}$.

You will get good at recognizing which transformations are linear. If the output involves squares or products or lengths, v_1^2 or $v_1 v_2$ or $\|\boldsymbol{v}\|$, then T is not linear.

Example 2 The length $T(\boldsymbol{v}) = \|\boldsymbol{v}\|$ is not linear. Requirement (a) for linearity would be $\|\boldsymbol{v} + \boldsymbol{w}\| = \|\boldsymbol{v}\| + \|\boldsymbol{w}\|$. Requirement (b) would be $\|c\boldsymbol{v}\| = c\|\boldsymbol{v}\|$. Both are false!

Not (a): The sides of a triangle satisfy an *inequality* $\|\boldsymbol{v} + \boldsymbol{w}\| \leq \|\boldsymbol{v}\| + \|\boldsymbol{w}\|$.
Not (b): The length $\| - \boldsymbol{v}\|$ is $\|\boldsymbol{v}\|$ and not $-\|\boldsymbol{v}\|$. For negative c, linearity fails.

Example 3 (Rotation) T is the transformation that *rotates every vector by* $30°$. The "*domain*" of T is the xy plane (all input vectors $\boldsymbol{v}$). The "*range*" of T is also the xy plane (all rotated vectors $T(\boldsymbol{v})$). We described T without a matrix: rotate the plane by $30°$.

Is rotation linear? *Yes it is.* We can rotate two vectors and add the results. The sum of rotations $T(\boldsymbol{v}) + T(\boldsymbol{w})$ is the same as the rotation $T(\boldsymbol{v} + \boldsymbol{w})$ of the sum. **The whole plane is turning together, in this linear transformation.**

Lines to Lines, Triangles to Triangles, Basis Tells All

Figure 8.1 shows the line from $\boldsymbol{v}$ to $\boldsymbol{w}$ in the input space. It also shows the line from $T(\boldsymbol{v})$ to $T(\boldsymbol{w})$ in the output space. Linearity tells us: Every point on the input line goes onto the output line. And more than that: ***Equally spaced points go to equally spaced points***. The middle point $\boldsymbol{u} = \frac{1}{2}\boldsymbol{v} + \frac{1}{2}\boldsymbol{w}$ goes to the middle point $T(\boldsymbol{u}) = \frac{1}{2}T(\boldsymbol{v}) + \frac{1}{2}T(\boldsymbol{w})$.

The second figure moves up a dimension. Now we have three corners $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, $\boldsymbol{v}_3$. Those inputs have three outputs $T(\boldsymbol{v}_1)$, $T(\boldsymbol{v}_2)$, $T(\boldsymbol{v}_3)$. *The input triangle goes onto the output triangle*. Equally spaced points stay equally spaced (along the edges, and then between the edges). The middle point $\boldsymbol{u} = \frac{1}{3}(\boldsymbol{v}_1 + \boldsymbol{v}_2 + \boldsymbol{v}_3)$ goes to the middle point $T(\boldsymbol{u}) = \frac{1}{3}(T(\boldsymbol{v}_1) + T(\boldsymbol{v}_2) + T(\boldsymbol{v}_3))$.

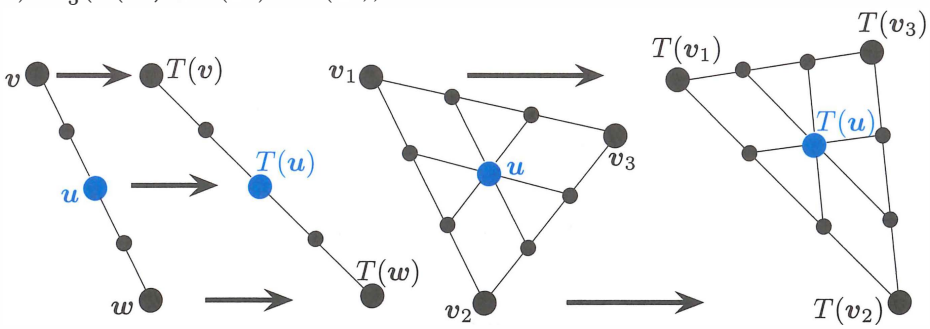


Figure 8.1: Lines to lines, equal spacing to equal spacing, $\boldsymbol{u} = \boldsymbol{0}$ to $T(\boldsymbol{u}) = \boldsymbol{0}$.

The rule of linearity extends to combinations of three vectors or n vectors:

Linearity $\boldsymbol{u} = c_1\boldsymbol{v}_1 + c_2\boldsymbol{v}_2 + \cdots + c_n\boldsymbol{v}_n$ **must transform to**

$$T(\boldsymbol{u}) = c_1T(\boldsymbol{v}_1) + c_2T(\boldsymbol{v}_2) + \cdots + c_nT(\boldsymbol{v}_n) \tag{1}$$

The 2-vector rule starts the 3-vector proof: $T(c\boldsymbol{u} + d\boldsymbol{v} + e\boldsymbol{w}) = T(c\boldsymbol{u}) + T(d\boldsymbol{v} + e\boldsymbol{w})$. Then linearity applies to both of those parts, to give $cT(\boldsymbol{u}) + dT(\boldsymbol{v}) + eT(\boldsymbol{w})$.

The n-vector rule (1) leads to the most important fact about linear transformations:

Suppose you know $T(v)$ for all vectors $v_1, \ldots, v_n$ in a basis

Then you know $T(u)$ for every vector u in the space.

You see the reason: Every $\boldsymbol{u}$ in the space is a combination of the basis vectors $\boldsymbol{v}_j$. Then linearity tells us that $T(\boldsymbol{u})$ is the same combination of the outputs $T(\boldsymbol{v}_j)$.

Example 4 **The transformation T takes the derivative of the input**: $T(\boldsymbol{u}) = \boldsymbol{du/dx}$.

How do you find the derivative of $u = 6 - 4x + 3x^2$? You start with the derivatives of 1, x, and x^2. Those are the basis vectors. Their derivatives are 0, 1, and $2x$. Then you use linearity for the derivative of any combination:

$$\frac{du}{dx} = 6\,(\text{derivative of } 1) - 4\,(\text{derivative of } x) + 3\,(\text{derivative of } x^2) = -4 + 6x.$$

All of calculus depends on linearity! Precalculus finds a few key derivatives, for x^n and $\sin x$ and $\cos x$ and e^x. Then linearity applies to all their combinations.

I would say that the only rule special to calculus is the *chain rule*. That produces the derivative of a chain of functions $f(g(x))$.

Nullspace of $T(u) = du/dx$. For the nullspace we solve $T(u) = 0$. The derivative is zero when u *is a constant function*. So the one-dimensional nullspace is a line in function space—all multiples of the special solution $u = 1$.

Column space of $T(u) = du/dx$. In our example the input space contains all quadratics $a + bx + cx^2$. The outputs (the column space) are all linear functions $b + 2cx$. Notice that the **Counting Theorem** is still true: $\boldsymbol{r} + (\boldsymbol{n} - \boldsymbol{r}) = \boldsymbol{n}$.

dimension (**column space**) + dimension (**nullspace**) = **2**+**1** = **3** = dimension (**input space**)

What is the matrix for d/dx? I can't leave derivatives without asking for a matrix. We have a linear transformation $T = d/dx$. We know what T does to the basis functions:

$$\boldsymbol{v_1, v_2, v_3 = 1, x, x^2} \qquad \frac{d\boldsymbol{v_1}}{dx} = \boldsymbol{0} \qquad \frac{d\boldsymbol{v_2}}{dx} = 1 = \boldsymbol{v_1} \qquad \frac{d\boldsymbol{v_3}}{dx} = 2x = \boldsymbol{2v_2}. \tag{2}$$

The 3-dimensional input space **V** (= quadratics) transforms to the 2-dimensional output space **W** (= linear functions). If $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$ were vectors, I would know the matrix.

$$\boxed{\boldsymbol{A} = \begin{bmatrix} 0 & \mathbf{1} & 0 \\ 0 & 0 & \mathbf{2} \end{bmatrix} = \text{matrix form of the derivative } T = \frac{d}{dx}.} \tag{3}$$

The linear transformation du/dx is perfectly copied by the matrix multiplication $A\boldsymbol{u}$.

Input $\boldsymbol{u}$ $a + bx + cx^2$ **Multiplication** $A\boldsymbol{u} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} b \\ 2c \end{bmatrix}$ **Output** $\boldsymbol{\frac{du}{dx}} = b + 2cx$.

The connection from T to A (we will connect every transformation to a matrix) depended on choosing an input basis $1, x, x^2$ and an output basis $1, x$.

Next we look at integrals. They give the pseudoinverse T^+ of the derivative! I can't write T^{-1} and I can't say "*inverse of* T" when the derivative of 1 is 0.

Example 5 **Integration T^+ is also linear**: $\int_0^x (D + Ex)\,dx = Dx + \frac{1}{2}Ex^2$.

The input basis is now $1, x$. The output basis is $1, x, x^2$. The matrix A^+ for T^+ is 3 by 2:

$$\textbf{Input } v \quad D + Ex \qquad \textbf{Multiplication } A^+v = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} D \\ E \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{D} \\ \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{E} \end{bmatrix} \qquad \textbf{Output} = \textbf{Integral of } v \quad T^+(v) = Dx + \tfrac{1}{2}Ex^2$$

The Fundamental Theorem of Calculus says that integration is the (pseudo)inverse of differentiation. For linear algebra, the matrix A^+ is the (pseudo)inverse of the matrix A:

$$A^+A = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} = \begin{bmatrix} \mathbf{0} & 0 & 0 \\ 0 & \mathbf{1} & 0 \\ 0 & 0 & \mathbf{1} \end{bmatrix} \quad \text{and} \quad AA^+ = \begin{bmatrix} \mathbf{1} & 0 \\ 0 & \mathbf{1} \end{bmatrix}. \tag{4}$$

The derivative of a constant function is zero. That zero is on the diagonal of A^+A. Calculus wouldn't be calculus without that 1-dimensional nullspace of $T = d/dx$.

Examples of Transformations (mostly linear)

Example 6 Project every 3-dimensional vector onto the horizontal plane $z = 1$. The vector $\boldsymbol{v} = (x, y, z)$ is transformed to $T(\boldsymbol{v}) = (x, y, 1)$. This transformation is not linear. Why not? It doesn't even transform $\boldsymbol{v} = \mathbf{0}$ into $T(\boldsymbol{v}) = \mathbf{0}$.

Example 7 Suppose A is an *invertible matrix*. Certainly $T(\boldsymbol{v} + \boldsymbol{w}) = A\boldsymbol{v} + A\boldsymbol{w} = T(\boldsymbol{v}) + T(\boldsymbol{w})$. Another linear transformation is multiplication by A^{-1}. This produces the ***inverse transformation*** T^{-1}, which brings every vector $T(\boldsymbol{v})$ back to $\boldsymbol{v}$:

$$T^{-1}(T(\boldsymbol{v})) = \boldsymbol{v} \quad \text{matches the matrix multiplication} \quad A^{-1}(A\boldsymbol{v}) = \boldsymbol{v}.$$

If $T(\boldsymbol{v}) = A\boldsymbol{v}$ and $S(\boldsymbol{u}) = B\boldsymbol{u}$, then the product $T(S(u))$ matches the product $AB\boldsymbol{u}$.

We are reaching an unavoidable question. ***Are all linear transformations from $\mathbf{V} = \mathbf{R}^n$ to $\mathbf{W} = \mathbf{R}^m$ produced by matrices?*** When a linear T is described as a "rotation" or "projection" or ". . .", is there always a matrix A hiding behind T? Is $T(\boldsymbol{v})$ always $A\boldsymbol{v}$?

The answer is *yes*! This is an approach to linear algebra that doesn't start with matrices. We still end up with matrices—*after we choose an input basis and output basis*.

Note Transformations have a language of their own. For a matrix, the column space contains all outputs $A\boldsymbol{v}$. The nullspace contains all inputs for which $A\boldsymbol{v} = \mathbf{0}$. Translate those words into "*range*" and "*kernel*":

Range of T = set of *all outputs* $T(\boldsymbol{v})$. Range corresponds to column space.

Kernel of T = set of *all inputs for which* $T(\boldsymbol{v}) = \mathbf{0}$. Kernel corresponds to nullspace.

The range is in the output space $\mathbf{W}$. The kernel is in the input space $\mathbf{V}$. When T is multiplication by a matrix, $T(\boldsymbol{v}) = A\boldsymbol{v}$, range is column space and kernel is nullspace.

Linear Transformations of the Plane

It is more interesting to *see* a transformation than to define it. When a 2 by 2 matrix A multiplies all vectors in $\mathbf{R}^2$, we can watch how it acts. Start with a "house" that has eleven endpoints. Those eleven vectors $\boldsymbol{v}$ are transformed into eleven vectors $A\boldsymbol{v}$. Straight lines between $\boldsymbol{v}$'s become straight lines between the transformed vectors $A\boldsymbol{v}$. (The transformation from house to house is linear!) Applying A to a standard house produces a new house—possibly stretched or rotated or otherwise unlivable.

This part of the book is visual, not theoretical. We will show four houses and the matrices that produce them. The columns of H are the eleven corners of the first house. (H is 2 by 12, so **plot2d** in Problem 25 will connect the 11th corner to the first.) A multiplies the 11 points in the house matrix H to produce the corners AH of the other houses.

House matrix
$$H = \begin{bmatrix} -6 & -6 & -7 & 0 & 7 & 6 & 6 & -3 & -3 & 0 & 0 & -6 \\ -7 & 2 & 1 & 8 & 1 & 2 & -7 & -7 & -2 & -2 & -7 & -7 \end{bmatrix}.$$

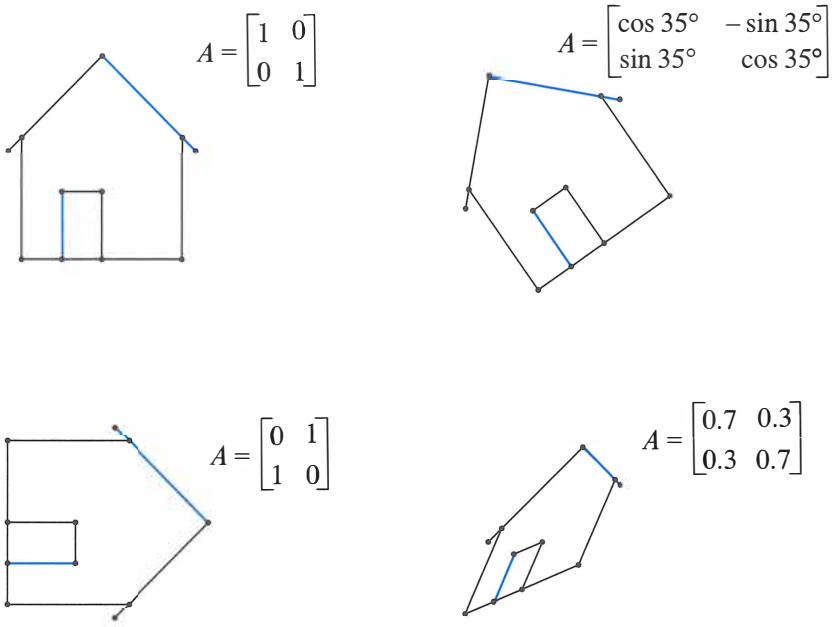


Figure 8.2: Linear transformations of a house drawn by **plot2d**$(A * H)$.

■ REVIEW OF THE KEY IDEAS ■

1. A transformation T takes each $\boldsymbol{v}$ in the input space to $T(\boldsymbol{v})$ in the output space.
2. T is **linear** if $T(\boldsymbol{v}+\boldsymbol{w}) = T(\boldsymbol{v}) + T(w)$ and $T(c\boldsymbol{v}) = cT(\boldsymbol{v})$: lines to lines.
3. Combinations to combinations: $T(c_1\boldsymbol{v}_1 + \cdots + c_n\boldsymbol{v}_n) = c_1\,T(\boldsymbol{v}_1) + \cdots + c_n\,T(\boldsymbol{v}_n)$.
4. $T =$ *derivative* and $T^+ =$ *integral* are linear. So is $T(\boldsymbol{v}) = A\boldsymbol{v}$ from $\mathbf{R}^n$ to $\mathbf{R}^m$.

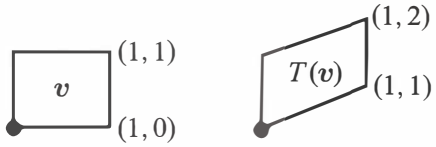
■ WORKED EXAMPLES ■

8.1 A The elimination matrix $\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ gives a ***shearing transformation*** from (x, y) to $\boldsymbol{T(x, y) = (x, x + y)}$. If the inputs fill a square, draw the transformed square.

Solution The points $(1, 0)$ and $(2, 0)$ on the x axis transform by T to $(1, 1)$ and $(2, 2)$ on the $45°$ line. Points on the y axis are *not moved*: $T(0, y) = (0, y) =$ eigenvectors with $\lambda = 1$.

Vertical lines slide up
This is the shearing
Squares go to parallelograms

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



8.1 B A **nonlinear transformation** T is invertible if every $\boldsymbol{b}$ in the output space comes from exactly one $\boldsymbol{x}$ in the input space: $T(\boldsymbol{x}) = \boldsymbol{b}$ always has exactly one solution. Which of these transformations (on real numbers x) is invertible and what is T^{-1}? ***None are linear, not even T_3.*** When you solve $T(\boldsymbol{x}) = \boldsymbol{b}$, you are inverting T:

$$T_1(x) = x^2 \quad T_2(x) = x^3 \quad T_3(x) = x + 9 \quad T_4(x) = e^x \quad T_5(x) = \frac{1}{x} \quad \textbf{for nonzero } x\textbf{'s}$$

Solution T_1 is not invertible: $x^2 = 1$ has *two* solutions and $x^2 = -1$ has *no* solution.
T_4 is not invertible because $e^x = -1$ has no solution. (If the output space changes to *positive* b's then the inverse of $e^x = b$ is $x = \ln b$.)

Notice $T_5^2 =$ identity. But $T_3^2(x) = x + 18$. What are $T_2^2(x)$ and $T_4^2(x)$?

T_2, T_3, T_5 are invertible: $x^3 = b$ and $x + 9 = b$ and $\frac{1}{x} = b$ have one solution x.

$$x = T_2^{-1}(b) = b^{1/3} \qquad x = T_3^{-1}(b) = b - 9 \qquad x = T_5^{-1}(b) = 1/b$$

Problem Set 8.1

1 A linear transformation must leave the zero vector fixed: $T(\mathbf{0}) = \mathbf{0}$. Prove this from $T(\boldsymbol{v} + \boldsymbol{w}) = T(\boldsymbol{v}) + T(\boldsymbol{w})$ by choosing $\boldsymbol{w} =$ _____ (and finish the proof). Prove it also from $T(c\boldsymbol{v}) = cT(\boldsymbol{v})$ by choosing $c =$ _____.

2 Requirement (b) gives $T(c\boldsymbol{v}) = cT(\boldsymbol{v})$ and also $T(d\boldsymbol{w}) = dT(\boldsymbol{w})$. Then by addition, requirement (a) gives $T(\quad) = (\quad)$. What is $T(c\boldsymbol{v} + d\boldsymbol{w} + e\boldsymbol{u})$?

3 Which of these transformations are not linear? The input is $\boldsymbol{v} = (v_1, v_2)$:

(a) $T(\boldsymbol{v}) = (v_2, v_1)$ (b) $T(\boldsymbol{v}) = (v_1, v_1)$ (c) $T(\boldsymbol{v}) = (0, v_1)$

(d) $T(\boldsymbol{v}) = (0, 1)$ (e) $T(\boldsymbol{v}) = v_1 - v_2$ (f) $T(\boldsymbol{v}) = v_1 v_2$.

4 If S and T are linear transformations, is $T(S(\boldsymbol{v}))$ linear or quadratic?

(a) (Special case) If $S(\boldsymbol{v}) = \boldsymbol{v}$ and $T(\boldsymbol{v}) = \boldsymbol{v}$, then $T(S(\boldsymbol{v})) = \boldsymbol{v}$ or $\boldsymbol{v}^2$?

(b) (General case) $S(\boldsymbol{v}_1+\boldsymbol{v}_2) = S(\boldsymbol{v}_1)+S(\boldsymbol{v}_2)$ and $T(\boldsymbol{v}_1+\boldsymbol{v}_2) = T(\boldsymbol{v}_1)+T(\boldsymbol{v}_2)$ combine into

$$T(S(\boldsymbol{v}_1 + \boldsymbol{v}_2)) = T(\_\_\_\_) = \_\_\_\_ + \_\_\_\_.$$

5 Suppose $T(\boldsymbol{v}) = \boldsymbol{v}$ except that $T(0, v_2) = (0, 0)$. Show that this transformation satisfies $T(c\boldsymbol{v}) = cT(\boldsymbol{v})$ but does not satisfy $T(\boldsymbol{v} + \boldsymbol{w}) = T(\boldsymbol{v}) + T(\boldsymbol{w})$.

6 Which of these transformations satisfy $T(\boldsymbol{v}+\boldsymbol{w}) = T(\boldsymbol{v})+T(\boldsymbol{w})$ and which satisfy $T(c\boldsymbol{v}) = cT(\boldsymbol{v})$?

(a) $T(\boldsymbol{v}) = \boldsymbol{v}/\|\boldsymbol{v}\|$ (b) $T(\boldsymbol{v}) = v_1+v_2+v_3$ (c) $T(\boldsymbol{v}) = (v_1, 2v_2, 3v_3)$

(d) $T(\boldsymbol{v}) =$ largest component of $\boldsymbol{v}$.

7 For these transformations of $\mathbf{V} = \mathbf{R}^2$ to $\mathbf{W} = \mathbf{R}^2$, find $T(T(\boldsymbol{v}))$. Show that when $T(\boldsymbol{v})$ is linear, then also $T(T(\boldsymbol{v}))$ is linear.

(a) $T(\boldsymbol{v}) = -\boldsymbol{v}$ (b) $T(\boldsymbol{v}) = \boldsymbol{v} + (1,1)$

(c) $T(\boldsymbol{v}) = 90°$ rotation $= (-v_2, v_1)$

(d) $T(\boldsymbol{v}) =$ projection $= \frac{1}{2}(v_1 + v_2,\ v_1 + v_2)$.

8 Find the range and kernel (like the column space and nullspace) of T:

(a) $T(v_1, v_2) = (v_1 - v_2, 0)$ (b) $T(v_1, v_2, v_3) = (v_1, v_2)$

(c) $T(v_1, v_2) = (0,0)$ (d) $T(v_1, v_2) = (v_1, v_1)$.

9 The "cyclic" transformation T is defined by $T(v_1, v_2, v_3) = (v_2, v_3, v_1)$. What is $T(T(\boldsymbol{v}))$? What is $T^3(\boldsymbol{v})$? What is $T^{100}(\boldsymbol{v})$? Apply T a hundred times to $\boldsymbol{v}$.

10 A linear transformation from $\mathbf{V}$ to $\mathbf{W}$ has an *inverse* from $\mathbf{W}$ to $\mathbf{V}$ when the range is all of $\mathbf{W}$ and the kernel contains only $\boldsymbol{v} = \mathbf{0}$. Then $T(\boldsymbol{v}) = \boldsymbol{w}$ has one solution $\boldsymbol{v}$ for each $\boldsymbol{w}$ in $\mathbf{W}$. Why are these T's not invertible?

(a) $T(v_1, v_2) = (v_2, v_2)$ $\mathbf{W} = \mathbf{R}^2$

(b) $T(v_1, v_2) = (v_1, v_2, v_1 + v_2)$ $\mathbf{W} = \mathbf{R}^3$

(c) $T(v_1, v_2) = v_1$ $\mathbf{W} = \mathbf{R}^1$

11 If $T(\boldsymbol{v}) = A\boldsymbol{v}$ and A is m by n, then T is "multiplication by A."

(a) What are the input and output spaces $\mathbf{V}$ and $\mathbf{W}$?

(b) Why is range of T = column space of A?

(c) Why is kernel of T = nullspace of A?

12 Suppose a linear T transforms $(1,1)$ to $(2,2)$ and $(2,0)$ to $(0,0)$. Find $T(v)$:

(a) $v = (2,2)$ (b) $v = (3,1)$ (c) $v = (-1,1)$ (d) $v = (a,b)$.

Problems 13-19 may be harder. The input space V contains all 2 by 2 matrices M.

13 M is any 2 by 2 matrix and $A = \left[\begin{smallmatrix} 1 & 2 \\ 3 & 4 \end{smallmatrix}\right]$. The transformation T is defined by $T(M) = AM$. What rules of matrix multiplication show that T is linear?

14 Suppose $A = \left[\begin{smallmatrix} 1 & 2 \\ 3 & 5 \end{smallmatrix}\right]$. Show that the range of T is the whole matrix space V and the kernel is the zero matrix:

(1) If $AM = 0$ prove that M must be the zero matrix.

(2) Find a solution to $AM = B$ for any 2 by 2 matrix B.

15 Suppose $A = \left[\begin{smallmatrix} 1 & 2 \\ 3 & 6 \end{smallmatrix}\right]$. Show that the identity matrix I is not in the range of T. Find a nonzero matrix M such that $T(M) = AM$ is zero.

16 Suppose T transposes every 2 by 2 matrix M. Try to find a matrix A which gives $AM = M^{\mathrm{T}}$. *Show that no matrix A will do it. To professors:* Is this a linear transformation that doesn't come from a matrix? The matrix should be 4 by 4!

17 The transformation T that transposes every 2 by 2 matrix is definitely linear. Which of these extra properties are true?

(a) $T^2 =$ identity transformation.

(b) The kernel of T is the zero matrix.

(c) Every 2 by 2 matrix is in the range of T.

(d) $T(M) = -M$ is impossible.

18 Suppose $T(M) = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right] M \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right]$. Find a matrix with $T(M) \neq 0$. Describe all matrices with $T(M) = 0$ (the kernel) and all output matrices $T(M)$ (the range).

19 If A and B are invertible and $T(M) = AMB$, find $T^{-1}(M)$ in the form $(\quad)M(\quad)$.

Questions 20–26 are about house transformations. The output is $T(H) = AH$.

20 How can you tell from the picture of T (house) that A is

(a) a diagonal matrix?

(b) a rank-one matrix?

(c) a lower triangular matrix?

21 Draw a picture of T (house) for these matrices:

$$D = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} .7 & .7 \\ .3 & .3 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

22 What are the conditions on $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ to ensure that T (house) will

(a) sit straight up?

(b) expand the house by 3 in all directions?

(c) rotate the house with no change in its shape?

23 Describe T (house) when $T(\boldsymbol{v}) = -\boldsymbol{v} + (1, 0)$. This T is "affine".

24 Change the house matrix H to add a chimney.

25 The standard house is drawn by **plot2d**(H). Circles from o and lines from −:

$$x = H(1,:)'; y = H(2,:)';$$
$$\text{axis}([-10\,10\,-10\,10]), \text{axis('square')}$$
$$\text{plot}(x, y, 'o', x, y, '-');$$

Test **plot2d**(A′ ∗ H) and **plot2d**(A′ ∗ A ∗ H) with the matrices in Figure 8.1.

26 Without a computer sketch the houses $A * H$ for these matrices A:

$$\begin{bmatrix} 1 & 0 \\ 0 & .1 \end{bmatrix} \text{ and } \begin{bmatrix} .5 & .5 \\ .5 & .5 \end{bmatrix} \text{ and } \begin{bmatrix} .5 & .5 \\ -.5 & .5 \end{bmatrix} \text{ and } \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}.$$

27 This code creates a vector theta of 50 angles. It draws the unit circle and then it draws T (circle) = ellipse. $T(\boldsymbol{v}) = A\boldsymbol{v}$ **takes circles to ellipses**.

```
A = [2 1;1 2]    % You can change A
theta = [0:2 * pi/50:2 * pi];
circle = [cos(theta); sin(theta)];
ellipse = A * circle;
axis([-4 4 -4 4]); axis('square')
plot(circle(1,:), circle(2,:), ellipse(1,:), ellipse(2,:))
```

28 Add two eyes and a smile to the circle in Problem 27. (If one eye is dark and the other is light, you can tell when the face is reflected across the y axis.) Multiply by matrices A to get new faces.

29 What conditions on $\det A = ad - bc$ ensure that the output house AH will

(a) be squashed onto a line?

(b) keep its endpoints in clockwise order (not reflected)?

(c) have the same area as the original house?

30 Why does every linear transformation T from $\mathbf{R}^2$ to $\mathbf{R}^2$ take squares to parallelograms? Rectangles also go to parallelograms (squashed if T is not invertible).

8.2 The Matrix of a Linear Transformation

1 We know all $T(v)$ if we know $T(v_1), \ldots, T(v_n)$ for an input basis $v_1, \ldots, v_n$: use **linearity**.

2 Column j in the "matrix for T" comes from applying T to the input basis vector v_j.

3 Write $T(v_j) = a_{1j}w_1 + \cdots + a_{mj}w_m$ in the output basis of w's. Those a_{ij} go into column j.

4 The matrix for $T(x) = Ax$ is A, *if* the input and output bases = columns of $I_{n\times n}$ and $I_{m\times m}$.

5 When the bases change to v's and w's, the matrix for the same T changes from A to $W^{-1}AV$.

6 Best bases: $V = W =$ eigenvectors and $V, W =$ singular vectors give diagonal Λ and Σ.

The next pages assign a matrix A to every linear transformation T. For ordinary column vectors, the input v is in $\mathbf{V} = \mathbf{R}^n$ and the output $T(v)$ is in $\mathbf{W} = \mathbf{R}^m$. The matrix A for this transformation will be m by n. Our choice of bases in $\mathbf{V}$ and $\mathbf{W}$ will decide A.

The standard basis vectors for $\mathbf{R}^n$ and $\mathbf{R}^m$ are the columns of I. That choice leads to a standard matrix. Then $T(v) = Av$ in the normal way. But these spaces also have other bases, so *the same transformation T is represented by other matrices.* A main theme of linear algebra is to choose the bases that give the best matrix (a diagonal matrix) for T.

All vector spaces $\mathbf{V}$ and $\mathbf{W}$ have bases. Each choice of those bases leads to a matrix for T. When the input basis is different from the output basis, the matrix for $T(v) = v$ will not be the identity I. It will be the "change of basis matrix". Here is the key idea:

Suppose we know $T(v)$ for the input basis vectors v_1 to v_n.
Columns 1 to n of the matrix will contain those outputs $T(v_1)$ to $T(v_n)$.
A times c = matrix times vector = combination of those n columns.
Ac is the correct combination $c_1T(v_1) + \cdots + c_nT(v_n) = T(v)$.

Reason Every v is a unique combination $c_1v_1 + \cdots + c_nv_n$ of the basis vectors v_j. Since T is a linear transformation (here is the moment for linearity), $T(v)$ must be **the same combination $c_1T(v_1) + \cdots + c_nT(v_n)$ of the outputs $T(v_j)$ in the columns.**

Our first example gives the matrix A for the standard basis vectors in $\mathbf{R}^2$ and $\mathbf{R}^3$.

Example 1 Suppose T transforms $v_1 = (1,0)$ to $T(v_1) = (2,3,4)$. Suppose the second basis vector $v_2 = (0,1)$ goes to $T(v_2) = (5,5,5)$. If T is linear from $\mathbf{R}^2$ to $\mathbf{R}^3$ then its "standard matrix" is 3 by 2. Those outputs $T(v_1)$ and $T(v_2)$ go into the columns of A:

$$A = \begin{bmatrix} 2 & 5 \\ 3 & 5 \\ 4 & 5 \end{bmatrix} \qquad c_1 = 1 \text{ and } c_2 = 1 \text{ give } T(v_1 + v_2) = \begin{bmatrix} 2 & 5 \\ 3 & 5 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 7 \\ 8 \\ 9 \end{bmatrix}.$$

Change of Basis

Example 2 Suppose the input space $\mathbf{V} = \mathbf{R}^2$ is also the output space $\mathbf{W} = \mathbf{R}^2$. *Suppose that* $T(\boldsymbol{v}) = \boldsymbol{v}$ *is the identity transformation.* You might expect its matrix to be I, but that only happens when the input basis is the same as the output basis. I will choose different bases to see how the matrix is constructed.

For this special case $T(\boldsymbol{v}) = \boldsymbol{v}$, I will call the matrix B instead of A. We are just changing basis from the $\boldsymbol{v}$'s to the $\boldsymbol{w}$'s. Each $\boldsymbol{v}$ is a combination of $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$.

$$\textbf{Input basis } \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 3 & 8 \end{bmatrix} \quad \textbf{Output basis } \begin{bmatrix} \boldsymbol{w}_1 & \boldsymbol{w}_2 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 1 & 2 \end{bmatrix} \quad \textbf{Change of basis } \begin{matrix} \boldsymbol{v}_1 = \mathbf{1}\boldsymbol{w}_1 + \mathbf{1}\boldsymbol{w}_2 \\ \boldsymbol{v}_2 = \mathbf{2}\boldsymbol{w}_1 + \mathbf{3}\boldsymbol{w}_2 \end{matrix}$$

Please notice! I wrote the input basis $\boldsymbol{v}_1, \boldsymbol{v}_2$ in terms of the output basis $\boldsymbol{w}_1, \boldsymbol{w}_2$. That is because of our key rule. We apply the identity transformation T to each input basis vector: $T(\boldsymbol{v}_1) = \boldsymbol{v}_1$ and $T(\boldsymbol{v}_2) = \boldsymbol{v}_2$. **Then we write those outputs $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ in the output basis $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$.** Those bold numbers $\mathbf{1}, \mathbf{1}$ and $\mathbf{2}, \mathbf{3}$ tell us column 1 and column 2 of the matrix B (the change of basis matrix): $WB = V$ so $\boldsymbol{B} = \boldsymbol{W}^{-1}\boldsymbol{V}$.

$$\textbf{Matrix } B \textbf{ for change of basis } \begin{bmatrix} \boldsymbol{w}_1 & \boldsymbol{w}_2 \end{bmatrix} \begin{bmatrix} B \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{bmatrix} \text{ is } \begin{bmatrix} 3 & 0 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} \mathbf{1} & \mathbf{2} \\ \mathbf{1} & \mathbf{3} \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 3 & 8 \end{bmatrix}. \quad (1)$$

When the input basis is in the columns of a matrix $\boldsymbol{V}$, and the output basis is in the columns of $\boldsymbol{W}$, the change of basis matrix for $T = I$ is $\boldsymbol{B} = \boldsymbol{W}^{-1}\boldsymbol{V}$.

The key I see a clear way to understand that rule $B = W^{-1}V$. Suppose the same vector $\boldsymbol{u}$ is written in the input basis of $\boldsymbol{v}$'s and the output basis of $\boldsymbol{w}$'s. I will do that three ways:

$$\begin{matrix} \boldsymbol{u} = c_1\boldsymbol{v}_1 + \cdots + c_n\boldsymbol{v}_n \\ \boldsymbol{u} = d_1\boldsymbol{w}_1 + \cdots + d_n\boldsymbol{w}_n \end{matrix} \text{ is } \begin{bmatrix} \boldsymbol{v}_1 \cdots \boldsymbol{v}_n \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}_1 \cdots \boldsymbol{w}_n \end{bmatrix} \begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} \text{ and } \boldsymbol{V}\boldsymbol{c} = \boldsymbol{W}\boldsymbol{d}.$$

The coefficients $\boldsymbol{d}$ in the new basis of $\boldsymbol{w}$'s are $\boldsymbol{d} = \boldsymbol{W}^{-1}\boldsymbol{V}\boldsymbol{c}$. Then $\boldsymbol{B}$ is $\boldsymbol{W}^{-1}\boldsymbol{V}$. (2)

This formula $\boldsymbol{B} = \boldsymbol{W}^{-1}\boldsymbol{V}$ produces one of the world's greatest mysteries: When the *standard basis* $\boldsymbol{V} = \boldsymbol{I}$ is changed to a different basis $\boldsymbol{W}$, **the change of basis matrix is not $\boldsymbol{W}$ but $\boldsymbol{B} = \boldsymbol{W}^{-1}$.** Larger basis vectors have smaller coefficients!

$$\begin{bmatrix} x \\ y \end{bmatrix} \text{ in the standard basis has coefficients } \begin{bmatrix} \boldsymbol{w}_1 & \boldsymbol{w}_2 \end{bmatrix}^{-1} \begin{bmatrix} x \\ y \end{bmatrix} \text{ in the } \boldsymbol{w}_1, \boldsymbol{w}_2 \text{ basis.}$$

Construction of the Matrix

Now we construct a matrix for any linear transformation. Suppose T transforms the space $\mathbf{V}$ (n-dimensional) to the space $\mathbf{W}$ (m-dimensional). We choose a basis $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ for $\mathbf{V}$ and we choose a basis $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m$ for $\mathbf{W}$. The matrix A will be m by n. To find the first column of A, apply T to the first basis vector $\boldsymbol{v}_1$. The output $T(\boldsymbol{v}_1)$ is in $\mathbf{W}$.

$T(\boldsymbol{v}_1)$ ***is a combination*** $a_{11}\boldsymbol{w}_1 + \cdots + a_{m1}\boldsymbol{w}_m$ ***of the output basis for*** $\mathbf{W}$.

These numbers $a_{11}, \ldots, a_{m1}$ *go into the first column of* A. Transforming $\boldsymbol{v}_1$ to $T(\boldsymbol{v}_1)$ matches multiplying $(1, 0, \ldots, 0)$ by A. It yields that first column of the matrix. When T is the derivative and the first basis vector is 1, its derivative is $T(\boldsymbol{v}_1) = \mathbf{0}$. So for the derivative matrix below, the first column of A is all zero.

Example 3 The input basis of $\boldsymbol{v}$'s is $1, x, x^2, x^3$. The output basis of $\boldsymbol{w}$'s is $1, x, x^2$. Then $\boldsymbol{T}$ **takes the derivative**: $\boldsymbol{T(v) = \dfrac{dv}{dx}}$ and A = "derivative matrix".

If $\boldsymbol{v} = c_1 + c_2 x + c_3 x^2 + c_4 x^3$
then $\dfrac{d\boldsymbol{v}}{dx} = \mathbf{1}c_2 + \mathbf{2}c_3 x + \mathbf{3}c_4 x^2$

$$A\boldsymbol{c} = \begin{bmatrix} 0 & \mathbf{1} & 0 & 0 \\ 0 & 0 & \mathbf{2} & 0 \\ 0 & 0 & 0 & \mathbf{3} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} c_2 \\ 2c_3 \\ 3c_4 \end{bmatrix}$$

Key rule: The jth column of A is found by applying T to the jth basis vector $\boldsymbol{v}_j$

$$T(\boldsymbol{v}_j) = \text{combination of output basis vectors} = a_{1j}\boldsymbol{w}_1 + \cdots + a_{mj}\boldsymbol{w}_m. \tag{3}$$

These numbers a_{ij} go into A. *The matrix is constructed to get the basis vectors right.* ***Then linearity gets all other vectors right.*** Every $\boldsymbol{v}$ is a combination $c_1\boldsymbol{v}_1 + \cdots + c_n\boldsymbol{v}_n$, and $T(\boldsymbol{v})$ is a combination of the $\boldsymbol{w}$'s. When A multiplies the vector $\boldsymbol{c} = (c_1, \ldots, c_n)$ in the $\boldsymbol{v}$ combination, $A\boldsymbol{c}$ produces the coefficients in the $T(\boldsymbol{v})$ combination. This is because matrix multiplication (combining columns) is linear like T.

The matrix A tells us what T does. Every linear transformation from $\mathbf{V}$ to $\mathbf{W}$ can be converted to a matrix. This matrix depends on the bases.

Example 4 For the integral $T^+(\boldsymbol{v})$, the first basis function is again 1. Its integral is the second basis function x. So the first column of the "integral matrix" A^+ is $(0, 1, 0, 0)$.

The integral of $d_1 + d_2 x + d_3 x^2$
is $d_1 x + \dfrac{1}{2}d_2 x^2 + \dfrac{1}{3}d_3 x^3$

$$A^+\boldsymbol{d} = \begin{bmatrix} 0 & 0 & 0 \\ \mathbf{1} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{d_1} \\ \frac{1}{2}\boldsymbol{d_2} \\ \frac{1}{3}\boldsymbol{d_3} \end{bmatrix}$$

If you integrate a function and then differentiate, you get back to the start. So $AA^+ = I$. But if you differentiate before integrating, the constant term is lost. So A^+A is not I. ***The integral of the derivative of*** 1 ***is zero:***

$$T^+T(1) = \text{integral of zero function} = 0.$$

This matches A^+A, whose first column is all zero. The derivative T has a kernel (the constant functions). Its matrix A has a nullspace. Main idea again: $A\boldsymbol{v}$ copies $T(\boldsymbol{v})$.

The examples of the derivative and integral made three points. First, linear transformations T are everywhere—in calculus and differential equations and linear algebra. Second, spaces other than $\mathbf{R}^n$ are important—we had functions in $\mathbf{V}$ and $\mathbf{W}$. Third, **if we differentiate and then integrate, we can multiply their matrices $\boldsymbol{A^+A}$.**

Matrix Products AB Match Transformations TS

We have come to something important—the real reason for the rule to multiply matrices. *At last we discover why*! Two linear transformations T and S are represented by two matrices A and B. Now compare TS with the multiplication AB:

When we apply the transformation T to the output from S, we get TS by this rule: $(TS)(\boldsymbol{u})$ *is defined to be* $\boldsymbol{T(S(u))}$. The output $S(\boldsymbol{u})$ becomes the input to T.

When we apply the matrix A to the output from B, we multiply AB by this rule: $(AB)(\boldsymbol{x})$ is defined to be $\boldsymbol{A(Bx)}$. The output $B\boldsymbol{x}$ becomes the input to A.
Matrix multiplication gives the correct matrix $\boldsymbol{AB}$ to represent $\boldsymbol{TS}$.

The transformation S is from a space $\mathbf{U}$ to $\mathbf{V}$. Its matrix B uses a basis $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p$ for $\mathbf{U}$ and a basis $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ for $\mathbf{V}$. That matrix is n by p. The transformation T is from $\mathbf{V}$ to $\mathbf{W}$ as before. *Its matrix A must use the same basis $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ for* $\mathbf{V}$—this is the output space for S and the input space for T. ***Then the matrix $\boldsymbol{AB}$ matches $\boldsymbol{TS}$.***

Multiplication The linear transformation TS starts with any vector $\boldsymbol{u}$ in $\mathbf{U}$, goes to $S(\boldsymbol{u})$ in $\mathbf{V}$ and then to $T(S(\boldsymbol{u}))$ in $\mathbf{W}$. The matrix AB starts with any $\boldsymbol{x}$ in $\mathbf{R}^p$, goes to $B\boldsymbol{x}$ in $\mathbf{R}^n$ and then to $AB\boldsymbol{x}$ in $\mathbf{R}^m$. **The matrix $\boldsymbol{AB}$ correctly represents $\boldsymbol{TS}$:**

TS: $\mathbf{U} \to \mathbf{V} \to \mathbf{W}$ $\qquad$ AB: $(m \text{ by } n)(n \text{ by } p) = (m \text{ by } p)$.

The input is $\boldsymbol{u} = x_1\boldsymbol{u}_1 + \cdots + x_p\boldsymbol{u}_p$. The output $T(S(\boldsymbol{u}))$ matches the output $AB\boldsymbol{x}$. ***Product of transformations $\boldsymbol{TS}$ matches product of matrices $\boldsymbol{AB}$.***

The most important cases are when the spaces $\mathbf{U}, \mathbf{V}, \mathbf{W}$ are the same and their bases are the same. With $m = n = p$ we have square matrices that we can multiply.

Example 5 S rotates the plane by θ and T also rotates by θ. Then TS rotates by 2θ. This transformation T^2 corresponds to the rotation matrix A^2 through 2θ:

$$T = S \qquad A = B \qquad T^2 = \text{rotation by } 2\theta \qquad A^2 = \begin{bmatrix} \cos 2\theta & -\sin 2\theta \\ \sin 2\theta & \cos 2\theta \end{bmatrix}. \tag{4}$$

By matching (transformation)^2 with (matrix)^2, we pick up the formulas for $\cos 2\theta$ and $\sin 2\theta$. Multiply A times A:

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} \cos^2\theta - \sin^2\theta & -2\sin\theta\cos\theta \\ 2\sin\theta\cos\theta & \cos^2\theta - \sin^2\theta \end{bmatrix}. \quad (5)$$

Comparing (4) with (5) produces $\cos 2\theta = \cos^2\theta - \sin^2\theta$ and $\sin 2\theta = 2\sin\theta\cos\theta$. Trigonometry (the double angle rule) comes from linear algebra.

Example 6 S rotates by the angle θ and T rotates by $-\theta$. Then $TS = I$ leads to $AB = I$. In this case $T(S(\boldsymbol{u}))$ is $\boldsymbol{u}$. We rotate forward and back. For the matrices to match, $AB\boldsymbol{x}$ must be $\boldsymbol{x}$. *The two matrices are inverses.* Check this by putting $\cos(-\theta) = \cos\theta$ and $\sin(-\theta) = -\sin\theta$ into the backward rotation matrix A:

$$AB = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} \cos^2\theta + \sin^2\theta & 0 \\ 0 & \cos^2\theta + \sin^2\theta \end{bmatrix} = I.$$

Choosing the Best Bases

Now comes the final step in this section of the book. **Choose bases that diagonalize the matrix.** With the standard basis (the columns of I) our transformation T produces some matrix A—probably not diagonal. That same T is represented by different matrices when we choose different bases. The two great choices are eigenvectors and singular vectors:

> **Eigenvectors** If T transforms $\mathbf{R}^n$ to $\mathbf{R}^n$, its matrix A is square. But using the standard basis, that matrix A is probably not diagonal. If there are n independent eigenvectors, *choose those as the input and output basis.* In this good basis, **the matrix for T is the diagonal eigenvalue matrix Λ.**

Example 7 **The projection matrix** T projects every $\boldsymbol{v} = (x, y)$ in $\mathbf{R}^2$ onto the line $y = -x$. Using the standard basis, $\boldsymbol{v}_1 = (1, 0)$ projects to $T(\boldsymbol{v}_1) = \left(\frac{1}{2}, -\frac{1}{2}\right)$. For $\boldsymbol{v}_2 = (0, 1)$ the projection is $T(\boldsymbol{v}_2) = \left(-\frac{1}{2}, \frac{1}{2}\right)$. Those are the columns of A:

Projection matrix
Standard bases
Not diagonal

$$A = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \text{ has } A^{\mathrm{T}} = A \text{ and } A^2 = A.$$

Now comes the main point of eigenvectors. Make them the basis vectors ! Diagonalize !

When the basis vectors are eigenvectors, the matrix becomes diagonal.

$\boldsymbol{v}_1 = \boldsymbol{w}_1 = (1, -1)$ projects to itself : $T(\boldsymbol{v}_1) = \boldsymbol{v}_1$ and $\lambda_1 = 1$

$\boldsymbol{v}_2 = \boldsymbol{w}_2 = (1, \quad 1)$ projects to zero : $T(\boldsymbol{v}_2) = \boldsymbol{0}$ and $\lambda_2 = 0$

Eigenvector bases
Diagonal matrix

$$\text{The new matrix is } \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \Lambda. \tag{6}$$

Eigenvectors are the perfect basis vectors. They produce the eigenvalue matrix Λ.

What about other choices of *input basis* $=$ *output basis*? Put those basis vectors into the columns of B. We saw above that the change of basis matrices (between standard basis and new basis) are $\boldsymbol{B_{\text{in}} = B}$ and $\boldsymbol{B_{\text{out}} = B^{-1}}$. The new matrix for T is **similar** to A:

$\boldsymbol{A_{\text{new}} = B^{-1}AB}$ in the new basis of $\boldsymbol{b}$'s is similar to $\boldsymbol{A}$ in the standard basis :

$$\boldsymbol{A}_{\boldsymbol{b}\text{'s to }\boldsymbol{b}\text{'s}} = \boldsymbol{B}^{-1}_{\text{standard to }\boldsymbol{b}\text{'s}} \quad \boldsymbol{A}_{\text{standard}} \quad \boldsymbol{B}_{\boldsymbol{b}\text{'s to standard}} \tag{7}$$

I used the multiplication rule for the transformation ITI. *The matrices for* I, T, I *were* B^{-1}, A, B. The matrix B contains the input vectors $\boldsymbol{b}$ in the standard basis.

Finally we allow *different spaces* $\boldsymbol{V}$ *and* $\boldsymbol{W}$, *and different bases* $\boldsymbol{v}$'s and $\boldsymbol{w}$'s. When we know T and we choose bases, we get a matrix A. Probably A is not symmetric or even square. But we can always choose $\boldsymbol{v}$'s and $\boldsymbol{w}$'s that produce a diagonal matrix. This will be the *singular value matrix* $\boldsymbol{\Sigma = \text{diag}\,(\sigma_1, \ldots, \sigma_r)}$ in the decomposition $A = U\Sigma V^{\mathrm{T}}$.

Singular vectors The SVD says that $U^{-1}AV = \Sigma$. The right singular vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ will be the input basis. The left singular vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ will be the output basis. By the rule for matrix multiplication, the matrix for the same transformation in these new bases is $B_{\text{out}}^{-1}AB_{\text{in}} = U^{-1}AV = \Sigma$.

I can't say that Σ is "similar" to A. We are working now with two bases, input and output. But those are *orthonormal bases* and they preserve the lengths of vectors. Following a good suggestion by David Vogan, I propose that we say: **$\boldsymbol{\Sigma}$ is "isometric" to $\boldsymbol{A}$.**

Definition $\quad C = Q_1^{-1}AQ_2$ *is isometric to* A *if* Q_1 *and* Q_2 *are orthogonal.*

Example 8 To construct the matrix A for the transformation $T = \frac{d}{dx}$, we chose the input basis $1, x, x^2, x^3$ and the output basis $1, x, x^2$. The matrix A was simple but unfortunately it wasn't diagonal. But we can take each basis *in the opposite order.*

Now the input basis is $x^3, x^2, x, 1$ and the output basis is $x^2, x, 1$. The change of basis matrices B_{in} and B_{out} are permutations. The matrix for $T(\boldsymbol{u}) = d\boldsymbol{u}/dx$ with the new bases is **the diagonal singular value matrix $\boldsymbol{B_{\text{out}}^{-1}AB_{\text{in}} = \Sigma}$** with σ's $= 3, 2, 1$:

$$\boldsymbol{B_{\text{out}}^{-1}AB_{\text{in}}} = \begin{bmatrix} & & 1 \\ & 1 & \\ 1 & & \end{bmatrix} \begin{bmatrix} 0 & \mathbf{1} & 0 & 0 \\ 0 & 0 & \mathbf{2} & 0 \\ 0 & 0 & 0 & \mathbf{3} \end{bmatrix} \begin{bmatrix} & & & 1 \\ & & 1 & \\ & 1 & & \\ 1 & & & \end{bmatrix} = \begin{bmatrix} \mathbf{3} & 0 & 0 & 0 \\ 0 & \mathbf{2} & 0 & 0 \\ 0 & 0 & \mathbf{1} & 0 \end{bmatrix}. \tag{8}$$

Well, this was a tough section. We found that x^3, x^2, x, 1 have derivatives $3x^2$, $2x$, 1, 0.

■ REVIEW OF THE KEY IDEAS ■

1. If we know $T(\boldsymbol{v}_1), \ldots, T(\boldsymbol{v}_n)$ for a basis, linearity will determine all other $T(\boldsymbol{v})$.

2. $\left\{ \begin{array}{l} \text{Linear transformation } T \\ \text{Input basis } \boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \\ \text{Output basis } \boldsymbol{w}_1, \ldots, \boldsymbol{w}_m \end{array} \right\} \rightarrow \begin{array}{c} \text{Matrix } A \ (m \text{ by } n) \\ \text{represents } T \\ \text{in these bases} \end{array}$

3. The change of basis matrix $B = W^{-1}V = B_{\text{out}}^{-1} B_{\text{in}}$ represents the identity $T(\boldsymbol{v}) = \boldsymbol{v}$.

4. If A and B represent T and S, and the output basis for S is the input basis for T, then the matrix AB represents the transformation $T(S(\boldsymbol{u}))$.

5. The best input-output bases are eigenvectors and/or singular vectors of A. Then

$$B^{-1}AB = \Lambda = \text{eigenvalues} \qquad B_{\text{out}}^{-1} A B_{\text{in}} = \Sigma = \text{singular values.}$$

■ WORKED EXAMPLES ■

8.2 A The space of 2 by 2 matrices has these four "vectors" as a basis:

$$\boldsymbol{v}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \qquad \boldsymbol{v}_2 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \qquad \boldsymbol{v}_3 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \qquad \boldsymbol{v}_4 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

T is the linear transformation that *transposes* every 2 by 2 matrix. What is the matrix A that represents T in this basis (output basis = input basis)? What is the inverse matrix A^{-1}? What is the transformation T^{-1} that inverts the transpose operation ?

Solution Transposing those four "basis matrices" just reverses $\boldsymbol{v}_2$ and $\boldsymbol{v}_3$:

$$\begin{array}{l} T(\boldsymbol{v}_1) = \boldsymbol{v}_1 \\ T(\boldsymbol{v}_2) = \boldsymbol{v}_3 \\ T(\boldsymbol{v}_3) = \boldsymbol{v}_2 \\ T(\boldsymbol{v}_4) = \boldsymbol{v}_4 \end{array} \quad \text{gives the four columns of} \quad A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The inverse matrix A^{-1} is the same as A. The inverse transformation T^{-1} is the same as T. If we transpose and transpose again, the final matrix equals the original matrix.

Notice that the space of 2 by 2 matrices is 4-dimensional. So the matrix A (for the transpose T) is 4 by 4. The nullspace of A is $\boldsymbol{Z}$ and the kernel of T is the zero matrix—the only matrix that transposes to zero. The eigenvalues of A are $1, 1, 1, -1$.

Which line of matrices has $T(A) = A^{\mathrm{T}} = -A$ with that eigenvalue $\lambda = -1$?

Problem Set 8.2

Questions 1–4 extend the first derivative example to higher derivatives.

1 The transformation S takes the ***second derivative***. Keep $1, x, x^2, x^3$ as the input basis $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4$ and also as output basis $\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3, \boldsymbol{w}_4$. Write $S(\boldsymbol{v}_1), S(\boldsymbol{v}_2), S(\boldsymbol{v}_3), S(\boldsymbol{v}_4)$ in terms of the $\boldsymbol{w}$'s. Find the 4 by 4 matrix A_2 for S.

2 What functions have $S(\boldsymbol{v}) = \boldsymbol{0}$? They are in the kernel of the second derivative S. What vectors are in the nullspace of its matrix A_2 in Problem 1?

3 The second derivative A_2 is not the square of a rectangular first derivative matrix A_1:

$$A_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix} \text{ does not allow } A_1^2 = A_2.$$

Add a zero row 4 to A_1 so that output space = input space. Compare A_1^2 with A_2. Conclusion: We want output basis = _____ basis. Then $m = n$.

4 (a) The product TS of first and second derivatives produces the *third* derivative. Add zeros to make 4 by 4 matrices, then compute $A_1 A_2 = A_3$.

(b) The matrix A_2^2 corresponds to $S^2 =$ *fourth* derivative. Why is this zero?

Questions 5–9 are about a particular transformation T and its matrix A.

5 With bases $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$ and $\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3$, suppose $T(\boldsymbol{v}_1) = \boldsymbol{w}_2$ and $T(\boldsymbol{v}_2) = T(\boldsymbol{v}_3) = \boldsymbol{w}_1 + \boldsymbol{w}_3$. T is a linear transformation. Find the matrix A and multiply by the vector $(1, 1, 1)$. What is the output from T when the input is $\boldsymbol{v}_1 + \boldsymbol{v}_2 + \boldsymbol{v}_3$?

6 Since $T(\boldsymbol{v}_2) = T(\boldsymbol{v}_3)$, the solutions to $T(\boldsymbol{v}) = \boldsymbol{0}$ are $\boldsymbol{v} =$ _____. What vectors are in the nullspace of A? Find all solutions to $T(\boldsymbol{v}) = \boldsymbol{w}_2$.

7 Find a vector that is not in the column space of A. Find a combination of $\boldsymbol{w}$'s that is not in the range of the transformation T.

8 You don't have enough information to determine T^2. Why is its matrix not necessarily A^2? What more information do you need?

9 Find the *rank* of A. The rank is not the dimension of the whole output space $\mathbf{W}$. It is the dimension of the _____ of T.

Questions 10–13 are about invertible linear transformations.

10 Suppose $T(\boldsymbol{v}_1) = \boldsymbol{w}_1 + \boldsymbol{w}_2 + \boldsymbol{w}_3$ and $T(\boldsymbol{v}_2) = \boldsymbol{w}_2 + \boldsymbol{w}_3$ and $T(\boldsymbol{v}_3) = \boldsymbol{w}_3$. Find the matrix A for T using these basis vectors. What input vector $\boldsymbol{v}$ gives $T(\boldsymbol{v}) = \boldsymbol{w}_1$?

11 Invert the matrix A in Problem 10. Also invert the transformation T—what are $T^{-1}(\boldsymbol{w}_1)$ and $T^{-1}(\boldsymbol{w}_2)$ and $T^{-1}(\boldsymbol{w}_3)$?

12 Which of these are true and why is the other one ridiculous?

(a) $T^{-1}T = I$ (b) $T^{-1}(T(\boldsymbol{v}_1)) = \boldsymbol{v}_1$ (c) $T^{-1}(T(\boldsymbol{w}_1)) = \boldsymbol{w}_1$.

13 Suppose the spaces $\mathbf{V}$ and $\mathbf{W}$ have the same basis $\boldsymbol{v}_1, \boldsymbol{v}_2$.

(a) Describe a transformation T (not I) that is its own inverse.

(b) Describe a transformation T (not I) that equals T^2.

(c) Why can't the same T be used for both (a) and (b)?

Questions 14–19 are about changing the basis.

14 (a) What matrix B transforms $(1,0)$ into $(2,5)$ and transforms $(0,1)$ to $(1,3)$?

(b) What matrix C transforms $(2,5)$ to $(1,0)$ and $(1,3)$ to $(0,1)$?

(c) Why does no matrix transform $(2,6)$ to $(1,0)$ and $(1,3)$ to $(0,1)$?

15 (a) What matrix M transforms $(1,0)$ and $(0,1)$ to (r,t) and (s,u)?

(b) What matrix N transforms (a,c) and (b,d) to $(1,0)$ and $(0,1)$?

(c) What condition on a,b,c,d will make part (b) impossible?

16 (a) How do M and N in Problem 15 yield the matrix that transforms (a,c) to (r,t) and (b,d) to (s,u)?

(b) What matrix transforms $(2,5)$ to $(1,1)$ and $(1,3)$ to $(0,2)$?

17 If you keep the same basis vectors but put them in a different order, the change of basis matrix B is a _____ matrix. If you keep the basis vectors in order but change their lengths, B is a _____ matrix.

18 The matrix that rotates the axis vectors $(1,0)$ and $(0,1)$ through an angle θ is Q. What are the coordinates (a,b) of the original $(1,0)$ using the new (rotated) axes? This *inverse* can be tricky. Draw a figure or solve for a and b:

$$Q = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \qquad \begin{bmatrix} 1 \\ 0 \end{bmatrix} = a \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} + b \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix}.$$

19 The matrix that transforms $(1,0)$ and $(0,1)$ to $(1,4)$ and $(1,5)$ is $B =$ _____. The combination $a(1,4) + b(1,5)$ that equals $(1,0)$ has $(a,b) = (\ ,\ )$. How are those new coordinates of $(1,0)$ related to B or B^{-1}?

Questions 20–23 are about the space of quadratic polynomials $\boldsymbol{y = A + Bx + Cx^2}$.

20 The parabola $\boldsymbol{w}_1 = \frac{1}{2}(x^2 + x)$ equals one at $x = 1$, and zero at $x = 0$ and $x = -1$. Find the parabolas $\boldsymbol{w}_2, \boldsymbol{w}_3$, and then find $\boldsymbol{y}(x)$ by linearity.

(a) $\boldsymbol{w}_2$ equals one at $x = 0$ and zero at $x = 1$ and $x = -1$.

(b) $\boldsymbol{w}_3$ equals one at $x = -1$ and zero at $x = 0$ and $x = 1$.

(c) $\boldsymbol{y}(x)$ equals 4 at $x = 1$ and 5 at $x = 0$ and 6 at $x = -1$. Use $\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3$.

21 One basis for second-degree polynomials is $\boldsymbol{v}_1 = 1$ and $\boldsymbol{v}_2 = x$ and $\boldsymbol{v}_3 = x^2$. Another basis is $\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3$ from Problem 20. Find two change of basis matrices, from the $\boldsymbol{w}$'s to the $\boldsymbol{v}$'s and from the $\boldsymbol{v}$'s to the $\boldsymbol{w}$'s.

22 What are the three equations for A, B, C if the parabola $y = A + Bx + Cx^2$ equals 4 at $x = a$ and 5 at $x = b$ and 6 at $x = c$? Find the determinant of the 3 by 3 matrix. That matrix transforms values like $4, 5, 6$ to parabolas y—or is it the other way?

23 Under what condition on the numbers $m_1, m_2, \ldots, m_9$ do these three parabolas give a basis for the space of all parabolas $a + bx + cx^2$?

$$\boldsymbol{v}_1 = m_1 + m_2 x + m_3 x^2, \ \boldsymbol{v}_2 = m_4 + m_5 x + m_6 x^2, \ \boldsymbol{v}_3 = m_7 + m_8 x + m_9 x^2.$$

24 The Gram-Schmidt process changes a basis $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$ to an orthonormal basis $\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{q}_3$. These are columns in $A = QR$. Show that R is the change of basis matrix from the $\boldsymbol{a}$'s to the $\boldsymbol{q}$'s ($\boldsymbol{a}_2$ is what combination of $\boldsymbol{q}$'s when $A = QR$?).

25 Elimination changes the rows of A to the rows of U with $A = LU$. Row 2 of A is what combination of the rows of U? Writing $A^{\mathrm{T}} = U^{\mathrm{T}} L^{\mathrm{T}}$ to work with columns, the change of basis matrix is $B = L^{\mathrm{T}}$. We have *bases* if the matrices are _____.

26 Suppose $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$ are **eigenvectors** for T. This means $T(\boldsymbol{v}_i) = \lambda_i \boldsymbol{v}_i$ for $i = 1, 2, 3$. What is the matrix for T when the input and output bases are the $\boldsymbol{v}$'s?

27 Every invertible linear transformation can have I as its matrix! Choose any input basis $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$. For output basis choose $\boldsymbol{w}_i = T(\boldsymbol{v}_i)$. Why must T be invertible?

28 Using $\boldsymbol{v}_1 = \boldsymbol{w}_1$ and $\boldsymbol{v}_2 = \boldsymbol{w}_2$ find the standard matrix for these T's:

(a) $T(\boldsymbol{v}_1) = \boldsymbol{0}$ and $T(\boldsymbol{v}_2) = 3\boldsymbol{v}_1$ (b) $T(\boldsymbol{v}_1) = \boldsymbol{v}_1$ and $T(\boldsymbol{v}_1 + \boldsymbol{v}_2) = \boldsymbol{v}_1$.

29 Suppose T reflects the xy plane across the x axis and S is reflection across the y axis. If $\boldsymbol{v} = (x, y)$ what is $S(T(\boldsymbol{v}))$? Find a simpler description of the product ST.

30 Suppose T is reflection across the $45°$ line, and S is reflection across the y axis. If $\boldsymbol{v} = (2, 1)$ then $T(\boldsymbol{v}) = (1, 2)$. Find $S(T(\boldsymbol{v}))$ and $T(S(\boldsymbol{v}))$. Usually $ST \neq TS$.

31 **The product of two reflections is a rotation**. Multiply these reflection matrices to find the rotation angle:

$$\begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix} \quad \begin{bmatrix} \cos 2\alpha & \sin 2\alpha \\ \sin 2\alpha & -\cos 2\alpha \end{bmatrix}.$$

32 Suppose A is a 3 by 4 matrix of rank $r = 2$, and $T(\boldsymbol{v}) = A\boldsymbol{v}$. Choose input basis vectors $\boldsymbol{v}_1, \boldsymbol{v}_2$ from the row space of A and $\boldsymbol{v}_3, \boldsymbol{v}_4$ from the nullspace. Choose output basis vectors $\boldsymbol{w}_1 = A\boldsymbol{v}_1$, $\boldsymbol{w}_2 = A\boldsymbol{v}_2$ in the column space and $\boldsymbol{w}_3$ from the nullspace of A^{T}. What specially simple matrix represents T in these special bases?

33 The space $\mathbf{M}$ of 2 by 2 matrices has the basis $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \boldsymbol{v}_4$ in Worked Example **8.2 A**. Suppose T multiplies each matrix by $\left[\begin{smallmatrix} \mathbf{a} & \mathbf{b} \\ \mathbf{c} & \mathbf{d} \end{smallmatrix}\right]$. With $\boldsymbol{w}$'s equal to $\boldsymbol{v}$'s, what 4 by 4 matrix A represents this transformation T on matrix space?

34 True or False: If we know $T(\boldsymbol{v})$ for n different nonzero vectors in $\mathbf{R}^n$, then we know $T(\boldsymbol{v})$ for every vector $\boldsymbol{v}$ in $\mathbf{R}^n$.

8.3 The Search for a Good Basis

1 With a new input basis B_{in} and output basis B_{out}, every matrix A becomes $\boldsymbol{B_{\text{out}}^{-1}AB_{\text{in}}}$.

2 $B_{\text{in}} = B_{\text{out}} =$ "**generalized eigenvectors of A**" produces the **Jordan form** $J = B^{-1}AB$.

3 The **Fourier matrix** $\boldsymbol{F} = B_{\text{in}} = B_{\text{out}}$ diagonalizes every circulant matrix (use the **FFT**).

4 Sines and cosines, Legendre and Chebyshev: those are great bases for **function space**.

This is an important section of the book. I am afraid that most readers will skip it—or won't get this far. The first chapters prepared the way by explaining the idea of a **basis**. Chapter 6 introduced the eigenvectors $\boldsymbol{x}$ and Chapter 7 found singular vectors $\boldsymbol{v}$ and $\boldsymbol{u}$. Those are two winners but many other choices are very valuable.

First comes the pure algebra from Section 8.2 and then come good bases. The input basis vectors will be the columns of B_{in}. The output basis vectors will be the columns of B_{out}. Always B_{in} and B_{out} are *invertible*—basis vectors are independent!

Pure algebra If A is the matrix for a transformation T in the standard basis, then

$$\boldsymbol{B_{\text{out}}^{-1}AB_{\text{in}}} \text{ is the matrix in the new bases.} \tag{1}$$

The standard basis vectors are the *columns of the identity*: $B_{\text{in}} = I_{n\times n}$ and $B_{\text{out}} = I_{m\times m}$. Now we are choosing special bases to make the matrix clearer and simpler than A. When $B_{\text{in}} = B_{\text{out}} = B$, the square matrix $B^{-1}AB$ is *similar* to A.

Applied algebra Applications are all about choosing good bases. Here are four important choices for vectors and three choices for functions. Eigenvectors and singular vectors led to Λ and Σ in Section 8.2. The Jordan form is new.

1 $\boldsymbol{B_{\text{in}} = B_{\text{out}} =}$ **eigenvector matrix** $\boldsymbol{X}$. Then $\boldsymbol{X^{-1}AX =}$ **eigenvalues in** $\boldsymbol{\Lambda}$.
This choice requires A to be a square matrix with n independent eigenvectors. "A must be diagonalizable." We get Λ when $B_{\text{in}} = B_{\text{out}}$ is the eigenvector matrix X.

2 $\boldsymbol{B_{\text{in}} = V}$ and $\boldsymbol{B_{\text{out}} = U}$: **singular vectors of** $\boldsymbol{A}$. Then $\boldsymbol{U^{-1}AV =}$ **diagonal** $\boldsymbol{\Sigma}$.
Σ is the singular value matrix (with $\sigma_1, \ldots, \sigma_r$ on its diagonal) when B_{in} and B_{out} are the singular vector matrices V and U. Recall that those columns of B_{in} and B_{out} are orthonormal eigenvectors of $A^{\mathrm{T}}A$ and AA^{T}. Then $A = U\Sigma V^{\mathrm{T}}$ gives $\Sigma = U^{-1}AV$.

3 $\boldsymbol{B_{\text{in}} = B_{\text{out}} =}$ **generalized eigenvectors of** $\boldsymbol{A}$. Then $\boldsymbol{B^{-1}AB =}$ **Jordan form** $\boldsymbol{J}$.
A is a square matrix but it may only have s independent eigenvectors. (If $s = n$ then B is X and J is Λ.) In all cases Jordan constructed $n - s$ additional "generalized" eigenvectors, aiming to make the Jordan form J *as diagonal as possible*:

i) There are s square blocks along the diagonal of J.

ii) Each block has one eigenvalue λ, one eigenvector, and 1's above the diagonal.

The good case has n 1×1 blocks, each containing an eigenvalue. Then J is Λ (diagonal).

Example 1 This Jordan matrix J has eigenvalues $\lambda = 2, 2, 3, 3$ (two double eigenvalues). Those eigenvalues lie along the diagonal because J is triangular. There are two independent eigenvectors for $\lambda = 2$, but there is only *one line of eigenvectors for* $\lambda = 3$. This will be true for every matrix $C = BJB^{-1}$ that is similar to J.

$$\textbf{Jordan matrix} \quad J = \begin{bmatrix} 2 & & & \\ & 2 & & \\ & & \begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix} \end{bmatrix} \quad \begin{array}{l} \text{Two 1 by 1 blocks} \\ \text{One 2 by 2 block} \\ \text{Three eigenvectors} \\ \text{Eigenvalues } 2, 2, 3, 3 \end{array}$$

Two eigenvectors for $\lambda = 2$ are $\boldsymbol{x}_1 = (1, 0, 0, 0)$ and $\boldsymbol{x}_2 = (0, 1, 0, 0)$. One eigenvector for $\lambda = 3$ is $\boldsymbol{x}_3 = (0, 0, 1, 0)$. The "generalized eigenvector" for this Jordan matrix is the fourth standard basis vector $\boldsymbol{x}_4 = (0, 0, 0, 1)$. The eigenvectors for J (normal and generalized) are just the columns $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4$ of the identity matrix I.

Notice $\boldsymbol{(J - 3I)x_4 = x_3}$. **The generalized eigenvector $\boldsymbol{x}_4$ connects to the true eigenvector $\boldsymbol{x}_3$.** A true $\boldsymbol{x}_4$ would have $(J - 3I)\boldsymbol{x}_4 = \boldsymbol{0}$, but that doesn't happen here.

Every matrix $C = BJB^{-1}$ that is similar to this J will have true eigenvectors $\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3$ in the first three columns of B. The fourth column of B will be a generalized eigenvector $\boldsymbol{b}_4$ of C, tied to the true $\boldsymbol{b}_3$. Here is a quick proof that uses $B\boldsymbol{x}_3 = \boldsymbol{b}_3$ and $B\boldsymbol{x}_4 = \boldsymbol{b}_4$ to show: The fourth column $\boldsymbol{b}_4$ is tied to $\boldsymbol{b}_3$ by $\boldsymbol{(C - 3I)b_4 = b_3}$.

$$(BJB^{-1} - 3I)\,\boldsymbol{b}_4 = BJ\,\boldsymbol{x}_4 - 3B\,\boldsymbol{x}_4 = B(J - 3I)\,\boldsymbol{x}_4 = B\,\boldsymbol{x}_3 = \boldsymbol{b}_3. \qquad (2)$$

The point of Jordan's theorem is that every square matrix A has a complete set of eigenvectors and generalized eigenvectors. When those go into the columns of B, the matrix $B^{-1}AB = J$ is in Jordan form. Based on Example 1, here is a description of J.

The Jordan Form

For every A, we want to choose B so that $B^{-1}AB$ is as ***nearly diagonal as possible***. When A has a full set of n eigenvectors, they go into the columns of B. Then $B = X$. The matrix $X^{-1}AX$ is diagonal, period. This is the Jordan form of A—when A can be diagonalized. In the general case, eigenvectors are missing and Λ can't be reached.

Suppose A has s independent eigenvectors. Then it is similar to a Jordan matrix with s blocks. Each block has an *eigenvalue on the diagonal with* 1's *just above it.* This block accounts for exactly one eigenvector of A. Then B contains generalized eigenvectors as well as ordinary eigenvectors.

When there are n eigenvectors, all n blocks will be 1 by 1. In that case $J = \Lambda$.

The Jordan form solves the differential equation $d\boldsymbol{u}/dt = A\boldsymbol{u}$ for **any square matrix $\boldsymbol{A = BJB^{-1}}$**. The solution $e^{At}\boldsymbol{u}(0)$ becomes $\boldsymbol{u}(t) = Be^{Jt}B^{-1}\boldsymbol{u}(0)$. J is triangular and its matrix exponential e^{Jt} involves $e^{\lambda t}$ times powers $1, t, \ldots, t^{s-1}$.

(Jordan form) If A has s independent eigenvectors, it is similar to a matrix J that has s Jordan blocks $J_1 \ldots, J_s$ on its diagonal. Some matrix B puts A into Jordan form:

Jordan form
$$B^{-1}AB = \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_s \end{bmatrix} = J. \tag{3}$$

Each block J_i has one eigenvalue λ_i, one eigenvector, and 1's just above the diagonal:

Jordan block
$$J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix}. \tag{4}$$

Matrices are similar if they share the same Jordan form J—not otherwise.

The Jordan form J has an off-diagonal 1 for each missing eigenvector (and the 1's are next to the eigenvalues). In every family of similar matrices, we are picking one outstanding member called J. It is nearly diagonal (or if possible completely diagonal). We can quickly solve $d\boldsymbol{u}/dt = J\boldsymbol{u}$ and take powers J^k. Every other matrix in the family has the form BJB^{-1}.

Jordan's Theorem is proved in my textbook *Linear Algebra and Its Applications*. Please refer to that book (or more advanced books) for the proof. The reasoning is rather intricate and in actual computations the Jordan form is not at all popular—its calculation is not stable. A slight change in A will separate the repeated eigenvalues and remove the off-diagonal 1's—switching Jordan to a diagonal Λ.

Proved or not, you have caught the central idea of similarity—to make A as simple as possible while preserving its essential properties. The best basis B gives $B^{-1}AB = J$.

Question Find the eigenvalues and all possible Jordan forms if $\boldsymbol{A^2} =$ zero matrix.

Answer The eigenvalues must all be zero, because $A\boldsymbol{x} = \lambda\boldsymbol{x}$ leads to $A^2\boldsymbol{x} = \lambda^2\boldsymbol{x} = 0\boldsymbol{x}$. The Jordan form of A has $J^2 = 0$ because $J^2 = (B^{-1}AB)(B^{-1}AB) = B^{-1}A^2B = 0$. Every block in J has $\lambda = 0$ on the diagonal. Look at J_k^2 for block sizes $1, 2, 3$:

$$\begin{bmatrix} 0 \end{bmatrix}^2 = \begin{bmatrix} 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}^2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}^2 = \begin{bmatrix} 0 & 0 & \mathbf{1} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Conclusion: If $J^2 = 0$ then all block sizes must be 1 or 2. J^2 is not zero for 3 by 3.

The rank of J (and A) will be the total number of 1's. **The maximum rank is $\boldsymbol{n/2}$.** This happens when there are $n/2$ blocks, each of size 2 and rank 1.

Now come the great bases of applied mathematics. Their discrete forms are vectors in $\mathbf{R}^n$. Their continuous forms are functions in a function space. Since they are chosen once and for all, *without knowing the matrix* A, these bases $B_{\text{in}} = B_{\text{out}}$ probably don't diagonalize A. But for many important matrices A in applied mathematics, the matrices $B^{-1}AB$ are *close to diagonal.*

4 $B_{\text{in}} = B_{\text{out}}$ = Fourier matrix F Then Fx is a Discrete Fourier Transform of x.

Those words are telling us: The Fourier matrix with columns $(1, \lambda, \lambda^2, \lambda^3)$ in equation (6) is important. Those are good basis vectors to work with.

We ask: Which matrices are diagonalized by F? This time we are starting with the eigenvectors $(1, \lambda, \lambda^2, \lambda^3)$ and finding the matrices that have those eigenvectors:

$$\text{If } \boldsymbol{\lambda^4 = 1} \text{ then } \quad P\boldsymbol{x} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ \lambda \\ \lambda^2 \\ \lambda^3 \end{bmatrix} = \lambda \begin{bmatrix} 1 \\ \lambda \\ \lambda^2 \\ \lambda^3 \end{bmatrix} = \lambda \boldsymbol{x}. \tag{5}$$

P is a permutation matrix. The equation $P\boldsymbol{x} = \lambda \boldsymbol{x}$ says that $\boldsymbol{x}$ is an eigenvector and λ is an eigenvalue of P. Notice how the fourth row of this vector equation is $1 = \lambda^4$. That rule for λ makes everything work.

Does this give four different eigenvalues λ? *Yes.* The four numbers $\lambda = \mathbf{1}, \boldsymbol{i}, \boldsymbol{-1}, \boldsymbol{-i}$ all satisfy $\lambda^4 = 1$. (You know $i^2 = -1$. Squaring both sides gives $i^4 = 1$.) So those four numbers are the eigenvalues of P, each with its eigenvector $\boldsymbol{x} = (1, \lambda, \lambda^2, \lambda^3)$. **The eigenvector matrix $\boldsymbol{F}$ diagonalizes the permutation matrix $\boldsymbol{P}$:**

$$\begin{matrix} \textbf{Eigenvalue} \\ \textbf{matrix } \boldsymbol{\Lambda} \end{matrix} \begin{bmatrix} 1 & & & \\ & i & & \\ & & -1 & \\ & & & -i \end{bmatrix} \qquad \begin{matrix} \textbf{Eigenvector} \\ \textbf{matrix is} \\ \textbf{Fourier} \\ \textbf{matrix } \boldsymbol{F} \end{matrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & i^2 & 1 & (-i)^2 \\ 1 & i^3 & -1 & (-i)^3 \end{bmatrix} \tag{6}$$

Those columns of F are orthogonal because they are eigenvectors of P (an orthogonal matrix). Unfortunately this Fourier matrix F is complex (it is the most important complex matrix in the world). Multiplications $F\boldsymbol{x}$ are done millions of times very quickly, by the Fast Fourier Transform. The FFT comes in Section 9.3.

Key question: What other matrices beyond P have this same eigenvector matrix F? We know that P^2 and P^3 and P^4 have the same eigenvectors as P. The same matrix F diagonalizes all powers of P. And the eigenvalues of P^2 and P^3 and P^4 are the numbers λ^2 and λ^3 and λ^4. For example $P^2\boldsymbol{x} = \lambda^2 \boldsymbol{x}$:

$$P^2\boldsymbol{x} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ \lambda \\ \lambda^2 \\ \lambda^3 \end{bmatrix} = \lambda^2 \begin{bmatrix} 1 \\ \lambda \\ \lambda^2 \\ \lambda^3 \end{bmatrix} = \lambda^2 \boldsymbol{x} \text{ when } \boldsymbol{\lambda^4 = 1}.$$

The fourth power is special because $\boldsymbol{P^4 = I}$. When we do the "cyclic permutation" four times, $P^4\boldsymbol{x}$ is the same vector $\boldsymbol{x}$ that we started with. The eigenvalues of $P^4 = I$ are just $1, 1, 1, 1$. And that number 1 agrees with the fourth power of all the eigenvalues of P: $1^4 = 1$ and $i^4 = 1$ and $(-1)^4 = 1$ and $(-i)^4 = 1$.

One more step brings in many more matrices. If P and P^2 and P^3 and $P^4 = I$ have the same eigenvector matrix F, so does any combination $C = c_1P + c_2P^2 + c_3P^3 + c_0I$:

Circulant matrix $C = \begin{bmatrix} c_0 & \boldsymbol{c_1} & c_2 & c_3 \\ c_3 & c_0 & \boldsymbol{c_1} & c_2 \\ c_2 & c_3 & c_0 & \boldsymbol{c_1} \\ \boldsymbol{c_1} & c_2 & c_3 & c_0 \end{bmatrix}$ has eigenvectors in the Fourier matrix F
has four eigenvalues $c_0 + c_1\lambda + c_2\lambda^2 + c_3\lambda^3$
from the four numbers $\lambda = 1, i, -1, -i$
The eigenvalue from $\lambda = 1$ is $c_0 + c_1 + c_2 + c_3$

That was a big step. We have found all the matrices (circulant matrices C) whose eigenvectors are the Fourier vectors in F. We also know the four eigenvalues of C, but we haven't given them a good formula or a name until now:

The four eigenvalues of C are given by the Fourier transform Fc

$$\boldsymbol{Fc} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{matrix} c_0 + c_1 + c_2 + c_3 \\ c_0 + ic_1 - c_2 - ic_3 \\ c_0 - c_1 + c_2 - c_3 \\ c_0 - ic_1 - c_2 + ic_3 \end{matrix}$$

Example 2 The same ideas work for a Fourier matrix F and a circulant matrix C of any size. Two by two matrices look trivial but they are very useful. Now eigenvalues of P have $\lambda^2 = 1$ instead of $\lambda^4 = 1$ and the complex number i is not needed: $\boldsymbol{\lambda = \pm 1}$.

Fourier matrix F from eigenvectors of P and C $\quad F = \begin{bmatrix} \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{-1} \end{bmatrix} \quad P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ Circulant $c_0I + c_1P$ $\quad C = \begin{bmatrix} \boldsymbol{c_0} & \boldsymbol{c_1} \\ \boldsymbol{c_1} & \boldsymbol{c_0} \end{bmatrix}$.

The eigenvalues of C are $c_0 + c_1$ and $c_0 - c_1$. Those are given by the Fourier transform $F\boldsymbol{c}$ when the vector $\boldsymbol{c}$ is (c_0, c_1). This transform $F\boldsymbol{c}$ gives the eigenvalues of C for any size n.

Notice that **circulant matrices have constant diagonals**. The same number c_0 goes down the main diagonal. The number c_1 is on the diagonal above, and that diagonal "wraps around" or "circles around" to the southwest corner of C. This explains the name *circulant* and it indicates that these matrices are *periodic* or *cyclic*. Even the powers of λ cycle around because $\lambda^4 = 1$ leads to $\lambda^5, \lambda^6, \lambda^7, \lambda^8 = \lambda, \lambda^2, \lambda^3, \lambda^4$.

Constancy down the diagonals is a crucial property of C. It corresponds to *constant coefficients* in a differential equation. This is exactly when Fourier works perfectly!

The equation $\dfrac{d^2u}{dt^2} = -u$ is solved by $u = c_0 \cos t + c_1 \sin t$.

The equation $\dfrac{d^2u}{dt^2} = tu$ cannot be solved by elementary functions.

These equations are linear. The first is the oscillation equation for a simple spring. It is Newton's Law $f = ma$ with mass $m = 1$, $a = d^2u/dt^2$, and force $f = -u$. Constant coefficients produce the differential equations that you can really solve.

The equation $u'' = tu$ has a variable coefficient t. This is Airy's equation in physics and optics (it was derived to explain a rainbow). The solutions change completely when t passes through zero, and those solutions require infinite series. *We won't go there.*

The point is that equations with constant coefficients have simple solutions like $e^{\lambda t}$. You discover λ by substituting $e^{\lambda t}$ into the differential equation. That number λ is like an eigenvalue. For $u = \cos t$ and $u = \sin t$ the number is $\lambda = i$. Euler's great formula $e^{it} = \cos t + i \sin t$ introduces complex numbers as we saw in the eigenvalues of P and C.

Bases for Function Space

For functions of x, the first basis I would think of contains the powers $1, x, x^2, x^3, \ldots$ Unfortunately this is a terrible basis. Those functions x^n are just barely independent. x^{10} is *almost* a combination of other basis vectors $1, x, \ldots, x^9$. It is virtually impossible to compute with this poor "ill-conditioned" basis.

If we had vectors instead of functions, the test for a good basis would look at $B^{\mathrm{T}}B$. This matrix contains all inner products between the basis vectors (columns of B). *The basis is orthonormal when* $B^{\mathrm{T}}B = I$. That is best possible. But the basis $1, x, x^2, \ldots$ produces the evil **Hilbert matrix**: $B^{\mathrm{T}}B$ has an enormous ratio between its largest and smallest eigenvalues. A large condition number signals an unhappy choice of basis.

Note Now the columns of B are functions instead of vectors. We still use $B^{\mathrm{T}}B$ to test for independence. So we need to know the dot product (inner product is a better name) of two functions—those are the numbers in $B^{\mathrm{T}}B$.

The dot product of vectors is just $\boldsymbol{x}^{\mathbf{T}}\boldsymbol{y} = x_1y_1 + \cdots + x_ny_n$. The inner product of functions will integrate instead of adding, but the idea is completely parallel:

$$\begin{aligned}
\text{Inner product } (\boldsymbol{f}, \boldsymbol{g}) &= \int f(x)g(x)\,dx \\
\text{Complex inner product } (\boldsymbol{f}, \boldsymbol{g}) &= \int \overline{f(x)}\,g(x)\,dx, \ \overline{f} = \text{ complex conjugate} \\
\text{Weighted inner product } (\boldsymbol{f}, \boldsymbol{g})_w &= \int w(x)\,\overline{f(x)}\,g(x)\,dx, \ w = \text{ weight function}
\end{aligned}$$

When the integrals go from $x = 0$ to $x = 1$, the inner product of x^i with x^j is

$$\int_0^1 x^i\,x^j\,dx = \frac{x^{i+j+1}}{i+j+1}\Bigg]_{x=0}^{x=1} = \frac{1}{i+j+1} = \text{entries of Hilbert matrix } B^{\mathrm{T}}B$$

By changing to the symmetric interval from $x = -1$ to $x = 1$, we immediately have *orthogonality between all even functions and all odd functions*:

$$\textbf{Interval } [-1,\ 1] \qquad \int_{-1}^{1} x^2\,x^5\,dx = 0 \qquad \int_{-1}^{1} \textbf{even}(x)\,\textbf{odd}(x)\,dx = 0.$$

This change makes half of the basis functions orthogonal to the other half. It is so simple that we continue using the symmetric interval -1 to 1 (or $-\pi$ to π). But we want a better basis than the powers x^n—hopefully an orthogonal basis.

Orthogonal Bases for Function Space

Here are the three leading even-odd bases for theoretical and numerical computations:

5. The **Fourier basis**	$1, \sin x, \cos x, \sin 2x, \cos 2x, \ldots$
6. The **Legendre basis**	$1, x, x^2 - \frac{1}{3}, x^3 - \frac{3}{5}x, \ldots$
7. The **Chebyshev basis**	$1, x, 2x^2 - 1, 4x^3 - 3x, \ldots$

The Fourier basis functions (sines and cosines) are all *periodic*. They repeat over every 2π interval because $\cos(x+2\pi) = \cos x$ and $\sin(x+2\pi) = \sin x$. So this basis is especially good for functions $f(x)$ that are themselves periodic: $f(x + 2\pi) = f(x)$.

This basis is also *orthogonal*. Every sine and cosine is orthogonal to every other sine and cosine. Of course we don't expect the basis function $\cos nx$ to be orthogonal to itself.

Most important, the sine-cosine basis is also *excellent for approximation*. If we have a smooth periodic function $f(x)$, then a few sines and cosines (low frequencies) are all we need. Jumps in $f(x)$ and noise in the signal are seen in higher frequencies (larger n). We hope and expect that the signal is not drowned by the noise.

The *Fourier transform* connects $f(x)$ to the coefficients a_k and b_k in its Fourier series:

Fourier series $f(x) = a_0 + b_1 \sin x + a_1 \cos x + b_2 \sin 2x + a_2 \cos 2x + \cdots$

We see that **function space is infinite-dimensional**. It takes infinitely many basis functions to capture perfectly a typical $f(x)$. But the formula for each coefficient (for example a_3) is just like the formula $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{a}/\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}$ for projecting a vector $\boldsymbol{b}$ onto the line through $\boldsymbol{a}$.

Here we are projecting the function $f(x)$ onto the line in function space through $\cos 3x$:

$$\textbf{Fourier coefficient} \quad a_3 = \frac{(f(x), \cos 3x)}{(\cos 3x, \cos 3x)} = \frac{\int f(x) \cos 3x\, dx}{\int \cos 3x \cos 3x\, dx}. \tag{7}$$

Example 3 The double angle formula in trigonometry is $\cos 2x = 2\cos^2 x - 1$. This tells us that $\cos^2 x = \frac{1}{2} + \frac{1}{2}\cos 2x$. A very short Fourier series. So is $\sin^2 x = \frac{1}{2} - \frac{1}{2}\cos 2x$.

Fourier series is just linear algebra in function space. Let me explain that properly as a highlight of Chapter 10 about applications.

Legendre Polynomials and Chebyshev Polynomials

The Legendre polynomials are the result of applying the Gram-Schmidt idea (Section 4.4). The plan is to orthogonalize the powers $1, x, x^2, \ldots$ To start, the odd function $\boldsymbol{x}$ is already orthogonal to the even function 1 over the interval from -1 to 1. Their product $(x)(1) = x$ integrates to zero. But the inner product between x^2 and 1 is $\int x^2\,dx = 2/3$:

$$\frac{(x^2, 1)}{(1, 1)} = \frac{\int x^2\,dx}{\int 1\,dx} = \frac{2/3}{2} = \frac{1}{3} \qquad \text{Gram-Schmidt gives } \boldsymbol{x^2 - \frac{1}{3} = \text{Legendre}}$$

Similarly the odd power x^3 has a component $3x/5$ in the direction of the odd function x:

$$\frac{(x^3, x)}{(x, x)} = \frac{\int x^4\,dx}{\int x^2\,dx} = \frac{2/5}{2/3} = \frac{3}{5} \qquad \text{Gram-Schmidt gives } \boldsymbol{x^3 - \frac{3}{5}x = \text{Legendre}}$$

Continuing Gram-Schmidt for $x^4, x^5, \ldots$ produces every Legendre function—a good basis.

Finally we turn to the Chebyshev polynomials $1, x, 2x^2 - 1, 4x^3 - 3x$. They don't come from Gram-Schmidt. Instead they are connected to $1, \cos\theta, \cos 2\theta, \cos 3\theta$. This gives a giant computational advantage—we can use the Fast Fourier Transform. The connection of Chebyshev to Fourier appears when we set $\boldsymbol{x = \cos\theta}$:

Chebyshev to Fourier	$\boldsymbol{2x^2 - 1} = 2(\cos\theta)^2 - 1 = \boldsymbol{\cos 2\theta}$ $\boldsymbol{4x^3 - 3x} = 4(\cos\theta)^3 - 3(\cos\theta) = \boldsymbol{\cos 3\theta}$

The n^{th} degree Chebyshev polynomial $\boldsymbol{T_n(x)}$ converts to Fourier's $\boldsymbol{\cos\, n\theta = T_n(\cos\theta)}$.

Note These polynomials are the basis for a big software project called **"chebfun"**. Every function $f(x)$ is replaced by a super-accurate Chebyshev approximation. Then you can integrate $f(x)$, and solve $f(x) = 0$, and find its maximum or minimum. More than that, you can solve differential equations involving $f(x)$—fast and to high accuracy.

When **chebfun** replaces $f(x)$ by a polynomial, you are ready to solve problems.

■ REVIEW OF THE KEY IDEAS ■

1. A basis is good if its matrix B is well-conditioned. Orthogonal bases are best.
2. Also good if $\Lambda = B^{-1}AB$ is diagonal. But the Jordan form J can be very unstable.
3. The Fourier matrix diagonalizes constant-coefficient periodic equations: perfection.
4. The basis $1, x, x^2, \ldots$ leads to $B^{\mathrm{T}}B =$ Hillbert matrix: Terrible for computations.
5. Legendre and Chebyshev polynomials are excellent bases for function space.

Problem Set 8.3

1 In Example 1, what is the rank of $J - 3I$? What is the dimension of its nullspace? This dimension gives the number of independent eigenvectors for $\lambda = 3$.

The algebraic multiplicity is 2, because $\det(J - \lambda I)$ has the repeated factor $(\lambda - 3)^2$. The geometric multiplicity is 1, because there is only 1 independent eigenvector.

2 These matrices A_1 and A_2 are similar to J. Solve $A_1 B_1 = B_1 J$ and $A_2 B_2 = B_2 J$ to find the basis matrices B_1 and B_2 with $J = B_1^{-1} A_1 B_1$ and $J = B_2^{-1} A_2 B_2$.

$$J = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad A_1 = \begin{bmatrix} 0 & 4 \\ 0 & 0 \end{bmatrix} \quad A_2 = \begin{bmatrix} 4 & -8 \\ 2 & -4 \end{bmatrix}$$

3 This transpose block J^{T} has the same triple eigenvalue 2 (with only one eigenvector) as J. Find the basis change B so that $J = B^{-1} J^{\mathrm{T}} B$ (which means $BJ = J^{\mathrm{T}} B$):

$$J = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{bmatrix} \qquad J^{\mathrm{T}} = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 2 & 0 \\ 0 & 1 & 2 \end{bmatrix}$$

4 J and K are Jordan forms with the same zero eigenvalues and the same rank 2. But show that no invertible B solves $BK = JB$, so K *is not similar to* J:

$$J = \begin{bmatrix} 0 & 1 & 0 & 0 \\ & 0 & 0 & 0 \\ & & 0 & 1 \\ & & & 0 \end{bmatrix} \qquad K = \begin{bmatrix} 0 & 1 & 0 & 0 \\ & 0 & 1 & 0 \\ & & 0 & 0 \\ & & & 0 \end{bmatrix}$$

5 If $A^3 = 0$ show that all $\lambda = 0$, and all Jordan blocks with $J^3 = 0$ have size $1, 2$, or 3. It follows that rank $(A) \leq 2n/3$. If $A^n = 0$ why is rank $(A) < n$?

6 Show that $\boldsymbol{u}(t) = \begin{bmatrix} te^{\lambda t} \\ e^{\lambda t} \end{bmatrix}$ solves $\dfrac{d\boldsymbol{u}}{dt} = J\boldsymbol{u}$ with $J = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}$ and $\boldsymbol{u}(0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

J is not diagonalizable so $te^{\lambda t}$ enters the solution.

7 Show that the difference equation $v_{k+2} - 2\lambda v_{k+1} + \lambda^2 v_k = 0$ is solved by $v_k = \lambda^k$ and also by $v_k = k\lambda^k$. Those correspond to $e^{\lambda t}$ and $te^{\lambda t}$ in Problem 6.

8 What are the 3 solutions to $\lambda^3 = 1$? They are complex numbers $\lambda = \cos\theta + i \sin\theta = e^{i\theta}$. Then $\lambda^3 = e^{3i\theta} = 1$ when the angle 3θ is 0 or 2π or 4π. Write the 3 by 3 Fourier matrix F with columns $(1, \lambda, \lambda^2)$.

9 Check that any 3 by 3 circulant C has eigenvectors $(1, \lambda, \lambda^2)$ from Problem 8. If the diagonals of your matrix C contain c_0, c_1, c_2 then its eigenvalues are in Fc.

10 Using formula (7) find $a_3 \cos 3x$ in the Fourier series of $f(x) = \begin{cases} 1 \text{ for } -L \leq x \leq L \\ 0 \text{ for } \quad L \leq |x| \leq 2\pi \end{cases}$

Chapter 9

Complex Vectors and Matrices

Real versus Complex

$\mathbf{R}$ = line of all real numbers $-\infty < x < \infty$ $\leftrightarrow$ $\mathbf{C}$ = plane of all complex numbers $z = x + iy$

$|x|$ = absolute value of x $\leftrightarrow$ $\boldsymbol{|z| = \sqrt{x^2 + y^2}} = \boldsymbol{r}$ = absolute value (or modulus) of z

1 and -1 solve $x^2 = 1$ $\leftrightarrow$ $z = 1, w, \ldots, w^{n-1}$ solve $z^n = 1$ where $w = e^{2\pi i/n}$

The **complex conjugate** of $z = x + iy$ is $\boldsymbol{\overline{z} = x - iy}$. $|z|^2 = x^2 + y^2 = z\overline{z}$ and $\dfrac{1}{z} = \dfrac{\overline{z}}{|z|^2}$.

The **polar form** of $z = x + iy$ is $|z|e^{i\theta} = \boldsymbol{re^{i\theta}} = \boldsymbol{r\cos\theta + ir\sin\theta}$. The angle has $\tan\theta = \dfrac{y}{x}$.

$\mathbf{R}^n$: vectors with n real components	$\leftrightarrow$	$\mathbf{C}^n$: vectors with n complex components
length: $\|\boldsymbol{x}\|^2 = x_1^2 + \cdots + x_n^2$	$\leftrightarrow$	length: $\|\boldsymbol{z}\|^2 = \|z_1\|^2 + \cdots + \|z_n\|^2$
transpose: $(A^{\mathrm{T}})_{ij} = A_{ji}$	$\leftrightarrow$	conjugate transpose: $(A^{\mathrm{H}})_{ij} = \overline{A_{ji}}$
dot product: $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = x_1y_1 + \cdots + x_ny_n$	$\leftrightarrow$	inner product: $\boldsymbol{u}^{\mathrm{H}}\boldsymbol{v} = \overline{u}_1v_1 + \cdots + \overline{u}_nv_n$
reason for A^{T}: $(A\boldsymbol{x})^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{x}^{\mathrm{T}}(A^{\mathrm{T}}\boldsymbol{y})$	$\leftrightarrow$	reason for A^{H}: $(A\boldsymbol{u})^{\mathrm{H}}\boldsymbol{v} = \boldsymbol{u}^{\mathrm{H}}(A^{\mathrm{H}}\boldsymbol{v})$
orthogonality: $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = 0$	$\leftrightarrow$	orthogonality: $\boldsymbol{u}^{\mathrm{H}}\boldsymbol{v} = 0$
symmetric matrices: $S = S^{\mathrm{T}}$	$\leftrightarrow$	Hermitian matrices: $S = S^{\mathrm{H}}$
$\boldsymbol{S = Q\Lambda Q^{-1} = Q\Lambda Q^{\mathrm{T}}}$(real Λ)	$\leftrightarrow$	$\boldsymbol{S = U\Lambda U^{-1} = U\Lambda U^{\mathrm{H}}}$ (real Λ)
skew-symmetric matrices: $K^{\mathrm{T}} = -K$	$\leftrightarrow$	skew-Hermitian matrices $K^{\mathrm{H}} = -K$
orthogonal matrices: $Q^{\mathrm{T}} = Q^{-1}$	$\leftrightarrow$	unitary matrices: $U^{\mathrm{H}} = U^{-1}$
orthonormal columns: $Q^{\mathrm{T}}Q = I$	$\leftrightarrow$	orthonormal columns: $U^{\mathrm{H}}U = I$
$(Q\boldsymbol{x})^{\mathrm{T}}(Q\boldsymbol{y}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}$ and $\|Q\boldsymbol{x}\| = \|\boldsymbol{x}\|$	$\leftrightarrow$	$(U\boldsymbol{x})^{\mathrm{H}}(U\boldsymbol{y}) = \boldsymbol{x}^{\mathrm{H}}\boldsymbol{y}$ and $\|U\boldsymbol{z}\| = \|\boldsymbol{z}\|$

A complete presentation of linear algebra must include complex numbers $z = x + iy$. Even when the matrix is real, ***the eigenvalues and eigenvectors are often complex***. Example: A 2 by 2 rotation matrix has complex eigenvectors $\boldsymbol{x} = (1, i)$ and $\overline{\boldsymbol{x}} = (1, -i)$. I will summarize Sections 9.1 and 9.2 in these few unforgettable words: When you transpose a vector $\boldsymbol{v}$ or a matrix A, *take the conjugate of every entry* ($\boldsymbol{i}$ **changes to** $\boldsymbol{-i}$). Section 9.3 is about the most important complex matrix of all—*the Fourier matrix* F.

9.1 Complex Numbers

Start with the imaginary number i. Everybody knows that $x^2 = -1$ has no real solution. When you square a real number, the answer is never negative. So the world has agreed on a solution called i. (Except that electrical engineers call it j.) Imaginary numbers follow the normal rules of addition and multiplication, with one difference. ***Replace*** i^2 ***by*** -1.

This section gives the main facts about complex numbers. It is a review for some students and a reference for everyone. Everything comes from $i^2 = -1$ and $e^{2\pi i} = 1$.

A complex number (say $3 + 2i$) ***is a real number*** (3) ***plus an imaginary number*** $(2i)$. Addition keeps the real and imaginary parts separate. Multiplication uses $i^2 = -1$:

Add: $(3 + 2i) + (3 + 2i) = 6 + 4i$

Multiply: $(3 + 2i)(1 - i) = 3 + 2i - 3i - 2i^2 = 5 - i.$

If I add $3 + i$ to $1 - i$, the answer is 4. The real numbers $3 + 1$ stay separate from the imaginary numbers $i - i$. We are adding the vectors $(3, 1)$ and $(1, -1)$ to get $(4, 0)$.

The number $(1 + i)^2$ is $1 + i$ times $1 + i$. The rules give the surprising answer $2i$:

$$(1 + i)(1 + i) = 1 + i + i + i^2 = 2i.$$

In the complex plane, $1+i$ is at an angle of $45°$. It is like the vector $(1, 1)$. When we square $1 + i$ to get $2i$, the angle doubles to $90°$. If we square again, the answer is $(2i)^2 = -4$. The $90°$ angle doubled to $180°$, the direction of a negative real number.

A real number is just a complex number $z = a + bi$, with zero imaginary part: $b = 0$.

The ***real part*** is $a = \text{Re}\,(a + bi)$. The ***imaginary part*** is $b = \text{Im}\,(a + bi)$.

The Complex Plane

Complex numbers correspond to points in a plane. Real numbers go along the x axis. Pure imaginary numbers are on the y axis. ***The complex number*** $\mathbf{3 + 2i}$ ***is at the point with coordinates*** $\mathbf{(3, 2)}$. The number zero, which is $0 + 0i$, is at the origin.

Adding and subtracting complex numbers is like adding and subtracting vectors in the plane. The real component stays separate from the imaginary component. The vectors go head-to-tail as usual. The complex plane $\mathbf{C}^1$ is like the ordinary two-dimensional plane $\mathbf{R}^2$, except that we multiply complex numbers and we didn't multiply vectors.

Now comes an important idea. ***The complex conjugate of*** $\mathbf{3 + 2i}$ ***is*** $\mathbf{3 - 2i}$. The complex conjugate of $z = 1 - i$ is $\overline{z} = 1 + i$. In general the conjugate of $z = a + bi$ is $\overline{z} = a - bi$. (**Some writers use a "*bar*" on the number and others use a "*star*":** $\overline{z} = z^*$.) The imaginary parts of z and "z bar" have opposite signs. In the complex plane, $\overline{z}$ is the image of z on the other side of the real axis.

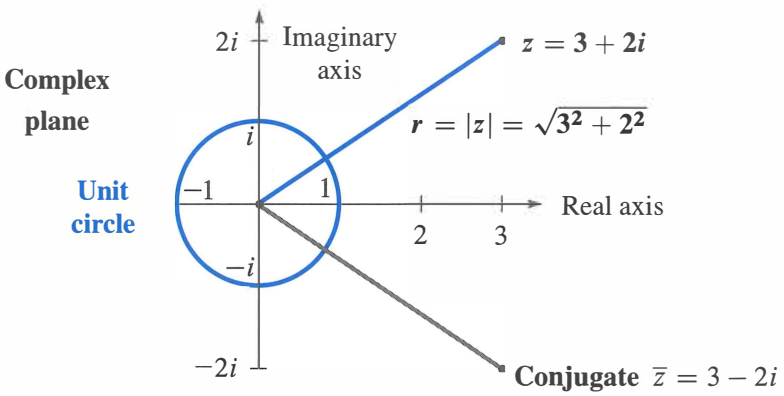


Figure 9.1: The number $z = a + bi$ corresponds to the point (a, b) and the vector $\begin{bmatrix} a \\ b \end{bmatrix}$.

Two useful facts. ***When we multiply conjugates $\overline{z}_1$ and $\overline{z}_2$, we get the conjugate of $z_1 z_2$.*** And when we add $\overline{z}_1$ and $\overline{z}_2$, we get the conjugate of $z_1 + z_2$:

$\overline{z}_1 + \overline{z}_2 = (3 - 2i) + (1 + i) = \mathbf{4 - i}$. This is the conjugate of $z_1 + z_2 = \mathbf{4 + i}$.

$\overline{z}_1 \times \overline{z}_2 = (3 - 2i) \times (1 + i) = \mathbf{5 + i}$. This is the conjugate of $z_1 \times z_2 = \mathbf{5 - i}$.

Adding and multiplying is exactly what linear algebra needs. By taking conjugates of $A\boldsymbol{x} = \lambda \boldsymbol{x}$, when A is real, we have another eigenvalue $\overline{\lambda}$ and its eigenvector $\overline{\boldsymbol{x}}$:

$$\textit{Eigenvalues } \boldsymbol{\lambda} \textit{ and } \overline{\boldsymbol{\lambda}} \qquad \textit{If } A\boldsymbol{x} = \lambda \boldsymbol{x} \textit{ and } A \textit{ is real then } \boldsymbol{A\overline{x} = \overline{\lambda}\overline{x}}. \tag{1}$$

Something special happens when $z = 3 + 2i$ combines with its own complex conjugate $\overline{z} = 3 - 2i$. The result from adding $z + \overline{z}$ or multiplying $z\overline{z}$ is always real:

$$\begin{aligned} z + \overline{z} &= \textbf{real} \quad (3 + 2i) + (3 - 2i) = 6 \quad \text{(real)} \\ z\overline{z} &= \textbf{real} \quad (3 + 2i) \times (3 - 2i) = 9 + 6i - 6i - 4i^2 = 13 \quad \text{(real)}. \end{aligned}$$

The sum of $z = a + bi$ and its conjugate $\overline{z} = a - bi$ is the real number $2a$. The product of z times $\overline{z}$ is the real number $a^2 + b^2$:

$$\textbf{Multiply } z \textbf{ times } \overline{z} \textbf{ to get } |z|^2 = r^2 \qquad (a + bi)(a - bi) = a^2 + b^2. \tag{2}$$

The next step with complex numbers is $1/z$. How to divide by $a + ib$? The best idea is to multiply first by $\overline{z}/\overline{z} = 1$. That produces $z\overline{z}$ in the denominator, which is $a^2 + b^2$:

$$\frac{1}{a + ib} = \frac{1}{a + ib}\frac{a - ib}{a - ib} = \frac{a - ib}{a^2 + b^2} \qquad \frac{1}{3 + 2i} = \frac{1}{3 + 2i}\frac{3 - 2i}{3 - 2i} = \frac{3 - 2i}{13}.$$

In case $a^2 + b^2 = 1$, this says that $(a + ib)^{-1}$ is $a - ib$. ***On the unit circle, $1/z$ equals $\overline{z}$.*** Later we will say: $1/e^{i\theta}$ is $e^{-i\theta}$. Use distance r and angle θ to multiply and divide.

The Polar Form $re^{i\theta}$

The square root of $a^2 + b^2$ is $|z|$. This is the ***absolute value*** (or ***modulus***) of the number $z = a + ib$. The square root $|z|$ is also written r, because it is the distance from 0 to z. ***The real number r in the polar form gives the size of the complex number z:***

The absolute value of $z = a + ib$ is $|z| = \sqrt{a^2 + b^2}$. **This is called r.**

The absolute value of $z = 3 + 2i$ is $|z| = \sqrt{3^2 + 2^2}$. This is $r = \sqrt{13}$.

The other part of the polar form is the angle θ. The angle for $z = 5$ is $\theta = 0$ (because this z is real and positive). The angle for $z = 3i$ is $\pi/2$ radians. The angle for a negative $z = -9$ is π radians. ***The angle doubles when the number is squared.*** The polar form is excellent for multiplying complex numbers (not good for addition).

When the distance is r and the angle is θ, trigonometry gives the other two sides of the triangle. The real part (along the bottom) is $a = r\cos\theta$. The imaginary part (up or down) is $b = r\sin\theta$. Put those together, and the rectangular form becomes the polar form $re^{i\theta}$.

The number $z = a + ib$ is also $z = r\cos\theta + ir\sin\theta$. This is $re^{i\theta}$

Note: **$\cos\theta + i\sin\theta$ *has absolute value* $r = 1$ *because* $\cos^2\theta + \sin^2\theta = 1$.** Thus $\cos\theta + i\sin\theta$ lies on the circle of radius 1—*the unit circle*.

Example 1 Find r and θ for $z = 1 + i$ and also for the conjugate $\overline{z} = 1 - i$.

Solution The absolute value is the same for z and $\overline{z}$. It is $r = \sqrt{1+1} = \sqrt{2}$:

$$|z|^2 = 1^2 + 1^2 = 2 \qquad \text{and also} \qquad |\overline{z}|^2 = 1^2 + (-1)^2 = 2.$$

The distance from the center is $r = \sqrt{2}$. What about the angle θ? The number $1 + i$ is at the point $(1, 1)$ in the complex plane. The angle to that point is $\pi/4$ radians or $45°$. The cosine is $1/\sqrt{2}$ and the sine is $1/\sqrt{2}$. Combining r and θ brings back $z = 1 + i$:

$$r\cos\theta + ir\sin\theta = \sqrt{2}\left(\frac{1}{\sqrt{2}}\right) + i\sqrt{2}\left(\frac{1}{\sqrt{2}}\right) = 1 + i.$$

The angle to the conjugate $1 - i$ can be positive or negative. We can go to $7\pi/4$ radians which is $315°$. Or we can go *backwards through a negative angle*, to $-\pi/4$ radians or $-45°$. ***If z is at angle θ, its conjugate $\overline{z}$ is at $2\pi - \theta$ and also at $-\theta$.***

We can freely add 2π or 4π or -2π to any angle! Those go full circles so the final point is the same. This explains why there are infinitely many choices of θ. Often we select the angle between 0 and 2π. But $-\theta$ is very useful for the conjugate $\overline{z}$. And $1 = e^0 = e^{2\pi i}$.

Powers and Products: Polar Form

Computing $(1+i)^2$ and $(1+i)^8$ is quickest in polar form. That form has $r = \sqrt{2}$ and $\theta = \pi/4$ (or $45°$). If we square the absolute value to get $r^2 = 2$, and double the angle to get $2\theta = \pi/2$ (or $90°$), we have $(1+i)^2$. For the eighth power we need r^8 and 8θ:

$$\textbf{(1+i)}^{\mathbf{8}} \qquad r^8 = 2 \cdot 2 \cdot 2 \cdot 2 = 16 \text{ and } 8\theta = 8 \cdot \frac{\pi}{4} = 2\pi.$$

This means: $(1+i)^8$ has absolute value 16 and angle 2π. *So* $(1+i)^8 = 16$.

Powers are easy in polar form. So is multiplication of complex numbers.

The nth power of $z = r(\cos\theta + i\sin\theta)$ ***is*** $z^n = r^n(\cos n\theta + i\sin n\theta)$. (3)

In that case z multiplies itself. To multiply z times z', ***multiply r's and add angles***:

$$r(\cos\theta + i\sin\theta) \text{ times } r'(\cos\theta' + i\sin\theta') = \boldsymbol{rr'}\big(\cos(\boldsymbol{\theta+\theta'}) + i\sin(\boldsymbol{\theta+\theta'})\big). \quad (4)$$

One way to understand this is by trigonometry. Why do we get the double angle 2θ for z^2?

$$(\cos\theta + i\sin\theta) \times (\cos\theta + i\sin\theta) = \cos^2\theta + i^2\sin^2\theta + 2i\sin\theta\cos\theta.$$

The real part $\cos^2\theta - \sin^2\theta$ is $\cos 2\theta$. The imaginary part $2\sin\theta\cos\theta$ is $\sin 2\theta$. Those are the "double angle" formulas. They show that θ in z becomes 2θ in z^2.

There is a second way to understand the rule for z^n. It uses the only amazing formula in this section. Remember that $\cos\theta + i\sin\theta$ has absolute value 1. The cosine is made up of even powers, starting with $1 - \frac{1}{2}\theta^2$. The sine is made up of odd powers, starting with $\theta - \frac{1}{6}\theta^3$. The beautiful fact is that $e^{i\theta}$ combines both of those series into $\cos\theta + i\sin\theta$:

$$e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \cdots \quad \text{becomes} \quad e^{i\theta} = 1 + i\theta + \frac{1}{2}i^2\theta^2 + \frac{1}{6}i^3\theta^3 + \cdots$$

Write -1 for i^2 to see $1 - \frac{1}{2}\theta^2$. ***The complex number*** $\boldsymbol{e^{i\theta}}$ ***is*** $\boldsymbol{\cos\theta + i\sin\theta}$***:***

Euler's Formula $e^{i\theta} = \cos\theta + i\sin\theta$ gives $z = r\cos\theta + ir\sin\theta = re^{i\theta}$ (5)

The special choice $\theta = 2\pi$ gives $\cos 2\pi + i\sin 2\pi$ which is 1. Somehow the infinite series $e^{2\pi i} = 1 + 2\pi i + \frac{1}{2}(2\pi i)^2 + \cdots$ adds up to 1.

Now multiply $e^{i\theta}$ times $e^{i\theta'}$. Angles add for the same reason that exponents add:

e^2 times e^3 is e^5 | $e^{i\theta}$ times $e^{i\theta}$ is $e^{2i\theta}$ | $e^{i\theta}$ times $e^{i\theta'}$ is $e^{i(\theta+\theta')}$

The powers $(re^{i\theta})^n$ are equal to $r^n e^{in\theta}$. They stay on the unit circle when $r = 1$ and $r^n = 1$. Then we find n different numbers whose nth powers equal 1:

Set $w = e^{2\pi i/n}$. ***The nth powers of*** $1, w, w^2, \ldots, w^{n-1}$ ***all equal 1.***

Those are the "nth roots of 1." They solve the equation $z^n = 1$. They are equally spaced around the unit circle in Figure 9.2b, where the full 2π is divided by n. Multiply their angles by n to take nth powers. That gives $w^n = e^{2\pi i}$ which is 1. Also $(w^2)^n = e^{4\pi i} = 1$. Each of those numbers, to the nth power, comes around the unit circle to 1.

These n roots of 1 are the key numbers for signal processing. The Discrete Fourier Transform uses $w = e^{2\pi i/n}$ and its powers. Section 9.3 shows how to decompose a vector (a signal) into n frequencies by the Fast Fourier Transform.

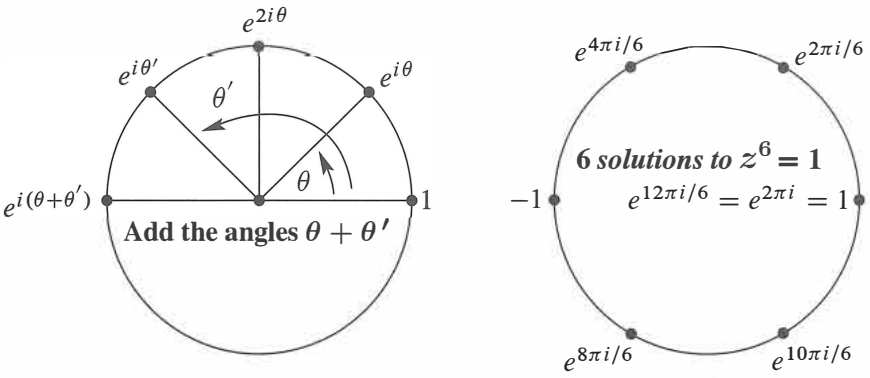


Figure 9.2: (a) $e^{i\theta}$ times $e^{i\theta'}$ is $e^{i(\theta+\theta')}$. (b) The nth power of $e^{2\pi i/n}$ is $e^{2\pi i} = 1$.

■ REVIEW OF THE KEY IDEAS ■

1. Adding $a + ib$ to $c + id$ is like adding $(a, b) + (c, d)$. Use $i^2 = -1$ to multiply.
2. The conjugate of $z = a + bi = re^{i\theta}$ is $\overline{z} = z^* = a - bi = re^{-i\theta}$.
3. z times $\overline{z}$ is $re^{i\theta}$ times $re^{-i\theta}$. This is $r^2 = |z|^2 = a^2 + b^2$ (real).
4. Powers and products are easy in polar form $z = re^{i\theta}$. *Multiply* r's *and add* θ's.

Problem Set 9.1

Questions 1–8 are about operations on complex numbers.

1 Add and multiply each pair of complex numbers:

(a) $2+i, 2-i$ (b) $-1+i, -1+i$ (c) $\cos\theta + i\sin\theta, \cos\theta - i\sin\theta$

2 Locate these points on the complex plane. Simplify them if necessary:

(a) $2+i$ (b) $(2+i)^2$ (c) $\frac{1}{2+i}$ (d) $|2+i|$

3 Find the absolute value $r = |z|$ of these four numbers. If θ is the angle for $6-8i$, what are the angles for the other three numbers?

(a) $6-8i$ (b) $(6-8i)^2$ (c) $\frac{1}{6-8i}$ (d) $(6+8i)^2$

4 If $|z| = 2$ and $|w| = 3$ then $|z \times w| =$ _____ and $|z+w| \leq$ _____ and $|z/w| =$ _____ and $|z-w| \leq$ _____.

5 Find $a+ib$ for the numbers at angles $30°, 60°, 90°, 120°$ on the unit circle. If w is the number at $30°$, check that w^2 is at $60°$. What power of w equals 1?

6 If $z = r\cos\theta + ir\sin\theta$ then $1/z$ has absolute value _____ and angle _____. Its polar form is _____. Multiply $z \times 1/z$ to get 1.

7 The complex multiplication $M = (a+bi)(c+di)$ is a 2 by 2 real multiplication

$$\begin{bmatrix} a & -b \\ b & a \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} \quad \\ \quad \end{bmatrix}.$$

The right side contains the real and imaginary parts of M. Test $M = (1+3i)(1-3i)$.

8 $A = A_1 + iA_2$ is a complex n by n matrix and $\boldsymbol{b} = \boldsymbol{b}_1 + i\boldsymbol{b}_2$ is a complex vector. The solution to $A\boldsymbol{x} = \boldsymbol{b}$ is $\boldsymbol{x}_1 + i\boldsymbol{x}_2$. Write $A\boldsymbol{x} = \boldsymbol{b}$ as a real system of size $2n$:

Complex n by n	$\begin{bmatrix} \quad \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix}.$
Real $2n$ by $2n$	

Questions 9–16 are about the conjugate $\overline{z} = a - ib = re^{-i\theta} = z^*$.

9 Write down the complex conjugate of each number by changing i to $-i$:

(a) $2-i$ (b) $(2-i)(1-i)$ (c) $e^{i\pi/2}$ (which is i)

(d) $e^{i\pi} = -1$ (e) $\frac{1+i}{1-i}$ (which is also i) (f) $i^{103} =$ _____.

10 The sum $z + \overline{z}$ is always _____. The difference $z - \overline{z}$ is always _____. Assume $z \neq 0$. The product $z \times \overline{z}$ is always _____. The ratio $z/\overline{z}$ has absolute value _____.

11 For a real matrix, the conjugate of $A\boldsymbol{x} = \lambda\boldsymbol{x}$ is $A\overline{\boldsymbol{x}} = \overline{\lambda}\overline{\boldsymbol{x}}$. This proves two things: $\overline{\lambda}$ is another eigenvalue and $\overline{\boldsymbol{x}}$ is its eigenvector. Find the eigenvalues $\lambda, \overline{\lambda}$ and eigenvectors $\boldsymbol{x}, \overline{\boldsymbol{x}}$ of $A = [\,a \;\; b;\;\; -b \;\; a\,]$.

12 The eigenvalues of a real 2 by 2 matrix come from the quadratic formula:

$$\det \begin{bmatrix} a-\lambda & b \\ c & d-\lambda \end{bmatrix} = \lambda^2 - (a+d)\lambda + (ad-bc) = 0$$

gives the two eigenvalues $\lambda = \left[a + d \pm \sqrt{(a+d)^2 - 4(ad-bc)}\right]/2$.

(a) If $a = b = d = 1$, the eigenvalues are complex when c is _____.

(b) What are the eigenvalues when $ad = bc$?

13 In Problem 12 the eigenvalues are not real when $(\text{trace})^2 = (a+d)^2$ is smaller than _____. Show that the λ's *are* real when $bc > 0$.

14 A real skew-symmetric matrix ($A^{\mathrm{T}} = -A$) has pure imaginary eigenvalues. First proof: If $A\boldsymbol{x} = \lambda\boldsymbol{x}$ then block multiplication gives

$$\begin{bmatrix} 0 & A \\ -A & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ i\boldsymbol{x} \end{bmatrix} = i\lambda \begin{bmatrix} \boldsymbol{x} \\ i\boldsymbol{x} \end{bmatrix}.$$

This block matrix is symmetric. Its eigenvalues must be _____! So λ is _____.

Questions 15–22 are about the form $re^{i\theta}$ of the complex number $r\cos\theta + ir\sin\theta$.

15 Write these numbers in Euler's form $re^{i\theta}$. Then square each number:

(a) $1 + \sqrt{3}i$ (b) $\cos 2\theta + i\sin 2\theta$ (c) $-7i$ (d) $5 - 5i$.

16 (A favorite) Find the absolute value and the angle for $z = \sin\theta + i\cos\theta$ (careful). Locate this z in the complex plane. Multiply z by $\cos\theta + i\sin\theta$ to get _____.

17 Draw all eight solutions of $z^8 = 1$ in the complex plane. What is the rectangular form $a + ib$ of the root $z = \overline{w} = \exp(-2\pi i/8)$?

18 Locate the cube roots of 1 in the complex plane. Locate the cube roots of -1. Together these are the sixth roots of _____.

19 By comparing $e^{3i\theta} = \cos 3\theta + i\sin 3\theta$ with $(e^{i\theta})^3 = (\cos\theta + i\sin\theta)^3$, find the "triple angle" formulas for $\cos 3\theta$ and $\sin 3\theta$ in terms of $\cos\theta$ and $\sin\theta$.

20 Suppose the conjugate $\overline{z}$ is equal to the reciprocal $1/z$. What are all possible z's?

21 (a) Why do e^i and i^e both have absolute value 1?

(b) In the complex plane put stars near the points e^i and i^e.

(c) The number i^e could be $(e^{i\pi/2})^e$ or $(e^{5i\pi/2})^e$. Are those equal?

22 Draw the paths of these numbers from $t = 0$ to $t = 2\pi$ in the complex plane:

(a) e^{it} (b) $e^{(-1+i)t} = e^{-t}e^{it}$ (c) $(-1)^t = e^{t\pi i}$.

9.2 Hermitian and Unitary Matrices

The main message of this section can be presented in one sentence: ***When you transpose a complex vector z or matrix A, take the complex conjugate too***. Don't stop at z^{T} or A^{T}. Reverse the signs of all imaginary parts. From a column vector with $z_j = a_j + ib_j$, the good row vector $\overline{z}^{\mathrm{T}}$ is the *conjugate transpose* with components $a_j - ib_j$:

$$\textbf{Conjugate transpose} \qquad \overline{\boldsymbol{z}}^{\mathrm{T}} = \begin{bmatrix} \overline{z}_1 & \cdots & \overline{z}_n \end{bmatrix} = \begin{bmatrix} a_1 - ib_1 & \cdots & a_n - ib_n \end{bmatrix}. \tag{1}$$

Here is one reason to go to $\overline{z}$. The length squared of a real vector is $x_1^2 + \cdots + x_n^2$. The length squared of a complex vector is *not* $z_1^2 + \cdots + z_n^2$. With that wrong definition, the length of $(1, i)$ would be $1^2 + i^2 = 0$. A nonzero vector would have zero length—not good. Other vectors would have complex lengths. Instead of $(a + bi)^2$ we want $a^2 + b^2$, the *absolute value squared*. This is $(a + bi)$ times $(a - bi)$.

For each component we want z_j times $\overline{z}_j$, which is $|z_j|^2 = a_j^2 + b_j^2$. That comes when the components of $\boldsymbol{z}$ multiply the components of $\overline{\boldsymbol{z}}$:

$$\textbf{Length squared} \quad \begin{bmatrix} \overline{z}_1 & \cdots & \overline{z}_n \end{bmatrix} \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix} = |z_1|^2 + \cdots + |z_n|^2. \quad \textbf{This is} \quad \overline{\boldsymbol{z}}^{\mathrm{T}}\boldsymbol{z} = \|\boldsymbol{z}\|^2. \tag{2}$$

Now the squared length of $(1, i)$ is $1^2 + |i|^2 = 2$. The length is $\sqrt{2}$. The squared length of $(1 + i, 1 - i)$ is 4. The only vectors with zero length are zero vectors.

The length $\|z\|$ is the square root of $\overline{\boldsymbol{z}}^{\mathrm{T}}\boldsymbol{z} = \boldsymbol{z}^{\mathrm{H}}\boldsymbol{z} = |z_1|^2 + \cdots + |z_n|^2$

Before going further we replace two symbols by one symbol. Instead of a bar for the conjugate and T for the transpose, we just use a superscript H. Thus $\overline{\boldsymbol{z}}^{\mathbf{T}} = \boldsymbol{z}^{\mathbf{H}}$. This is "$\boldsymbol{z}$ Hermitian," the *conjugate transpose* of $\boldsymbol{z}$. The new word is pronounced "Hermeeshan." The new symbol applies also to matrices: The conjugate transpose of a matrix A is A^{H}.

Another popular notation is A^*. The MATLAB transpose command ' automatically takes complex conjugates ($\boldsymbol{z}'$ is $\boldsymbol{z}^{\mathrm{H}} = \overline{\boldsymbol{z}}^{\mathrm{T}}$ and A' is $A^{\mathrm{H}} = \overline{A}^{\mathrm{T}}$).

$$\boldsymbol{A^{\mathrm{H}}}\textbf{ is "}\boldsymbol{A}\textbf{ Hermitian"} \quad \text{If} \quad A = \begin{bmatrix} 1 & i \\ 0 & 1+i \end{bmatrix} \quad \text{then} \quad A^{\mathrm{H}} = \begin{bmatrix} 1 & 0 \\ -i & 1-i \end{bmatrix}$$

Complex Inner Products

For real vectors, the length squared is $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$—*the inner product of $\boldsymbol{x}$ with itself.* For complex vectors, the length squared is $\boldsymbol{z}^{\mathrm{H}}\boldsymbol{z}$. It will be very desirable if $\boldsymbol{z}^{\mathrm{H}}\boldsymbol{z}$ is the inner product of $\boldsymbol{z}$ with itself. To make that happen, the complex inner product should use the conjugate transpose (not just the transpose). This has no effect on real vectors.

DEFINITION The inner product of real or complex vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ is $\boldsymbol{u}^{\mathrm{H}}\boldsymbol{v}$:

$$\boldsymbol{u}^{\mathrm{H}}\boldsymbol{v} = \begin{bmatrix} \overline{u}_1 & \cdots & \overline{u}_n \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \overline{u}_1 v_1 + \cdots + \overline{u}_n v_n. \tag{3}$$

With complex vectors, $\boldsymbol{u}^{\mathrm{H}}\boldsymbol{v}$ is different from $\boldsymbol{v}^{\mathrm{H}}\boldsymbol{u}$. *The order of the vectors is now important.* In fact $\boldsymbol{v}^{\mathrm{H}}\boldsymbol{u} = \overline{v}_1 u_1 + \cdots + \overline{v}_n u_n$ is the complex conjugate of $\boldsymbol{u}^{\mathrm{H}}\boldsymbol{v}$. We have to put up with a few inconveniences for the greater good.

Example 1 The inner product of $\boldsymbol{u} = \begin{bmatrix} 1 \\ i \end{bmatrix}$ with $\boldsymbol{v} = \begin{bmatrix} i \\ 1 \end{bmatrix}$ is $\begin{bmatrix} 1 & -i \end{bmatrix} \begin{bmatrix} i \\ 1 \end{bmatrix} = 0.$

Example 1 is surprising. Those vectors $(1, i)$ and $(i, 1)$ don't look perpendicular. But they are. ***A zero inner product still means that the*** (complex) ***vectors are orthogonal.*** Similarly the vector $(1, i)$ is orthogonal to the vector $(1, -i)$. Their inner product is $1 - 1$. We are correctly getting zero for the inner product—where we would be incorrectly getting zero for the length of $(1, i)$ if we forgot to take the conjugate.

Note We have chosen to conjugate the first vector $\boldsymbol{u}$. Some authors choose the second vector $\boldsymbol{v}$. Their complex inner product would be $\boldsymbol{u}^{\mathrm{T}}\overline{\boldsymbol{v}}$. I think it is a free choice.

The inner product of $A\boldsymbol{u}$ with $\boldsymbol{v}$ equals the inner product of $\boldsymbol{u}$ with $A^{\mathrm{H}}\boldsymbol{v}$:

A^{H} is also called the "adjoint" of A

$$(A\boldsymbol{u})^{\mathrm{H}}\boldsymbol{v} = \boldsymbol{u}^{\mathrm{H}}(A^{\mathrm{H}}\boldsymbol{v}). \tag{4}$$

The conjugate of $A\boldsymbol{u}$ is $\overline{A\boldsymbol{u}}$. Transposing $\overline{A\boldsymbol{u}}$ gives $\overline{\boldsymbol{u}}^{\mathrm{T}}\overline{A}^{\mathrm{T}}$ as usual. This is $\boldsymbol{u}^{\mathrm{H}}A^{\mathrm{H}}$. Everything that should work, does work. The rule for $^{\mathrm{H}}$ comes from the rule for $^{\mathrm{T}}$.
We constantly use the fact that $(a - ib)(c - id)$ is the conjugate of $(a + ib)(c + id)$.

The conjugate transpose of AB is $(AB)^{\mathrm{H}} = B^{\mathrm{H}}A^{\mathrm{H}}$.

Hermitian Matrices $S = S^{\mathrm{H}}$

Among real matrices, *symmetric matrices* form the most important special class: $S = S^{\mathrm{T}}$. They have real eigenvalues and the orthogonal eigenvectors in an orthogonal matrix Q. Every real symmetric matrix can be written as $S = Q\Lambda Q^{-1}$ and also as $S = Q\Lambda Q^{\mathrm{T}}$ (because $Q^{-1} = Q^{\mathrm{T}}$). All this follows from $S^{\mathrm{T}} = S$, when S is real.

Among complex matrices, the special class contains the ***Hermitian matrices***: $S = S^{\mathrm{H}}$. The condition on the entries is $s_{ij} = \overline{s_{ji}}$. In this case we say that "S ***is*** Hermitian." *Every real symmetric matrix is Hermitian*, because taking its conjugate has no effect. The next matrix is also Hermitian, $S = S^{\mathrm{H}}$:

Example 2 $S = \begin{bmatrix} 2 & 3-3i \\ 3+3i & 5 \end{bmatrix}$ The main diagonal must be real since $s_{ii} = \overline{s_{ii}}$. Across it are conjugates $3+3i$ and $3-3i$.

This example will illustrate the three crucial properties of all Hermitian matrices.

If $S = S^{\mathrm{H}}$ and z is any real or complex column vector, the number $z^{\mathrm{H}}Sz$ is real.

Quick proof: $z^{\mathrm{H}}Sz$ is certainly 1 by 1. Take its conjugate transpose:

$$(z^{\mathrm{H}}Sz)^{\mathrm{H}} = z^{\mathrm{H}}S^{\mathrm{H}}(z^{\mathrm{H}})^{\mathrm{H}} \quad \text{which is } z^{\mathrm{H}}Sz \text{ again.}$$

So the number $z^{\mathrm{H}}Sz$ equals its conjugate and must be real. Here is that "energy" $z^{\mathrm{H}}Sz$:

$$\begin{bmatrix} \overline{z}_1 & \overline{z}_2 \end{bmatrix} \begin{bmatrix} 2 & 3-3i \\ 3+3i & 5 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \underset{\textit{diagonal}}{2\overline{z}_1 z_1 + 5\overline{z}_2 z_2} + \underset{\textit{off-diagonal}}{(3-3i)\overline{z}_1 z_2 + (3+3i) z_1 \overline{z}_2}.$$

The terms $2|z_1|^2$ and $5|z_2|^2$ from the diagonal are both real. The off-diagonal terms are conjugates of each other—so their sum is real. (The imaginary parts cancel when we add.) The whole expression $z^{\mathrm{H}}Sz$ is real, and this will make λ real.

Every eigenvalue of a Hermitian matrix is real.

Proof Suppose $Sz = \lambda z$. *Multiply both sides by z^{H} to get $z^{\mathrm{H}}Sz = \lambda z^{\mathrm{H}}z$.* On the left side, $z^{\mathrm{H}}Sz$ is real. On the right side, $z^{\mathrm{H}}z$ is the length squared, real and positive. So the ratio $\lambda = z^{\mathrm{H}}Sz/z^{\mathrm{H}}z$ is a real number. Q.E.D.

The example above has eigenvalues $\lambda = 8$ and $\lambda = -1$, real because $S = S^{\mathrm{H}}$:

$$\begin{vmatrix} 2-\lambda & 3-3i \\ 3+3i & 5-\lambda \end{vmatrix} = \lambda^2 - 7\lambda + 10 - |3+3i|^2$$
$$= \lambda^2 - 7\lambda + 10 - 18 = (\lambda - 8)(\lambda + 1).$$

The eigenvectors of a Hermitian matrix are orthogonal (when they correspond to different eigenvalues). If $Sz = \lambda z$ and $Sy = \beta y$ then $y^{\mathrm{H}}z = 0$.

Proof Multiply $Sz = \lambda z$ on the left by y^{H}. Multiply $y^{\mathrm{H}}S^{\mathrm{H}} = \beta y^{\mathrm{H}}$ on the right by z:

$$y^{\mathrm{H}}Sz = \lambda y^{\mathrm{H}}z \quad \text{and} \quad y^{\mathrm{H}}S^{\mathrm{H}}z = \beta y^{\mathrm{H}}z. \tag{5}$$

The left sides are equal so $\lambda y^{\mathrm{H}}z = \beta y^{\mathrm{H}}z$. Then $y^{\mathrm{H}}z$ **must be zero**.

The eigenvectors are orthogonal in our example with $\lambda = 8$ and $\beta = -1$:

$$(S - 8I)\boldsymbol{z} = \begin{bmatrix} -6 & 3-3i \\ 3+3i & -3 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{z} = \begin{bmatrix} 1 \\ 1+i \end{bmatrix}$$

$$(S + I)\boldsymbol{y} = \begin{bmatrix} 3 & 3-3i \\ 3+3i & 6 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \boldsymbol{y} = \begin{bmatrix} 1-i \\ -1 \end{bmatrix}.$$

Orthogonal eigenvectors $\qquad \boldsymbol{y}^{\mathrm{H}}\boldsymbol{z} = \begin{bmatrix} 1+i & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 1+i \end{bmatrix} = 0.$

These eigenvectors have squared length $1^2 + 1^2 + 1^2 = 3$. After division by $\sqrt{3}$ they are unit vectors. They were orthogonal, now they are ***orthonormal***. They go into the columns of the *eigenvector matrix* X, which diagonalizes S.

When S is real and symmetric, X is Q—an orthogonal matrix. Now S is complex and Hermitian. Its eigenvectors are complex and orthonormal. ***The eigenvector matrix X is like Q, but complex***: $\boldsymbol{Q^{\mathrm{H}}Q = I}$. We assign Q a new name "*unitary*" but still call it Q.

Unitary Matrices

A ***unitary matrix*** Q is a (complex) square matrix that has ***orthonormal columns***.

Unitary matrix that diagonalizes S: $\qquad Q = \dfrac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1-i \\ 1+i & -1 \end{bmatrix}$

This Q is also a Hermitian matrix. I didn't expect that! The example is almost too perfect. We will see that the eigenvalues of this Q must be 1 and -1.

The matrix test for real orthonormal columns was $Q^{\mathrm{T}}Q = I$. The zero inner products appear off the diagonal. In the complex case, Q^{T} becomes Q^{H}. The columns show themselves as orthonormal when Q^{H} multiplies Q. The inner products fill up $\boldsymbol{Q^{\mathrm{H}}Q = I}$:

Every matrix Q with orthonormal columns has $Q^{\mathrm{H}}Q = I$.

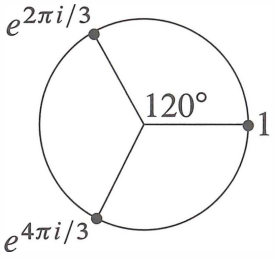
If Q is square, it is a unitary matrix. Then $\boldsymbol{Q^{\mathrm{H}} = Q^{-1}}$.

Suppose Q (with orthonormal columns) multiplies any $\boldsymbol{z}$. The vector length stays the same, because $\boldsymbol{z}^{\mathrm{H}}Q^{\mathrm{H}}Q\boldsymbol{z} = \boldsymbol{z}^{\mathrm{H}}\boldsymbol{z}$. If $\boldsymbol{z}$ is an eigenvector of Q we learn something more: ***The eigenvalues of unitary (and orthogonal) matrices Q all have absolute value*** $|\lambda| = 1$.

If Q is unitary then $\boldsymbol{\|Q\boldsymbol{z}\| = \|\boldsymbol{z}\|}$. ***Therefore $Q\boldsymbol{z} = \lambda\boldsymbol{z}$ leads to*** $\boldsymbol{|\lambda| = 1}$.

Our 2 by 2 example is both Hermitian ($Q = Q^{\mathrm{H}}$) and unitary ($Q^{-1} = Q^{\mathrm{H}}$). That means real eigenvalues and it means $|\lambda| = 1$. A real number with $|\lambda| = 1$ has only two possibilities: *The eigenvalues are* 1 *or* -1. The trace of Q is zero so $\lambda = 1$ and $\lambda = -1$.

Example 3 The 3 by 3 ***Fourier matrix*** is in Figure 9.3. Is it Hermitian? Is it unitary? F_3 is certainly symmetric. It equals its transpose. But it doesn't equal its conjugate transpose—*it is not Hermitian*. If you change i to $-i$, you get a different matrix.



Fourier matrix

$$F = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1 & 1 \\ 1 & e^{2\pi i/3} & e^{4\pi i/3} \\ 1 & e^{4\pi i/3} & e^{2\pi i/3} \end{bmatrix}.$$

Figure 9.3: The cube roots of 1 go into the Fourier matrix $F = F_3$.

Is F unitary? *Yes*. The squared length of every column is $\frac{1}{3}(1+1+1)$ (unit vector). The first column is orthogonal to the second column because $1 + e^{2\pi i/3} + e^{4\pi i/3} = 0$. This is the sum of the three numbers marked in Figure 9.3.

Notice the symmetry of the figure. If you rotate it by 120°, the three points are in the same position. Therefore their sum S also stays in the same position! The only possible sum in the same position after 120° rotation is $S = 0$.

Is column 2 of F orthogonal to column 3? Their dot product looks like

$$\tfrac{1}{3}(1 + e^{6\pi i/3} + e^{6\pi i/3}) = \tfrac{1}{3}(1+1+1).$$

This is not zero. The answer is wrong because we forgot to take complex conjugates. The complex inner product uses $^{\mathrm{H}}$ not $^{\mathrm{T}}$:

$$\begin{aligned}(\text{column 2})^{\mathrm{H}}(\text{column 3}) &= \tfrac{1}{3}(1 \cdot 1 + e^{-2\pi i/3}e^{4\pi i/3} + e^{-4\pi i/3}e^{2\pi i/3}) \\ &= \tfrac{1}{3}(1 + e^{2\pi i/3} + e^{-2\pi i/3}) = 0.\end{aligned}$$

So we do have orthogonality. ***Conclusion:*** F ***is a unitary matrix.***

The next section will study the n by n Fourier matrices. Among all complex unitary matrices, these are the most important. When we multiply a vector by F, we are computing its ***Discrete Fourier Transform***. When we multiply by F^{-1}, we are computing the *inverse transform*. The special property of unitary matrices is that $F^{-1} = F^{\mathrm{H}}$. The inverse transform only differs by changing i to $-i$:

Change i to $-i$

$$F^{-1} = F^{\mathrm{H}} = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1 & 1 \\ 1 & e^{-2\pi i/3} & e^{-4\pi i/3} \\ 1 & e^{-4\pi i/3} & e^{-2\pi i/3} \end{bmatrix}.$$

Everyone who works with F recognizes its value. The last section of this chapter will bring together Fourier analysis and complex numbers and linear algebra.

Problem Set 9.2

1 Find the lengths of $\boldsymbol{u} = (1+i, 1-i, 1+2i)$ and $\boldsymbol{v} = (i, i, i)$. Find $\boldsymbol{u}^{\mathrm{H}}\boldsymbol{v}$ and $\boldsymbol{v}^{\mathrm{H}}\boldsymbol{u}$.

2 Compute $A^{\mathrm{H}}A$ and AA^{H}. Those are both _____ matrices:

$$A = \begin{bmatrix} i & 1 & i \\ 1 & i & i \end{bmatrix}.$$

3 Solve $A\boldsymbol{z} = \boldsymbol{0}$ to find a vector $\boldsymbol{z}$ in the nullspace of A in Problem 2. Show that $\boldsymbol{z}$ is orthogonal to the columns of A^{H}. Show that $\boldsymbol{z}$ is *not* orthogonal to the columns of A^{T}. ***The good row space is no longer*** $\boldsymbol{C}(A^{\mathrm{T}})$. **Now it is** $\boldsymbol{C(A^{\mathrm{H}})}$.

4 Problem 3 indicates that the four fundamental subspaces are $\boldsymbol{C}(A)$ and $\boldsymbol{N}(A)$ and _____ and _____. Their dimensions are still r and $n-r$ and r and $m-r$. They are still orthogonal subspaces. *The symbol* $^{\mathrm{H}}$ *takes the place of* $^{\mathrm{T}}$.

5 (a) Prove that $A^{\mathrm{H}}A$ is always a Hermitian matrix.

(b) If $A\boldsymbol{z} = \boldsymbol{0}$ then $A^{\mathrm{H}}A\boldsymbol{z} = \boldsymbol{0}$. If $A^{\mathrm{H}}A\boldsymbol{z} = \boldsymbol{0}$, multiply by $\boldsymbol{z}^{\mathrm{H}}$ to prove that $A\boldsymbol{z} = \boldsymbol{0}$. The nullspaces of A and $A^{\mathrm{H}}A$ are _____. Therefore $A^{\mathrm{H}}A$ is an invertible Hermitian matrix when the nullspace of A contains only $\boldsymbol{z} = 0$.

6 True or false (give a reason if true or a counterexample if false):

(a) If A is a real matrix then $A + iI$ is invertible.

(b) If S is a Hermitian matrix then $S + iI$ is invertible.

(c) If Q is a unitary matrix then $Q + iI$ is invertible.

7 When you multiply a Hermitian matrix by a real number c, is cS still Hermitian? Show that iS is skew-Hermitian when S is Hermitian. The 3 by 3 Hermitian matrices are a subspace provided the "scalars" are real numbers.

8 Which classes of matrices does P belong to: invertible, Hermitian, unitary?

$$P = \begin{bmatrix} 0 & i & 0 \\ 0 & 0 & i \\ i & 0 & 0 \end{bmatrix}.$$

Compute P^2, P^3, and P^{100}. What are the eigenvalues of P?

9 Find the unit eigenvectors of P in Problem 8, and put them into the columns of a unitary matrix Q. What property of P makes these eigenvectors orthogonal?

10 Write down the 3 by 3 circulant matrix $C = 2I + 5P$. It has the same eigenvectors as P in Problem 8. Find its eigenvalues.

11 If Q and U are unitary matrices, show that Q^{-1} is unitary and also QU is unitary. Start from $Q^{\mathrm{H}}Q = I$ and $U^{\mathrm{H}}U = I$.

12 How do you know that the determinant of every Hermitian matrix is real?

13 The matrix $A^{\mathrm{H}}A$ is not only Hermitian but also positive definite, when the columns of A are independent. Proof: $\boldsymbol{z}^{\mathrm{H}}A^{\mathrm{H}}A\boldsymbol{z}$ is positive if $\boldsymbol{z}$ is nonzero because _____.

14 Diagonalize these Hermitian matrices to reach $S = Q\Lambda Q^{\mathrm{H}}$:

$$S = \begin{bmatrix} 0 & 1-i \\ i+1 & 1 \end{bmatrix} \qquad \text{and} \qquad S = \begin{bmatrix} 2 & 1+i \\ i-1 & 3 \end{bmatrix}.$$

15 Diagonalize this skew-Hermitian matrix to reach $K = Q\Lambda Q^{\mathrm{H}}$. All λ's are _____:

$$K = \begin{bmatrix} 0 & -1+i \\ 1+i & i \end{bmatrix}.$$

16 Diagonalize this orthogonal matrix to reach $U = Q\Lambda Q^{\mathrm{H}}$. Now all λ's are _____:

$$U = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

17 Diagonalize this unitary matrix to reach $U = Q\Lambda Q^{\mathrm{H}}$. Again all λ's are _____:

$$U = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1-i \\ 1+i & -1 \end{bmatrix}.$$

18 If $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ is an orthonormal basis for $\mathbf{C}^n$, the matrix with those columns is a _____ matrix. Show that any vector $\boldsymbol{z}$ equals $(\boldsymbol{v}_1^{\mathrm{H}}\boldsymbol{z})\boldsymbol{v}_1 + \cdots + (\boldsymbol{v}_n^{\mathrm{H}}\boldsymbol{z})\boldsymbol{v}_n$.

19 $\boldsymbol{v} = (1, i, 1), \boldsymbol{w} = (i, 1, 0)$ and $\boldsymbol{z} =$ _____ are an orthogonal basis for _____.

20 If $S = A + iB$ is a Hermitian matrix, are its real and imaginary parts symmetric?

21 The (complex) dimension of $\mathbf{C}^n$ is _____. Find a non-real basis for $\mathbf{C}^n$.

22 Describe all 1 by 1 and 2 by 2 Hermitian matrices and unitary matrices.

23 How are the eigenvalues of A^{H} related to the eigenvalues of the square matrix A?

24 If $\boldsymbol{u}^{\mathrm{H}}\boldsymbol{u} = 1$ show that $I - 2\boldsymbol{u}\boldsymbol{u}^{\mathrm{H}}$ is Hermitian and also unitary. The rank-one matrix $\boldsymbol{u}\boldsymbol{u}^{\mathrm{H}}$ is the projection onto what line in $\mathbf{C}^n$?

25 If $A + iB$ is a unitary matrix (A and B are real) show that $Q = \begin{bmatrix} \mathbf{A} & \mathbf{-B} \\ \mathbf{B} & \mathbf{A} \end{bmatrix}$ is an orthogonal matrix.

26 If $A + iB$ is Hermitian (A and B are real) show that $\begin{bmatrix} \mathbf{A} & \mathbf{-B} \\ \mathbf{B} & \mathbf{A} \end{bmatrix}$ is symmetric.

27 Prove that the inverse of a Hermitian matrix is also Hermitian (transpose $S^{-1}S = I$).

28 A matrix with orthonormal eigenvectors has the form $N = Q\Lambda Q^{-1} = Q\Lambda Q^{\mathrm{H}}$. *Prove that* $NN^{\mathrm{H}} = N^{\mathrm{H}}N$. These N are exactly the **normal matrices**. Examples are Hermitian, skew-Hermitian, and unitary matrices. Construct a 2 by 2 normal matrix from $Q\Lambda Q^{\mathrm{H}}$ by choosing complex eigenvalues in Λ.

9.3 The Fast Fourier Transform

Many applications of linear algebra take time to develop. It is not easy to explain them in an hour. The teacher and the author must choose between completing the theory and adding new applications. Often the theory wins, but this section is an exception. It explains the most valuable numerical algorithm in the last century.

We want to multiply quickly by F and F^{-1}, the Fourier matrix and its inverse. This is achieved by the Fast Fourier Transform. An ordinary product Fc uses n^2 multiplications (F has n^2 entries). The FFT needs only n times $\frac{1}{2}\log_2 n$. We will see how.

The FFT has revolutionized signal processing. Whole industries are speeded up by this one idea. Electrical engineers are the first to know the difference—they take your Fourier transform as they meet you (if you are a function). Fourier's idea is to represent f as a sum of harmonics $c_k e^{ikx}$. The function is seen in *frequency space* through the coefficients c_k, instead of *physical space* through its values $f(x)$. The passage backward and forward between c's and f's is by the Fourier transform. Fast passage is by the FFT.

Roots of Unity and the Fourier Matrix

Quadratic equations have two roots (or one repeated root). Equations of degree n have n roots (counting repetitions). This is the Fundamental Theorem of Algebra, and to make it true we must allow complex roots. This section is about the very special equation $z^n = 1$. The solutions z are the "nth roots of unity." They are n evenly spaced points around the unit circle in the complex plane.

Figure 9.4 shows the eight solutions to $z^8 = 1$. Their spacing is $\frac{1}{8}(360°) = 45°$. The first root is at $45°$ or $\theta = 2\pi/8$ radians. ***It is the complex number*** $w = e^{i\theta} = \boldsymbol{e^{i2\pi/8}}$. We call this number w_8 to emphasize that it is an 8th root. You could write it in terms of $\cos\frac{2\pi}{8}$ and $\sin\frac{2\pi}{8}$, but don't do it. The seven other 8th roots are $w^2, w^3, \ldots, w^8$, going around the circle. Powers of w are best in polar form, because we work only with the angles $\frac{2\pi}{8}, \frac{4\pi}{8}, \ldots, \frac{16\pi}{8} = 2\pi$. Those 8 angles in degrees are $45°, 90°, 135°, \ldots, 360°$.

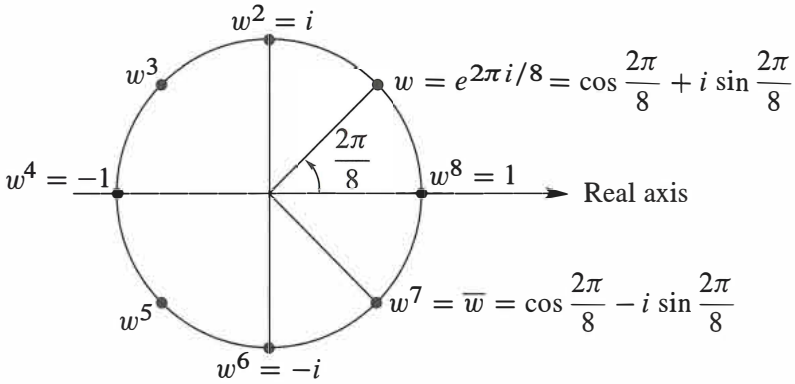


Figure 9.4: The eight solutions to $z^8 = 1$ are $1, w, w^2, \ldots, w^7$ with $w = (1+i)/\sqrt{2}$.

The fourth roots of 1 are also in the figure. They are $i, -1, -i, 1$. The angle is now $2\pi/4$ or $90°$. The first root $w_4 = e^{2\pi i/4}$ is nothing but i. Even the square roots of 1 are seen, with $w_2 = e^{i2\pi/2} = -1$. Do not despise those square roots 1 and -1. The idea behind the FFT is to go from an **8 by 8** Fourier matrix (containing powers of w_8) to the **4 by 4** matrix below (with powers of $w_4 = i$). The same idea goes from 4 to 2. By exploiting the connections of F_8 down to F_4 and up to F_{16} (and beyond), the FFT makes multiplication by F_{1024} very quick.

We describe the *Fourier matrix*, first for $n = 4$. Its rows contain powers of 1 and w and w^2 and w^3. These are the fourth roots of 1, and their powers come in a special order.

Fourier matrix $n = 4$ $\boldsymbol{w = i}$

$$F = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & w & w^2 & w^3 \\ 1 & w^2 & w^4 & w^6 \\ 1 & w^3 & w^6 & w^9 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & i^2 & i^3 \\ 1 & i^2 & i^4 & i^6 \\ 1 & i^3 & i^6 & i^9 \end{bmatrix}.$$

The matrix is symmetric ($F = F^{\mathrm{T}}$). It is *not* Hermitian. Its main diagonal is not real. But $\frac{1}{2}F$ is a ***unitary matrix***, which means that $(\frac{1}{2}F^{\mathrm{H}})(\frac{1}{2}F) = I$:

The columns of F give $\boldsymbol{F^{\mathrm{H}}F = 4I}$. ***Its inverse is*** $\boldsymbol{\frac{1}{4}F^{\mathrm{H}}}$ ***which is*** $\boldsymbol{F^{-1} = \frac{1}{4}\overline{F}}$.

The inverse changes from $w = i$ to $\overline{w} = -i$. That takes us from F to $\overline{F}$. When the Fast Fourier Transform gives a quick way to multiply by F, it does the same for $\overline{F}$ and F^{-1}.

Every column has length $\sqrt{n}$. So the unitary matrices are $Q = F/\sqrt{n}$ and $Q^{-1} = \overline{F}/\sqrt{n}$. We avoid $\sqrt{n}$ and just use F and $\boldsymbol{F^{-1} = \overline{F}/n}$. The main point is to multiply F times c_0, c_1, c_2, c_3:

4-point Fourier series

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \end{bmatrix} = F\boldsymbol{c} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & w & w^2 & w^3 \\ 1 & w^2 & w^4 & w^6 \\ 1 & w^3 & w^6 & w^9 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix}. \tag{1}$$

The input is four complex coefficients c_0, c_1, c_2, c_3. The output is four function values y_0, y_1, y_2, y_3. The first output $y_0 = c_0 + c_1 + c_2 + c_3$ is the value of the Fourier series $\sum c_k e^{ikx}$ at $x = 0$. *The second output is the value of that series* $\sum c_k e^{ikx}$ *at* $x = 2\pi/4$:

$$y_1 = c_0 + c_1 e^{i2\pi/4} + c_2 e^{i4\pi/4} + c_3 e^{i6\pi/4} = c_0 + c_1 w + c_2 w^2 + c_3 w^3.$$

The third and fourth outputs y_2 and y_3 are the values of $\sum c_k e^{ikx}$ at $x = 4\pi/4$ and $x = 6\pi/4$. These are *finite* Fourier series! *They contain* $n = 4$ *terms and they are evaluated at* $n = 4$ *points*. Those points $x = 0, 2\pi/4, 4\pi/4, 6\pi/4$ are equally spaced.

The next point would be $x = 8\pi/4$ which is 2π. Then the series is back to y_0, because $e^{2\pi i}$ is the same as $e^0 = 1$. Everything cycles around with period 4. In this world $2 + 2$ is 0 because $(w^2)(w^2) = w^0 = 1$. We follow the convention that ***j and k go from*** $\boldsymbol{0}$ ***to*** $\boldsymbol{n-1}$ (instead of 1 to n). The "zeroth row" and "zeroth column" of F contain all ones.

The n by n Fourier matrix contains powers of $w = e^{2\pi i/n}$:

$$F_n \boldsymbol{c} = \begin{bmatrix} 1 & 1 & 1 & \cdot & 1 \\ 1 & w & w^2 & \cdot & w^{n-1} \\ 1 & w^2 & w^4 & \cdot & w^{2(n-1)} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & w^{n-1} & w^{2(n-1)} & \cdot & w^{(n-1)^2} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \cdot \\ c_{n-1} \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \cdot \\ y_{n-1} \end{bmatrix} = \boldsymbol{y}. \tag{2}$$

F_n is symmetric but not Hermitian. ***Its columns are orthogonal***, and $F_n \overline{F}_n = nI$. *Then* F_n^{-1} *is* $\overline{F}_n/n$. The inverse contains powers of $\overline{w}_n = e^{-2\pi i/n}$. Look at the pattern in F:

The entry in row j, column k ***is*** w^{jk}. ***Row zero and column zero contain*** $w^0 = 1$.

When we multiply $\boldsymbol{c}$ by F_n, we sum the series at n points. *When we multiply* $\boldsymbol{y}$ *by* F_n^{-1}, *we find the coefficients* $\boldsymbol{c}$ *from the function values* $\boldsymbol{y}$. In MATLAB that command is $\boldsymbol{c} = \mathsf{fft}(\boldsymbol{y})$. The matrix F passes from "frequency space" to "physical space."

Important note. Many authors prefer to work with $\omega = e^{-2\pi i/N}$, which is the *complex conjugate* of our w. (They often use the Greek omega, and I will do that to keep the two options separate.) With this choice, their DFT matrix contains powers of ω not w. It is $\overline{F}$, the conjugate of our F. $\overline{F}$ goes from physical space to frequency space.

$\overline{F}$ is a completely reasonable choice! MATLAB uses $\omega = e^{-2\pi i/N}$. The DFT matrix $\mathsf{fft}(\mathsf{eye}(N))$ contains powers of this number $\omega = \overline{w}$. **The Fourier matrix F with w's reconstructs $\boldsymbol{y}$ from $\boldsymbol{c}$. The matrix $\overline{F}$ with ω's computes Fourier coefficients as $\mathsf{fft}(\boldsymbol{y})$.**

Also important. When a function $f(x)$ has period 2π, and we change x to $e^{i\theta}$, the function is defined around the unit circle (where $z = e^{i\theta}$). The Discrete Fourier Transform is the same as interpolation. Find the polynomial $p(z) = c_0 + c_1 z + \cdots + c_{n-1} z^{n-1}$ that matches n values $f_0, \ldots, f_{n-1}$:

Interpolation Find $c_0, \ldots, c_{n-1}$ so that $p(z) = f$ at n points $z = 1, \ldots, w^{n-1}$

The Fourier matrix is the Vandermonde matrix for interpolation at those n special points.

One Step of the Fast Fourier Transform

We want to multiply F times $\boldsymbol{c}$ as quickly as possible. Normally a matrix times a vector takes n^2 separate multiplications—the matrix has n^2 entries. You might think it is impossible to do better. (If the matrix has zero entries then multiplications can be skipped. But the Fourier matrix has no zeros!) By using the special pattern w^{jk} for its entries, F can be factored in a way that produces many zeros. This is the **FFT**.

The key idea is to connect $\boldsymbol{F_n}$ ***with the half-size Fourier matrix*** $\boldsymbol{F_{n/2}}$. Assume that n is a power of 2 (say $n = 2^{10} = 1024$). We will connect F_{1024} to ***two copies of*** $\boldsymbol{F_{512}}$.

When $n = 4$, the key is in the relation between F_4 and two copies of F_2 :

$$\boldsymbol{F_4} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & i^2 & i^3 \\ 1 & i^2 & i^4 & i^6 \\ 1 & i^3 & i^6 & i^9 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \boldsymbol{F_2} & \\ & \boldsymbol{F_2} \end{bmatrix} = \begin{bmatrix} 1 & 1 & & \\ 1 & i^2 & & \\ & & 1 & 1 \\ & & 1 & i^2 \end{bmatrix}.$$

On the left is F_4, with no zeros. On the right is a matrix that is half zero. The work is cut in half. But wait, those matrices are not the same. We need two sparse and simple matrices to complete the FFT factorization:

Factors for FFT

$$F_4 = \begin{bmatrix} 1 & & 1 & \\ & 1 & & i \\ 1 & & -1 & \\ & 1 & & -i \end{bmatrix} \begin{bmatrix} 1 & 1 & & \\ 1 & i^2 & & \\ & & 1 & 1 \\ & & 1 & i^2 \end{bmatrix} \begin{bmatrix} 1 & & & \\ & & 1 & \\ & 1 & & \\ & & & 1 \end{bmatrix}. \tag{3}$$

The last matrix is a permutation. It puts the even c's (c_0 and c_2) ahead of the odd c's (c_1 and c_3). The middle matrix performs half-size transforms F_2 and F_2 on the even c's and odd c's. The matrix at the left combines the two half-size outputs—in a way that produces the correct full-size output $\boldsymbol{y} = F_4\boldsymbol{c}$.

The same idea applies when $n = 1024$ and $m = \frac{1}{2}n = 512$. The number w is $e^{2\pi i/1024}$. It is at the angle $\theta = 2\pi/1024$ on the unit circle. The Fourier matrix F_{1024} is full of powers of w. The first stage of the FFT is the great factorization discovered by Cooley and Tukey (and foreshadowed in 1805 by Gauss):

$$\boldsymbol{F_{1024}} = \begin{bmatrix} I_{512} & D_{512} \\ I_{512} & -D_{512} \end{bmatrix} \begin{bmatrix} \boldsymbol{F_{512}} & \\ & \boldsymbol{F_{512}} \end{bmatrix} \begin{bmatrix} \text{even-odd} \\ \text{permutation} \end{bmatrix}. \tag{4}$$

I_{512} is the identity matrix. D_{512} is the diagonal matrix with entries $(1, w, \ldots, w^{511})$. The two copies of F_{512} are what we expected. Don't forget that they use the 512th root of unity (which is nothing but w^2!!) The permutation matrix separates the incoming vector $\boldsymbol{c}$ into its even and odd parts $\boldsymbol{c'} = (c_0, c_2, \ldots, c_{1022})$ and $\boldsymbol{c''} = (c_1, c_3, \ldots, c_{1023})$.

Here are the algebra formulas which say the same thing as that factorization of F_{1024}:

(One step of the FFT) Set $\boldsymbol{m = \frac{1}{2}n}$. The first m and last m components of $\boldsymbol{y} = F_n\boldsymbol{c}$ combine the half-size transforms $\boldsymbol{y'} = F_m\boldsymbol{c'}$ and $\boldsymbol{y''} = F_m\boldsymbol{c''}$. Equation (4) shows this step from n to $m = n/2$ as $I\boldsymbol{y'} + D\boldsymbol{y''}$ and $I\boldsymbol{y'} - D\boldsymbol{y''}$:

$$\begin{aligned} y_j &= y'_j + (w_n)^j y''_j, \quad j = 0, \ldots, m-1 \\ y_{j+m} &= y'_j - (w_n)^j y''_j, \quad j = 0, \ldots, m-1. \end{aligned} \tag{5}$$

Split $\boldsymbol{c}$ into $\boldsymbol{c'}$ and $\boldsymbol{c''}$, transform them by F_m into $\boldsymbol{y'}$ and $\boldsymbol{y''}$, then (5) reconstructs $\boldsymbol{y}$.

Those formulas come from separating $c_0 \ldots, c_{n-1}$ into even c_{2k} and odd c_{2k+1} : w is w_n.

$$\boldsymbol{y = Fc} \qquad y_j = \sum_0^{n-1} w^{jk} c_k = \sum_0^{m-1} w^{2jk} c_{2k} + \sum_0^{m-1} w^{j(2k+1)} c_{2k+1} \text{ with } m = \tfrac{1}{2}n. \tag{6}$$

The even c's go into $\boldsymbol{c'} = (c_0, c_2, \ldots)$ and the odd c's go into $\boldsymbol{c''} = (c_1, c_3, \ldots)$. Then come the transforms $F_m c'$ and $F_m c''$. **The key is $\boldsymbol{w_n^2 = w_m}$.** This gives $w_n^{2jk} = w_m^{jk}$.

$$\textbf{Rewrite (6)} \quad y_j = \sum (w_m)^{jk} c_k' \; + \; (w_n)^j \sum (w_m)^{jk} c_k'' = y_j' \; + \; (w_n)^j y_j'' \,. \tag{7}$$

For $j \geq m$, the minus sign in (5) comes from factoring out $(w_n)^m = -1$ from $(w_n)^j$.

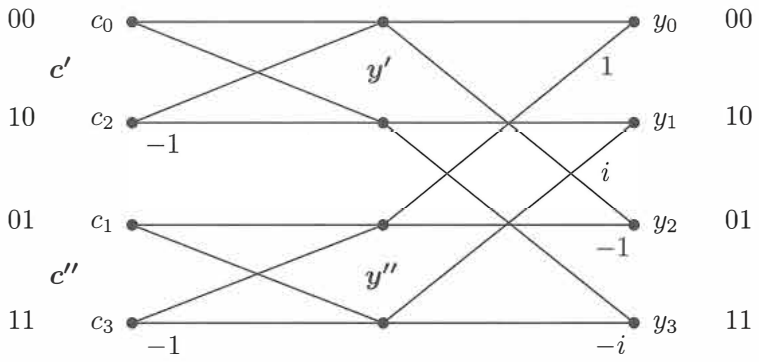
MATLAB easily separates even c's from odd c's and multiplies by w_n^j. We use conj(F) or equivalently MATLAB's inverse transform ifft, because fft is based on $\omega = \overline{w} = e^{-2\pi i/n}$. Problem 16 shows that F and conj(F) are linked by permuting rows.

FFT step from n to $n/2$ in MATLAB

```
y' = ifft(c(0:2:n-2))*n/2;
y'' = ifft(c(1:2:n-1))*n/2;
d = w.^(0:n/2-1)';
y = [y' + d.*y'; y' - d.*y''];
```

The flow graph shows c' and c'' going through the half-size F_2. Those steps are called "*butterflies*," from their shape. Then the outputs y' and y'' are combined (multiplying y'' by $1, i$ from D and also by $-1, -i$ from $-D$) to produce $y = F_4 c$.

This reduction from F_n to two F_m's almost cuts the work in half—you see the zeros in the matrix factorization. That reduction is good but not great. The full idea of the **FFT** is much more powerful. It saves much more than half the time.



The Full FFT by Recursion

If you have read this far, you probably guessed what comes next. We reduced F_n to $F_{n/2}$. ***Keep going to $F_{n/4}$.*** Every F_{512} leads to F_{256}. Then 256 leads to 128. *That is recursion.*

Recursion is a basic principle of many fast algorithms. Here is step 2 with four copies of F_{256} and D (256 powers of ω_{512}). Evens of evens $c_0, c_4, c_8, \ldots$ come first:

$$\begin{bmatrix} F_{512} & \\ & F_{512} \end{bmatrix} = \begin{bmatrix} I & D & & \\ I & -D & & \\ & & I & D \\ & & I & -D \end{bmatrix} \begin{bmatrix} F & & & \\ & F & & \\ & & F & \\ & & & F \end{bmatrix} \begin{bmatrix} \text{pick} & 0,4,8,\ldots \\ \text{pick} & 2,6,10,\ldots \\ \text{pick} & 1,5,9,\ldots \\ \text{pick} & 3,7,11,\ldots \end{bmatrix}.$$

We will count the individual multiplications, to see how much is saved. Before the **FFT** was invented, the count was the usual $n^2 = (1024)^2$. This is about a million multiplications. I am not saying that they take a long time. The cost becomes large when we have many, many transforms to do—which is typical. Then the saving by the FFT is also large:

The final count for size $n = 2^\ell$ is reduced from n^2 to $\frac{1}{2}n\ell$.

The number 1024 is 2^{10}, so $\ell = 10$. The original count of $(1024)^2$ is reduced to (5)(1024). The saving is a factor of 200. A million is reduced to five thousand. That is why the FFT has revolutionized signal processing.

Here is the reasoning behind $\frac{1}{2}n\ell$. There are ℓ levels, going from $n = 2^\ell$ down to $n = 1$. Each level has $n/2$ multiplications from the diagonal D's, to reassemble the half-size outputs from the lower level. This yields the final count $\frac{1}{2}n\ell$, which is $\frac{1}{2}n\log_2 n$.

One last note about this remarkable algorithm. There is an amazing rule for the order that the c's enter the FFT, after all the even-odd permutations. Write the numbers 0 to $n-1$ in binary (like $00, 01, 10, 11$ for $n = 4$). Reverse the order of those digits: $00, 10, 01, 11$. That gives the **bit-reversed order $\mathbf{0, 2, 1, 3}$** with evens before odds (See Problem 17.) The complete picture shows the c's in bit-reversed order, the $\ell = \log_2 n$ steps of the recursion, and the final output $y_0, \ldots, y_{n-1}$ which is F_n times $\boldsymbol{c}$.

The chapter ends with that very fundamental idea, a matrix multiplying a vector.

Problem Set 9.3

1 Multiply the three matrices in equation (3) and compare with F. In which six entries do you need to know that $i^2 = -1$?

2 Invert the three factors in equation (3) to find a fast factorization of F^{-1}.

3 F is symmetric. So transpose equation (3) to find a new Fast Fourier Transform!

4 All entries in the factorization of F_6 involve powers of $w_6 =$ sixth root of 1:

$$F_6 = \begin{bmatrix} I & D \\ I & -D \end{bmatrix} \begin{bmatrix} F_3 & \\ & F_3 \end{bmatrix} \begin{bmatrix} P \end{bmatrix}.$$

Write down these matrices with $1, w_6, w_6^2$ in D and $w_3 = w_6^2$ in F_3. Multiply!

5 If $\boldsymbol{v} = (1, 0, 0, 0)$ and $\boldsymbol{w} = (1, 1, 1, 1)$, show that $F\boldsymbol{v} = \boldsymbol{w}$ and $F\boldsymbol{w} = 4\boldsymbol{v}$. Therefore $F^{-1}\boldsymbol{w} = \boldsymbol{v}$ and $F^{-1}\boldsymbol{v} =$ _____.

6 What is F^2 and what is F^4 for the 4 by 4 Fourier matrix?

7 Put the vector $\boldsymbol{c} = (1, 0, 1, 0)$ through the three steps of the FFT to find $\boldsymbol{y} = F\boldsymbol{c}$. Do the same for $\boldsymbol{c} = (0, 1, 0, 1)$.

8 Compute $\boldsymbol{y} = F_8\boldsymbol{c}$ by the three FFT steps for $\boldsymbol{c} = (1, 0, 1, 0, 1, 0, 1, 0)$. Repeat the computation for $\boldsymbol{c} = (0, 1, 0, 1, 0, 1, 0, 1)$.

9 If $w = e^{2\pi i/64}$ then w^2 and $\sqrt{w}$ are among the _____ and _____ roots of 1.

10 (a) Draw all the sixth roots of 1 on the unit circle. Prove they add to zero.

(b) What are the three cube roots of 1? Do they also add to zero?

11 The columns of the Fourier matrix F are the *eigenvectors* of the cyclic permutation P (see Section 8.3). Multiply PF to find the eigenvalues $\lambda_1, \lambda_2, \lambda_3, \lambda_4$:

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & i^2 & i^3 \\ 1 & i^2 & i^4 & i^6 \\ 1 & i^3 & i^6 & i^9 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & i^2 & i^3 \\ 1 & i^2 & i^4 & i^6 \\ 1 & i^3 & i^6 & i^9 \end{bmatrix} \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \lambda_3 & \\ & & & \lambda_4 \end{bmatrix}.$$

This is $PF = F\Lambda$ or $P = F\Lambda F^{-1}$. The eigenvector matrix (usually X) is F.

12 The equation $\det(P - \lambda I) = 0$ is $\lambda^4 = 1$. This shows again that the eigenvalues are $\lambda =$ _____. Which permutation P has eigenvalues = cube roots of 1?

13 (a) Two eigenvectors of C are $(1,1,1,1)$ and $(1, i, i^2, i^3)$. Find the eigenvalues e.

$$\begin{bmatrix} c_0 & c_1 & c_2 & c_3 \\ c_3 & c_0 & c_1 & c_2 \\ c_2 & c_3 & c_0 & c_1 \\ c_1 & c_2 & c_3 & c_0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = e_1 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad \text{and} \quad C \begin{bmatrix} 1 \\ i \\ i^2 \\ i^3 \end{bmatrix} = e_2 \begin{bmatrix} 1 \\ i \\ i^2 \\ i^3 \end{bmatrix}.$$

(b) $P = F\Lambda F^{-1}$ immediately gives $P^2 = F\Lambda^2 F^{-1}$ and $P^3 = F\Lambda^3 F^{-1}$. Then $\boldsymbol{C = c_0 I + c_1 P + c_2 P^2 + c_3 P^3 = F(c_0 I + c_1 \Lambda + c_2 \Lambda^2 + c_3 \Lambda^3) F^{-1} = FEF^{-1}}$. That matrix E in parentheses is diagonal. It contains the _____ of C.

14 Find the eigenvalues of the "periodic" $-1, 2, -1$ matrix from $E = 2I - \Lambda - \Lambda^3$, with the eigenvalues of P in Λ. The -1's in the corners make this matrix periodic:

$$C = \begin{bmatrix} 2 & -1 & 0 & -1 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ -1 & 0 & -1 & 2 \end{bmatrix} \quad \text{has } c_0 = 2, c_1 = -1, c_2 = 0, c_3 = -1.$$

15 ***Fast convolution = Fast multiplication by C:*** To multiply C times a vector $\boldsymbol{x}$, we can multiply $F(E(F^{-1}\boldsymbol{x}))$ instead. The direct way uses n^2 separate multiplications. Knowing E and F, the second way uses only $n \log_2 n + n$ multiplications. How many of those come from E, how many from F, and how many from F^{-1}?

16 **Notice.** Why is row i of $\overline{F}$ the same as row $N - i$ of F (numbered 0 to $N - 1$)?

17 What is the *bit-reversed order* of the numbers $0, 1, \ldots, 7$? Write them all in binary (base 2) as $000, 001, \ldots, 111$ and reverse each order. The 8 numbers are now _____.

Chapter 10

Applications

10.1 Graphs and Networks

Over the years I have seen one model so often, and I found it so basic and useful, that I always put it first. The model consists of ***nodes connected by edges***. This is called a ***graph***.

Graphs of the usual kind display functions $f(x)$. Graphs of this node-edge kind lead to matrices. This section is about the ***incidence matrix*** of a graph—which tells how the n nodes are connected by the m edges. Normally $m > n$, there are more edges than nodes.

For any m by n matrix there are two fundamental subspaces in $\mathbf{R}^n$ and two in $\mathbf{R}^m$. They are the row spaces and nullspaces of A and A^{T}. Their *dimensions* $r, n-r$ and $r, m-r$ come from the most important theorem in linear algebra. The second part of that theorem is the *orthogonality* of the row space and nullspace. Our goal is to show how examples from graphs illuminate this Fundamental Theorem of Linear Algebra.

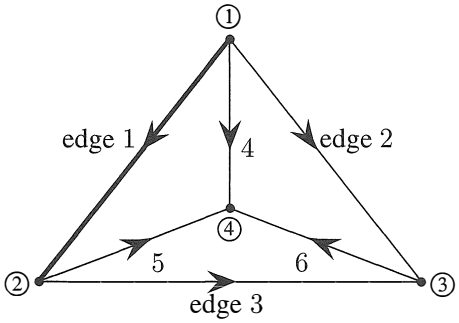
When I construct a ***graph*** and its ***incidence matrix***, the subspace dimensions will be easy to discover. But we want the subspaces themselves—and orthogonality helps. It is essential to connect the subspaces to the graph they come from. By specializing to incidence matrices, **the laws of linear algebra become Kirchhoff's laws**. Please don't be put off by the words "current" and "voltage." These rectangular matrices are the best.

Every entry of an incidence matrix is 0 or 1 or -1. This continues to hold during elimination. All pivots and multipliers are ± 1. Therefore both factors in $A = LU$ also contain $0, 1, -1$. So do the nullspace matrices! All four subspaces have basis vectors with these exceptionally simple components. The matrices are not concocted for a textbook, they come from a model that is absolutely essential in pure and applied mathematics.

The Incidence Matrix

Figure 10.1 displays a graph with $m = 6$ edges and $n = 4$ nodes. The 6 by 4 matrix A tells which nodes are connected by which edges. The first row $-1, 1, 0, 0$ shows that the first edge goes *from node* 1 *to node* 2 (-1 for node 1 because the arrow goes out, $+1$ for node 2 with arrow in).

Row numbers in A are edge numbers, column numbers $1, 2, 3, 4$ are node numbers!

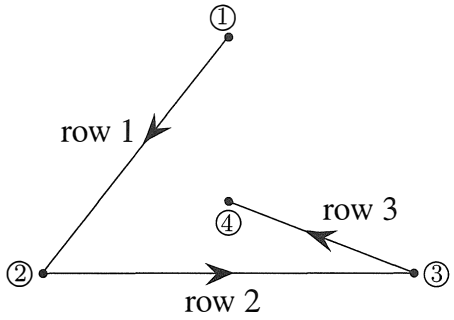


$$A = \begin{matrix} \\ \textcircled{1}\ \textcircled{2}\ \textcircled{3}\ \textcircled{4} \end{matrix} \text{ node}$$

$$\begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{bmatrix} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} \text{ edge}$$

Figure 10.1: Complete graph with $m = 6$ edges and $n = 4$ nodes: 6 by 4 incidence matrix A.

You can write down the matrix by looking at the graph. The second graph has the same four nodes but only three edges. Its incidence matrix B is 3 by 4.



$$B = \begin{matrix} \textcircled{1}\ \textcircled{2}\ \textcircled{3}\ \textcircled{4} \\ \\ \\ \\ \end{matrix}$$

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} \text{ edge}$$

Figure 10.1*: Tree with 3 edges and 4 nodes and no loops. Then B has independent rows.

The first graph is *complete*—every pair of nodes is connected by an edge. The second graph is a ***tree***—the graph has ***no closed loops***. Those are the two extremes. The maximum number of edges is $\frac{1}{2}n(n-1) = 6$ and the minimum to stay connected is $n - 1 = 3$.

Elimination reduces every graph to a tree. Loops produce dependent rows in A and zero rows in the echelon forms U and R. Look at the large loop from edges 1, 2, 3 in the first graph, which leads to a zero row in U:

$$\begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 1 & 0 \end{bmatrix} \longrightarrow \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 1 & 0 \end{bmatrix} \longrightarrow \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

Those steps are typical. When edges 1 and 2 share node 1, elimination produces the "shortcut edge" without node 1. If the graph already has this shortcut edge making a loop, then elimination gives a row of zeros. When the dust clears we have a tree.

An idea suggests itself: ***Rows are dependent when edges form a loop***. Independent rows come from trees. This is the key to the row space. We are assuming that the graph is connected, and the arrows could go either way. On each edge, *flow with the arrow is "positive."* Flow in the opposite direction counts as negative. The flow might be a current or a signal or a force—or even oil or gas or water.

When x_1, x_2, x_3, x_4 are voltages at the nodes, Ax gives voltage differences:

$$A\boldsymbol{x} = \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} x_2 - x_1 \\ x_3 - x_1 \\ x_3 - x_2 \\ x_4 - x_1 \\ x_4 - x_2 \\ x_4 - x_3 \end{bmatrix}. \tag{1}$$

Let me say that again. The incidence matrix A is a difference matrix. The input vector $\boldsymbol{x}$ gives voltages, the output vector $A\boldsymbol{x}$ gives voltage differences (along edges 1 to 6). If the voltages are equal, the differences are zero. This tells us the nullspace of A.

1 The ***nullspace*** contains the solutions to $A\boldsymbol{x} = \mathbf{0}$. All six voltage differences are zero. This means: *All four voltages are equal.* Every $\boldsymbol{x}$ in the nullspace is a **constant vector**: $\boldsymbol{x} = (c, c, c, c)$. The nullspace of A is a line in $\mathbf{R}^n$—its dimension is $n - r = 1$.

The second incidence matrix B has the same nullspace. It contains $(\mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1})$:

1−dimensional nullspace: same for the tree

$$B\boldsymbol{x} = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

We can raise or lower all voltages by the same amount c, without changing the differences. There is an "arbitrary constant" in the voltages. Compare this with the same statement for functions. We can raise or lower a function by C, without changing its derivative.

Calculus adds "$+C$" to indefinite integrals. Graph theory adds (c, c, c, c) to the vector $\boldsymbol{x}$. Linear algebra adds any vector $\boldsymbol{x}_n$ in the nullspace to one particular solution of $A\boldsymbol{x} = \boldsymbol{b}$.

The "$+C$" disappears in calculus when a definite integral starts at a known point. Similarly the nullspace disappears when we fix $\boldsymbol{x_4} = \mathbf{0}$. The unknown x_4 is removed and so are the fourth columns of A and B (those columns multiplied x_4). Electrical engineers would say that node 4 has been "grounded."

2 The ***row space*** contains all combinations of the six rows. Its dimension is certainly not 6. The equation $r + (n - r) = n$ must be $3 + 1 = 4$. The rank is $r = 3$, as we saw from elimination. After 3 edges, we start forming loops! The new rows are not independent.

How can we tell if $\boldsymbol{v} = (v_1, v_2, v_3, v_4)$ is in the row space? The slow way is to combine rows. The quick way is by orthogonality:

$\boldsymbol{v}$ is in the row space if and only if it is perpendicular to $(\mathbf{1}, \mathbf{1}, \mathbf{1}, \mathbf{1})$ in the nullspace.

The vector $\boldsymbol{v} = (0, 1, 2, 3)$ fails this test—its components add to 6. The vector $(-6, 1, 2, 3)$ is in the row space: $-6+1+2+3 = 0$. That vector equals $6(\text{row } 1) + 5(\text{row } 3) + 3(\text{row } 6)$.

Each row of A adds to zero. This must be true for every vector in the row space.

3 The ***column space*** contains all combinations of the four columns. We expect three independent columns, since there were three independent rows. The first three columns of A are independent (so are any three). But the four columns add to the zero vector, which says again that $(1,1,1,1)$ is in the nullspace. ***How can we tell if a particular vector $\boldsymbol{b}$ is in the column space of an incidence matrix***?

First answer Try to solve $A\boldsymbol{x} = \boldsymbol{b}$. That misses all the insight. As before, orthogonality gives a better answer. We are now coming to Kirchhoff's two famous laws of circuit theory—the voltage law and current law (**KVL** and **KCL**). Those are natural expressions of "laws" of linear algebra. It is especially pleasant to see the key role of the left nullspace.

Second answer $A\boldsymbol{x}$ is the vector of voltage differences $x_i - x_j$. If we add differences around a closed loop in the graph, they cancel to leave zero. Around the big triangle formed by edges $1, 3, -2$ (*the arrow goes backward on edge* 2) the differences cancel:

Sum of differences is 0 $$(x_2 - x_1) + (x_3 - x_2) - (x_3 - x_1) = \mathbf{0}.$$

Kirchhoff's Voltage Law: The components of $A\boldsymbol{x} = \boldsymbol{b}$ add to zero around every loop.

Around the big triangle: $b_1 + b_3 - b_2 = 0.$

By testing each loop, the Voltage Law decides whether $\boldsymbol{b}$ is in the column space. $A\boldsymbol{x} = \boldsymbol{b}$ can be solved exactly when the components of $\boldsymbol{b}$ satisfy all the same dependencies as the rows of A. Then elimination leads to $0 = 0$, and $A\boldsymbol{x} = \boldsymbol{b}$ is consistent.

4 The ***left nullspace*** contains the solutions to $A^{\mathrm{T}}\boldsymbol{y} = \mathbf{0}$. Its dimension is $m - r = 6 - 3$:

Current Law
$$A^{\mathrm{T}}\boldsymbol{y} = \begin{bmatrix} -1 & -1 & 0 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 & 0 \\ 0 & 1 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{2}$$

The true number of equations is $r = 3$ and not $n = 4$. Reason: The four equations add to $0 = 0$. The fourth equation follows automatically from the first three.

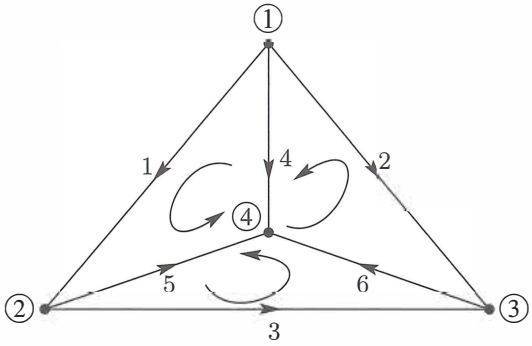
What do the equations mean? The first equation says that $-y_1 - y_2 - y_4 = 0$. ***The net flow into node 1 is zero***. The fourth equation says that $y_4 + y_5 + y_6 = 0$. ***Flow into node 4 minus flow out is zero***. The equations $A^{\mathrm{T}}\boldsymbol{y} = \mathbf{0}$ are famous and fundamental:

Kirchhoff's Current Law: $A^{\mathrm{T}}\boldsymbol{y} = 0$ ***Flow in equals flow out at each node.***

This law deserves first place among the equations of applied mathematics. It expresses "*conservation*" and "*continuity*" and "*balance*." Nothing is lost, nothing is gained. When currents or forces are balanced, the equation to solve is $A^{\mathrm{T}}\boldsymbol{y} = \mathbf{0}$. Notice the beautiful fact that the matrix in this balance equation is the transpose of the incidence matrix A.

What are the actual solutions to $A^{\mathrm{T}}\boldsymbol{y} = \mathbf{0}$? The currents must balance themselves. The easiest way is to **flow around a loop**. If a unit of current goes around the big triangle (forward on edge 1 and 3, backward on 2), the six currents are $\boldsymbol{y} = (1, -1, 1, 0, 0, 0)$. This satisfies $A^{\mathrm{T}}\boldsymbol{y} = \mathbf{0}$. *Every loop current is a solution to the Current Law.* Flow in equals flow out at every node. A smaller loop goes forward on edge 1, forward on 5, back on 4. Then $\boldsymbol{y} = (\mathbf{1}, 0, 0, \mathbf{-1}, \mathbf{1}, 0)$ is also in the left nullspace.

We expect three independent $\boldsymbol{y}$'s: $m - r = 6 - 3 = 3$. The three small loops in the graph are independent. The big triangle seems to give a fourth $\boldsymbol{y}$, but that flow is the sum of flows around the small loops. *Flows around the 3 small loops are a basis for the left nullspace.*



$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ -1 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \\ 0 \\ 1 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

3 small loops — big loop

The incidence matrix A comes from a connected graph with n nodes and m edges. The row space and column space have dimensions $r = n - 1$. The nullspaces of A and A^{T} have dimensions 1 and $m - n + 1$:

$\boldsymbol{N}(A)$ The constant vectors $(c, c, \ldots, c)$ make up the nullspace of A : dim $= 1$.

$\boldsymbol{C}(A^{\mathrm{T}})$ The edges of any tree give r independent rows of A : $r = n - 1$.

$\boldsymbol{C}(A)$ ***Voltage Law***: The components of $A\boldsymbol{x}$ add to zero around all loops: dim $= n - 1$.

$\boldsymbol{N}(A^{\mathrm{T}})$ ***Current Law***: $A^{\mathrm{T}}\boldsymbol{y} =$ **(flow in)** $-$ **(flow out)** $= \mathbf{0}$ is solved by loop currents.

There are $m - r = m - n + 1$ independent small loops in the graph.

For every graph in a plane, linear algebra yields ***Euler's formula***: Theorem 1 in topology!

(number of nodes) − (number of edges) + (number of small loops) = 1.

This is $(n) - (m) + (m - n + 1) = 1$. The graph in our example has $4 - 6 + 3 = 1$.

A single triangle has (3 nodes) $-$ (3 edges) $+$ (1 loop). On a 10-node tree with 9 edges and no loops, Euler's count is $10 - 9 + 0$. All planar graphs lead to the answer 1.

The next figure shows a network with a current source. Kirchhoff's Current Law changes from $A^{\mathrm{T}}\boldsymbol{y} = \mathbf{0}$ to $A^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{f}$, to balance the source $\boldsymbol{f}$ from outside. *Flow into each node still equals flow out.* The six edges would have conductances $c_1, \ldots, c_6$, and the current source goes into node 1. The source comes out from node 4 to keep the overall balance (**in** $=$ **out**). The problem is: ***Find the currents $y_1, \ldots, y_6$ on the six edges.***
Flows in networks now lead us from the incidence matrix A to the Laplacian matrix $A^{\mathrm{T}}A$.

Voltages and Currents and $A^{\mathrm{T}}Ax = f$

We started with voltages $\boldsymbol{x} = (x_1, \ldots, x_n)$ at the nodes. So far we have $A\boldsymbol{x}$ to find voltage differences $x_i - x_j$ along edges. And we have the Current Law $A^{\mathrm{T}}\boldsymbol{y} = \mathbf{0}$ to find edge currents $\boldsymbol{y} = (y_1, \ldots y_m)$. If all resistances in the network are 1, Ohm's Law will match $\boldsymbol{y} = A\boldsymbol{x}$. Then $A^{\mathrm{T}}\boldsymbol{y} = A^{\mathrm{T}}A\boldsymbol{x} = \mathbf{0}$. We are close but not quite there.

Without any sources, the solution to $A^{\mathrm{T}}A\boldsymbol{x} = \mathbf{0}$ will just be no flow: $\boldsymbol{x} = \mathbf{0}$ and $\boldsymbol{y} = \mathbf{0}$. I can see three ways to produce $\boldsymbol{x} \neq \mathbf{0}$ and $\boldsymbol{y} \neq \mathbf{0}$.

1 Assign fixed voltages x_i to one or more nodes.

2 Add batteries (voltage sources) in one or more edges.

3 Add current sources going into one or more nodes. See Figure 10.2

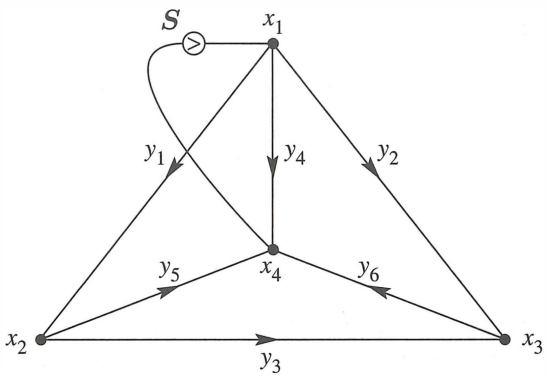


Figure 10.2: The currents y_1 to y_6 in a network with a source S from node 4 to node 1.

Example Figure 10.2 includes a current source S from node 4 to node 1. That current will trickle back through the network to node 4. Some current y_4 will go directly on edge 4. Other current will go the long way from node 1 to 2 to 4, or 1 to 3 to 4. By symmetry I expect no current ($y_3 = 0$) from node 2 to node 3. Solving the network equations will confirm this. **The matrix in those equations is $\boldsymbol{A^{\mathrm{T}}A}$, *the graph Laplacian matrix*:**

$$\begin{bmatrix} -1 & -1 & 0 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 & 0 \\ 0 & 1 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{bmatrix} = \underbrace{\begin{bmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{bmatrix}}_{\boldsymbol{A^{\mathrm{T}}A}}$$

That Laplacian matrix is not invertible! We cannot solve for all four potentials because $(1,1,1,1)$ is in the nullspace of A and $A^{\mathrm{T}}A$. *One node has to be grounded*. Setting $x_4 = 0$ removes the fourth row and column, and this leaves a 3 by 3 invertible matrix. Now we solve $A^{\mathrm{T}}A\boldsymbol{x} = \boldsymbol{f}$ for the unknown potentials x_1, x_2, x_3, with source S into node 1:

Voltages $A^{\mathrm{T}}A\boldsymbol{x} = \boldsymbol{f}$

$$\begin{bmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} S \\ 0 \\ 0 \end{bmatrix} \quad \text{gives} \quad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} S/2 \\ S/4 \\ S/4 \end{bmatrix}.$$

Currents $\boldsymbol{y} = -A\boldsymbol{x}$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = - \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} S/2 \\ S/4 \\ S/4 \\ 0 \end{bmatrix} = \begin{bmatrix} S/4 \\ S/4 \\ 0 \\ S/2 \\ S/4 \\ S/4 \end{bmatrix}.$$

Half the current goes directly on edge 4. That is $y_4 = S/2$. No current crosses from node 2 to node 3. Symmetry indicated $y_3 = 0$ and now the solution proves it.

Admission of error I remembered that current flows from high voltage to low voltage. That produces the minus sign in $\boldsymbol{y} = -A\boldsymbol{x}$. And the correct form of Ohm's Law will be $R\boldsymbol{y} = -A\boldsymbol{x}$ when the resistances on the edges are not all 1. *Conductances* are neater than resistances: $C = R^{-1} =$ diagonal matrix. **We now present Ohm's Law $\boldsymbol{y} = -C\boldsymbol{A}\boldsymbol{x}$.**

Networks and $A^{\mathrm{T}}CA$

In a real network, the current $\boldsymbol{y}$ along an edge is the product of two numbers. One number is the difference between the potentials $\boldsymbol{x}$ at the ends of the edge. This voltage difference is $A\boldsymbol{x}$ and it drives the flow. The other number c is the "***conductance***"—which measures how easily flow gets through.

In physics and engineering, c is decided by the material. For electrical currents, c is high for metal and low for plastics. For a superconductor, c is nearly infinite. If we consider elastic stretching, c might be low for metal and higher for plastics. In economics, c measures the capacity of an edge or its cost.

To summarize, the graph is known from its incidence matrix A. This tells the node-edge connections. A ***network*** goes further, and assigns a conductance c to each edge. *These numbers $c_1, \ldots, c_m$ go into the "conductance matrix" C—which is diagonal.*

For a network of resistors, the conductance is $c = 1/(\text{resistance})$. In addition to Kirchhoff's Laws for the whole system of currents, we have Ohm's Law for each current. Ohm's Law connects the current y_1 on edge 1 to the voltage difference $x_2 - x_1$:

Ohm's Law: Current along edge = conductance times voltage difference.

Ohm's Law for all m currents is $\boldsymbol{y} = -CA\boldsymbol{x}$. The vector $A\boldsymbol{x}$ gives the potential differences, and C multiplies by the conductances. Combining Ohm's Law with Kirchhoff's

Current Law $A^{\mathrm{T}}\boldsymbol{y} = \mathbf{0}$, we get $A^{\mathrm{T}}CA\boldsymbol{x} = \mathbf{0}$. This is *almost* the central equation for network flows. The only thing wrong is the zero on the right side! The network needs power from outside—a voltage source or a current source—to make something happen.

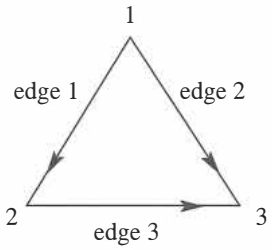
Note about signs In circuit theory we change from $A\boldsymbol{x}$ to $-A\boldsymbol{x}$. The flow is from higher potential to lower potential. There is (positive) current from node 1 to node 2 when $x_1 - x_2$ is positive—whereas $A\boldsymbol{x}$ was constructed to yield $x_2 - x_1$. The minus sign in physics and electrical engineering is a plus sign in mechanical engineering and economics. $A\boldsymbol{x}$ versus $-A\boldsymbol{x}$ is a general headache but unavoidable.

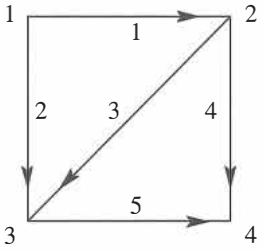
Note about applied mathematics Every new application has its own form of Ohm's Law. For springs it is Hooke's Law. The stress $\boldsymbol{y}$ is (elasticity C) times (stretching $A\boldsymbol{x}$). For heat conduction, $A\boldsymbol{x}$ is a temperature gradient. For oil flows it is a pressure gradient. For least squares regression in statistics (Chapter 12) C^{-1} is the covariance matrix.

My textbooks *Introduction to Applied Mathematics* and *Computational Science and Engineering* (Wellesley-Cambridge Press) are practically built on $A^{\mathrm{T}}CA$. This is the key to equilibrium in matrix equations and also in differential equations. Applied mathematics is more organized than it looks! *In new problems I have learned to watch for* $A^{\mathrm{T}}CA$.

Problem Set 10.1

Problems 1–7 and 8–14 are about the incidence matrices for these graphs.





1 Write down the 3 by 3 incidence matrix A for the triangle graph. The first row has -1 in column 1 and $+1$ in column 2. What vectors (x_1, x_2, x_3) are in its nullspace? How do you know that $(1, 0, 0)$ is not in its row space?

2 Write down A^{T} for the triangle graph. Find a vector $\boldsymbol{y}$ in its nullspace. The components of $\boldsymbol{y}$ are currents on the edges—how much current is going around the triangle?

3 Eliminate x_1 and x_2 from the third equation to find the echelon matrix U. What tree corresponds to the two nonzero rows of U?

$$\begin{aligned} -x_1 + x_2 &= b_1 \\ -x_1 + x_3 &= b_2 \\ -x_2 + x_3 &= b_3. \end{aligned}$$

4 Choose a vector (b_1, b_2, b_3) for which $A\boldsymbol{x} = \boldsymbol{b}$ can be solved, and another vector $\boldsymbol{b}$ that allows no solution. How are those $\boldsymbol{b}$'s related to $\boldsymbol{y} = (1, -1, 1)$?

5 Choose a vector (f_1, f_2, f_3) for which $A^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{f}$ can be solved, and a vector $\boldsymbol{f}$ that allows no solution. How are those $\boldsymbol{f}$'s related to $\boldsymbol{x} = (1, 1, 1)$? The equation $A^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{f}$ is Kirchhoff's _____ law.

6 Multiply matrices to find $A^{\mathrm{T}}A$. Choose a vector $\boldsymbol{f}$ for which $A^{\mathrm{T}}A\boldsymbol{x} = \boldsymbol{f}$ can be solved, and solve for $\boldsymbol{x}$. Put those potentials $\boldsymbol{x}$ and the currents $\boldsymbol{y} = -A\boldsymbol{x}$ and current sources $\boldsymbol{f}$ onto the triangle graph. Conductances are 1 because $C = I$.

7 With conductances $c_1 = 1$ and $c_2 = c_3 = 2$, multiply matrices to find $A^{\mathrm{T}}CA$. For $\boldsymbol{f} = (1, 0, -1)$ find a solution to $A^{\mathrm{T}}CA\boldsymbol{x} = \boldsymbol{f}$. Write the potentials $\boldsymbol{x}$ and currents $\boldsymbol{y} = -CA\boldsymbol{x}$ on the triangle graph, when the current source $\boldsymbol{f}$ goes into node 1 and out from node 3.

8 Write down the 5 by 4 incidence matrix A for the square graph with two loops. Find one solution to $A\boldsymbol{x} = \boldsymbol{0}$ and two solutions to $A^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{0}$.

9 Find two requirements on the b's for the five differences $x_2 - x_1, x_3 - x_1, x_3 - x_2$, $x_4 - x_2, x_4 - x_3$ to equal b_1, b_2, b_3, b_4, b_5. You have found Kirchhoff's _____ law around the two _____ in the graph.

10 Reduce A to its echelon form U. The three nonzero rows give the incidence matrix for what graph? You found one tree in the square graph—find the other seven trees.

11 Multiply matrices to find $A^{\mathrm{T}}A$ and guess how its entries come from the graph:

(a) The diagonal of $A^{\mathrm{T}}A$ tells how many _____ into each node.

(b) The off-diagonals -1 or 0 tell which pairs of nodes are _____.

12 Why is each statement true about $A^{\mathrm{T}}A$? *Answer for $A^{\mathrm{T}}A$ not A.*

(a) Its nullspace contains $(1, 1, 1, 1)$. Its rank is $n - 1$.

(b) It is positive semidefinite but not positive definite.

(c) Its four eigenvalues are real and their signs are _____.

13 With conductances $c_1 = c_2 = 2$ and $c_3 = c_4 = c_5 = 3$, multiply the matrices $A^{\mathrm{T}}CA$. Find a solution to $A^{\mathrm{T}}CA\boldsymbol{x} = \boldsymbol{f} = (1, 0, 0, -1)$. Write these potentials $\boldsymbol{x}$ and currents $\boldsymbol{y} = -CA\boldsymbol{x}$ on the nodes and edges of the square graph.

14 The matrix $A^{\mathrm{T}}CA$ is not invertible. What vectors $\boldsymbol{x}$ are in its nullspace? Why does $A^{\mathrm{T}}CA\boldsymbol{x} = \boldsymbol{f}$ have a solution if and only if $f_1 + f_2 + f_3 + f_4 = 0$?

15 A connected graph with 7 nodes and 7 edges has how many loops?

16 For the graph with 4 nodes, 6 edges, and 3 loops, add a new node. If you connect it to one old node, Euler's formula becomes $(\quad) - (\quad) + (\quad) = 1$. If you connect it to two old nodes, Euler's formula becomes $(\quad) - (\quad) + (\quad) = 1$.

17 Suppose A is a 12 by 9 incidence matrix from a connected (but unknown) graph.

(a) How many columns of A are independent?

(b) What condition on $\boldsymbol{f}$ makes it possible to solve $A^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{f}$?

(c) The diagonal entries of $A^{\mathrm{T}}A$ give the number of edges into each node. What is the sum of those diagonal entries?

18 Why does a complete graph with $n = 6$ nodes have $m = 15$ edges? A tree connecting 6 nodes has _____ edges.

Note The ***stoichiometric matrix*** in chemistry is an important "generalized" incidence matrix. Its entries show how much of each chemical species (each column) goes into each reaction (each row).

10.2 Matrices in Engineering

This section will show how engineering problems produce symmetric matrices K (often K is positive definite). The "linear algebra reason" for symmetry and positive definiteness is their form $K = A^{\mathrm{T}}A$ and $K = A^{\mathrm{T}}CA$. The "physical reason" is that the expression $\frac{1}{2}\boldsymbol{u}^{\mathrm{T}}K\boldsymbol{u}$ represents *energy*—and energy is never negative. The matrix C, often diagonal, contains positive physical constants like conductance or stiffness or diffusivity.

Our best examples come from mechanical and civil and aeronautical engineering. K is the ***stiffness matrix***, and $K^{-1}\boldsymbol{f}$ is the structure's response to forces $\boldsymbol{f}$ from outside. Section 10.1 turned to electrical engineering—the matrices came from networks and circuits. The exercises involve chemical engineering and I could go on! Economics and management and engineering design come later in this chapter (the key is optimization).

Engineering leads to linear algebra in two ways, directly and indirectly:

Direct way The physical problem has only a finite number of pieces. The laws connecting their position or velocity are *linear* (movement is not too big or too fast). The laws are expressed by *matrix equations*.

Indirect way The physical system is "continuous". Instead of individual masses, the mass density and the forces and the velocities are functions of x or x, y or x, y, z. The laws are expressed by *differential equations*. ***To find accurate solutions we approximate by finite difference equations or finite element equations***.

Both ways produce matrix equations and linear algebra. I really believe that you cannot do modern engineering without matrices.

Here we present equilibrium equations $K\boldsymbol{u} = \boldsymbol{f}$. With motion, $Md^2\boldsymbol{u}/dt^2 + K\boldsymbol{u} = \boldsymbol{f}$ becomes dynamic. Then we would use eigenvalues from $K\boldsymbol{x} = \lambda M\boldsymbol{x}$, or finite differences.

Differential Equation to Matrix Equation

Differential equations are continuous. Our basic example will be $-d^2u/dx^2 = f(x)$. Matrix equations are discrete. Our basic example will be $K_0\boldsymbol{u} = \boldsymbol{f}$. By taking the step from second derivatives to second differences, you will see the big picture in a very short space. *Start with fixed boundary conditions at both ends $x = 0$ and $x = 1$:*

Fixed-fixed boundary value problem

$$-\frac{d^2u}{dx^2} = 1 \text{ with } u(0) = 0 \text{ and } u(1) = 0. \tag{1}$$

That differential equation is linear. A particular solution is $u_p = -\frac{1}{2}x^2$ (then $d^2u/dx^2 = -1$). We can add any function "in the nullspace". Instead of solving $A\boldsymbol{x} = \boldsymbol{0}$ for a vector $\boldsymbol{x}$, we solve $-d^2u/dx^2 = 0$ for a function $u_n(x)$. (Main point: The right side is zero.)

The nullspace solutions are $u_n(x) = C + Dx$ (a 2-dimensional nullspace for a second order differential equation). The complete solution is $u_p + u_n$:

Complete solution to $-\dfrac{d^2u}{dx^2} = 1$ $\quad u(x) = -\dfrac{1}{2}x^2 + C + Dx.$ (2)

Now find C and D from the two boundary conditions: Set $x = 0$ and then $x = 1$. At $x = 0, u(0) = 0$ forces $\boldsymbol{C = 0}$. At $x = 1, u(1) = 0$ forces $-\frac{1}{2} + D = 0$. Then $\boldsymbol{D = \frac{1}{2}}$:

$$u(x) = -\frac{1}{2}x^2 + \frac{1}{2}x = \frac{\boldsymbol{1}}{\boldsymbol{2}}(\boldsymbol{x - x^2}) \text{ solves the fixed-fixed boundary value problem.} \quad (3)$$

Differences Replace Derivatives

To get matrices instead of derivatives, we have three basic choices—*forward or backward or centered differences*. Start with first derivatives and first differences:

$$\frac{\boldsymbol{du}}{\boldsymbol{dx}} \approx \frac{u(x+\Delta x) - u(x)}{\Delta x} \text{ or } \frac{u(x) - u(x-\Delta x)}{\Delta x} \text{ or } \frac{u(x+\Delta x) - u(x-\Delta x)}{2\Delta x}.$$

Between $x = 0$ and $x = 1$, we divide the interval into $n + 1$ equal pieces. The pieces have width $\Delta x = 1/(n + 1)$. The values of u at the n breakpoints $\Delta x, 2\Delta x, \ldots$ will be the unknowns u_1 to u_n in our matrix equation $K\boldsymbol{u} = \boldsymbol{f}$:

Solution to compute: $\boldsymbol{u} = (u_1, u_2, \ldots, u_n) \approx (u(\Delta x), u(2\Delta x), \ldots, u(n\Delta x))$.
Zero values $u_0 = u_{n+1} = 0$ come from the boundary conditions $u(0) = u(1) = 0$.

Replace the derivatives in $-\dfrac{d}{dx}\left(\dfrac{du}{dx}\right) = 1$ *by forward and backward differences*:

$$\frac{1}{(\Delta x)^2}\begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & -1 \\ 0 & 0 & -1 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \quad (4)$$

This is our matrix equation when $n = 3$ and $\Delta x = \frac{1}{4}$. The two first differences are transposes of each other! The equation is $A^{\mathrm{T}}A\boldsymbol{u} = (\Delta x)^2\boldsymbol{f}$. When we multiply $A^{\mathrm{T}}A$, we get the positive definite second difference matrix K_0:

$$\boldsymbol{K_0 u = (\Delta x)^2 f} \quad \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \frac{1}{16}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \text{ gives } \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \frac{1}{32}\begin{bmatrix} 3 \\ 4 \\ 3 \end{bmatrix}. \quad (5)$$

The wonderful fact in this example is that those numbers u_1, u_2, u_3 are exactly correct! They agree with the true solution $u = \frac{1}{2}(x - x^2)$ at the three meshpoints $x = \frac{1}{4}, \frac{2}{4}, \frac{3}{4}$. Figure 10.3 shows the true solution (continuous curve) and the approximations u_1, u_2, u_3 (lying exactly on the curve). This curve is a parabola.

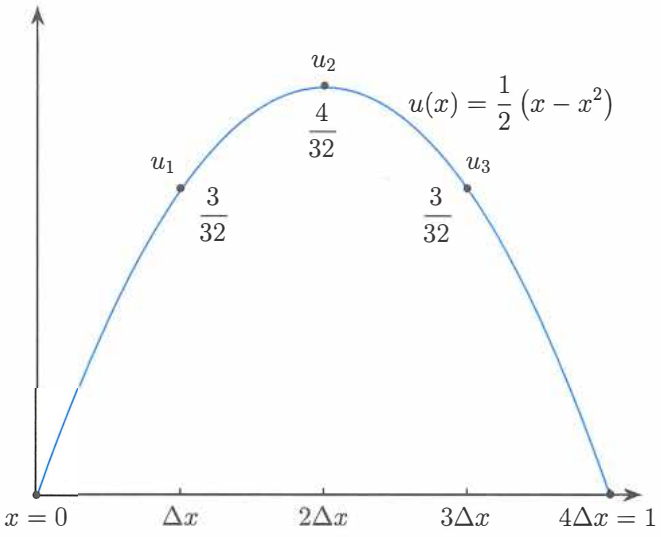


Figure 10.3: Solutions to $-\dfrac{d^2u}{dx^2} = 1$ and $K_0\boldsymbol{u} = \boldsymbol{(\Delta x)^2 f}$ with fixed-fixed boundaries.

How to explain this perfect answer, lying right on the graph of $u(x)$? In the matrix equation, $K_0 = A^{\mathrm{T}}A$ is a "second difference matrix." It gives a centered approximation to $-d^2u/dx^2$. I included the minus sign because the first derivative is *antisymmetric*. The second derivative by itself is *negative*:

The "transpose" of $\dfrac{d}{dx}$ is $-\dfrac{d}{dx}$. Then $\left(-\dfrac{d}{dx}\right)\left(\dfrac{d}{dx}\right)$ is positive definite.

You can see that in the matrices A and A^{T}. The transpose of $A =$ *forward difference* is $A^{\mathrm{T}} = -$ *backward difference*. I don't want to choose a centered $u(x+\Delta x) - u(x-\Delta x)$. Centered is the best for a first difference, but then the second difference $A^{\mathrm{T}}A$ would stretch from $u(x + \mathbf{2}\Delta x)$ to $u(x - \mathbf{2}\Delta x)$: not good.

Now we can explain the perfect answers, exactly on the true curve $u(x) = \frac{1}{2}\left(x - x^2\right)$. Second differences $-1, 2, -1$ are exactly correct for straight lines $y = x$ and parabolas !

$$\boldsymbol{y = x} \qquad -\frac{d^2y}{dx^2} = \mathbf{0} \qquad -(x + \Delta x) \ +2x \ -(x - \Delta x) \ = \ \mathbf{0}(\Delta x)^2$$

$$\boldsymbol{y = x^2} \qquad -\frac{d^2y}{dx^2} = \mathbf{-2} \qquad -(x + \Delta x)^2 + 2x^2 - (x - \Delta x)^2 = \ \mathbf{-2}(\Delta x)^2$$

The miracle continues to $y = x^3$. The correct $-d^2y/dx^2 = -6x$ is produced by second differences. But for $y = x^4$ we return to earth. Second differences don't exactly match $-y'' = -12x^2$. The approximations u_1, u_2, u_3 won't fall on the graph of $u(x)$.

Fixed End and Free End and Variable Coefficient $c(x)$

To see two new possibilities, I will change the equation and also one boundary condition:

$$-\frac{d}{dx}\left((1+x)\frac{du}{dx}\right) = f(x) \text{ with } u(0) = 0 \text{ and } \frac{\boldsymbol{du}}{\boldsymbol{dx}}(\mathbf{1}) = \mathbf{0}. \tag{6}$$

The end $x = 1$ is now **free**. There is no support at that end. "A hanging bar is fixed only at the top." There is no force at the free end $x = 1$. That translates to $du/dx = 0$ instead of the fixed condition $u = 0$ at $x = 1$.

The other change is in the coefficient $c(x) = 1 + x$. The stiffness of the bar is varying as you go from $x = 0$ to $x = 1$. Maybe its width is changing, or the material changes. This coefficient $1 + x$ will bring a new matrix C into the difference equation.

Since u_4 is no longer fixed at 0, it becomes a new unknown. The backward difference A is 4 by 4. And the multiplication by $c(x) = 1 + x$ becomes a diagonal matrix C—which multiplies by $1 + \Delta x, \ldots, 1 + 4\Delta x$ at the meshpoints. Here are A^{T}, C, and A:

$$A^{\mathrm{T}}CA = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1.25 & & & \\ & 1.5 & & \\ & & 1.75 & \\ & & & 2.0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}. \tag{7}$$

This matrix $K = A^{\mathrm{T}}CA$ will be symmetric and positive definite! Symmetric because $(A^{\mathrm{T}}CA)^{\mathrm{T}} = A^{\mathrm{T}}C^{\mathrm{T}}A^{\mathrm{TT}} = A^{\mathrm{T}}CA$. Positive definite because it passes the energy test: A has independent columns, so $A\boldsymbol{x} \neq \mathbf{0}$ when $\boldsymbol{x} \neq \mathbf{0}$.

$$\text{Energy } = \boldsymbol{x}^{\mathrm{T}}A^{\mathrm{T}}CA\boldsymbol{x} = (A\boldsymbol{x})^{\mathrm{T}}C(A\boldsymbol{x}) > 0 \text{ for every } \boldsymbol{x} \neq \mathbf{0}, \text{ because } A\boldsymbol{x} \neq \mathbf{0}.$$

When you multiply the matrices $A^{\mathrm{T}}A$ and $A^{\mathrm{T}}CA$ for this fixed-free combination, watch how 1 replaces 2 in the last corner of $A^{\mathrm{T}}A$. That fourth equation has $u_4 - u_3$, a first (not second) difference coming from the free boundary condition $du/dx = 0$.

Notice in $A^{\mathrm{T}}CA$ how c_1, c_2, c_3, c_4 come from $c(x) = 1 + x$ in equation (7). Previously the c's were simply $1, 1, 1, 1$. Here are the **fixed-free** matrices:

$$A^{\mathrm{T}}A = \begin{bmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & -1 & 2 & -1 \\ & & -1 & \mathbf{1} \end{bmatrix} \quad A^{\mathrm{T}}CA = \begin{bmatrix} c_1 + c_2 & -c_2 & & \\ -c_2 & c_2 + c_3 & -c_3 & \\ & -c_3 & c_3 + c_4 & -c_4 \\ & & -c_4 & \boldsymbol{c_4} \end{bmatrix}. \tag{8}$$

Free-free Boundary Conditions

Suppose both ends of the bar are free. Now $du/dx = 0$ at both $x = 0$ and $x = 1$. Nothing is holding the bar in place! Physically it is unstable—it can move with no force. Mathematically all constant functions like $u = 1$ satisfy these free conditions. **Algebraically our matrices $A^{\mathrm{T}}A$ and $A^{\mathrm{T}}CA$ will not be invertible**:

Free-free examples
Unknown u_0, u_1, u_2
$\Delta x = 0.5$

$$A^{\mathrm{T}}A = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & \mathbf{1} \end{bmatrix} \qquad A^{\mathrm{T}}CA = \begin{bmatrix} \boldsymbol{c_0} & -c_0 & \\ -c_0 & c_0 + c_1 & -c_1 \\ & -c_1 & \boldsymbol{c_1} \end{bmatrix}.$$

The vector $(1, 1, 1)$ is in both nullspaces. This matches $u(x) = 1$ in the continuous problem. Free-free $A^{\mathrm{T}}A\boldsymbol{u} = \boldsymbol{f}$ and $A^{\mathrm{T}}CA\boldsymbol{u} = \boldsymbol{f}$ are generally unsolvable.

Before explaining more physical examples, may I write down six of the matrices? The tridiagonal K_0 appears many times in this textbook. Now we are seeing its applications. These matrices are all symmetric, and the first four are positive definite:

$$K_0 = A_0^{\mathrm{T}}A_0 = \begin{bmatrix} \mathbf{2} & -1 & \\ -1 & 2 & -1 \\ & -1 & \mathbf{2} \end{bmatrix} \qquad A_0^{\mathrm{T}}C_0A_0 = \begin{bmatrix} c_1 + c_2 & -c_2 & \\ -c_2 & c_2 + c_3 & -c_3 \\ & -c_3 & c_3 + c_4 \end{bmatrix}$$

Fixed-fixed **Spring constants included**

$$K_1 = A_1^{\mathrm{T}}A_1 = \begin{bmatrix} \mathbf{2} & -1 & \\ -1 & 2 & -1 \\ & -1 & \mathbf{1} \end{bmatrix} \qquad A_1^{\mathrm{T}}C_1A_1 = \begin{bmatrix} c_1 + c_2 & -c_2 & \\ -c_2 & c_2 + c_3 & -c_3 \\ & -c_3 & c_3 \end{bmatrix}$$

Fixed-free **Spring constants included**

$$K_{\textbf{singular}} = \begin{bmatrix} \mathbf{1} & -1 & \\ -1 & 2 & -1 \\ & -1 & \mathbf{1} \end{bmatrix} \qquad K_{\textbf{circular}} = \begin{bmatrix} 2 & -1 & \mathbf{-1} \\ -1 & 2 & -1 \\ \mathbf{-1} & -1 & 2 \end{bmatrix}$$

Free-free **Periodic $\boldsymbol{u(0) = u(1)}$**

The matrices $K_0, K_1, K_{\textbf{singular}}$, and $K_{\textbf{circular}}$ have $C = I$ for simplicity. This means that all the "spring constants" are $c_i = 1$. We included $A_0^{\mathrm{T}}C_0A_0$ and $A_1^{\mathrm{T}}C_1A_1$ to show how the spring constants enter the matrix (without changing its positive definiteness). Our next goal is to see these same stiffness matrices in other engineering problems.

A Line of Springs and Masses

Figure 10.4 shows three masses m_1, m_2, m_3 connected by a line of springs. The fixed-fixed case has four springs, with top and bottom fixed. That leads to K_0 and $A_0^{\mathrm{T}} C_0 A_0$. The fixed-free case has only three springs; the lowest mass hangs freely. That will lead to K_1 and $A_1^{\mathrm{T}} C_1 A_1$. A **free-free** problem produces $K_{\textbf{singular}}$.

We want equations for the mass movements $\boldsymbol{u}$ and the spring tensions $\boldsymbol{y}$:

$$\begin{aligned} \boldsymbol{u} &= (u_1, u_2, u_3) = \textbf{movements of the masses} \text{ (down is positive)} \\ \boldsymbol{y} &= (y_1, y_2, y_3, y_4) \text{ or } (y_1, y_2, y_3) = \textbf{tensions in the springs} \end{aligned}$$

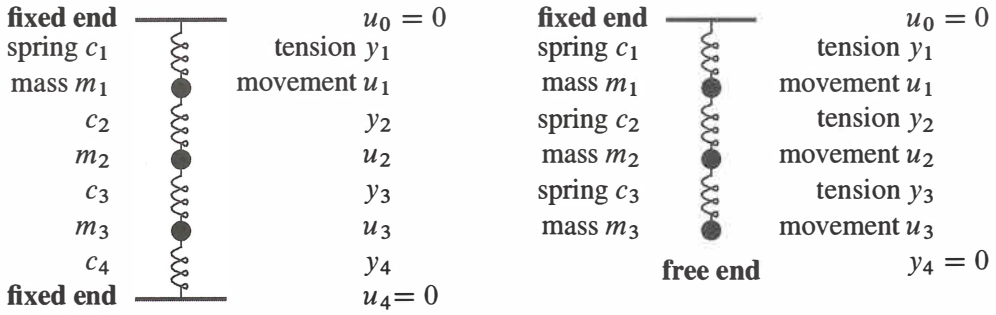


Figure 10.4: Lines of springs and masses: **fixed-fixed** and **fixed-free** ends.

When a mass moves downward, its displacement is positive ($u_j > 0$). For the springs, tension is positive and compression is negative ($y_i < 0$). In tension, the spring is stretched so it pulls the masses inward. Each spring is controlled by its own Hooke's Law $y = ce$: (*stretching force* $\boldsymbol{y}$) = (*spring constant* $\boldsymbol{c}$) times (*stretching distance* $\boldsymbol{e}$).

Our job is to link these one-spring equations $y = ce$ into a vector equation $K\boldsymbol{u} = \boldsymbol{f}$ for the whole system. The force vector $\boldsymbol{f}$ comes from gravity. The gravitational constant g will multiply each mass to produce downward forces $\boldsymbol{f} = (m_1 g, m_2 g, m_3 g)$.

The real problem is to find the stiffness matrix (**fixed-fixed** and **fixed-free**). The best way to create K is in three steps, not one. Instead of connecting the movements $\boldsymbol{u}_j$ directly to the forces f_i, it is much better to connect each vector to the next in this list:

$$\begin{aligned} \boldsymbol{u} &= \textit{Movements} \text{ of } n \text{ masses} &&= (u_1, \ldots, u_n) \\ \boldsymbol{e} &= \textit{Elongations} \text{ of } m \text{ springs} &&= (e_1, \ldots, e_m) \\ \boldsymbol{y} &= \textit{Internal forces} \text{ in } m \text{ springs} &&= (y_1, \ldots, y_m) \\ \boldsymbol{f} &= \textit{External forces} \text{ on } n \text{ masses} &&= (f_1, \ldots, f_n) \end{aligned}$$

A great framework for applied mathematics connects $\boldsymbol{u}$ to $\boldsymbol{e}$ to $\boldsymbol{y}$ to $\boldsymbol{f}$. Then $\boldsymbol{A^{\mathrm{T}} C A u} = \boldsymbol{f}$:

$$\begin{array}{ccc} \boxed{\boldsymbol{u}} & & \boxed{\boldsymbol{f}} \\ \boldsymbol{A}\downarrow & & \uparrow \boldsymbol{A^{\mathrm{T}}} \\ \boxed{\boldsymbol{e}} & \xrightarrow{\boldsymbol{C}} & \boxed{\boldsymbol{y}} \end{array} \qquad \begin{aligned} \boldsymbol{e} &= A\boldsymbol{u} \\ \boldsymbol{y} &= C\boldsymbol{e} \\ \boldsymbol{f} &= A^{\mathrm{T}}\boldsymbol{y} \end{aligned} \qquad \begin{aligned} &A \text{ is } m \text{ by } n \\ &C \text{ is } m \text{ by } m \\ &A^{\mathrm{T}} \text{ is } n \text{ by } m \end{aligned}$$

We will write down the matrices A and C and A^{T} for the two examples, first with fixed ends and then with the lower end free. Forgive the simplicity of these matrices, it is their form that is so important. Especially the appearance of A together with A^{T}.

The elongation $\boldsymbol{e}$ is the stretching distance—how far the springs are extended. Originally there is no stretching—the system is lying on a table. When it becomes vertical and upright, gravity acts. The masses move down by distances u_1, u_2, u_3. Each spring is stretched or compressed by $e_i = u_i - u_{i-1}$, *the difference in displacements of its ends*:

	First spring:	$\boldsymbol{e_1 = u_1}$	(the top is fixed so $u_0 = 0$)
Stretching of	Second spring:	$\boldsymbol{e_2 = u_2 - u_1}$	
each spring	Third spring:	$\boldsymbol{e_3 = u_3 - u_2}$	
	Fourth spring:	$\boldsymbol{e_4 = \quad - u_3}$	(the bottom is fixed so $u_4 = 0$)

If both ends move the same distance, that spring is not stretched: $u_j = u_{j-1}$ and $e_j = 0$. The matrix in those four equations is a 4 by 3 *difference matrix* A, and $\boldsymbol{e = Au}$:

$$\textbf{Stretching distances (elongations)} \qquad \boldsymbol{e = Au} \quad \text{is} \quad \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}. \tag{9}$$

The next equation $\boldsymbol{y = Ce}$ connects spring elongation $\boldsymbol{e}$ with spring tension $\boldsymbol{y}$. *This is Hooke's Law $y_i = c_i e_i$ for each separate spring*. It is the "constitutive law" that depends on the material in the spring. A soft spring has small c, so a moderate force y can produce a large stretching e. Hooke's linear law is nearly exact for real springs, before they are overstretched and the material becomes plastic.

Since each spring has its own law, the matrix in $\boldsymbol{y = Ce}$ is a diagonal matrix C:

$$\textbf{Hooke's Law } \boldsymbol{y = Ce} \qquad \begin{aligned} y_1 &= \boldsymbol{c_1 e_1} \\ y_2 &= \boldsymbol{c_2 e_2} \\ y_3 &= \boldsymbol{c_3 e_3} \\ y_4 &= \boldsymbol{c_4 e_4} \end{aligned} \quad \text{is} \quad \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} c_1 & & & \\ & c_2 & & \\ & & c_3 & \\ & & & c_4 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix} \tag{10}$$

Combining $\boldsymbol{e = Au}$ with $\boldsymbol{y = Ce}$, the spring forces (tension forces) are $\boldsymbol{y = CAu}$.

Finally comes the balance equation, the most fundamental law of applied mathematics. The internal forces from the springs balance the external forces on the masses. Each mass is pulled or pushed by the spring force y_j above it. From below it feels the spring force y_{j+1} plus f_j from gravity. Thus $y_j = y_{j+1} + f_j$ or $f_j = y_j - y_{j+1}$:

$$\textbf{Force balance } \boldsymbol{f = A^{\mathrm{T}} y} \qquad \begin{aligned} \boldsymbol{f_1} &= \boldsymbol{y_1 - y_2} \\ \boldsymbol{f_2} &= \boldsymbol{y_2 - y_3} \\ \boldsymbol{f_3} &= \boldsymbol{y_3 - y_4} \end{aligned} \quad \text{is} \quad \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \tag{11}$$

That matrix is A^{T}! *The equation for balance of forces is $\boldsymbol{f = A^{\mathrm{T}} y}$*. Nature transposes the rows and columns of the $\boldsymbol{e - u}$ matrix to produce the $\boldsymbol{f - y}$ matrix. This is the beauty of the framework, that A^{T} appears along with A. The three equations combine into $\boldsymbol{Ku = f}$.

$$\left\{\begin{array}{ccc} e & = & Au \\ y & = & Ce \\ f & = & A^{\mathrm{T}}y \end{array}\right\} \quad \begin{array}{l} \text{combine into} \quad A^{\mathrm{T}}CAu = f \quad \text{or} \quad Ku = f \\ K = A^{\mathrm{T}}CA \text{ is the } \textbf{stiffness matrix} \text{ (mechanics)} \\ K = A^{\mathrm{T}}CA \text{ is the } \textbf{conductance matrix} \text{ (networks)} \end{array}$$

Finite element programs spend major effort on assembling $K = A^{\mathrm{T}}CA$ from thousands of smaller pieces. We find K for four springs (**fixed-fixed**) by multiplying A^{T} times CA:

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} c_1 & 0 & 0 \\ -c_2 & c_2 & 0 \\ 0 & -c_3 & c_3 \\ 0 & 0 & -c_4 \end{bmatrix} = \begin{bmatrix} c_1 + c_2 & -c_2 & 0 \\ -c_2 & c_2 + c_3 & -c_3 \\ 0 & -c_3 & c_3 + c_4 \end{bmatrix}$$

If all springs are identical, with $c_1 = c_2 = c_3 = c_4 = 1$, then $C = I$. The stiffness matrix reduces to $A^{\mathrm{T}}A$. It becomes the special $-1, 2, -1$ matrix K_0.

Note the difference between $A^{\mathrm{T}}A$ from engineering and LU from linear algebra. The matrix A from four springs is 4 by 3. The triangular matrices from elimination are square. The stiffness matrix K is assembled from $A^{\mathrm{T}}A$, and then broken up into LU. One step is applied mathematics, the other is computational mathematics. Each K is built from rectangular matrices and factored into square matrices.

May I list some properties of $K = A^{\mathrm{T}}CA$? You know almost all of them:

1. K is **tridiagonal**, because mass 3 is not connected to mass 1.
2. K is **symmetric**, because C is symmetric and A^{T} comes with A.
3. K is **positive definite**, because $c_i > 0$ and A has **independent columns**.
4. K^{-1} is a **full matrix** (not sparse) with **all positive entries**.

Property 4 leads to an important fact about $u = K^{-1}f$: *If all forces act downwards* ($f_j > 0$) *then all movements are downwards* ($u_j > 0$). Notice that "positive" is different from "positive definite". K^{-1} is positive (K is not). Both are positive definite.

Example 1 Suppose all $c_i = c$ and $m_j = m$. Find the movements u and tensions y.

All springs are the same and all masses are the same. But all movements and elongations and tensions will *not* be the same. K^{-1} includes $\frac{1}{c}$ because $A^{\mathrm{T}}CA$ includes c:

$$\textbf{Movements} \qquad u = K^{-1}f = \frac{1}{4c}\begin{bmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{bmatrix}\begin{bmatrix} mg \\ mg \\ mg \end{bmatrix} = \frac{mg}{c}\begin{bmatrix} \mathbf{3/2} \\ \mathbf{2} \\ \mathbf{3/2} \end{bmatrix}$$

The displacement u_2, for the mass in the middle, is greater than u_1 and u_3. The units are correct: the force mg divided by force per unit length c gives a length u. Then

$$\textbf{Elongations} \qquad e = Au = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}\frac{mg}{c}\begin{bmatrix} \frac{3}{2} \\ 2 \\ \frac{3}{2} \end{bmatrix} = \frac{mg}{c}\begin{bmatrix} \mathbf{3/2} \\ \mathbf{1/2} \\ \mathbf{-1/2} \\ \mathbf{-3/2} \end{bmatrix}.$$

Warning: *Normally you cannot write* $K^{-1} = A^{-1}C^{-1}(A^{\mathrm{T}})^{-1}$.

The three matrices are mixed together by $A^{\mathrm{T}}CA$, and they cannot easily be untangled. In general, $A^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{f}$ has many solutions. And four equations $A\boldsymbol{u} = \boldsymbol{e}$ would usually have no solution with three unknowns. But $A^{\mathrm{T}}CA$ gives the correct solution to all three equations in the framework. Only when $m = n$ and the matrices are square can we go from $\boldsymbol{y} = (A^{\mathrm{T}})^{-1}\boldsymbol{f}$ to $\boldsymbol{e} = C^{-1}\boldsymbol{y}$ to $\boldsymbol{u} = A^{-1}\boldsymbol{e}$. We will see that now.

Fixed End and Free End

Remove the fourth spring. All matrices become 3 by 3. The pattern does not change! The matrix A loses its fourth row and (of course) A^{T} loses its fourth column. The new stiffness matrix K_1 becomes a product of square matrices:

$$A_1^{\mathrm{T}}C_1A_1 = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} c_1 & & \\ & c_2 & \\ & & c_3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}.$$

The missing column of A^{T} and row of A multiplied the missing c_4. So the quickest way to find the new $A^{\mathrm{T}}CA$ is to set $c_4 = 0$ in the old one:

FIXED FREE
$$A_1^{\mathrm{T}}C_1A_1 = \begin{bmatrix} c_1 + c_2 & -c_2 & 0 \\ -c_2 & c_2 + c_3 & -c_3 \\ 0 & -c_3 & c_3 \end{bmatrix}. \tag{12}$$

Example 2 If $c_1 = c_2 = c_3 = 1$ and $C = I$, this is the $-1, 2, -1$ tridiagonal matrix K_1. The last entry of K_1 is 1 instead of 2 because the spring at the bottom is free. Suppose all $m_j = m$:

Fixed-free
$$\boldsymbol{u} = K_1^{-1}\boldsymbol{f} = \frac{1}{c}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} mg \\ mg \\ mg \end{bmatrix} = \frac{mg}{c}\begin{bmatrix} 3 \\ 5 \\ 6 \end{bmatrix}.$$

Those movements are greater than the free-free case. The number 3 appears in u_1 because all three masses are pulling the first spring down. The next mass moves by that 3 plus an additional 2 from the masses below it. The third mass drops even more $(3 + 2 + 1 = 6)$. The elongations $\boldsymbol{e} = A\boldsymbol{u}$ in the springs display those numbers $3, 2, 1$:

$$\boldsymbol{e} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \frac{mg}{c}\begin{bmatrix} 3 \\ 5 \\ 6 \end{bmatrix} = \frac{mg}{c}\begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}.$$

Two Free Ends: K is Singular

Freedom at *both ends* means trouble. The whole line can move. A is 2 by 3:

$$\textbf{FREE-FREE} \quad \boldsymbol{e = Au} \qquad \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} u_2 - u_1 \\ u_3 - u_2 \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}. \tag{13}$$

Now there is a nonzero solution to $A\boldsymbol{u} = \mathbf{0}$. ***The masses can move with no stretching of the springs***. The whole line can shift by $\boldsymbol{u} = (1, 1, 1)$ and this leaves $\boldsymbol{e} = (0, 0)$:

$$A\boldsymbol{u} = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \textbf{no stretching}. \tag{14}$$

$A\boldsymbol{u} = \mathbf{0}$ certainly leads to $A^{\mathrm{T}}CA\boldsymbol{u} = \mathbf{0}$. Then $A^{\mathrm{T}}CA$ is only *positive semidefinite*, without c_1 and c_4. The pivots will be c_2 and c_3 and *no third pivot*. The rank is only 2:

$$\begin{bmatrix} -1 & 0 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} c_2 & \\ & c_3 \end{bmatrix} \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} = \begin{bmatrix} c_2 & -c_2 & 0 \\ -c_2 & c_2 + c_3 & -c_3 \\ 0 & -c_3 & c_3 \end{bmatrix} \tag{15}$$

Two eigenvalues will be positive but $\boldsymbol{x} = (1, 1, 1)$ is an eigenvector for $\lambda = 0$. We can solve $A^{\mathrm{T}}CA\boldsymbol{u} = \boldsymbol{f}$ only for special vectors $\boldsymbol{f}$. The forces have to add to $f_1 + f_2 + f_3 = 0$, or the whole line of springs (with both ends free) will take off like a rocket.

Circle of Springs

A third spring will complete the circle from mass 3 back to mass 1. This doesn't make K invertible—the stiffness matrix $K_{circular}$ matrix is still singular:

$$A^{\mathrm{T}}_{\textbf{circular}} A_{\textbf{circular}} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}. \tag{16}$$

The only pivots are 2 and $\frac{3}{2}$. The eigenvalues are 3 and 3 and 0. The determinant is zero. The nullspace still contains $\boldsymbol{x} = (1, 1, 1)$, when all the masses move together. This movement vector $(1, 1, 1)$ is in the nullspace of $A_{\textbf{circular}}$ and $K_{\textbf{circular}} = A^{\mathrm{T}}CA$.

May I summarize this section? I hope the example will help you connect calculus with linear algebra, replacing differential equations by difference equations. If your step Δx is small enough, you will have a totally satisfactory solution.

$$\textbf{The equation is} \quad -\frac{d}{dx}\left(c(x)\frac{du}{dx}\right) = f(x) \quad \textbf{with} \quad u(0) = 0 \quad \textbf{and} \quad \left[u(1) \text{ or } \frac{du}{dx}(1)\right] = 0$$

Divide the bar into N pieces of length Δx. Replace du/dx by $A\boldsymbol{u}$ and $-dy/dx$ by $A^{\mathrm{T}}\boldsymbol{y}$. Now A and A^{T} include $1/\Delta x$. The end conditions are $u_0 = 0$ and $[u_N = 0$ or $y_N = 0]$.

The three steps $-d/dx$ and $c(x)$ and d/dx correspond to A^{T} and C and A:

$$\boldsymbol{f} = A^{\mathrm{T}}\boldsymbol{y} \text{ and } \boldsymbol{y} = C\boldsymbol{e} \text{ and } \boldsymbol{e} = A\boldsymbol{u} \text{ give } A^{\mathrm{T}}CA\boldsymbol{u} = \boldsymbol{f}.$$

This is a fundamental example in computational science and engineering.

1. Model the problem by a differential equation
2. Discretize the differential equation to a difference equation
3. Understand and solve the difference equation (and boundary conditions!)
4. Interpret the solution; visualize it; redesign if needed.

Numerical simulation has become a third branch of science, beside experiment and deduction. Computer design of the Boeing 777 was much less expensive than a wind tunnel.

The two texts *Introduction to Applied Mathematics* and *Computational Science and Engineering* (Wellesley-Cambridge Press) develop this whole subject further—see the course page **math.mit.edu/18085** with video lectures (The lectures are also on **ocw.mit.edu** and **YouTube**). I hope this book helps you to see the framework behind the computations.

Problem Set 10.2

1 Show that $\det A_0^{\mathrm{T}}C_0A_0 = c_1c_2c_3+c_1c_3c_4+c_1c_2c_4+c_2c_3c_4$. Find also $\det A_1^{\mathrm{T}}C_1A_1$ in the fixed-free example.

2 Invert $A_1^{\mathrm{T}}C_1A_1$ in the fixed-free example by multiplying $A_1^{-1}C_1^{-1}(A_1^{\mathrm{T}})^{-1}$.

3 In the free-free case when $A^{\mathrm{T}}CA$ in equation (15) is singular, add the three equations $A^{\mathrm{T}}CA\boldsymbol{u} = \boldsymbol{f}$ to show that we need $f_1+f_2+f_3 = 0$. Find a solution to $A^{\mathrm{T}}CA\boldsymbol{u} = \boldsymbol{f}$ when the forces $\boldsymbol{f} = (-1, 0, 1)$ balance themselves. Find all solutions!

4 Both end conditions for the free-free differential equation are $du/dx = 0$:

$$-\frac{d}{dx}\left(c(x)\frac{du}{dx}\right) = f(x) \quad \text{with} \quad \frac{du}{dx} = 0 \text{ at both ends.}$$

Integrate both sides to show that the force $f(x)$ must balance itself, $\int f(x)\,dx = 0$, or there is no solution. The complete solution is one particular solution $u(x)$ plus any constant. The constant corresponds to $\boldsymbol{u} = (1, 1, 1)$ in the nullspace of $A^{\mathrm{T}}CA$.

5 In the fixed-free problem, the matrix A is square and invertible. We can solve $A^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{f}$ separately from $A\boldsymbol{u} = \boldsymbol{e}$. Do the same for the differential equation:

$$\text{Solve } -\frac{dy}{dx} = f(x) \text{ with } y(1) = 0. \quad \text{Graph } y(x) \text{ if } f(x) = 1.$$

6 The 3 by 3 matrix $K_1 = A_1^{\mathrm{T}} C_1 A_1$ in equation (6) splits into three "element matrices" $c_1 E_1 + c_2 E_2 + c_3 E_3$. Write down those pieces, one for each c. Show how they come from *column times row* multiplication of $A_1^{\mathrm{T}} C_1 A_1$. This is how finite element stiffness matrices are actually assembled.

7 For five springs and four masses with both ends fixed, what are the matrices A and C and K? With $C = I$ solve $K\boldsymbol{u} = \mathsf{ones}(4)$.

8 Compare the solution $\boldsymbol{u} = (u_1, u_2, u_3, u_4)$ in Problem 7 to the solution of the continuous problem $-u'' = 1$ with $u(0) = 0$ and $u(1) = 0$. The parabola $u(x)$ should correspond at $x = \frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}$ to $\boldsymbol{u}$—is there a $(\Delta x)^2$ factor to account for?

9 Solve the fixed-free problem $-u'' = mg$ with $u(0) = 0$ and $u'(1) = 0$. Compare $u(x)$ at $x = \frac{1}{3}, \frac{2}{3}, \frac{3}{3}$ with the vector $\boldsymbol{u} = (3mg, 5mg, 6mg)$ in Example 2.

10 Suppose $c_1 = c_2 = c_3 = c_4 = 1$, $m_1 = 2$ and $m_2 = m_3 = 1$. Solve $A^{\mathrm{T}} C A\,\boldsymbol{u} = (2, 1, 1)$ for this fixed-fixed line of springs. Which mass moves the most (largest u)?

11 (MATLAB) Find the displacements $u(1), \ldots, u(100)$ of 100 masses connected by springs all with $c = 1$. Each force is $f(i) = .01$. Print graphs of $\boldsymbol{u}$ with **fixed-fixed** and **fixed-free** ends. Note that $\mathsf{diag}(\mathsf{ones}(n, 1), d)$ is a matrix with n ones along diagonal d. This print command will graph a vector u:

$\mathsf{plot}(u, \text{'+'})$; $\mathsf{xlabel}(\text{'mass number'})$; $\mathsf{ylabel}(\text{'movement'})$; print

12 (MATLAB) Chemical engineering has a first derivative du/dx from fluid velocity as well as d^2u/dx^2 from diffusion. Replace du/dx by a *forward* difference, then a *centered* difference, then a *backward* difference, with $\Delta x = \frac{1}{8}$. Graph your three numerical solutions of

$$-\frac{d^2u}{dx^2} + 10\,\frac{du}{dx} = 1 \quad \text{with} \quad u(0) = u(1) = 0.$$

This ***convection-diffusion equation*** appears everywhere. It transforms to the Black-Scholes equation for option prices in mathematical finance.

Problem 12 is developed into the first MATLAB homework in my 18.085 course on Computational Science and Engineering at MIT. Videos on *ocw.mit.edu.*

10.3 Markov Matrices, Population, and Economics

This section is about ***positive matrices***: every $a_{ij} > 0$. The key fact is quick to state: ***The largest eigenvalue is real and positive and so is its eigenvector.*** In economics and ecology and population dynamics and random walks, that fact leads a long way:

Markov $\lambda_{\max} = 1$ **Population** $\lambda_{\max} > 1$ **Consumption** $\lambda_{\max} < 1$

$\lambda_{\max}$ controls the powers of A. We will see this first for $\lambda_{\max} = 1$.

Markov Matrices

Multiply a positive vector $\boldsymbol{u}_0$ again and again by this matrix A:

Markov matrix $A = \begin{bmatrix} .8 & .3 \\ .2 & .7 \end{bmatrix}$ $\quad \boldsymbol{u}_1 = A\boldsymbol{u}_0 \quad \boldsymbol{u}_2 = A\boldsymbol{u}_1 = A^2\boldsymbol{u}_0$

After k steps we have $A^k\boldsymbol{u}_0$. The vectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3, \ldots$ will approach a "*steady state*" $\boldsymbol{u}_\infty = (.6, .4)$. This final outcome does not depend on the starting vector $\boldsymbol{u}_0$. ***For every*** $\boldsymbol{u}_0 = (a, 1-a)$ ***we converge to the same*** $\boldsymbol{u}_{\infty(.6,.4)}$. The question is why.

The steady state equation $A\boldsymbol{u}_\infty = \boldsymbol{u}_\infty$ makes $\boldsymbol{u}_\infty$ ***an eigenvector with eigenvalue*** $\mathbf{1}$:

Steady state $$\begin{bmatrix} .8 & .3 \\ .2 & .7 \end{bmatrix} \begin{bmatrix} .6 \\ .4 \end{bmatrix} = \begin{bmatrix} .6 \\ .4 \end{bmatrix} = \boldsymbol{u}_\infty.$$

Multiplying by A does not change $\boldsymbol{u}_\infty$. But this does not explain why so many vectors $\boldsymbol{u}_0$ lead to $\boldsymbol{u}_\infty$. Other examples might have a steady state, but it is not necessarily attractive:

Not Markov $B = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ has the unattractive steady state $B \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

In this case, the starting vector $\boldsymbol{u}_0 = (0, 1)$ will give $\boldsymbol{u}_1 = (0, 2)$ and $\boldsymbol{u}_2 = (0, 4)$. The second components are doubled. In the language of eigenvalues, B has $\lambda = 1$ but also $\lambda = 2$— this produces instability. The component of $\boldsymbol{u}$ along that unstable eigenvector is multiplied by λ, and $|\lambda| > 1$ means blowup.

This section is about two special properties of A that guarantee a *stable steady state*. These properties define a positive ***Markov matrix***, and A above is one particular example:

Markov matrix

1. ***Every entry of*** A ***is positive:*** $a_{ij} > 0$.
2. ***Every column of*** A ***adds to*** $\mathbf{1}$.

Column 2 of B adds to 2, not 1. When A is a Markov matrix, two facts are immediate:
Because of $\mathbf{1}$: Multiplying $\boldsymbol{u}_0 \geq 0$ by A produces a nonnegative $\boldsymbol{u}_1 = A\boldsymbol{u}_0 \geq 0$.
Because of $\mathbf{2}$: If the components of $\boldsymbol{u}_0$ add to 1, so do the components of $\boldsymbol{u}_1 = A\boldsymbol{u}_0$.

Reason: The components of $\boldsymbol{u}_0$ add to 1 when $[1 \ \ldots \ 1]\boldsymbol{u}_0 = 1$. This is true for each column of A by Property 2. Then by matrix multiplication $[1 \ \ldots \ 1]A = [1 \ \ldots \ 1]$:

Components of $A\boldsymbol{u}_0$ add to 1 $\quad [1 \ \ldots \ 1]A\boldsymbol{u}_0 = [1 \ \ldots \ 1]\boldsymbol{u}_0 = 1.$

The same facts apply to $\boldsymbol{u}_2 = A\boldsymbol{u}_1$ and $\boldsymbol{u}_3 = A\boldsymbol{u}_2$. ***Every vector $A^k\boldsymbol{u}_0$ is nonnegative with components adding to*** 1. These are "***probability vectors***." The limit $\boldsymbol{u}_\infty$ is also a probability vector—but we have to prove that there is a limit. We will show that $\lambda_{\max} = 1$ for a positive Markov matrix.

Example 1 The fraction of rental cars in Denver starts at $\frac{1}{50} = .02$. The fraction outside Denver is .98. Every month, 80% of the Denver cars stay in Denver (and 20% leave). Also 5% of the outside cars come in (95% stay outside). This means that the fractions $\boldsymbol{u}_0 = (.02, .98)$ are multiplied by A:

First month $\quad A = \begin{bmatrix} \mathbf{.80} & \mathbf{.05} \\ \mathbf{.20} & \mathbf{.95} \end{bmatrix}$ leads to $\boldsymbol{u}_1 = A\boldsymbol{u}_0 = A\begin{bmatrix} .02 \\ .98 \end{bmatrix} = \begin{bmatrix} .065 \\ .935 \end{bmatrix}$.

Notice that $.065 + .935 = 1$. All cars are accounted for. Each step multiplies by A:

Next month $\quad \boldsymbol{u}_2 = A\boldsymbol{u}_1 = (.09875, \ .90125)$. This is $A^2\boldsymbol{u}_0$.

All these vectors are positive because A is positive. Each vector $\boldsymbol{u}_k$ will have its components adding to 1. The first component has grown from .02 and cars are moving toward Denver. What happens in the long run?

This section involves powers of matrices. The understanding of A^k was our first and best application of diagonalization. Where A^k can be complicated, the diagonal matrix Λ^k is simple. The eigenvector matrix X connects them: A^k equals $X\Lambda^k X^{-1}$. The new application to Markov matrices uses the eigenvalues (in Λ) and the eigenvectors (in X). We will show that $\boldsymbol{u}_\infty$ **is an eigenvector of $\boldsymbol{A}$ corresponding to** $\lambda = 1$.

Since every column of A adds to 1, nothing is lost or gained. We are moving rental cars or populations, and no cars or people suddenly appear (or disappear). The fractions add to 1 and the matrix A keeps them that way. The question is how they are distributed after k time periods—which leads us to A^k.

Solution $A^k\boldsymbol{u}_0$ gives the fractions in and out of Denver after k steps. We diagonalize A to understand A^k. The eigenvalues are $\lambda = \mathbf{1}$ and $\mathbf{.75}$ (the trace is 1.75).

$A\boldsymbol{x} = \lambda\boldsymbol{x}$ $\qquad A\begin{bmatrix} .2 \\ .8 \end{bmatrix} = 1\begin{bmatrix} .2 \\ .8 \end{bmatrix}$ **and** $A\begin{bmatrix} -1 \\ 1 \end{bmatrix} = .75\begin{bmatrix} -1 \\ 1 \end{bmatrix}$.

The starting vector $\boldsymbol{u}_0$ combines $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, in this case with coefficients 1 and .18:

Combination of eigenvectors $\qquad \boldsymbol{u}_0 = \begin{bmatrix} .02 \\ .98 \end{bmatrix} = \begin{bmatrix} .2 \\ .8 \end{bmatrix} + .18\begin{bmatrix} -1 \\ 1 \end{bmatrix}$.

Now multiply by A to find $\boldsymbol{u}_1$. The eigenvectors are multiplied by $\lambda_1 = 1$ and $\lambda_2 = .75$:

Each x is multiplied by λ $\qquad \boldsymbol{u}_1 = 1\begin{bmatrix} .2 \\ .8 \end{bmatrix} + (.75)(.18)\begin{bmatrix} -1 \\ 1 \end{bmatrix}$.

Every month, another $\lambda = .75$ multiplies the vector $\boldsymbol{x}_2$. The eigenvector $\boldsymbol{x}_1$ is unchanged:

After k steps $$\boldsymbol{u}_k = A^k \boldsymbol{u}_0 = 1^k \begin{bmatrix} .2 \\ .8 \end{bmatrix} + (.75)^k (.18) \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

This equation reveals what happens. ***The eigenvector $\boldsymbol{x}_1$ with $\lambda = 1$ is the steady state.*** The other eigenvector $\boldsymbol{x}_2$ disappears because $|\lambda| < 1$. The more steps we take, the closer we come to $\boldsymbol{u}_\infty = (.2, .8)$. In the limit, $\frac{2}{10}$ of the cars are in Denver and $\frac{8}{10}$ are outside. This is the pattern for Markov chains, even starting from $\boldsymbol{u}_0 = (0, 1)$:

If A is a *positive* Markov matrix (entries $a_{ij} > 0$, each column adds to 1), then $\lambda_1 = 1$ is larger than any other eigenvalue. The eigenvector $\boldsymbol{x}_1$ is the ***steady state***:

$$\boldsymbol{u}_k = \boldsymbol{x}_1 + c_2(\lambda_2)^k \boldsymbol{x}_2 + \cdots + c_n(\lambda_n)^k \boldsymbol{x}_n \quad \textbf{\textit{always approaches}} \quad \boldsymbol{u}_\infty = \boldsymbol{x}_1.$$

The first point is to see that $\lambda = 1$ is an eigenvalue of A. *Reason*: Every column of $A - I$ adds to $1 - 1 = 0$. The rows of $A - I$ add up to the zero row. Those rows are linearly dependent, so $A - I$ is singular. Its determinant is zero and $\lambda = 1$ is an eigenvalue.

The second point is that no eigenvalue can have $|\lambda| > 1$. With such an eigenvalue, the powers A^k would grow. But A^k is also a Markov matrix! A^k has positive entries still adding to 1—and that leaves no room to get large.

A lot of attention is paid to the possibility that another eigenvalue has $|\lambda| = 1$.

Example 2 $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ has no steady state because $\lambda_2 = -1$.

This matrix sends all cars from inside Denver to outside, and vice versa. The powers A^k alternate between A and I. The second eigenvector $\boldsymbol{x}_2 = (-1, 1)$ will be multiplied by $\lambda_2 = -1$ at every step—and does not become smaller: No steady state.

Suppose the entries of A or any power of A are all *positive*—zero is not allowed. In this "regular" or "primitive" case, $\lambda = 1$ is strictly larger than any other eigenvalue. The powers A^k approach the rank one matrix that has the steady state in every column.

Example 3 ("**Everybody moves**") Start with three groups. At each time step, half of group 1 goes to group 2 and the other half goes to group 3. The other groups also *split in half and move*. Take one step from the starting populations p_1, p_2, p_3:

New populations $$\boldsymbol{u}_1 = A\boldsymbol{u}_0 = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2}p_2 + \frac{1}{2}p_3 \\ \frac{1}{2}p_1 + \frac{1}{2}p_3 \\ \frac{1}{2}p_1 + \frac{1}{2}p_2 \end{bmatrix}.$$

A is a Markov matrix. Nobody is born or lost. A contains zeros, which gave trouble in Example 2. But after two steps in this new example, the zeros disappear from A^2:

Two-step matrix $$\boldsymbol{u}_2 = A^2\boldsymbol{u}_0 = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix}.$$

The eigenvalues of A are $\lambda_1 = 1$ (because A is Markov) and $\lambda_2 = \lambda_3 = -\frac{1}{2}$. For $\lambda = 1$, ***the eigenvector*** $\boldsymbol{x}_1 = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ ***will be the steady state***. When three equal populations split in half and move, the populations are again equal. Starting from $\boldsymbol{u}_0 = (8, 16, 32)$, the Markov chain approaches its steady state:

$$\boldsymbol{u}_0 = \begin{bmatrix} 8 \\ 16 \\ 32 \end{bmatrix} \quad \boldsymbol{u}_1 = \begin{bmatrix} 24 \\ 20 \\ 12 \end{bmatrix} \quad \boldsymbol{u}_2 = \begin{bmatrix} 16 \\ 18 \\ 22 \end{bmatrix} \quad \boldsymbol{u}_3 = \begin{bmatrix} 20 \\ 19 \\ 17 \end{bmatrix}.$$

The step to $\boldsymbol{u}_4$ will split some people in half. This cannot be helped. The total population is $8 + 16 + 32 = 56$ at every step. The steady state is 56 times $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. You can see the three populations approaching, but never reaching, their final limits $56/3$.

Challenge Problem 6.7.16 created a Markov matrix A from the number of links between websites. The steady state $\boldsymbol{u}$ will give the Google rankings. ***Google finds $\boldsymbol{u}_\infty$ by a random walk that follows links*** (*random surfing*). That eigenvector comes from counting the fraction of visits to each website—a quick way to compute the steady state.

The size $|\lambda_2|$ of the second eigenvalue controls the speed of convergence to steady state.

Perron-Frobenius Theorem

One matrix theorem dominates this subject. The Perron-Frobenius Theorem applies when all $a_{ij} \geq 0$. There is no requirement that columns add to 1. We prove the neatest form, when all $a_{ij} > 0$: any positive matrix A (not necessarily positive definite!).

Perron-Frobenius for $A > 0$	***All numbers in $A\boldsymbol{x} = \lambda_{\max}\boldsymbol{x}$ are strictly positive.***

Proof The key idea is to look at all numbers t such that $A\boldsymbol{x} \geq t\boldsymbol{x}$ for some nonnegative vector $\boldsymbol{x}$ (other than $\boldsymbol{x} = \boldsymbol{0}$). We are allowing inequality in $A\boldsymbol{x} \geq t\boldsymbol{x}$ in order to have many small positive candidates t. For the largest value $t_{\max}$ (which is attained), we will show that ***equality holds***: $A\boldsymbol{x} = t_{\max}\boldsymbol{x}$.

Otherwise, if $A\boldsymbol{x} \geq t_{\max}\boldsymbol{x}$ is not an equality, multiply by A. Because A is positive that produces a strict inequality $A^2\boldsymbol{x} > t_{\max}A\boldsymbol{x}$. Therefore the positive vector $\boldsymbol{y} = A\boldsymbol{x}$ satisfies $A\boldsymbol{y} > t_{\max}\boldsymbol{y}$, and $t_{\max}$ could be increased. This contradiction forces the equality $A\boldsymbol{x} = t_{\max}\boldsymbol{x}$, and *we have an eigenvalue*. Its eigenvector $\boldsymbol{x}$ is positive because on the left side of that equality, $A\boldsymbol{x}$ is sure to be positive.

To see that no eigenvalue can be larger than $t_{\max}$, suppose $A\boldsymbol{z} = \lambda\boldsymbol{z}$. Since λ and $\boldsymbol{z}$ may involve negative or complex numbers, we take absolute values: $|\lambda||\boldsymbol{z}| = |A\boldsymbol{z}| \leq A|\boldsymbol{z}|$ by the "triangle inequality." This $|\boldsymbol{z}|$ is a nonnegative vector, so this $|\lambda|$ is one of the possible candidates t. Therefore $|\lambda|$ cannot exceed $t_{\max}$—which must be $\lambda_{\max}$.

Population Growth

Divide the population into three age groups: age < 20, age 20 to 39, and age 40 to 59. At year T the sizes of those groups are n_1, n_2, n_3. Twenty years later, the sizes have changed for three reasons: births, deaths, and getting older.

1. **Reproduction** $n_1^{\text{new}} = F_1\, n_1 + F_2\, n_2 + F_3\, n_3$ gives a new generation
2. **Survival** $n_2^{\text{new}} = P_1 n_1$ and $n_3^{\text{new}} = P_2 n_2$ gives the older generations

The fertility rates are F_1, F_2, F_3 (F_2 largest). The *Leslie matrix* A might look like this:

$$\begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix}^{\text{new}} = \begin{bmatrix} F_1 & F_2 & F_3 \\ P_1 & 0 & 0 \\ 0 & P_2 & 0 \end{bmatrix} \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix} = \begin{bmatrix} .04 & \mathbf{1.1} & .01 \\ .98 & 0 & 0 \\ 0 & \mathbf{.92} & 0 \end{bmatrix} \begin{bmatrix} n_1 \\ \boldsymbol{n_2} \\ n_3 \end{bmatrix}.$$

This is population projection in its simplest form, the same matrix A at every step. In a realistic model, A will change with time (from the environment or internal factors). Professors may want to include a fourth group, age ≥ 60, but we don't allow it.

The matrix has $A \geq 0$ but not $A > 0$. The Perron-Frobenius theorem still applies because $A^3 > 0$. The largest eigenvalue is $\lambda_{\max} \approx 1.06$. You can watch the generations move, starting from $n_2 = 1$ in the middle generation:

$$\textbf{eig}(A) = \begin{matrix} \mathbf{1.06} \\ -1.01 \\ -0.01 \end{matrix} \quad A^2 = \begin{bmatrix} 1.08 & \mathbf{0.05} & .00 \\ 0.04 & \mathbf{1.08} & .01 \\ 0.90 & 0 & 0 \end{bmatrix} \quad A^3 = \begin{bmatrix} 0.10 & \mathbf{1.19} & .01 \\ 0.06 & \mathbf{0.05} & .00 \\ 0.04 & \mathbf{0.99} & .01 \end{bmatrix}.$$

A fast start would come from $\boldsymbol{u}_0 = (0, 1, 0)$. That middle group will reproduce 1.1 and also survive .92. The newest and oldest generations are in $\boldsymbol{u}_1 = (1.1, 0, .92) =$ column 2 of A. Then $\boldsymbol{u}_2 = A\boldsymbol{u}_1 = A^2\boldsymbol{u}_0$ is the second column of A^2. The early numbers (transients) depend a lot on $\boldsymbol{u}_0$, but ***the asymptotic growth rate*** $\lambda_{\max}$ ***is the same from every start***. Its eigenvector $\boldsymbol{x} = (.63, .58, .51)$ shows all three groups growing steadily together.

Caswell's book on *Matrix Population Models* emphasizes sensitivity analysis. The model is never exactly right. If the F's or P's in the matrix change by 10%, does $\lambda_{\max}$ go below 1 (which means extinction)? Problem 19 will show that a matrix change ΔA produces an eigenvalue change $\Delta\lambda = \boldsymbol{y}^{\mathrm{T}}(\Delta A)\boldsymbol{x}$. Here $\boldsymbol{x}$ and $\boldsymbol{y}^{\mathrm{T}}$ are the right and left eigenvectors of A, with $A\boldsymbol{x} = d\boldsymbol{x}$ and $A^{\mathrm{T}}\boldsymbol{y} = \lambda\boldsymbol{y}$.

Linear Algebra in Economics: The Consumption Matrix

A long essay about linear algebra in economics would be out of place here. A short note about one matrix seems reasonable. The ***consumption matrix*** tells how much of each input goes into a unit of output. This describes the manufacturing side of the economy.

Consumption matrix We have n industries like chemicals, food, and oil. To produce a unit of chemicals may require .2 units of chemicals, .3 units of food, and .4 units of oil. Those numbers go into row 1 of the consumption matrix A:

$$\begin{bmatrix}\text{chemical output}\\ \text{food output}\\ \text{oil output}\end{bmatrix} = \begin{bmatrix}.2 & .3 & .4\\ .4 & .4 & .1\\ .5 & .1 & .3\end{bmatrix}\begin{bmatrix}\text{chemical input}\\ \text{food input}\\ \text{oil input}\end{bmatrix}.$$

Row 2 shows the inputs to produce food—a heavy use of chemicals and food, not so much oil. Row 3 of A shows the inputs consumed to refine a unit of oil. The real consumption matrix for the United States in 1958 contained 83 industries. The models in the 1990's are much larger and more precise. We chose a consumption matrix that has a convenient eigenvector.

Now comes the question: Can this economy meet demands y_1, y_2, y_3 for chemicals, food, and oil? To do that, the inputs p_1, p_2, p_3 will have to be higher—because part of $\boldsymbol{p}$ is consumed in producing $\boldsymbol{y}$. The input is $\boldsymbol{p}$ and the consumption is $A\boldsymbol{p}$, which leaves the output $\boldsymbol{p} - A\boldsymbol{p}$. This net production is what meets the demand $\boldsymbol{y}$:

Problem Find a vector $\boldsymbol{p}$ such that $\boldsymbol{p} - A\boldsymbol{p} = \boldsymbol{y}$ or $\boldsymbol{p} = (I - A)^{-1}\boldsymbol{y}$.

Apparently the linear algebra question is whether $I - A$ is invertible. But there is more to the problem. The vector $\boldsymbol{y}$ of required outputs is nonnegative, and so is A. *The production levels in $\boldsymbol{p} = (I - A)^{-1}\boldsymbol{y}$ must also be nonnegative.* The real question is:

When is $(\boldsymbol{I - A})^{-1}$ a nonnegative matrix?

This is the test on $(I - A)^{-1}$ for a productive economy, which can meet any demand. If A is small compared to I, then $A\boldsymbol{p}$ is small compared to $\boldsymbol{p}$. There is plenty of output. If A is too large, then production consumes too much and the demand $\boldsymbol{y}$ cannot be met.

"Small" or "large" is decided by the largest eigenvalue λ_1 of A (which is positive):

If $\lambda_1 > 1$ then $(I - A)^{-1}$ has negative entries
If $\lambda_1 = 1$ then $(I - A)^{-1}$ fails to exist
If $\lambda_1 < 1$ then $(I - A)^{-1}$ is nonnegative as desired.

The main point is that last one. The reasoning uses a nice formula for $(I - A)^{-1}$, which we give now. The most important infinite series in mathematics is the ***geometric series*** $1 + x + x^2 + \cdots$. This series adds up to $1/(1 - x)$ provided x lies between -1 and 1. When $x = 1$ the series is $1 + 1 + 1 + \cdots = \infty$. When $|x| \geq 1$ the terms x^n don't go to zero and the series has no chance to converge.

The nice formula for $(I - A)^{-1}$ is the ***geometric series of matrices***:

Geometric series $(I - A)^{-1} = I + A + A^2 + A^3 + \cdots$.

If you multiply the series $S = I + A + A^2 + \cdots$ by A, you get the same series except for I. Therefore $S - AS = I$, which is $(I - A)S = I$. The series adds to $S = (I - A)^{-1}$ if it converges. ***And it converges if all eigenvalues of*** A ***have*** $|\lambda| < 1$.

In our case $A \geq 0$. All terms of the series are nonnegative. Its sum is $(I - A)^{-1} \geq 0$.

Example 4 $A = \begin{bmatrix} .2 & .3 & .4 \\ .4 & .4 & .1 \\ .5 & .1 & .3 \end{bmatrix}$ has $\lambda_{\max} = \mathbf{.9}$ and $(I - A)^{-1} = \frac{1}{93}\begin{bmatrix} 41 & 25 & 27 \\ 33 & 36 & 24 \\ 34 & 23 & 36 \end{bmatrix}$.

This economy is productive. A is small compared to I, because $\lambda_{\max}$ is .9. To meet the demand $\boldsymbol{y}$, start from $\boldsymbol{p} = (I - A)^{-1}\boldsymbol{y}$. Then $A\boldsymbol{p}$ is consumed in production, leaving $\boldsymbol{p} - A\boldsymbol{p}$. This is $(I - A)\boldsymbol{p} = \boldsymbol{y}$, and the demand is met.

Example 5 $A = \begin{bmatrix} 0 & 4 \\ 1 & 0 \end{bmatrix}$ has $\lambda_{\max} = \mathbf{2}$ and $(I - A)^{-1} = -\frac{1}{3}\begin{bmatrix} 1 & 4 \\ 1 & 1 \end{bmatrix}$.

This consumption matrix A is too large. Demands can't be met, because production consumes more than it yields. The series $I + A + A^2 + \ldots$ does not converge to $(I - A)^{-1}$ because $\lambda_{\max} > 1$. The series is growing while $(I - A)^{-1}$ is actually negative.

In the same way $1 + 2 + 4 + \cdots$ is not really $1/(1 - 2) = -1$. But not entirely false !

Problem Set 10.3

Questions 1–12 are about Markov matrices and their eigenvalues and powers.

1 Find the eigenvalues of this Markov matrix (their sum is the trace):

$$A = \begin{bmatrix} .90 & .15 \\ .10 & .85 \end{bmatrix}.$$

What is the steady state eigenvector for the eigenvalue $\lambda_1 = 1$?

2 Diagonalize the Markov matrix in Problem 1 to $A = X\Lambda X^{-1}$ by finding its other eigenvector:

$$A = \begin{bmatrix} & \\ & \end{bmatrix}\begin{bmatrix} 1 & \\ & .75 \end{bmatrix}\begin{bmatrix} & \\ & \end{bmatrix}.$$

What is the limit of $A^k = X\Lambda^k X^{-1}$ when $\Lambda^k = \begin{bmatrix} 1 & 0 \\ 0 & .75^k \end{bmatrix}$ approaches $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$?

3 What are the eigenvalues and steady state eigenvectors for these Markov matrices?

$$A = \begin{bmatrix} 1 & .2 \\ 0 & .8 \end{bmatrix} \quad A = \begin{bmatrix} .2 & 1 \\ .8 & 0 \end{bmatrix} \quad A = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}.$$

4 For every 4 by 4 Markov matrix, what eigenvector of A^{T} corresponds to the (known) eigenvalue $\lambda = 1$?

5 Every year 2% of young people become old and 3% of old people become dead. (No births.) Find the steady state for

$$\begin{bmatrix} \text{young} \\ \text{old} \\ \text{dead} \end{bmatrix}_{k+1} = \begin{bmatrix} .98 & .00 & 0 \\ .02 & .97 & 0 \\ .00 & .03 & 1 \end{bmatrix} \begin{bmatrix} \text{young} \\ \text{old} \\ \text{dead} \end{bmatrix}_{k}.$$

6 For a Markov matrix, the sum of the components of $\boldsymbol{x}$ equals the sum of the components of $A\boldsymbol{x}$. If $A\boldsymbol{x} = \lambda\boldsymbol{x}$ with $\lambda \neq 1$, prove that the components of this non-steady eigenvector $\boldsymbol{x}$ add to zero.

7 Find the eigenvalues and eigenvectors of A. Explain why A^k approaches A^∞:

$$A = \begin{bmatrix} .8 & .3 \\ .2 & .7 \end{bmatrix} \qquad A^\infty = \begin{bmatrix} .6 & .6 \\ .4 & .4 \end{bmatrix}.$$

Challenge problem: Which Markov matrices produce that steady state $(.6, .4)$?

8 The steady state eigenvector of a permutation matrix is $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. This is *not* approached when $\boldsymbol{u}_0 = (0, 0, 0, 1)$. What are $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ and $\boldsymbol{u}_3$ and $\boldsymbol{u}_4$? What are the four eigenvalues of P, which solve $\lambda^4 = 1$?

Permutation matrix = Markov matrix $$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

9 Prove that the square of a Markov matrix is also a Markov matrix.

10 If $A = \begin{bmatrix} \mathbf{a} & \mathbf{b} \\ \mathbf{c} & \mathbf{d} \end{bmatrix}$ is a Markov matrix, its eigenvalues are 1 and _____. The steady state eigenvector is $\boldsymbol{x}_1 =$ _____.

11 Complete A to a Markov matrix and find the steady state eigenvector. When A is a symmetric Markov matrix, why is $\boldsymbol{x}_1 = (1, \ldots, 1)$ its steady state?

$$A = \begin{bmatrix} .7 & .1 & .2 \\ .1 & .6 & .3 \\ - & - & - \end{bmatrix}.$$

12 A Markov differential equation is not $d\boldsymbol{u}/dt = A\boldsymbol{u}$ but $d\boldsymbol{u}/dt = (A - I)\boldsymbol{u}$. The diagonal is negative, the rest of $A - I$ is positive. The columns add to zero, not 1.

Find λ_1 and λ_2 for $B = A - I = \begin{bmatrix} -.2 & .3 \\ .2 & -.3 \end{bmatrix}$. Why does $A - I$ have $\lambda_1 = 0$?

When $e^{\lambda_1 t}$ and $e^{\lambda_2 t}$ multiply $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, what is the steady state as $t \to \infty$?

Questions 13–15 are about linear algebra in economics.

13 Each row of the consumption matrix in Example 4 adds to .9. Why does that make $\lambda = .9$ an eigenvalue, and what is the eigenvector?

14 Multiply $I + A + A^2 + A^3 + \cdots$ by $I - A$ to get I. The series adds to $(I - A)^{-1}$. For $A = \begin{bmatrix} 0 & \frac{1}{2} \\ 1 & 0 \end{bmatrix}$, find A^2 and A^3 and use the pattern to add up the series.

15 For which of these matrices does $I + A + A^2 + \cdots$ yield a nonnegative matrix $(I - A)^{-1}$? Then the economy can meet any demand:

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \qquad A = \begin{bmatrix} 0 & 4 \\ .2 & 0 \end{bmatrix} \qquad A = \begin{bmatrix} .5 & 1 \\ .5 & 0 \end{bmatrix}.$$

If the demands are $\boldsymbol{y} = (2, 6)$, what are the vectors $\boldsymbol{p} = (I - A)^{-1}\boldsymbol{y}$?

16 (Markov again) This matrix has zero determinant. What are its eigenvalues?

$$A = \begin{bmatrix} .4 & .2 & .3 \\ .2 & .4 & .3 \\ .4 & .4 & .4 \end{bmatrix}.$$

Find the limits of $A^k\boldsymbol{u}_0$ starting from $\boldsymbol{u}_0 = (1, 0, 0)$ and then $\boldsymbol{u}_0 = (100, 0, 0)$.

17 If A is a Markov matrix, why doesn't $I + A + A^2 + \cdots$ add up to $(I - A)^{-1}$?

18 For the Leslie matrix show that $\det(A - \lambda I) = 0$ gives $F_1\lambda^2 + F_2P_1\lambda + F_3P_1P_2 = \lambda^3$. The right side λ^3 is larger as $\lambda \longrightarrow \infty$. The left side is larger at $\lambda = 1$ if $F_1 + F_2P_1 + F_3P_1P_2 > 1$. In that case the two sides are equal at an eigenvalue $\lambda_{\max} > 1$: *growth.*

19 **Sensitivity of eigenvalues**: A matrix change ΔA produces eigenvalue changes $\Delta\Lambda$. *Those changes* $\Delta\lambda_1, \ldots, \Delta\lambda_n$ *are on the diagonal of* $(X^{-1}\,\Delta A\,X)$. **Challenge**: Start from $AX = X\Lambda$. The eigenvectors and eigenvalues change by ΔX and $\Delta\Lambda$:

$(A+\Delta A)(X+\Delta X) = (X+\Delta X)(\Lambda+\Delta\Lambda)$ becomes $A(\Delta X)+(\Delta A)X = X(\Delta\Lambda)+(\Delta X)\Lambda$.

Small terms $(\Delta A)(\Delta X)$ and $(\Delta X)(\Delta\Lambda)$ are ignored. *Multiply the last equation by* X^{-1}. From the inner terms, the diagonal part of $X^{-1}(\Delta A)X$ gives $\Delta\Lambda$ as we want. *Why do the outer terms* $X^{-1}\,A\,\Delta X$ *and* $X^{-1}\,\Delta X\,\Lambda$ *cancel on the diagonal*?

Explain $X^{-1}A = \Lambda X^{-1}$ and then $\quad$ **diag**$(\Lambda\,X^{-1}\,\Delta X) =$ **diag**$(X^{-1}\,\Delta X\,\Lambda)$.

20 Suppose $B > A > 0$, meaning that each $b_{ij} > a_{ij} > 0$. How does the Perron-Frobenius discussion show that $\lambda_{\max}(B) > \lambda_{\max}(A)$?

10.4 Linear Programming

Linear programming is linear algebra plus two new ideas: ***inequalities*** and ***minimization***. The starting point is still a matrix equation $A\boldsymbol{x} = \boldsymbol{b}$. But the only acceptable solutions are *nonnegative*. We require $\boldsymbol{x} \geq 0$ (meaning that no component of $\boldsymbol{x}$ can be negative). The matrix has $n > m$, more unknowns than equations. If there are any solutions $\boldsymbol{x} \geq \mathbf{0}$ to $A\boldsymbol{x} = \boldsymbol{b}$, there are probably a lot. Linear programming picks the solution $\boldsymbol{x}^* \geq \mathbf{0}$ that minimizes the cost:

The cost is $c_1x_1 + \cdots + c_nx_n$. The winning vector x^* is the nonnegative solution of $Ax = b$ that has smallest cost.

Thus a linear programming problem starts with a matrix A and two vectors $\boldsymbol{b}$ and $\boldsymbol{c}$:

i) A has $n > m$: for example $A = [1 \quad 1 \quad 2]$ (one equation, three unknowns)

ii) $\boldsymbol{b}$ has m components for m equations $A\boldsymbol{x} = \boldsymbol{b}$: for example $\boldsymbol{b} = [4]$

iii) The ***cost vector*** $\boldsymbol{c}$ has n components: for example $\boldsymbol{c} = [5 \quad 3 \quad 8]$.

Then the problem is to minimize $\boldsymbol{c} \cdot \boldsymbol{x}$ subject to the requirements $A\boldsymbol{x} = \boldsymbol{b}$ and $\boldsymbol{x} \geq 0$:

Minimize $5x_1 + 3x_2 + 8x_3$ ***subject to*** $x_1 + x_2 + 2x_3 = 4$ ***and*** $x_1, x_2, x_3 \geq 0$.

We jumped right into the problem, without explaining where it comes from. Linear programming is actually the most important application of mathematics to management. Development of the fastest algorithm and fastest code is highly competitive. You will see that finding $\boldsymbol{x}^*$ is harder than solving $A\boldsymbol{x} = \boldsymbol{b}$, because of the extra requirements: $\boldsymbol{x}^* \geq 0$ and minimum cost $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}^*$. We will explain the background, and the famous *simplex method*, and *interior point methods*, after solving the example.

Look first at the "constraints": $A\boldsymbol{x} = \boldsymbol{b}$ and $\boldsymbol{x} \geq 0$. The equation $x_1 + x_2 + 2x_3 = 4$ gives a plane in three dimensions. The nonnegativity $x_1 \geq 0, x_2 \geq 0, x_3 \geq 0$ chops the plane down to a triangle. The solution $\boldsymbol{x}^*$ must lie in the triangle PQR in Figure 8.6.

Inside that triangle, all components of $\boldsymbol{x}$ are positive. On the edges of PQR, one component is zero. At the corners P and Q and R, two components are zero. ***The optimal solution x^* will be one of those corners!*** We will now show why.

The triangle contains all vectors $\boldsymbol{x}$ that satisfy $A\boldsymbol{x} = \boldsymbol{b}$ and $\boldsymbol{x} \geq 0$. Those $\boldsymbol{x}$'s are called *feasible points*, and the triangle is the ***feasible set***. These points are the allowed candidates in the minimization of $\boldsymbol{c} \cdot \boldsymbol{x}$, which is the final step:

Find x^* in the triangle PQR to minimize the cost $5x_1 + 3x_2 + 8x_3$.

The vectors that have *zero* cost lie on the plane $5x_1 + 3x_2 + 8x_3 = 0$. That plane does not meet the triangle. We cannot achieve zero cost, while meeting the requirements on $\boldsymbol{x}$. So increase the cost C until the plane $5x_1 + 3x_2 + 8x_3 = C$ does meet the triangle. As C increases, we have *parallel planes moving toward the triangle*.

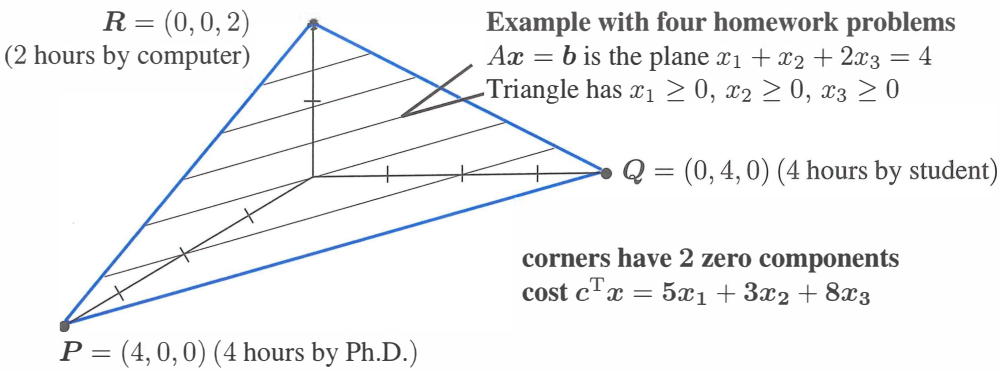


Figure 10.5: The triangle contains all nonnegative solutions: $A\boldsymbol{x} = \boldsymbol{b}$ and $\boldsymbol{x} \geq 0$. The lowest cost solution $\boldsymbol{x}^*$ is a corner $\boldsymbol{P}$, $\boldsymbol{Q}$, or $\boldsymbol{R}$ of this feasible set.

The first plane $5x_1 + 3x_2 + 8x_3 = C$ to touch the triangle has minimum cost C. ***The point where it touches is the solution*** $\boldsymbol{x}^*$. This touching point must be one of the corners $\boldsymbol{P}$ or $\boldsymbol{Q}$ or $\boldsymbol{R}$. A moving plane could not reach the inside of the triangle before it touches a corner! So check the cost $5x_1 + 3x_2 + 8x_3$ at each corner:

$\boldsymbol{P} = (4, 0, 0)$ costs 20	$\boldsymbol{Q} = (0, 4, 0)$ costs 12	$\boldsymbol{R} = (0, 0, 2)$ costs 16.

The winner is $\boldsymbol{Q}$. Then $\boldsymbol{x}^* = (0, 4, 0)$ solves the linear programming problem.

If the cost vector $\boldsymbol{c}$ is changed, the parallel planes are tilted. For small changes, $\boldsymbol{Q}$ is still the winner. For the cost $\boldsymbol{c} \cdot \boldsymbol{x} = 5x_1 + 4x_2 + 7x_3$, the optimum $\boldsymbol{x}^*$ moves to $\boldsymbol{R} = (0, 0, 2)$. The minimum cost is now $7 \cdot 2 = 14$.

Note 1 Some linear programs *maximize profit* instead of minimizing cost. The mathematics is almost the same. The parallel planes start with a large value of C, instead of a small value. They move toward the origin (instead of away), as C gets smaller. *The first touching point is still a corner.*

Note 2 The requirements $A\boldsymbol{x} = \boldsymbol{b}$ and $\boldsymbol{x} \geq 0$ could be impossible to satisfy. The equation $x_1 + x_2 + x_3 = -1$ cannot be solved with $\boldsymbol{x} \geq 0$. *That feasible set is empty.*

Note 3 It could also happen that the feasible set is *unbounded*. If the requirement is $x_1 + x_2 - 2x_3 = 4$, the large positive vector $(100, 100, 98)$ is now a candidate. So is the larger vector $(1000, 1000, 998)$. The plane $A\boldsymbol{x} = \boldsymbol{b}$ is no longer chopped off to a triangle. The two corners $\boldsymbol{P}$ and $\boldsymbol{Q}$ are still candidates for $\boldsymbol{x}^*$, but $\boldsymbol{R}$ moved to infinity.

Note 4 With an unbounded feasible set, the minimum cost could be $-\infty$ (*minus infinity*). Suppose the cost is $-x_1 - x_2 + x_3$. Then the vector $(100, 100, 98)$ costs $C = -102$. The vector $(1000, 1000, 998)$ costs $C = -1002$. We are being paid to include x_1 and x_2, instead of paying a cost. In realistic applications this will not happen. But it is theoretically possible that A, $\boldsymbol{b}$, and $\boldsymbol{c}$ can produce unexpected triangles and costs.

The Primal and Dual Problems

This first problem will fit $\boldsymbol{A}, \boldsymbol{b}, \boldsymbol{c}$ in that example. The unknowns x_1, x_2, x_3 represent hours of work by a Ph.D. and a student and a machine. The costs per hour are \$5, \$3, and \$8. (*I apologize for such low pay.*) The number of hours cannot be negative: $x_1 \geq 0, x_2 \geq 0, x_3 \geq 0$. The Ph.D. and the student get through one homework problem per hour. *The machine solves two problems in one hour.* In principle they can share out the homework, which has four problems to be solved: $x_1 + x_2 + 2x_3 = 4$.

The problem is to finish the four problems at minimum cost $\boldsymbol{c^{\mathrm{T}}x}$.

If all three are working, the job takes one hour: $x_1 = x_2 = x_3 = 1$. The cost is $5 + 3 + 8 = 16$. But certainly the Ph.D. should be put out of work by the student (who is just as fast and costs less—this problem is getting realistic). When the student works two hours and the machine works one, the cost is $6 + 8$ and all four problems get solved. We are on the edge $\boldsymbol{QR}$ of the triangle because the Ph.D. is not working: $x_1 = 0$. But the best point is all work by student (at $\boldsymbol{Q}$) or all work by machine (at $\boldsymbol{R}$). In this example the student solves four problems in four hours for \$12—the minimum cost.

With only one equation in $A\boldsymbol{x} = \boldsymbol{b}$, the corner $(0, 4, 0)$ has only one nonzero component. ***When*** $A\boldsymbol{x} = \boldsymbol{b}$ ***has*** m ***equations, corners have*** m ***nonzeros***. We solve $A\boldsymbol{x} = \boldsymbol{b}$ for those m variables, with $n - m$ free variables set to zero. But unlike Chapter 3, ***we don't know which*** $\boldsymbol{m}$ ***variables to choose***.

The number of possible corners is the number of ways to choose m components out of n. This number "n choose m" is heavily involved in gambling and probability. With $n = 20$ unknowns and $m = 8$ equations (still small numbers), the "feasible set" can have $20!/8!12!$ corners. That number is $(20)(19)\cdots(13) = 5{,}079{,}110{,}400$.

Checking three corners for the minimum cost was fine. Checking five billion corners is not the way to go. The simplex method described below is much faster.

The Dual Problem In linear programming, problems come in pairs. There is a minimum problem and a maximum problem—the original and its "dual." The original problem was specified by a matrix A and two vectors $\boldsymbol{b}$ and $\boldsymbol{c}$. The dual problem transposes A and switches $\boldsymbol{b}$ and $\boldsymbol{c}$: ***Maximize*** $\boldsymbol{b \cdot y}$. Here is the dual to our example:

A cheater offers to solve homework problems by selling the answers. The charge is y dollars per problem, or $4y$ altogether. (Note how $\boldsymbol{b} = 4$ has gone into the cost.) The cheater must be as cheap as the Ph.D. or student or machine: $y \leq 5$ and $y \leq 3$ and $2y \leq 8$. (Note how $\boldsymbol{c} = (5, 3, 8)$ has gone into inequality constraints). The cheater maximizes the income $4y$.

Dual Problem ***Maximize*** $\boldsymbol{b \cdot y}$ ***subject to*** $A^{\mathrm{T}}\boldsymbol{y} \leq \boldsymbol{c}$.

The maximum occurs when $y = 3$. The income is $4y = 12$. The maximum in the dual problem (\$12) equals the minimum in the original (\$12). *Max = min* is duality.

If either problem has a best vector (x^* or y^*) then so does the other.
Minimum cost $c \cdot x^*$ equals maximum income $b \cdot y^*$

This book started with a row picture and a column picture. The first "duality theorem" was about rank: The number of independent rows equals the number of independent columns. That theorem, like this one, was easy for small matrices. Minimum cost = maximum income is proved in our text *Linear Algebra and Its Applications*. One line will establish the easy half of the theorem: ***The cheater's income $b^{\mathrm{T}}y$ cannot exceed the honest cost***:

$$\textit{If } Ax = b, x \geq 0, A^{\mathrm{T}}y \leq c \textit{ then } b^{\mathrm{T}}y = (Ax)^{\mathrm{T}}y = x^{\mathrm{T}}(A^{\mathrm{T}}y) \leq x^{\mathrm{T}}c. \quad (1)$$

The full duality theorem says that when $b^{\mathrm{T}}y$ reaches its maximum and $x^{\mathrm{T}}c$ reaches its minimum, they are equal: $b \cdot y^* = c \cdot x^*$. Look at the last step in (1), with $\leq$ sign:

The dot product of $x \geq 0$ and $s = c - A^{\mathrm{T}}y \geq 0$ gave $x^{\mathrm{T}}s \geq 0$. This is $x^{\mathrm{T}}A^{\mathrm{T}}y \leq x^{\mathrm{T}}c$.

Equality needs $x^{\mathrm{T}}s = 0$ So the optimal solution has $x_j^* = 0$ or $s_j^* = 0$ for each j.

The Simplex Method

Elimination is the workhorse for linear equations. The simplex method is the workhorse for linear inequalities. We cannot give the simplex method as much space as elimination, but the idea can be clear. *The simplex method goes from one corner to a neighboring corner of lower cost.* Eventually (and quite soon in practice) it reaches the corner of minimum cost.

A ***corner*** is a vector $x \geq 0$ that satisfies the m equations $Ax = b$ with at most m positive components. *The other $n - m$ components are zero.* (Those are the free variables. Back substitution gives the m basic variables. All variables must be nonnegative or x is a false corner.) For a *neighboring corner*, one zero component of x becomes positive and one positive component becomes zero.

The simplex method must decide which component "enters" by becoming positive, and which component "leaves" by becoming zero. That exchange is chosen so as to lower the total cost. This is one step of the simplex method, moving toward x^*.

Here is the overall plan. Look at each zero component at the current corner. If it changes from 0 to 1, the other nonzeros have to adjust to keep $Ax = b$. Find the new x by back substitution and compute the change in the total cost $c \cdot x$. This change is the "reduced cost" r of the new component. The ***entering variable*** is the one that gives the ***most negative*** r. This is the greatest cost reduction for a single unit of a new variable.

Example 1 Suppose the current corner is $P = (4, 0, 0)$, with the Ph.D. doing all the work (the cost is \$20). If the student works one hour, the cost of $x = (3, 1, 0)$ is down to \$18. The reduced cost is $r = -2$. If the machine works one hour, then $x = (2, 0, 1)$ also

costs \$18. The reduced cost is also $r = -2$. In this case the simplex method can choose either the student or the machine as the entering variable.

Even in this small example, the first step may not go immediately to the best $\boldsymbol{x}^*$. The method chooses the entering variable before it knows how much of that variable to include. We computed r when the entering variable changes from 0 to 1, but one unit may be too much or too little. The method now chooses the leaving variable (the Ph.D.). It moves to corner $\boldsymbol{Q}$ or $\boldsymbol{R}$ in the figure.

The more of the entering variable we include, the lower the cost. This has to stop when one of the positive components (which are adjusting to keep $A\boldsymbol{x} = \boldsymbol{b}$) hits zero. *The* ***leaving variable*** *is the first positive x_i to reach zero.* When that happens, a neighboring corner has been found. Then start again (from the new corner) to find the next variables to enter and leave.

When all reduced costs are positive, the current corner is the optimal $\boldsymbol{x}^*$. No zero component can become positive without increasing $\boldsymbol{c} \cdot \boldsymbol{x}$. No new variable should enter. The problem is solved (and we can show that $\boldsymbol{y}^*$ is found too).

Note Generally $\boldsymbol{x}^*$ is reached in αn steps, where α is not large. But examples have been invented which use an exponential number of simplex steps. Eventually a different approach was developed, which is guaranteed to reach $\boldsymbol{x}^*$ in fewer (but more difficult) steps. The new methods travel through the *interior* of the feasible set.

Example 2 Minimize the cost $\boldsymbol{c} \cdot \boldsymbol{x} = 3x_1 + x_2 + 9x_3 + x_4$. The constraints are $\boldsymbol{x} \geq 0$ and two equations $A\boldsymbol{x} = \boldsymbol{b}$:

$$\begin{aligned} x_1 + 2x_3 + x_4 &= 4 \qquad m = 2 \quad \text{equations} \\ x_2 + x_3 - x_4 &= 2 \qquad n = 4 \quad \text{unknowns.} \end{aligned}$$

A starting corner is $\boldsymbol{x} = (4, 2, 0, 0)$ which costs $\boldsymbol{c} \cdot \boldsymbol{x} = 14$. It has $m = 2$ nonzeros and $n - m = 2$ zeros. The zeros are x_3 and x_4. The question is whether x_3 or x_4 should enter (become nonzero). Try one unit of each of them:

If $x_3 = 1$ and $x_4 = 0$,	then $\boldsymbol{x} = (2, 1, 1, 0)$ costs 16.
If $x_4 = 1$ and $x_3 = 0$,	then $\boldsymbol{x} = (3, 3, 0, 1)$ costs 13.

Compare those costs with 14. The reduced cost of x_3 is $r = 2$, positive and useless. The reduced cost of x_4 is $r = -1$, negative and helpful. *The entering variable is x_4.*

How much of x_4 can enter? One unit of x_4 made x_1 drop from 4 to 3. Four units will make x_1 drop from 4 to zero (while x_2 increases all the way to 6). *The leaving variable is x_1.* The new corner is $\boldsymbol{x} = (0, 6, 0, 4)$, which costs only $\boldsymbol{c} \cdot \boldsymbol{x} = 10$. This is the optimal $\boldsymbol{x}^*$, but to know that we have to try another simplex step from $(0, 6, 0, 4)$. Suppose x_1 or x_3 tries to enter:

Start from the corner $(0, 6, 0, 4)$	If $x_1 = 1$ and $x_3 = 0$,	then $\boldsymbol{x} = (1, 5, 0, 3)$ costs 11.
	If $x_3 = 1$ and $x_1 = 0$,	then $\boldsymbol{x} = (0, 3, 1, 2)$ costs 14.

Those costs are higher than 10. Both r's are positive—it does not pay to move. The current corner $(0, 6, 0, 4)$ is the solution $\boldsymbol{x}^*$.

These calculations can be streamlined. Each simplex step solves three linear systems with the same matrix B. (This is the m by m matrix that keeps the m basic columns of A.) When a column enters and an old column leaves, there is a quick way to update B^{-1}. That is how most codes organize the simplex method.

Our text on *Computational Science and Engineering* includes a short code with comments. (The code is also on **math.mit.edu/cse**) The best $\boldsymbol{y}^*$ solves m equations $A^{\mathrm{T}}\boldsymbol{y}^* = \boldsymbol{c}$ in the m components that are nonzero in $\boldsymbol{x}^*$. Then we have optimality $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{s} = 0$ and this is duality: *Either* $x_j^* = 0$ *or the "slack" in* $\boldsymbol{s}^* = \boldsymbol{c} - A^{\mathrm{T}}y^*$ *has* $s_j^* = 0$.

When $\boldsymbol{x}^* = (0, 4, 0)$ was the optimal corner $\boldsymbol{Q}$, the cheater's price was set by $y^* = 3$.

Interior Point Methods

The simplex method moves along the edges of the feasible set, eventually reaching the optimal corner $\boldsymbol{x}^*$. ***Interior point methods move inside the feasible set*** (where $\boldsymbol{x} > \boldsymbol{0}$). These methods hope to go more directly to $\boldsymbol{x}^*$. They work well.

One way to stay inside is to put a barrier at the boundary. Add extra cost as a *logarithm that blows up* when any variable x_j touches zero. The best vector has $\boldsymbol{x} > \boldsymbol{0}$. The number θ is a small parameter that we move toward zero.

Barrier problem **Minimize** $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} - \theta\,(\log x_1 + \cdots + \log x_n)$ **with** $A\boldsymbol{x} = \boldsymbol{b}$ (2)

This cost is nonlinear (but linear programming is already nonlinear from inequalities). The constraints $x_j \geq 0$ are not needed because $\log x_j$ becomes infinite at $x_j = 0$.

The barrier gives an *approximate problem* for each θ. The m constraints $A\boldsymbol{x} = \boldsymbol{b}$ have Lagrange multipliers $y_1, \ldots, y_m$. This is the good way to deal with constraints.

$\boldsymbol{y}$ from Lagrange $$L(x, y, \theta) = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} - \theta\,(\textstyle\sum \log x_i) - \boldsymbol{y}^{\mathrm{T}}(A\boldsymbol{x} - \boldsymbol{b}) \tag{3}$$

$\partial L/\partial y = 0$ brings back $A\boldsymbol{x} = \boldsymbol{b}$. The derivatives $\partial L/\partial x_j$ are interesting !

Optimality in barrier pbm $\dfrac{\partial L}{\partial x_j} = c_j - \dfrac{\theta}{x_j} - (A^{\mathrm{T}}y)_j = 0$ which is $\boldsymbol{x_j s_j = \theta}$. (4)

The true problem has $x_j s_j = 0$. The barrier problem has $x_j s_j = \theta$. The solutions $\boldsymbol{x}^*(\theta)$ lie on the ***central path*** to $\boldsymbol{x}^*(0)$. Those n optimality equations $x_j s_j = \theta$ are nonlinear, and we solve them iteratively by Newton's method.

The current $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{s}$ will satisfy $A\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}$ and $A^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{s} = \boldsymbol{c}$, *but not* $x_j s_j = \theta$. Newton's method takes a step $\Delta\boldsymbol{x}, \Delta\boldsymbol{y}, \Delta\boldsymbol{s}$. By ignoring the second-order term $\Delta\boldsymbol{x}\Delta\boldsymbol{s}$

in $(\boldsymbol{x}+\Delta\boldsymbol{x})(\boldsymbol{s}+\Delta\boldsymbol{s})=\boldsymbol{\theta}$, the corrections in $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{s}$ come from linear equations:

Newton step

$$\begin{aligned} A\,\Delta\boldsymbol{x} &= 0 \\ A^{\mathrm{T}}\Delta\boldsymbol{y}+\Delta\boldsymbol{s} &= 0 \\ s_j\Delta x_j + x_j\Delta s_j &= \theta - x_j s_j \end{aligned} \tag{5}$$

Newton iteration has quadratic convergence for each θ, and then θ approaches zero. The duality gap $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{s}$ generally goes below 10^{-8} after 20 to 60 steps. The explanation in my *Computational Science and Engineering* textbook takes one Newton step in detail, for the example with four homework problems. I didn't intend that the student should end up doing all the work, but $\boldsymbol{x}^*$ turned out that way.

This interior point method is used almost "as is" in commercial software, for a large class of linear and nonlinear optimization problems.

Problem Set 10.4

1 Draw the region in the xy plane where $x+2y=6$ and $x\ge 0$ and $y\ge 0$. Which point in this "feasible set" minimizes the cost $c=x+3y$? Which point gives maximum cost? Those points are at corners.

2 Draw the region in the xy plane where $x+2y\le 6,\ 2x+y\le 6,\ x\ge 0,\ y\ge 0$. It has four corners. Which corner minimizes the cost $c=2x-y$?

3 What are the corners of the set $x_1+2x_2-x_3=4$ with x_1, x_2, x_3 all ≥ 0? Show that the cost x_1+2x_3 can be very negative in this feasible set. This is an example of unbounded cost: no minimum.

4 Start at $\boldsymbol{x}=(0,0,2)$ where the machine solves all four problems for \$16. Move to $\boldsymbol{x}=(0,1,\ \ )$ to find the reduced cost r (the savings per hour) for work by the student. Find r for the Ph.D. by moving to $\boldsymbol{x}=(1,0,\ \ )$ with 1 hour of Ph.D. work.

5 Start Example 1 from the Ph.D. corner $(4,0,0)$ with $\boldsymbol{c}$ changed to $[5\ \ 3\ \ 7]$. Show that r is better for the machine even when the total cost is lower for the student. The simplex method takes two steps, first to the machine and then to the student for $\boldsymbol{x}^*$.

6 Choose a different cost vector $\boldsymbol{c}$ so the Ph.D. gets the job. Rewrite the dual problem (maximum income to the cheater).

7 A six-problem homework on which the Ph.D. is fastest gives a second constraint $2x_1+x_2+x_3=6$. Then $\boldsymbol{x}=(2,2,0)$ shows two hours of work by Ph.D. and student on each homework. Does this $\boldsymbol{x}$ minimize the cost $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}$ with $\boldsymbol{c}=(5,3,8)$?

8 These two problems are also dual. Prove weak duality, that always $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{b}\le\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}$:

Primal problem Minimize $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}$ with $A\boldsymbol{x}\ge\boldsymbol{b}$ and $\boldsymbol{x}\ge\boldsymbol{0}$.

Dual problem Maximize $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{b}$ with $A^{\mathrm{T}}\boldsymbol{y}\le\boldsymbol{c}$ and $\boldsymbol{y}\ge\boldsymbol{0}$.

10.5 Fourier Series: Linear Algebra for Functions

This section goes from finite dimensions to *infinite* dimensions. I want to explain linear algebra in infinite-dimensional space, and to show that it still works. First step: look back. This book began with vectors and dot products and linear combinations. We begin by converting those basic ideas to the infinite case—then the rest will follow.

What does it mean for a vector to have infinitely many components? There are two different answers, both good:

1. The vector is infinitely long: $\boldsymbol{v} = (v_1, v_2, v_3, \ldots)$. It could be $(1, \frac{1}{2}, \frac{1}{4}, \ldots)$.
2. The vector is a function $f(x)$. It could be $\boldsymbol{v} = \sin x$.

We will go both ways. Then the idea of a Fourier series will connect them.

After vectors come *dot products*. The natural dot product of two infinite vectors $(v_1, v_2, \ldots)$ and $(w_1, w_2, \ldots)$ is an infinite series:

$$\textbf{Dot product} \qquad \boldsymbol{v} \cdot \boldsymbol{w} = v_1 w_1 + v_2 w_2 + \cdots . \tag{1}$$

This brings a new question, which never occurred to us for vectors in $\mathbf{R}^n$. Does this infinite sum add up to a finite number? Does the series converge? Here is the first and biggest difference between finite and infinite.

When $\boldsymbol{v} = \boldsymbol{w} = (1, 1, 1, \ldots)$, the sum certainly does not converge. In that case $\boldsymbol{v} \cdot \boldsymbol{w} = 1 + 1 + 1 + \cdots$ is infinite. Since $\boldsymbol{v}$ equals $\boldsymbol{w}$, we are really computing $\boldsymbol{v} \cdot \boldsymbol{v} = \|\boldsymbol{v}\|^2$, the length squared. The vector $(1, 1, 1, \ldots)$ has infinite length. *We don't want that vector.* Since we are making the rules, we don't have to include it. The only vectors to be allowed are those with finite length:

DEFINITION The vector $\boldsymbol{v} = (v_1, v_2, \ldots)$ and the function $f(x)$ are in our infinite-dimensional "***Hilbert spaces***" if and only if their lengths $\|\boldsymbol{v}\|$ and $\|f\|$ are finite:

$$\|\boldsymbol{v}\|^2 = \boldsymbol{v} \cdot \boldsymbol{v} = v_1^2 + v_2^2 + v_3^2 + \cdots \quad \text{must add to a finite number.}$$

$$\|f\|^2 = (f, f) = \int_0^{2\pi} |f(x)|^2 \, dx \quad \text{must be a finite integral.}$$

Example 1 The vector $\boldsymbol{v} = (1, \frac{1}{2}, \frac{1}{4}, \ldots)$ is included in Hilbert space, because its length is $2/\sqrt{3}$. We have a geometric series that adds to $4/3$. The length of $\boldsymbol{v}$ is the square root:

$$\textbf{Length squared} \qquad \boldsymbol{v} \cdot \boldsymbol{v} = 1 + \tfrac{1}{4} + \tfrac{1}{16} + \cdots = \frac{1}{1 - \frac{1}{4}} = \tfrac{4}{3}.$$

Question If $\boldsymbol{v}$ and $\boldsymbol{w}$ have finite length, how large can their dot product be?

Answer The sum $\boldsymbol{v} \cdot \boldsymbol{w} = v_1 w_1 + v_2 w_2 + \cdots$ also adds to a finite number. We can safely take dot products. The Schwarz inequality is still true:

$$\textbf{Schwarz inequality} \qquad |\boldsymbol{v} \cdot \boldsymbol{w}| \le \|\boldsymbol{v}\| \, \|\boldsymbol{w}\|. \tag{2}$$

The ratio of $\boldsymbol{v} \cdot \boldsymbol{w}$ to $\|\boldsymbol{v}\| \, \|\boldsymbol{w}\|$ is still the cosine of θ (the angle between $\boldsymbol{v}$ and $\boldsymbol{w}$). Even in infinite-dimensional space, $|\cos\theta|$ is not greater than 1.

Now change over to functions. Those are the "vectors." The space of functions $f(x)$, $g(x)$, $h(x)$, . . . defined for $0 \le x \le 2\pi$ must be somehow bigger than $\mathbf{R}^n$. ***What is the dot product of $f(x)$ and $g(x)$? What is the length of $f(x)$?***

Key point in the continuous case: *Sums are replaced by integrals.* Instead of a sum of v_j times w_j, the dot product is an integral of $f(x)$ times $g(x)$. Change the "dot" to parentheses with a comma, and change the words "dot product" to *inner product*:

DEFINITION The ***inner product*** of $f(x)$ and $g(x)$, and the ***length squared*** of $f(x)$, are

$$(f,g) = \int_0^{2\pi} f(x)g(x)\,dx \quad \text{and} \quad \|f\|^2 = \int_0^{2\pi} \big(f(x)\big)^2\,dx. \tag{3}$$

The interval $[0, 2\pi]$ where the functions are defined could change to a different interval like $[0, 1]$ or $(-\infty, \infty)$. We chose 2π because our first examples are $\sin x$ and $\cos x$.

Example 2 The length of $f(x) = \sin x$ comes from its inner product with itself:

$$(f,f) = \int_0^{2\pi} (\sin\ x)^2\,dx = \pi. \quad \text{The length of } \sin x \text{ is } \sqrt{\pi}.$$

That is a standard integral in calculus—not part of linear algebra. By writing $\sin^2 x$ as $\frac{1}{2} - \frac{1}{2}\cos 2x$, we see it go above and below its average value $\frac{1}{2}$. Multiply that average by the interval length 2π to get the answer π.

More important: $\sin x$ ***and*** $\cos x$ ***are orthogonal in function space***: $(f, g) = 0$

Inner product is zero
$$\int_0^{2\pi} \sin x \cos x\,dx = \int_0^{2\pi} \tfrac{1}{2}\sin 2x\,dx = \big[-\tfrac{1}{4}\cos 2x\big]_0^{2\pi} = 0. \tag{4}$$

This zero is no accident. It is highly important to science. The orthogonality goes beyond the two functions $\sin x$ and $\cos x$, to an infinite list of sines and cosines. The list contains $\cos 0x$ (which is 1), $\sin x, \cos x, \sin 2x, \cos 2x, \sin 3x, \cos 3x, \ldots$.

Every function in that list is orthogonal to every other function in the list.

Fourier Series

The Fourier series of a function $f(x)$ is its expansion into sines and cosines:

$$f(x) = a_0 + a_1\cos x + b_1 \sin x + a_2 \cos 2x + b_2 \sin 2x + \cdots. \tag{5}$$

We have an orthogonal basis! The vectors in "function space" are combinations of the sines and cosines. On the interval from $x = 2\pi$ to $x = 4\pi$, all our functions repeat what they did from 0 to 2π. They are "*periodic*." The distance between repetitions is the period 2π.

Remember: The list is infinite. The Fourier series is an infinite series. We avoided the vector $\boldsymbol{v} = (1, 1, 1, \ldots)$ because its length is infinite, now we avoid a function like $\frac{1}{2} + \cos x + \cos 2x + \cos 3x + \cdots$. (*Note*: This is π times the famous **delta function** $\delta(x)$. It is an infinite "spike" above a single point. At $x = 0$ its height $\frac{1}{2} + 1 + 1 + \cdots$ is infinite. At all points inside $0 < x < 2\pi$ the series adds in some average way to zero.) The integral of $\delta(x)$ is 1. But $\int \delta^2(x) = \infty$, so delta functions are not allowed into Hilbert space.

Compute the length of a typical sum $f(x)$:

$$
\begin{aligned}
(f, f) &= \int_0^{2\pi} (a_0 + a_1 \cos x + b_1 \sin x + a_2 \cos 2x + \cdots)^2 \, dx \\
&= \int_0^{2\pi} (a_0^2 + a_1^2 \cos^2 x + b_1^2 \sin^2 x + a_2^2 \cos^2 2x + \cdots) \, dx \\
\|\boldsymbol{f}\|^2 &= \boldsymbol{2\pi a_0^2 + \pi(a_1^2 + b_1^2 + a_2^2 + \cdots).} \qquad (6)
\end{aligned}
$$

The step from line 1 to line 2 used orthogonality. All products like $\cos x \cos 2x$ integrate to give zero. Line 2 contains what is left—the integrals of each sine and cosine squared. Line 3 evaluates those integrals. (The integral of 1^2 is 2π, when all other integrals give π.) If we divide by their lengths, our functions become *orthonormal*:

$$
\frac{1}{\sqrt{2\pi}}, \frac{\cos x}{\sqrt{\pi}}, \frac{\sin x}{\sqrt{\pi}}, \frac{\cos 2x}{\sqrt{\pi}}, \ldots \textbf{ is an orthonormal basis for our function space.}
$$

These are unit vectors. We could combine them with coefficients $A_0, A_1, B_1, A_2, \ldots$ to yield a function $F(x)$. Then the 2π and the π's drop out of the formula for length.

Function length = vector length $\qquad \|F\|^2 = (F, F) = A_0^2 + A_1^2 + B_1^2 + A_2^2 + \cdots$. (7)

Here is the important point, for $f(x)$ as well as $F(x)$. *The function has finite length exactly when the vector of coefficients has finite length.* Fourier series gives us a perfect match between the Hilbert spaces for functions and for vectors. The function is in L^2, its Fourier coefficients are in ℓ^2.

The function space contains $f(x)$ exactly when the Hilbert space contains the vector $\boldsymbol{v} = (a_0, a_1, b_1, \ldots)$ of Fourier coefficients of $f(x)$. Both must have finite length.

Example 3 Suppose $f(x)$ is a "square wave," equal to 1 for $0 \le x < \pi$. Then $f(x)$ drops to -1 for $\pi \le x < 2\pi$. The $+1$ and -1 repeat forever. This $f(x)$ is an odd function like the sines, and all its cosine coefficients are zero. We will find its Fourier series, containing only sines:

$$
\textbf{Square wave} \qquad f(x) = \frac{4}{\pi}\left[\frac{\sin x}{1} + \frac{\sin 3x}{3} + \frac{\sin 5x}{5} + \cdots\right]. \qquad (8)
$$

The length of this function is $\sqrt{2\pi}$, because at every point $\big(f(x)\big)^2$ is $(-1)^2$ or $(+1)^2$:

$$
\|f\|^2 = \int_0^{2\pi} \big(f(x)\big)^2 \, dx = \int_0^{2\pi} 1 \, dx = 2\pi.
$$

At $x = 0$ the sines are zero and the Fourier series gives zero. This is half way up the jump from -1 to $+1$. The Fourier series is also interesting when $x = \frac{\pi}{2}$. At this point the square wave equals 1, and the sines in (8) alternate between $+1$ and -1:

$$\textbf{Formula for } \pi \qquad 1 = \frac{4}{\pi}\left(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots\right). \tag{9}$$

Multiply by π to find a magical formula $4(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots)$ for that famous number.

The Fourier Coefficients

How do we find the a's and b's which multiply the cosines and sines? For a given function $f(x)$, we are asking for its Fourier coefficients a_k and b_k:

$$\textbf{Fourier series} \qquad f(x) = a_0 + a_1 \cos x + b_1 \sin x + a_2 \cos 2x + \cdots .$$

Here is the way to find a_1. *Multiply both sides by* $\cos x$. *Then integrate from* 0 *to* 2π. The key is orthogonality! All integrals on the right side are zero, except for $\cos^2 x$:

$$\textbf{For coefficient } a_1 \qquad \int_0^{2\pi} f(x)\cos x\, dx = \int_0^{2\pi} a_1 \cos^2 x\, dx = \pi a_1. \tag{10}$$

Divide by π and you have a_1. To find any other a_k, multiply the Fourier series by $\cos kx$. Integrate from 0 to 2π. Use orthogonality, so only the integral of $a_k \cos^2 kx$ is left. That integral is πa_k, and divide by π:

$$\boldsymbol{a_k} = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos kx\, dx \qquad \text{and similarly} \qquad \boldsymbol{b_k} = \frac{1}{\pi}\int_0^{2\pi} f(x)\sin kx\, dx. \tag{11}$$

The exception is a_0. This time we multiply by $\cos 0x = 1$. The integral of 1 is 2π:

$$\textbf{Constant term} \qquad a_0 = \frac{1}{2\pi}\int_0^{2\pi} f(x)\cdot 1\, dx = \textbf{\textit{average value}} \text{ of } f(x). \tag{12}$$

I used those formulas to find the Fourier coefficients for the square wave in equation (8). The integral of $f(x)\cos kx$ was zero. The integral of $f(x)\sin kx$ was $4/k$ for odd k.

Compare Linear Algebra in $\mathbf{R}^n$

Infinite-dimensional Hilbert space is very much like the n-dimensional space $\mathbf{R}^n$. Suppose the nonzero vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ are orthogonal in $\mathbf{R}^n$. We want to write the vector $\boldsymbol{b}$ (instead of the function $f(x)$) as a combination of those $\boldsymbol{v}$'s:

$$\textbf{Finite orthogonal series} \qquad \boldsymbol{b} = c_1\boldsymbol{v}_1 + c_2\boldsymbol{v}_2 + \cdots + c_n\boldsymbol{v}_n. \tag{13}$$

Multiply both sides by $\boldsymbol{v}_1^{\mathrm{T}}$. Use orthogonality, so $\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{v}_2 = 0$. Only the c_1 term is left:

$$\textbf{Coefficient } c_1 \qquad \boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{b} = c_1\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{v}_1 + 0 + \cdots + 0. \quad \text{Therefore } c_1 = \boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{b}/\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{v}_1. \tag{14}$$

The denominator $\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{v}_1$ is the length squared, like π in equation (11). The numerator $\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{b}$ is the inner product like $\int f(x)\cos kx\, dx$. ***Coefficients are easy to find when the***

basis vectors are orthogonal. We are just doing one-dimensional projections, to find the components along each basis vector.

The formulas are even better when the vectors are orthonormal. Then we have unit vectors in Q. The denominators $\boldsymbol{v}_k^{\mathrm{T}}\boldsymbol{v}_k$ are all 1. You know $c_k = \boldsymbol{v}_k^{\mathrm{T}}\boldsymbol{b}$ in another form:

$$\textbf{Equation for } \boldsymbol{c}\textbf{'s} \quad c_1\boldsymbol{v}_1 + \cdots + c_n\boldsymbol{v}_n = \boldsymbol{b} \quad \text{or} \quad \begin{bmatrix} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_n \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = \boldsymbol{b}.$$

$$Q\boldsymbol{c} = \boldsymbol{b} \quad \text{yields} \quad \boldsymbol{c} = Q^{\mathrm{T}}\boldsymbol{b}. \quad \text{Row by row this is } c_k = \boldsymbol{q}_k^{\mathrm{T}}\boldsymbol{b}.$$

Fourier series is like having a matrix with infinitely many orthogonal columns. Those columns are the basis functions $1, \cos x, \sin x, \ldots$. After dividing by their lengths we have an "infinite orthogonal matrix." Its inverse is its transpose, Q^{T}. Orthogonality is what reduces a series of terms to one single term, when we integrate.

Problem Set 10.5

1 Integrate the trig identity $2\cos jx \cos kx = \cos(j+k)x + \cos(j-k)x$ to show that $\cos jx$ is orthogonal to $\cos kx$, provided $j \neq k$. What is the result when $j = k$?

2 Show that $1, x$, and $x^2 - \frac{1}{3}$ are orthogonal, when the integration is from $x = -1$ to $x = 1$. Write $f(x) = 2x^2$ as a combination of those orthogonal functions.

3 Find a vector $(w_1, w_2, w_3, \ldots)$ that is orthogonal to $\boldsymbol{v} = (1, \frac{1}{2}, \frac{1}{4}, \ldots)$. Compute its length $\|\boldsymbol{w}\|$.

4 The first three *Legendre polynomials* are $1, x$, and $x^2 - \frac{1}{3}$. Choose c so that the fourth polynomial $x^3 - cx$ is orthogonal to the first three. All integrals go from -1 to 1.

5 For the square wave $f(x)$ in Example 3 jumping from 1 to -1, show that

$$\int_0^{2\pi} f(x)\cos x\,dx = 0 \qquad \int_0^{2\pi} f(x)\sin x\,dx = 4 \qquad \int_0^{2\pi} f(x)\sin 2x\,dx = 0.$$

Which three Fourier coefficients come from those integrals?

6 The square wave has $\|f\|^2 = 2\pi$. Then (6) gives what remarkable sum for π^2?

7 Graph the square wave. Then graph by hand the sum of two sine terms in its series, or graph by machine the sum of 2, 3, and 10 terms. The famous ***Gibbs phenomenon*** is the oscillation that overshoots the jump (this doesn't die down with more terms).

8 Find the lengths of these vectors in Hilbert space:

(a) $\boldsymbol{v} = \left(\frac{1}{\sqrt{1}}, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{4}}, \frac{1}{\sqrt{8}}, \ldots\right)$

(b) $v = (1, a, a^2, \ldots)$

(c) $f(x) = 1 + \sin x$.

9 Compute the Fourier coefficients a_k and b_k for $f(x)$ defined from 0 to 2π:

(a) $f(x) = 1$ for $0 \le x \le \pi$, $f(x) = 0$ for $\pi < x < 2\pi$

(b) $f(x) = x$.

10 When $f(x)$ has period 2π, why is its integral from $-\pi$ to π the same as from 0 to 2π? If $f(x)$ is an *odd* function, $f(-x) = -f(x)$, show that $\int_{-\pi}^{\pi} f(x)\,dx$ is zero. Odd functions only have sine terms, even functions only have cosines.

11 Using trigonometric identities find the two terms in the Fourier series for $f(x)$:

(a) $f(x) = \cos^2 x$ (b) $f(x) = \cos\left(x + \frac{\pi}{3}\right)$ (c) $f(x) = \sin^3 x$

12 The functions $1, \cos x, \sin x, \cos 2x, \sin 2x, \ldots$ are a basis for Hilbert space. Write the derivatives of those first five functions as combinations of the same five functions. What is the 5 by 5 "differentiation matrix" for these functions?

13 Find the Fourier coefficients a_k and b_k of the square pulse $F(x)$ centered at $x = 0$: $F(x) = 1/h$ for $|x| \le h/2$ and $F(x) = 0$ for $h/2 < |x| \le \pi$.

As $h \to 0$, this $F(x)$ approaches a delta function. Find the limits of a_k and b_k.

Section 4.1 of *Computational Science and Engineering* explains the sine series, cosine series, complete series, and complex series $\Sigma c_k e^{ikx}$ on **math.mit.edu/cse**.

Section 9.3 of this book explains the *Discrete Fourier Transform*. This is "Fourier series for vectors" and it is computed by the **Fast Fourier Transform**. That fast algorithm comes quickly from special complex numbers $z = e^{i\theta} = \cos\theta + i\sin\theta$ when the angle is $\theta = 2\pi k/n$.

10.6 Computer Graphics

Computer graphics deals with images. The images are moved around. Their scale is changed. Three dimensions are projected onto two dimensions. All the main operations are done by matrices—but the shape of these matrices is surprising.

The transformations of three-dimensional space are done with 4 *by* 4 *matrices.* You would expect 3 by 3. The reason for the change is that one of the four key operations cannot be done with a 3 by 3 matrix multiplication. Here are the four operations:

Translation	**(shift the origin to another point** $P_0 = (x_0, y_0, z_0)$**)**
Rescaling	**(by** c **in all directions or by different factors** c_1, c_2, c_3**)**
Rotation	**(around an axis through the origin or an axis through** P_0**)**
Projection	**(onto a plane through the origin or a plane through** P_0**).**

Translation is the easiest—just add (x_0, y_0, z_0) to every point. But this is not linear! No 3 by 3 matrix can move the origin. So we change the coordinates of the origin to $(0, 0, 0, 1)$. This is why the matrices are 4 by 4. The "*homogeneous coordinates*" of the point (x, y, z) are $(x, y, z, 1)$ and we now show how they work.

1. Translation Shift the whole three-dimensional space along the vector $\boldsymbol{v}_0$. The origin moves to (x_0, y_0, z_0). This vector $\boldsymbol{v}_0$ is added to every point $\boldsymbol{v}$ in $\mathbf{R}^3$. Using homogeneous coordinates, the 4 by 4 matrix T shifts the whole space by $\boldsymbol{v}_0$:

Translation matrix $$T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ x_0 & y_0 & z_0 & 1 \end{bmatrix}.$$

Important: *Computer graphics works with row vectors.* We have row times matrix instead of matrix times column. You can quickly check that $[0\ 0\ 0\ 1]\,T = [x_0\ y_0\ z_0\ 1]$.

To move the points $(0, 0, 0)$ and (x, y, z) by $\boldsymbol{v}_0$, change to homogeneous coordinates $(0, 0, 0, 1)$ and $(x, y, z, 1)$. Then multiply by T. A row vector times T gives a row vector. ***Every*** $\boldsymbol{v}$ ***moves to*** $\boldsymbol{v} + \boldsymbol{v}_0$: $[x\ y\ z\ 1]\,T = [x + x_0\ y + y_0\ z + z_0\ 1]$.

The output tells where any $\boldsymbol{v}$ will move. (It goes to $\boldsymbol{v} + \boldsymbol{v}_0$.) Translation is now achieved by a matrix, which was impossible in $\mathbf{R}^3$.

2. Scaling To make a picture fit a page, we change its width and height. A copier will rescale a figure by 90%. In linear algebra, we multiply by .9 times the identity matrix. That matrix is normally 2 by 2 for a plane and 3 by 3 for a solid. In computer graphics, with homogeneous coordinates, the matrix is *one size larger:*

Rescale the plane: $$S = \begin{bmatrix} .9 & & \\ & .9 & \\ & & 1 \end{bmatrix}$$

Rescale a solid: $$S = \begin{bmatrix} c & 0 & 0 & 0 \\ 0 & c & 0 & 0 \\ 0 & 0 & c & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Important: S *is not* cI. We keep the "1" in the lower corner. Then $[x, y, 1]$ times S is the correct answer in homogeneous coordinates. The origin stays in its normal position because $[0\ 0\ 1]S = [0\ 0\ 1]$.

If we change that 1 to c, the result is strange. ***The point*** (cx, cy, cz, c) ***is the same as*** $(x, y, z, 1)$. The special property of homogeneous coordinates is that *multiplying by* cI *does not move the point.* The origin in $\mathbf{R}^3$ has homogeneous coordinates $(0, 0, 0, 1)$ and $(0, 0, 0, c)$ for every nonzero c. This is the idea behind the word "homogeneous."

Scaling can be different in different directions. To fit a full-page picture onto a half-page, scale the y direction by $\frac{1}{2}$. To create a margin, scale the x direction by $\frac{3}{4}$. The graphics matrix is diagonal but not 2 by 2. It is 3 by 3 to rescale a plane and 4 by 4 to rescale a space:

Scaling matrices $S = \begin{bmatrix} \frac{3}{4} & & \\ & \frac{1}{2} & \\ & & 1 \end{bmatrix}$ and $S = \begin{bmatrix} c_1 & & & \\ & c_2 & & \\ & & c_3 & \\ & & & 1 \end{bmatrix}$.

That last matrix S rescales the x, y, z directions by positive numbers c_1, c_2, c_3. The extra column in all these matrices leaves the extra 1 at the end of every vector.

Summary The scaling matrix S is the same size as the translation matrix T. They can be multiplied. To translate and then rescale, multiply $\boldsymbol{v}TS$. To rescale and then translate, multiply $\boldsymbol{v}ST$. Are those different? *Yes.*

The point (x, y, z) in $\mathbf{R}^3$ has homogeneous coordinates $(x, y, z, 1)$ in $\mathbf{P}^3$. This "projective space" is not the same as $\mathbf{R}^4$. It is still three-dimensional. To achieve such a thing, (cx, cy, cz, c) is the same point as $(x, y, z, 1)$. Those points of projective space $\mathbf{P}^3$ are really lines through the origin in $\mathbf{R}^4$.

Computer graphics uses ***affine*** transformations, *linear plus shift.* An affine transformation T is executed on $\mathbf{P}^3$ by a 4 by 4 matrix with a special fourth column:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & 0 \\ a_{21} & a_{22} & a_{23} & 0 \\ a_{31} & a_{32} & a_{33} & 0 \\ a_{41} & a_{42} & a_{43} & 1 \end{bmatrix} = \begin{bmatrix} T(1,0,0) & 0 \\ T(0,1,0) & 0 \\ T(0,0,1) & 0 \\ T(0,0,0) & 1 \end{bmatrix}.$$

The usual 3 by 3 matrix tells us three outputs, this tells four. The usual outputs come from the inputs $(1, 0, 0)$ and $(0, 1, 0)$ and $(0, 0, 1)$. When the transformation is linear, three outputs reveal everything. When the transformation is affine, the matrix also contains the output from $(0, 0, 0)$. Then we know the shift.

3. Rotation A rotation in $\mathbf{R}^2$ or $\mathbf{R}^3$ is achieved by an orthogonal matrix Q. The determinant is $+1$. (With determinant -1 we get an extra reflection through a mirror.) Include the extra column when you use homogeneous coordinates!

Plane rotation $Q = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ becomes $R = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

This matrix rotates the plane around the origin. ***How would we rotate around a different point*** $(4,5)$? The answer brings out the beauty of homogeneous coordinates. ***Translate*** $(4,5)$ ***to*** $(0,0)$***, then rotate by*** θ***, then translate*** $(0,0)$ ***back to*** $(4,5)$:

$$\boldsymbol{v}\,T_-RT_+ = \begin{bmatrix} x & y & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -4 & -5 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 4 & 5 & 1 \end{bmatrix}.$$

I won't multiply. The point is to apply the matrices one at a time: $\boldsymbol{v}$ translates to $\boldsymbol{v}T_-$, then rotates to $\boldsymbol{v}T_-R$, and translates back to $\boldsymbol{v}T_-RT_+$. Because each point $\begin{bmatrix} x & y & 1 \end{bmatrix}$ is a row vector, T_- acts first. The center of rotation $(4,5)$—otherwise known as $(4,5,1)$—moves first to $(0,0,1)$. Rotation doesn't change it. Then T_+ moves it back to $(4,5,1)$. All as it should be. The point $(4,6,1)$ moves to $(0,1,1)$, then turns by θ and moves back.

In three dimensions, every rotation Q turns around an axis. The axis doesn't move—it is a line of eigenvectors with $\lambda = 1$. Suppose the axis is in the z direction. The 1 in Q is to leave the z axis alone, the extra 1 in R is to leave the origin alone:

$$Q = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad R = \begin{bmatrix} & Q & & 0 \\ & & & 0 \\ & & & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Now suppose the rotation is around the unit vector $\boldsymbol{a} = (a_1, a_2, a_3)$. With this axis $\boldsymbol{a}$, the rotation matrix Q which fits into R has three parts:

$$Q = (\cos\theta)I + (1-\cos\theta)\begin{bmatrix} a_1^2 & a_1a_2 & a_1a_3 \\ a_1a_2 & a_2^2 & a_2a_3 \\ a_1a_3 & a_2a_3 & a_3^2 \end{bmatrix} - \sin\theta \begin{bmatrix} 0 & a_3 & -a_2 \\ -a_3 & 0 & a_1 \\ a_2 & -a_1 & 0 \end{bmatrix}. \quad (1)$$

The axis doesn't move because $\boldsymbol{a}Q = \boldsymbol{a}$. When $\boldsymbol{a} = (0,0,1)$ is in the z direction, this Q becomes the previous Q—for rotation around the z axis.

The linear transformation Q always goes in the upper left block of R. Below it we see zeros, because rotation leaves the origin in place. When those are not zeros, the transformation is affine and the origin moves.

4. Projection In a linear algebra course, most planes go through the origin. In real life, most don't. A plane through the origin is a vector space. The other planes are affine spaces, sometimes called "flats." An affine space is what comes from translating a vector space.

We want to project three-dimensional vectors onto planes. Start with a plane through the origin, whose unit normal vector is $\boldsymbol{n}$. (We will keep $\boldsymbol{n}$ as a column vector.) The vectors in the plane satisfy $\boldsymbol{n}^{\mathrm{T}}\boldsymbol{v} = 0$. ***The usual projection onto the plane is the matrix*** $\boldsymbol{I - nn^{\mathrm{T}}}$**.** To project a vector, multiply by this matrix. The vector $\boldsymbol{n}$ is projected to zero, and the in-plane vectors $\boldsymbol{v}$ are projected onto themselves:

$$(I - nn^{\mathrm{T}})n = n - n(n^{\mathrm{T}}n) = 0 \quad \text{and} \quad (I - nn^{\mathrm{T}})v = v - n(n^{\mathrm{T}}v) = v.$$

In homogeneous coordinates the projection matrix becomes 4 by 4 (but the origin doesn't move):

Projection onto the plane $\boldsymbol{n}^{\mathrm{T}}\boldsymbol{v} = 0$ $\qquad P = \begin{bmatrix} & & & 0 \\ & I - \boldsymbol{n}\boldsymbol{n}^{\mathrm{T}} & & 0 \\ & & & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$

Now project onto a plane $\boldsymbol{n}^{\mathrm{T}}(\boldsymbol{v} - \boldsymbol{v}_0) = 0$ that does *not* go through the origin. One point on the plane is $\boldsymbol{v}_0$. This is an affine space (or a ***flat***). It is like the solutions to $A\boldsymbol{v} = \boldsymbol{b}$ when the right side is not zero. One particular solution $\boldsymbol{v}_0$ is added to the nullspace—to produce a flat.

The projection onto the flat has three steps. Translate $\boldsymbol{v}_0$ to the origin by T_-. Project along the $\boldsymbol{n}$ direction, and translate back along the row vector $\boldsymbol{v}_0$:

Projection onto a flat $\qquad T_- P T_+ = \begin{bmatrix} I & 0 \\ -\boldsymbol{v}_0 & 1 \end{bmatrix} \begin{bmatrix} I - \boldsymbol{n}\boldsymbol{n}^{\mathrm{T}} & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} I & 0 \\ \boldsymbol{v}_0 & 1 \end{bmatrix}.$

I can't help noticing that T_- and T_+ are inverse matrices: translate and translate back. They are like the elementary matrices of Chapter 2.

The exercises will include reflection matrices, also known as *mirror matrices*. These are the fifth type needed in computer graphics. A reflection moves each point twice as far as a projection—***the reflection goes through the plane and out the other side***. So change the projection $I - \boldsymbol{n}\boldsymbol{n}^{\mathrm{T}}$ to $I - 2\boldsymbol{n}\boldsymbol{n}^{\mathrm{T}}$ for a mirror matrix.

The matrix P gave a "*parallel*" projection. All points move parallel to $\boldsymbol{n}$, until they reach the plane. The other choice in computer graphics is a "*perspective*" projection. This is more popular because it includes foreshortening. With perspective, an object looks larger as it moves closer. Instead of staying parallel to $\boldsymbol{n}$ (and parallel to each other), the lines of projection come *toward the eye*—the center of projection. This is how we perceive depth in a two-dimensional photograph.

The basic problem of computer graphics starts with a scene and a viewing position. Ideally, the image on the screen is what the viewer would see. The simplest image assigns just one bit to every small picture element—called a ***pixel***. It is light or dark. This gives a black and white picture with no shading. You would not approve. In practice, we assign shading levels between 0 and 2^8 for three colors like red, green, and blue. That means $8 \times 3 = 24$ bits for each pixel. Multiply by the number of pixels, and a lot of memory is needed!

Physically, a *raster frame buffer* directs the electron beam. It scans like a television set. The quality is controlled by the number of pixels and the number of bits per pixel. In this area, the standard text is *Computer Graphics*: *Principles and Practice* by Hughes, Van Dam, McGuire, Skylar, Foley, Feiner, and Akeley (3rd edition, Addison-Wesley, 2014). Notes by Ronald Goldman and by Tony DeRose were excellent references.

■ REVIEW OF THE KEY IDEAS ■

1. Computer graphics needs shift operations $T(\boldsymbol{v}) = \boldsymbol{v} + \boldsymbol{v}_0$ as well as linear operations $T(\boldsymbol{v}) = A\boldsymbol{v}$.

2. A shift in $\mathbf{R}^n$ can be executed by a matrix of order $n + 1$, using homogeneous coordinates.

3. The extra component 1 in $[x\, y\, z\, 1]$ is preserved when all matrices have the numbers $0, 0, 0, 1$ as last column.

Problem Set 10.6

1 A typical point in $\mathbf{R}^3$ is $x\boldsymbol{i} + y\boldsymbol{j} + z\boldsymbol{k}$. The coordinate vectors $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$ are $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1)$. The coordinates of the point are (x, y, z).

This point in computer graphics is $x\boldsymbol{i} + y\boldsymbol{j} + z\boldsymbol{k} +$ **origin**. Its homogeneous coordinates are (, , ,). Other coordinates for the same point are (, , ,).

2 A linear transformation T is determined when we know $T(\boldsymbol{i}), T(\boldsymbol{j}), T(\boldsymbol{k})$. For an affine transformation we also need $T($_____$)$. The input point $(x, y, z, 1)$ is transformed to $xT(\boldsymbol{i}) + yT(\boldsymbol{j}) + zT(\boldsymbol{k}) +$ _____.

3 Multiply the 4 by 4 matrix T for translation along $(1, 4, 3)$ and the matrix T_1 for translation along $(0, 2, 5)$. The product TT_1 is translation along _____.

4 Write down the 4 by 4 matrix S that scales by a constant c. Multiply ST and also TS, where T is translation by $(1, 4, 3)$. To blow up the picture around the center point $(1, 4, 3)$, would you use $\boldsymbol{v}ST$ or $\boldsymbol{v}TS$?

5 What scaling matrix S (in homogeneous coordinates, so 3 by 3) would produce a 1 by 1 square page from a standard 8.5 by 11 page?

6 What 4 by 4 matrix would move a corner of a cube to the origin and then multiply all lengths by 2? The corner of the cube is originally at $(1, 1, 2)$.

7 When the three matrices in equation 1 multiply the unit vector $\boldsymbol{a}$, show that they give $(\cos\theta)\boldsymbol{a}$ and $(1 - \cos\theta)\boldsymbol{a}$ and $\mathbf{0}$. Addition gives $\boldsymbol{a}Q = \boldsymbol{a}$ and the rotation axis is not moved.

8 If $\boldsymbol{b}$ is perpendicular to $\boldsymbol{a}$, multiply by the three matrices in 1 to get $(\cos\theta)\boldsymbol{b}$ and $\mathbf{0}$ and a vector perpendicular to $\boldsymbol{b}$. So $Q\boldsymbol{b}$ makes an angle θ with $\boldsymbol{b}$. ***This is rotation***.

9 What is the 3 by 3 projection matrix $I - \boldsymbol{n}\boldsymbol{n}^{\mathrm{T}}$ onto the plane $\frac{2}{3}x + \frac{2}{3}y + \frac{1}{3}z = 0$? In homogeneous coordinates add $0, 0, 0, 1$ as an extra row and column in P.

10 With the same 4 by 4 matrix P, multiply T_-PT_+ to find the projection matrix onto the plane $\frac{2}{3}x+\frac{2}{3}y+\frac{1}{3}z=1$. The translation T_- moves a point on that plane (choose one) to $(0,0,0,1)$. The inverse matrix T_+ moves it back.

11 Project $(3,3,3)$ onto those planes. Use P in Problem 9 and T_-PT_+ in Problem 10.

12 If you project a square onto a plane, what shape do you get?

13 If you project a cube onto a plane, what is the outline of the projection? Make the projection plane perpendicular to a diagonal of the cube.

14 The 3 by 3 mirror matrix that reflects through the plane $\boldsymbol{n}^{\mathrm{T}}\boldsymbol{v}=0$ is $M=I-2\boldsymbol{n}\boldsymbol{n}^{\mathrm{T}}$. Find the reflection of the point $(3,3,3)$ in the plane $\frac{2}{3}x+\frac{2}{3}y+\frac{1}{3}z=0$.

15 Find the reflection of $(3,3,3)$ in the plane $\frac{2}{3}x+\frac{2}{3}y+\frac{1}{3}z=1$. Take three steps T_-MT_+ using 4 by 4 matrices: translate by T_- so the plane goes through the origin, reflect the translated point $(3,3,3,1)T_-$ in that plane, then translate back by T_+.

16 The vector between the origin $(0,0,0,1)$ and the point $(x,y,z,1)$ is the difference $\boldsymbol{v}=$ _____. In homogeneous coordinates, vectors end in _____. So we add a _____ to a point, not a point to a point.

17 If you multiply only the *last* coordinate of each point to get (x,y,z,c), you rescale the whole space by the number _____. This is because the point (x,y,z,c) is the same as $(\ ,\ ,\ ,1)$.

10.7 Linear Algebra for Cryptography

1 Codes can use finite fields as alphabets: letters in the message become numbers $0, 1, \ldots, p-1$.

2 The numbers are added and multiplied $(mod\ p)$. Divide by p, keep the remainder.

3 A Hill Cipher multiplies blocks of the message by a secret matrix $E\ (mod\ p)$.

4 To decode, multiply each block by the inverse matrix $D\ (mod\ p)$. Not a very secure cipher!

Cryptography is about encoding and decoding messages. Banks do this all the time with financial information. Amazingly, modern algorithms can involve extremely deep mathematics. "Elliptic curves" play a part in cryptography, as they did in the sensational proof by Andrew Wiles of Fermat's Last Theorem.

This section will not go that far! But it will be our first experience with *finite fields* and *finite vector spaces*. The field for $\mathbf{R}^n$ contains all real numbers. The field for "modular arithmetic" contains only p integers $0, 1, \ldots, p-1$. There were infinitely many vectors in $\mathbf{R}^n$—now there will only be p^n messages of length n in message space. The alphabet from A to Z is finite (as in $p = 26$).

The codes in this section will be easily breakable—they are much too simple for practical security. The power of computers demands more complex cryptography, because that power would quickly detect a small encoding matrix. But a matrix code (the Hill Cipher) will allow us to see linear algebra at work in a new way.

All our calculations in encoding and decoding will be "**mod p**". But the central concepts of linear independence and bases and inverse matrices and determinants survive this change. We will be doing "linear algebra with finite fields". Here is the meaning of *mod* p:

$\mathbf{27 \equiv 2\ (mod\ 5)}$	means that **$27 - 2$ is divisible by 5**
$\boldsymbol{y \equiv x\ (mod\ p)}$	means that **$y - x$ is divisible by p**

Dividing y by 5 produces one of the five possible remainders $x = 0, 1, 2, 3, 4$. All the numbers $5, -5, 10, -10, \ldots$ with no remainder are congruent to zero $(mod\ 5)$. The numbers $y = 6, -4, 11, -9, \ldots$ are all congruent to $x = 1 (mod\ 5)$.

We use the word **congruent** for the symbol $\equiv$ and we call this "modular arithmetic". Every integer y produces one of the values $x = 0, 1, 2, \ldots, p-1$.

The theory is best if p is a prime number. With $p = 26$ letters from A to Z, we unfortunately don't start with a prime p. Cryptography can deal with this problem.

Modular Arithmetic

Linear algebra is based on linear combinations of vectors. Now our vectors $(x_1, \ldots, x_n)$ are strings of integers limited to $x = 0, 1, \ldots, p-1$. All calculations produce these integers when we work "$mod\ p$". This means: *Every integer y outside that range is divided by p and x is the remainder*:

$y = qp + x$	$y \equiv x\ (mod\ p)$	y **divided by** p **has remainder** x

Addition $mod\ 3$ $10 \equiv 1\ (mod\ 3)$ and $16 \equiv 1\ (mod\ 3)$ and $10 + 16 \equiv 1 + 1\ (mod\ 3)$

I could add $10 + 16$ and divide 26 by 3 to get the remainder 2.
Or I can just add remainders $1 + 1$ to reach the same answer 2.

Addition $mod\ 2$ $11 \equiv 1\ (mod\ 2)$ and $17 \equiv 1\ (mod\ 2)$ and $11 + 17 = 28 \equiv 0\ (mod\ 2)$

The remainders added to $1 + 1$ *but this is not* 2. The final step was $2 \equiv 0\ (mod\ 2)$.

Addition $mod\ p$ is completely reasonable. So is **multiplication $mod\ p$**. Here $p = 3$:

$10 \equiv 1\ (mod\ 3)$ times $16 \equiv 1\ (mod\ 3)$ gives 1 times $1 \equiv 1$ $\quad 160 \equiv 1\ (mod\ 3)$

$5 \equiv 2\ (mod\ 3)$ times $8 \equiv 2\ (mod\ 3)$ gives 2 times $2 \equiv 1$ $\quad 40 \equiv 1\ (mod\ 3)$

Conclusion: We can safely add and multiply modulo p. So we can take linear combinations. This is the key operation in linear algebra. **But can we divide ?**

In the real number field, the inverse is $1/y$ (for any number except $y = 0$). This means: We found another real number z so that $yz = 1$. Invertibility is a requirement for a field. **Is inversion always possible $mod\ p$?** For every number $y = 1, \ldots, p-1$ can we find another number $z = 1, \ldots, p-1$ so that $yz \equiv 1\ mod\ p$?

The examples $3^{-1} \equiv 4\ (mod\ 11)$ and $2^{-1} \equiv 6\ (mod\ 11)$ and $5^{-1} \equiv 9\ (mod\ 11)$ all succeed. Can you solve $7z \equiv 1\ (mod\ 11)$? Inverting numbers will be the key to inverting matrices.

Let me show that inversion $mod\ p$ has a problem when p is not a prime number. The example $p = 26$ factors into 2 times 13. **Then $y = 2$ cannot have an inverse $z\ (mod\ 26)$.** The requirement $2z \equiv 1\ (mod\ 26)$ is impossible to satisfy because $2z$ and 26 are even.

Similarly 5 has no inverse z when p is 25. We can't solve $5z \equiv 1\ (mod\ 25)$. The number $5z - 1$ is never going to be a multiple of 5, so it can't be a multiple of 25.

Inversion of every y $(0 < y < p)$ will be possible if and only if p is prime.

Inversion needs $y, 2y, 3y, \ldots, py$ to have different remainders when divided by p.

If my and ny had the same remainder x then $(m - n)y$ would be divisible by p.

The prime number p would have to divide either $m - n$ or y. Both are impossible.

So $y, \ldots, py$ have different remainders: **One of those remainders must be $x = 1$.**

The Enigma Machine and the Hill Cipher

Lester Hill published his cipher (his system for encoding and decoding) in the American Mathematical Monthly (1929). The idea was simple, but in some way it started the transition of cryptography from linguistics to mathematics. Codes up to that time mainly mixed up alphabets and rearranged messages. The **Enigma code** used by the German Navy in World War II was a giant advance—using machines that look to us like primitive computers. The English set up Bletchley Park to break Enigma. They hired puzzle solvers and language majors. And by good luck they also happened to get Alan Turing.

I don't know if you have seen the movie about him: *The Imitation Game*. A lot of it is unrealistic (like *Good Will Hunting* and *A Beautiful Mind* at MIT). But the core idea of breaking the Enigma code was correct, using human weaknesses in the encoding and broadcasting. The German naval command openly sent out their coded orders—knowing that the codes were too complicated to break (if it hadn't been for those weaknesses). The codebreaking required English electronics to undo the German electronics. It also required genius.

Alan Turing was surely a genius—England's most exceptional mathematician. His life was ultimately tragic and he ended it in 1954. The biography by Andrew Hodges is excellent. Turing arrived at Bletchley Park the day after Poland was invaded. It is to Winston Churchill's credit that he gave fast and full support when his support was needed.

The Enigma Machine had gears and wheels. The Hill Cipher only needs a matrix. That is the code to be explained now, using linear algebra. You will see how decoding involved inverse matrices. All steps use modular arithmetic, multiplying and inverting $mod\ p$.

I will follow the neat exposition of Professor Spickler of Salisbury State University, which he made available on the Web: **facultyfp.salisbury.edu/despickler/personal/index.asp**

Modular Arithmetic with Matrices

Addition, subtraction, and multiplication are all we need for $A\boldsymbol{x}$ (matrix times vector). To multiply $mod\ p$ we can multiply the integers in A times the integers in $\boldsymbol{x}$ as usual—and then replace every entry of $A\boldsymbol{x}$ by its value $mod\ p$.

Key questions: When can we solve $A\boldsymbol{x} \equiv \boldsymbol{b}\ (mod\ p)$? Do we still have the four subspaces $\boldsymbol{C}(A), \boldsymbol{N}(A), \boldsymbol{C}(A^{\mathrm{T}}), \boldsymbol{N}(A^{\mathrm{T}})$? Are they still orthogonal in pairs? Is there still an inverse matrix $mod\ p$ whenever the determinant of A is nonzero $mod\ p$? I am happy to say that the last three answers are *yes* (but the inverse question requires p to be a prime number).

We can find $A^{-1}\ (mod\ p)$ by Gauss-Jordan elimination, reducing $[A\ \ I]$ to $[I\ \ A^{-1}]$ as in Section 2.5. Or we can use determinants and the cofactor matrix C in the formula $A^{-1} = (\det A)^{-1}\, C^{\mathrm{T}}$. I will work $mod\ 3$ with a 2 by 2 integer matrix A :

$$\left[\begin{array}{cc|cc} A & & I & \end{array}\right] = \left[\begin{array}{cc|cc} \mathbf{2} & \mathbf{0} & 1 & 0 \\ \mathbf{2} & \mathbf{1} & 0 & 1 \end{array}\right] \rightarrow \left[\begin{array}{cc|cc} 2 & 0 & 1 & 0 \\ 0 & 1 & 2 & 1 \end{array}\right] \rightarrow \begin{array}{c}\text{multiply row 1} \\ \text{by } 2^{-1} \equiv 2\end{array} \rightarrow \left[\begin{array}{cc|cc} 1 & 0 & \mathbf{2} & \mathbf{0} \\ 0 & 1 & \mathbf{2} & \mathbf{1} \end{array}\right] = \left[\begin{array}{cc|cc} & I & & A^{-1} \end{array}\right]$$

By pure chance $A^{-1} \equiv A$! Multiplying A times A mod 3 does give the identity matrix:

$$A^2 = AA^{-1} = \begin{bmatrix} 2 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 6 & 1 \end{bmatrix} \equiv \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \ (mod\ 3).$$

The determinant of A is 2, and the cofactor formula from Section 5.3 also gives $A^{-1} \equiv A$:

$$\begin{bmatrix} 2 & 0 \\ 2 & 1 \end{bmatrix}^{-1} = 2^{-1} \begin{bmatrix} 1 & -0 \\ -2 & 2 \end{bmatrix} \equiv 2 \begin{bmatrix} 1 & -0 \\ -2 & 2 \end{bmatrix} \equiv \begin{bmatrix} 2 & 0 \\ 2 & 1 \end{bmatrix} \ (mod\ 3).$$

Theorem. A^{-1} exists $mod\ p$ if and only if $(\det A)^{-1}$ exists $mod\ p$.
The requirement is: $\det A$ and p have no common factors.

Encryption with the Hill Cipher

The original cipher used the letters A to Z with $p = 26$. Hill chose an n by n encryption matrix E so that $\det E$ is not divisible by 2 or 13. Then the number $\det E$ has an inverse $mod\ 26$ and so does the matrix E. The inverse matrix $E^{-1} \equiv D\ (mod\ 26)$ will be the decryption matrix that decodes the message.

Now convert each letter of the message into a number from 0 to 25. The obvious choice from $A = 0$ to $Z = 25$ is acceptable because the matrix will make this cipher stronger.

> Ignore spaces and divide the message into blocks $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots$ of size n.
> Then multiply each message block $(mod\ p)$ by the encryption matrix E.
> The coded message is $E\boldsymbol{v}_1, E\boldsymbol{v}_2, \ldots$ and you know what the decoder will do.

Spikler's example has $D = E^{-1} =$ $\det E = 583 \equiv 11\ (mod\ 26)$

$$\begin{bmatrix} 2 & 3 & 15 \\ 5 & 8 & 12 \\ 1 & 13 & 4 \end{bmatrix}^{-1} \equiv \begin{bmatrix} 10 & 19 & 16 \\ 4 & 23 & 7 \\ 17 & 5 & 19 \end{bmatrix} \ (mod\ 26).$$

Of course a codebreaker will not know E or D. And the block size n is generally unknown too. For the matrices Hill had in mind n would not be very large and a computer could quickly discover E and D.

I am not sure if Hill's Cipher could become seriously difficulty to break by choosing very large matrices and a large prime number p. And by encoding the coded message a second time, using a different block size n_2 and large matrix E_2 and large prime p_2.

Finite Fields and Finite Vector Spaces

In algebra, a field **F** is a set of scalars that can be added and multiplied and inverted (except 0 can't be inverted). Familiar examples are the real numbers **R** and the complex numbers **C** and the rational numbers **Q** (containing every ratio p/q of integers). From a field you build vectors $\boldsymbol{v} = (f_1, f_2, \ldots, f_n)$. From linear combinations of vectors you build vector spaces. *So linear algebra begins with a field* **F**.

I taught for ten years from a textbook that started with fields. On the way to $\mathbf{R}^n$, we lost a lot of students. That was a signal—the emphasis was misplaced if we wanted the

course to be useful. I believe the right way is to understand $\mathbf{R}^n$ and its subspaces first, as you do. Then you can look at other fields and vector spaces with a natural question in mind: *What is new when the field is not* $\mathbf{R}$?

These pages are asking that question for **finite fields**. The possibilities become more limited but also highly interesting. The starting point (and not quite the ending point) is the finite field $\mathbf{F}_p$. It contains only the numbers $0, 1, \ldots, p-1$ and p is a prime number. I will focus first on the field $\mathbf{F}_2$ with only 2 members "**0**" and "**1**". You could think of **0** and **1** as "even" and "odd" because the rules to add and multiply are obeyed by the even numbers and odd numbers: even + odd = *odd* and even × odd = *even*.

Addition table

	0	1
0	0	1
1	1	0

Multiplication table

	0	1
0	0	0
1	0	1

This is addition and multiplication "$mod\ 2$".

From this field $\mathbf{F}_2$ we can build vectors like $\boldsymbol{v} = (0, 0, 1)$ and $\boldsymbol{w} = (1, 0, 1)$. There are three components with two choices each: a total of $2^3 = 8$ different vectors in the vector space $(\mathbf{F}_2)^3$. You know the requirements on a subspace and the possibilities it opens up:

a) The zero-dimensional subspace containing only $\mathbf{0} = (0, 0, 0)$.

b) One-dimensional subspaces containing $\mathbf{0}$ and a vector like $\boldsymbol{v}$. Notice $\boldsymbol{v} + \boldsymbol{v} = \mathbf{0}$!

c) Two-dimensional subspaces with a basis like $\boldsymbol{v}$ and $\boldsymbol{w}$ and 4 vectors $\mathbf{0}, \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{v} + \boldsymbol{w}$.

d) The full three-dimensional subspace $(\mathbf{F}_2)^3$ with 8 vectors.

What are the possible bases for $(\mathbf{F}_2)^3$? The standard basis contains $(1, 0, 0)$ and $(0, 1, 0)$ and $(0, 0, 1)$. Those vectors are linearly independent and they span $(\mathbf{F}_2)^3$. Their eight combinations with coefficients 0 and 1 fill all of $(\mathbf{F}_2)^3$.

What about matrices that multiply those vectors? The matrices will be 1 by 3, or 2 by 3, or 3 by 3. When they are 3 by 3 we can ask if they are invertible. Their determinants can only be 0 (singular matrix) or 1 (invertible matrix). Let me leave you the pleasure of deciding whether these matrices are invertible. *And how would you find the inverse* ?

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \qquad B = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix} \qquad C = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

Out of 2^9 possible matrices over $\mathbf{F}_2$, I will guess that most are singular.

To conclude this discussion of $\mathbf{F}_2$, I mention a field with $2^2 = 4$ members. It will not come from multiplication $(mod\ 4)$, because 4 is not prime. The multiplication 2 times 2 will give 0 (and 2 has no inverse): *not a field.* But we can start with the numbers 0 and 1 in $\mathbf{F}_2$ and invent two more numbers a and $1 + a$—provided they follow these two rules: $(\boldsymbol{a} + \boldsymbol{a} = \mathbf{0})$ and $(\boldsymbol{a} \times \boldsymbol{a} = \mathbf{1} + \boldsymbol{a})$. Then a and $1 + a$ are inverses. Not obvious !

Add	0	1	a	$1+a$
0	0	1	a	$1+a$
1	1	0	$1+a$	a
a	a	$1+a$	0	1
$1+a$	$1+a$	a	1	0

Multiply	0	1	a	$1+a$
0	0	0	0	0
1	0	1	a	$1+a$
a	0	a	$1+a$	1
$1+a$	0	$1+a$	1	a

Beyond $p = 2$, we have the fields $\mathbf{F}_p$ for all prime numbers p. They use addition and multiplication $mod\ p$. They are alphabets for codes. They provide the components for vectors $\boldsymbol{v} = (f_1, \ldots, f_n)$ in the space $(\mathbf{F}_p)^n$. They provide the entries for matrices that multiply those vectors. These fields $\mathbf{F}_p$ are the most frequently used finite fields.

The only other finite fields have p^k members. The example above of $0, 1, a, 1 + a$ had $2^2 = 4$ members. We will leave it there and get back safely to $\mathbf{R}$.

Problem Set 10.7

1 If you multiply n whole numbers (even or odd) when is the answer odd? Translate into multiplication $(mod\ 2)$: If you multiply 0's and 1's when is the answer 1 ?

2 If you add n whole numbers (even or odd) when is the sum of the numbers odd? Translate into adding 0's and 1's $(mod\ 2)$. When do they add to 1 ?

3 (a) If $y_1 \equiv x_1$ and $y_2 \equiv x_2$, why is $y_1 + y_2 \equiv x_1 + x_2$? All are $mod\ p$.
Suggestion: $y_1 = p\,q_1 + x_1$ and $y_2 = p\,q_2 + x_2$. Now add $y_1 + y_2$.

(b) Can you be sure that $x_1 + x_2$ is smaller than p ? *No*. Give an example where there is a smaller x with $(y_1 + y_2) = x\ (mod\ p)$.

4 $p = 39$ is not prime. Find a number a that has no inverse $z\ (mod\ 39)$. This means that $az \equiv 1\ (mod\ 39)$ has no solution. Then find a 2 by 2 matrix A that has no inverse matrix $Z\ (mod\ 39)$. This means that $AZ \equiv I\ (mod\ 39)$ has no solution.

5 Show that $y \equiv x\ (mod\ p)$ *leads to* $-y \equiv -x\ (mod\ p)$.

6 Find a matrix that has independent columns in $\mathbf{R}^2$ but dependent columns $(mod\ 5)$.

7 What are all the 2 by 2 matrices of 0's and 1's that are invertible $(mod\ 2)$?

8 Is the row space of A still orthogonal to the nullspace in modular arithmetic $(mod\ 11)$? Are bases for those subspaces still bases $(mod\ 11)$?

9 (Hill's Cipher) Separate the message THISWHOLEBOOKISINCODE into blocks of 3 letters. Replace each letter by a number from 1 to 26 (normal order). Multiply each block by the 3 by 3 matrix L with 1's on and below the diagonal. What is the coded message (in numbers) and how would you decode it?

10 Suppose you know the original message (the plaintext). Suppose you also see the coded message. How would you start to discover the matrix in Hill's Cipher ? For a very long message do you expect success ?

Chapter 11

Numerical Linear Algebra

1 The goals of numerical linear algebra are **speed** and **accuracy** and **stability**: $n > 10^3$ or 10^6.

2 Matrices can be full or sparse or banded or structured: special algorithms for each.

3 Accuracy of elimination is controlled by the **condition number** $||\boldsymbol{A}||\ ||\boldsymbol{A}^{-1}||$.

4 Gram-Schmidt is often computed by using **Householder reflections** $H = I - 2\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$ to find Q.

5 Eigenvalues use $\boldsymbol{QR}$ **iterations** $\boldsymbol{A_0} = Q_0 R_0 \to R_0 Q_0 = \boldsymbol{A_1} = Q_1 R_1 \to \to \boldsymbol{A_n}$.

6 **Shifted** $\boldsymbol{QR}$ is even better: Shift to $A_k - c_k I = Q_k R_k$, shift back $A_{k+1} = R_k Q_k + c_k I$.

7 Iteration $S\boldsymbol{x}_{k+1} = \boldsymbol{b} - T\boldsymbol{x}_k$ solves $(S+T)\,\boldsymbol{x} = \boldsymbol{b}$ if all eigenvalues of $S^{-1}T$ have $|\lambda| < 1$.

8 Iterative methods often use **preconditioners** P. Change $A\boldsymbol{x} = \boldsymbol{b}$ to $PA\boldsymbol{x} = P\boldsymbol{b}$ with $PA \approx I$.

9 **Conjugate gradients** and **GMRES** are Krylov methods; see Trefethen-Bau (and other texts).

11.1 Gaussian Elimination in Practice

Numerical linear algebra is a struggle for *quick* solutions and also *accurate* solutions. We need efficiency but we have to avoid instability. In Gaussian elimination, the main freedom (always available) is to ***exchange equations***. This section explains when to exchange rows for the sake of speed, and when to do it for the sake of accuracy.

The key to accuracy is to avoid unnecessarily large numbers. Often that requires us to avoid small numbers! A small pivot generally means large multipliers (since we divide by the pivot). A good plan is "***partial pivoting***", to choose the *largest available pivot* in each new column. We will see why this pivoting strategy is built into computer programs.

Other row exchanges are done to save elimination steps. In practice, most large matrices are ***sparse***—almost all entries are zeros. Elimination is fastest when the equations are

ordered *to produce a narrow band of nonzeros*. Zeros inside the band "fill in" during elimination—those zeros are destroyed and don't save computing time.

Section 11.2 is about instability that can't be avoided. It is built into the problem, and this sensitivity is measured by the "***condition number***". Then Section 11.3 describes how to solve $A\boldsymbol{x} = \boldsymbol{b}$ by ***iterations***. Instead of direct elimination, the computer solves an easier equation many times. Each answer $\boldsymbol{x}_k$ leads to the next guess $\boldsymbol{x}_{k+1}$. For good iterations (the **conjugate gradient method** is extremely good), the $\boldsymbol{x}_k$ converge quickly to $\boldsymbol{x} = A^{-1}\boldsymbol{b}$.

The Fastest Supercomputer

A new supercomputing record was announced by IBM and Los Alamos on May 20, 2008. The Roadrunner was the first to achieve a quadrillion (10^{15}) floating-point operations per second: *a petaflop machine*. The benchmark for this world record was a large dense linear system $A\boldsymbol{x} = \boldsymbol{b}$: computer speed is tested by linear algebra.

That machine was shut down in 2013! The TOP500 project ranks the 500 most powerful computer systems in the world. As I write this page in October 2015, the first four are from NUDT in China, Cray and IBM in the US, and Fujitsu in Japan. They all use a LINUX-based system. And all vector processors have fallen out of the top 500.

Looking ahead, the Summit is expected to take first place with 150-300 petaflops. President Obama has just ordered the development of an exascale system (1000 petaflops). Up to now we are following Moore's Law of doubling every 14 months.

The LAPACK software does elimination with partial pivoting. The biggest difference from this book is to organize the steps to use large submatrices and never single numbers. And graphics processing units (GPU's) are now almost required for success. The market for video games dwarfs scientific computing and led to astonishing acceleration in the chips.

Before IBM's BlueGene, a key issue was to count the standard quad-core processors that a petaflop machine would need: 32,000. The new architecture uses much less power, but its hybrid design has a price: a code needs three separate compilers and explicit instructions to move all the data. Please see the excellent article in *SIAM News* (**siam.org**, July 2008) and the update on **www.lanl.gov/roadrunner**.

Our thinking about matrix calculations is reflected in the highly optimized **BLAS** (*Basic Linear Algebra Subroutines*). They come at levels 1, 2, and 3:

Level 1 Linear combinations of vectors $a\boldsymbol{u} + \boldsymbol{v}$: $O(n)$ work

Level 2 Matrix-vector multiplications $A\boldsymbol{u} + \boldsymbol{v}$: $O(n^2)$ work

Level 3 Matrix-matrix multiplications $AB + C$: $O(n^3)$ work

Level 1 is an elimination step (multiply row j by ℓ_{ij} and subtract from row i). Level 2 can eliminate a whole column at once. A high performance solver is rich in Level 3 BLAS (AB has $2n^3$ flops and $2n^2$ data, a good ratio of work to talk).

It is *data passing* and *storage retrieval* that limit the speed of parallel processing. The high-velocity cache between main memory and floating-point computation has to be fully used! Top speed demands a ***block matrix approach*** to elimination.

The big change, coming now, is parallel processing at the chip level.

Roundoff Error and Partial Pivoting

Up to now, any pivot (nonzero of course) was accepted. In practice a small pivot is dangerous. A catastrophe can occur when numbers of different sizes are added. Computers keep a fixed number of significant digits (say three decimals, for a very weak machine). The sum $10{,}000 + 1$ is rounded off to $10{,}000$. The "1" is completely lost. Watch how that changes the solution to this problem:

$$\begin{aligned} .0001u + v &= 1 \\ -u + v &= 0 \end{aligned} \qquad \text{starts with coefficient matrix} \qquad A = \begin{bmatrix} .0001 & 1 \\ -1 & 1 \end{bmatrix}.$$

If we accept $.0001$ as the pivot, elimination adds $10{,}000$ times row 1 to row 2. Roundoff leaves

$$10{,}000v = 10{,}000 \qquad \text{instead of} \qquad 10{,}001v = 10{,}000.$$

The computed answer $v = 1$ is near the true $v = .9999$. But then back substitution puts the wrong $v = 1$ into the equation for u:

$$.0001\,u + 1 = 1 \qquad \text{instead of} \qquad .0001\,u + .9999 = 1.$$

The first equation gives $u = 0$. The correct answer (look at the second equation) is $u = 1.000$. By losing the "1" in the matrix, we have lost the solution. ***The small change from 10,001 to 10,000 has changed the answer from $u = 1$ to $u = 0$*** (100% error!).
If we exchange rows, even this weak computer finds an answer that is correct to 3 places:

$$\begin{aligned} -u + v &= 0 \\ .0001u + v &= 1 \end{aligned} \quad \longrightarrow \quad \begin{aligned} -u + v &= 0 \\ v &= 1 \end{aligned} \quad \longrightarrow \quad \begin{aligned} u &= 1 \\ v &= 1. \end{aligned}$$

The original pivots were $.0001$ and $10{,}000$—badly scaled. After a row exchange the exact pivots are -1 and 1.0001—well scaled. The computed pivots -1 and 1 come close to the exact values. Small pivots bring numerical instability, and the remedy is ***partial pivoting***. Here is our strategy when we reach and search column k for the best available pivot:

Choose the largest number in row k or below. Exchange its row with row k.

The strategy of *complete pivoting* looks also in later columns for the largest pivot. It exchanges columns as well as rows. This expense is seldom justified, and all major codes use partial pivoting. Multiplying a row or column by a scaling constant can also be very worthwhile. *If the first equation above is $u + 10{,}000v = 10{,}000$ and we don't rescale, then 1 looks like a good pivot and we would miss the essential row exchange.*

For positive definite matrices, row exchanges are *not* required. It is safe to accept the pivots as they appear. Small pivots can occur, but the matrix is not improved by row exchanges. When its condition number is high, the problem is in the matrix and not in the code. In this case the output is unavoidably sensitive to the input.

The reader now understands how a computer actually solves $A\boldsymbol{x} = \boldsymbol{b}$***—by elimination with partial pivoting***. Compared with the theoretical description—***find A^{-1} and multiply $A^{-1}\boldsymbol{b}$***—the details took time. But in computer time, elimination is much faster. I believe that elimination is also the best approach to the algebra of row spaces and nullspaces.

Operation Counts: Full Matrices

Here is a practical question about cost. *How many separate operations are needed to solve* $A\boldsymbol{x} = \boldsymbol{b}$ *by elimination*? This decides how large a problem we can afford.

Look first at A, which changes gradually into U. When a multiple of row 1 is subtracted from row 2, we do n operations. The first is a division by the pivot, to find the multiplier ℓ. For the other $n-1$ entries along the row, the operation is a "multiply-subtract". For convenience, we count this as a single operation. If you regard multiplying by ℓ and subtracting from the existing entry as two separate operations, *multiply all our counts by* 2.

The matrix A is n by n. The operation count applies to all $n-1$ rows below the first. Thus it requires n times $n-1$ operations, or $n^2 - n$, to produce zeros below the first pivot. *Check*: *All* n^2 *entries are changed, except the* n *entries in the first row*.

When elimination is down to k equations, the rows are shorter. We need only $k^2 - k$ operations (instead of $n^2 - n$) to clear out the column below the pivot. This is true for $1 \le k \le n$. The last step requires no operations ($1^2 - 1 = 0$); forward elimination is complete. The total count to reach U is the sum of $k^2 - k$ over all values of k from 1 to n:

$$(1^2 + \cdots + n^2) - (1 + \cdots + n) = \frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)}{2} = \boxed{\frac{n^3 - n}{3}}.$$

Those are known formulas for the sum of the first n numbers and their squares. Substituting $n = 100$ gives a million minus a hundred—then divide by 3. (That translates into one second on a workstation.) We will ignore n in comparison with n^3, to reach our main conclusion:

The multiply-subtract count is $\frac{1}{3}n^3$ ***for forward elimination (*** A ***to*** U***, producing*** L***).***

That means $\frac{1}{3}n^3$ multiplications and subtractions. Doubling n increases this cost by eight (because n is cubed). 100 equations are easy, 1000 are more expensive,10000 dense equations are close to impossible. We need a faster computer or a lot of zeros or a new idea.

On the right side of the equations, the steps go much faster. We operate on single numbers, not whole rows. ***Each right side needs exactly*** n^2 ***operations***. Down and back up we are solving two triangular systems, $L\boldsymbol{c} = \boldsymbol{b}$ forward and $U\boldsymbol{x} = \boldsymbol{c}$ backward. In back substitution, the last unknown needs only division by the last pivot. The equation above it needs two operations—substituting $\boldsymbol{x}_n$ and dividing by *its* pivot. The kth step needs k multiply-subtract operations, and the total for back substitution is

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2} \approx \tfrac{1}{2}n^2 \quad \text{operations.}$$

The forward part is similar. *The* n^2 *total exactly equals the count for multiplying* $A^{-1}\boldsymbol{b}$! This leaves Gaussian elimination with two big advantages over $A^{-1}\boldsymbol{b}$:

1 Elimination requires $\frac{1}{3}n^3$ multiply-subtracts, compared to n^3 for A^{-1}.

2 If A is *banded* so are L and U: by comparison A^{-1} is full of nonzeros.

Band Matrices

These counts are improved when A has "*good zeros*". A good zero is an entry that remains zero in L and U. ***The best zeros are at the beginning of a row.*** They require no elimination steps (the multipliers are zero). So we also find those same good zeros in L. That is especially clear for this *tridiagonal matrix* A (and for band matrices in Figure 11.1):

Tridiagonal
Bidiagonal
times
bidiagonal

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

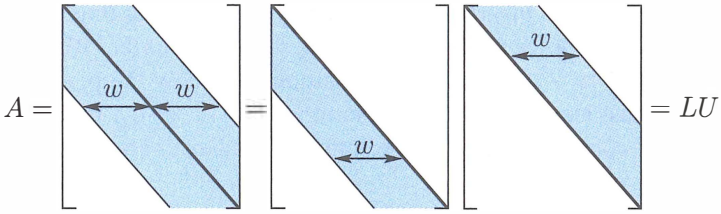


Figure 11.1: $A = LU$ for a band matrix. Good zeros in A *stay zero* in L and U.

These zeros lead to a complete change in the operation count, for "half-bandwidth" w:

A band matrix has $a_{ij} = 0$ when $|i - j| > w$.

Thus $w = 1$ for a diagonal matrix, $w = 2$ for tridiagonal, $w = n$ for dense. The length of the pivot row is at most w. There are no more than $w - 1$ nonzeros below any pivot. Each stage of elimination is complete after $w(w-1)$ operations, and *the band structure survives*. There are n columns to clear out. Therefore:

Elimination on a band matrix (A to L and U) needs less than $w^2 n$ operations.

For a band matrix, the count is proportional to n instead of n^3. It is also proportional to w^2. A full matrix has $w = n$ and we are back to n^3. For an exact count, remember that the bandwidth drops below w in the lower right corner (not enough space):

$$\textbf{Band} \quad \frac{w(w-1)(3n-2w+1)}{3} \qquad \textbf{Dense} \quad \frac{n(n-1)(n+1)}{3} = \frac{n^3-n}{3}$$

On the right side of $A\boldsymbol{x} = \boldsymbol{b}$, to find $\boldsymbol{x}$ from $\boldsymbol{b}$, the cost is about $2wn$ (compared to the usual n^2). *Main point*: *For a band matrix the operation counts are* ***proportional to n.*** This is extremely fast. A tridiagonal matrix of order 10,000 is very cheap, provided *we don't compute* A^{-1}. That inverse matrix has no zeros at all:

$$A = \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} \quad \text{has} \quad A^{-1} = U^{-1}L^{-1} = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 3 & 3 & 2 & 1 \\ 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

We are actually worse off knowing A^{-1} than knowing L and U. Multiplication by A^{-1} needs the full n^2 steps. Solving $L\boldsymbol{c} = \boldsymbol{b}$ and $U\boldsymbol{x} = \boldsymbol{c}$ needs only $2wn$.

A band structure is very common in practice, when the matrix reflects connections between near neighbors: $a_{13} = 0$ and $a_{14} = 0$ because 1 is not a neighbor of 3 and 4.

We close with counts for Gauss-Jordan and Gram-Schmidt-Householder:

A^{-1} costs n^3 multiply-subtract steps. ***QR costs $\frac{2}{3}n^3$ steps.***

In $AA^{-1} = I$, the jth column of A^{-1} solves $A\boldsymbol{x}_j = j$th column of I. The left side costs $\frac{1}{3}n^3$ as usual. (This is a one-time cost! L and U are not repeated.) The special saving for the jth column of I comes from its first $j-1$ zeros. No work is required on the right side until elimination reaches row j. The forward cost is $\frac{1}{2}(n-j)^2$ instead of $\frac{1}{2}n^2$. Summing over j, the total for forward elimination on the n right sides is $\frac{1}{6}n^3$. The final multiply-subtract count for A^{-1} is n^3 if we actually want the inverse:

$$\textbf{For } A^{-1} \qquad \frac{n^3}{3}\,(L \text{ and } U) + \frac{n^3}{6}\,(\text{forward}) + n\left(\frac{n^2}{2}\right)(\text{back substitutions}) = \boldsymbol{n^3}. \quad (1)$$

Orthogonalization *(A to Q)*: The key difference from elimination is that *each multiplier is decided by a dot product.* That takes n operations, where elimination just divides by the pivot. Then there are n "multiply-subtract" operations to remove from column k its projection along column $j < k$ (see Section 4.4). The combined cost is $2n$ where for elimination it is n. This factor 2 is the price of orthogonality. We are changing a dot product to zero where elimination changes an entry to zero.

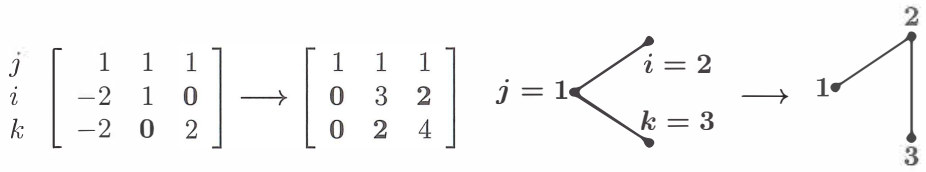
Caution To judge a numerical algorithm, it is **not enough** to count the operations. Beyond "flop counting" is a study of stability (Householder wins) and the flow of data.

Reordering Sparse Matrices

For band matrices with constant width w, the row ordering is optimal. But for most sparse matrices in real computations, the width of the band is *not constant* and there are many zeros inside the band. Those zeros can fill in as elimination proceeds—they are lost. We need to ***renumber the equations to reduce fill-in***, and thereby speed up elimination.

Generally speaking, we want to move zeros to early rows and columns. Later rows and columns are shorter anyway. The "**a**pproximate **m**inimum **d**egree" algorithm in sparse MATLAB is *greedy*—it chooses the row to eliminate without counting all the consequences. We may reach a nearly full matrix near the end, but the total operation count to reach LU is still much smaller. To find the absolute minimum of nonzeros in L and U is an NP-hard problem, much too expensive, and **amd** is a good compromise.

Fill-in is famous when each point on a square grid is connected to its four nearest neighbors. It is impossible to number all the gridpoints so that neighbors stay together! If we number by rows of the grid, there is a long wait to come around to the gridpoint above.



$a_{32} = 0$ $a_{32} = 2$ $a_{32} = 0$ **before** $a_{32} \neq 0$ **after**

We only need the *positions* of the nonzeros, not their exact values. Think of the graph of nonzeros: *Node i is connected to node j if $a_{ij} \neq 0$.* Watch to see how elimination can create nonzeros (new edges), which we are trying to avoid.

The command **nnz**(L) counts the nonzero multipliers in the lower triangular L, **find** (L) will list them, and **spy**(L) shows them all.

The goal of **colamd** and **symamd** is a better ordering (permutation P) that reduces fill-in for AP and $P^{\mathrm{T}}AP$—*by choosing the pivot with the fewest nonzeros below it.*

Fast Orthogonalization

There are three ways to reach the important factorization $A = QR$. Gram-Schmidt works to find the orthonormal vectors in Q. Then R is upper triangular because of the order of Gram-Schmidt steps. Now we look at better methods (Householder and Givens), which use a product of specially simple Q's that we *know* are orthogonal.

Elimination gives $A = LU$, orthogonalization gives $A = QR$. We don't want a triangular L, we want an orthogonal Q. L is a product of E's from elimination, with 1's on the diagonal and the multiplier ℓ_{ij} below. *Q will be a product of orthogonal matrices.*

There are two simple orthogonal matrices to take the place of the E's. The ***reflection matrices*** $I - 2\boldsymbol{u}\boldsymbol{u}^{\mathrm{T}}$ are named after Householder. The ***plane rotation matrices*** are named after Givens. The simple matrix that rotates the xy plane by θ is Q_{21}:

Givens rotation in the 1-2 plane

$$Q_{21} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Use Q_{21} the way you used E_{21}, to produce a zero in the $(2,1)$ position. That determines the angle θ. Bill Hager gives this example in *Applied Numerical Linear Algebra*:

$$Q_{21}A = \begin{bmatrix} .6 & .8 & 0 \\ -.8 & .6 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 90 & -153 & 114 \\ 120 & -79 & -223 \\ 200 & -40 & 395 \end{bmatrix} = \begin{bmatrix} 150 & -155 & -110 \\ 0 & 75 & -225 \\ 200 & -40 & 395 \end{bmatrix}.$$

The zero came from $-.8(90) + .6(120)$. No need to find θ, what we needed was $\cos\theta$:

$$\cos\theta = \frac{90}{\sqrt{90^2 + 120^2}} \quad \text{and} \quad \sin\theta = \frac{-120}{\sqrt{90^2 + 120^2}}. \tag{2}$$

Now we attack the $(3,1)$ entry. The rotation will be in rows and columns 3 and 1. The numbers $\cos\theta$ and $\sin\theta$ are determined from 150 and 200, instead of 90 and 120.

$$Q_{31}Q_{21}A = \begin{bmatrix} .6 & 0 & .8 \\ 0 & 1 & 0 \\ -.8 & 0 & .6 \end{bmatrix} \begin{bmatrix} 150 & \cdot & \cdot \\ 0 & \cdot & \cdot \\ 200 & \cdot & \cdot \end{bmatrix} = \begin{bmatrix} 250 & -125 & 250 \\ 0 & 75 & -225 \\ 0 & 100 & 325 \end{bmatrix}.$$

One more step to R. The $(3,2)$ entry has to go. The numbers $\cos\theta$ and $\sin\theta$ now come from 75 and 100. The rotation is now in rows and columns 2 and 3:

$$Q_{32}Q_{31}Q_{21}A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & .6 & .8 \\ 0 & -.8 & .6 \end{bmatrix} \begin{bmatrix} 250 & -125 & \cdot \\ 0 & 75 & \cdot \\ 0 & 100 & \cdot \end{bmatrix} = \begin{bmatrix} 250 & -125 & 250 \\ 0 & 125 & 125 \\ 0 & 0 & 375 \end{bmatrix}.$$

We have reached the upper triangular R. What is Q? Move the plane rotations Q_{ij} to the other side to find $A = QR$—just as you moved the elimination matrices E_{ij} to the other side to find $A = LU$:

$$Q_{32}Q_{31}Q_{21}A = R \qquad \text{means} \qquad A = (Q_{21}^{-1}Q_{31}^{-1}Q_{32}^{-1})R = QR. \tag{3}$$

The inverse of each Q_{ij} is Q_{ij}^{T} (rotation through $-\theta$). The inverse of E_{ij} was not an orthogonal matrix! LU and QR are similar but L and Q are not the same.

Householder reflections are faster than rotations because each one clears out a whole column below the diagonal. Watch how the first column $\boldsymbol{a}_1$ of A becomes column $\boldsymbol{r}_1$ of R:

Reflection by H_1

$$H_1 = I - 2\boldsymbol{u}_1\boldsymbol{u}_1^{\mathrm{T}} \qquad H_1\,\boldsymbol{a}_1 = \begin{bmatrix} \|\boldsymbol{a}_1\| \\ 0 \\ \cdot \\ 0 \end{bmatrix} \text{ or } \begin{bmatrix} -\|\boldsymbol{a}_1\| \\ 0 \\ \cdot \\ 0 \end{bmatrix} = \boldsymbol{r}_1. \tag{4}$$

The length was not changed, and $\boldsymbol{u}_1$ is in the direction of $\boldsymbol{a}_1 - \boldsymbol{r}_1$. We have $n-1$ entries in the unit vector $\boldsymbol{u}_1$ to get $n-1$ zeros in $\boldsymbol{r}_1$. (Rotations had one angle θ to get one zero.) When we reach column k, we have $n-k$ available choices in the unit vector $\boldsymbol{u}_k$. This leads to $n-k$ zeros in $\boldsymbol{r}_k$. *We just store the $\boldsymbol{u}$'s and $\boldsymbol{r}$'s to know the final Q and R:*

Inverse of H_i is H_i $\quad (H_{n-1}\dots H_1)A = R$ means $A = (H_1\dots H_{n-1})R = QR.$ (5)

This is how LAPACK improves on 19th century Gram-Schmidt. Q is *exactly* orthogonal.

Section 11.3 explains how $A = QR$ is used in the other big computation of linear algebra—***the eigenvalue problem***. The factors QR are reversed to give $A_1 = RQ$ which is $Q^{-1}AQ$. Since A_1 is similar to A, the eigenvalues are unchanged. Then A_1 is factored into Q_1R_1, and reversing the factors gives A_2. Amazingly, the entries below the diagonal get smaller in $A_1, A_2, A_3, \dots$ and we can identify the eigenvalues. This is the "QR method" for $A\boldsymbol{x} = \lambda\boldsymbol{x}$, a big success of numerical linear algebra.

Problem Set 11.1

1 Find the two pivots with and without row exchange to maximize the pivot:

$$A = \begin{bmatrix} .001 & 0 \\ 1 & 1000 \end{bmatrix}.$$

With row exchanges to maximize pivots, why are no entries of L larger than 1? Find a 3 by 3 matrix A with all $|a_{ij}| \le 1$ and $|\ell_{ij}| \le 1$ but third pivot $= 4$.

2 Compute the exact inverse of the Hilbert matrix A by elimination. Then compute A^{-1} again by rounding all numbers to three figures:

Ill-conditioned matrix $\qquad A = \text{hilb(3)} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix}.$

3 For the same A compute $\boldsymbol{b} = A\boldsymbol{x}$ for $\boldsymbol{x} = (1, 1, 1)$ and $\boldsymbol{x} = (0, 6, -3.6)$. A small change $\Delta\boldsymbol{b}$ produces a large change $\Delta\boldsymbol{x}$.

4 Find the eigenvalues (by computer) of the 8 by 8 Hilbert matrix $a_{ij} = 1/(i + j - 1)$. In the equation $A\boldsymbol{x} = \boldsymbol{b}$ with $\|\boldsymbol{b}\| = 1$, how large can $\|\boldsymbol{x}\|$ be? If $\boldsymbol{b}$ has roundoff error less than 10^{-16}, how large an error can this cause in $\boldsymbol{x}$? See Section 9.2.

5 For back substitution with a band matrix (width w), show that the number of multiplications to solve $U\boldsymbol{x} = \boldsymbol{c}$ is approximately wn.

6 If you know L and U and Q and R, is it faster to solve $LU\boldsymbol{x} = \boldsymbol{b}$ or $QR\boldsymbol{x} = \boldsymbol{b}$?

7 Show that the number of multiplications to invert an upper triangular n by n matrix is about $\frac{1}{6}n^3$. Use back substitution on the columns of I, upward from 1's.

8 Choosing the largest available pivot in each column (partial pivoting), factor each A into $PA = LU$:

$$A = \begin{bmatrix} 1 & 0 \\ 2 & 2 \end{bmatrix} \qquad \text{and} \qquad A = \begin{bmatrix} 1 & 0 & 1 \\ 2 & 2 & 0 \\ 0 & 2 & 0 \end{bmatrix}.$$

9 Put 1's on the three central diagonals of a 4 by 4 tridiagonal matrix. Find the cofactors of the six zero entries. Those entries are nonzero in A^{-1}.

10 (Suggested by C. Van Loan.) Find the LU factorization and solve by elimination when $\varepsilon = 10^{-3}, 10^{-6}, 10^{-9}, 10^{-12}, 10^{-15}$:

$$\begin{bmatrix} \varepsilon & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 + \varepsilon \\ 2 \end{bmatrix}.$$

The true $\boldsymbol{x}$ is $(1, 1)$. Make a table to show the error for each ε. Exchange the two equations and solve again—the errors should almost disappear.

11 (a) Choose $\sin\theta$ and $\cos\theta$ to triangularize A, and find R:

Givens rotation $$Q_{21}A = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 3 & 5 \end{bmatrix} = \begin{bmatrix} * & * \\ 0 & * \end{bmatrix} = R.$$

(b) Choose $\sin\theta$ and $\cos\theta$ to make QAQ^{-1} triangular. What are the eigenvalues?

12 When A is multiplied by a plane rotation Q_{ij}, which entries of A are changed? When $Q_{ij}A$ is multiplied on the right by Q_{ij}^{-1}, which entries are changed now?

13 How many multiplications and how many additions are used to compute $Q_{ij}A$? Careful organization of the whole sequence of rotations gives $\frac{2}{3}n^3$ multiplications and $\frac{2}{3}n^3$ additions—the same as for QR by reflectors and twice as many as for LU.

Challenge Problems

14 (**Turning a robot hand**) The robot produces any 3 by 3 rotation A from plane rotations around the x, y, z axes. Then $Q_{32}Q_{31}Q_{21}A = R$, where A is orthogonal so R is I! The three robot turns are in $A = Q_{21}^{-1}Q_{31}^{-1}Q_{32}^{-1}$. The three angles are "Euler angles" and $\det Q = 1$ to avoid reflection. Start by choosing $\cos\theta$ and $\sin\theta$ so that

$$Q_{21}A = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \frac{1}{3} \begin{bmatrix} -1 & 2 & 2 \\ 2 & -1 & 2 \\ 2 & 2 & -1 \end{bmatrix} \text{ is zero in the } (2,1) \text{ position.}$$

15 Create the 10 by 10 second difference matrix $K = \mathbf{toeplitz}([2\ -1\ \mathbf{zeros}(1,8)])$. Permute rows and columns randomly by $KK = K(\mathbf{randperm}(10), \mathbf{randperm}(10))$. Factor by $[L, U] = \mathbf{lu}(K)$ and $[LL, UU] = \mathbf{lu}(KK)$, and count nonzeros by $\mathbf{nnz}(L)$ and $\mathbf{nnz}(LL)$. In this case L is in perfect tridiagonal order, but not LL.

16 Another ordering for this matrix K colors the meshpoints alternately red and black. This permutation P changes the normal $1, \ldots, 10$ to $1, 3, 5, 7, 9, 2, 4, 6, 8, 10$:

Red-black ordering $$PKP^{\mathrm{T}} = \begin{bmatrix} 2I & D \\ D^{\mathrm{T}} & 2I \end{bmatrix}. \quad \textit{Find the matrix } D.$$

So many interesting experiments are possible. If you send good ideas they can go on the linear algebra website math.mit.edu/linearalgebra. I also recommend learning the command $B = \mathbf{sparse}(A)$, after which $\mathbf{find}(B)$ will list the nonzero entries and $\boldsymbol{\ell}\mathbf{u}(B)$ will factor B using that sparse format for L and U. Only the nonzeros are computed, where ordinary (dense) MATLAB computes all the zeros too.

17 Jeff Stuart has created a student activity that brilliantly demonstrates ill-conditioning:

$$\begin{bmatrix} 1 & 1.0001 \\ 1 & 1.0000 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3.0001 + \boldsymbol{e} \\ 3.0000 + \boldsymbol{E} \end{bmatrix} \quad \begin{matrix} \textbf{With errors} \\ \boldsymbol{e} \textbf{ and } \boldsymbol{E} \end{matrix} \quad \begin{matrix} x = 2 - 10000(\boldsymbol{e} - \boldsymbol{E}) \\ y = 1 + 10000(\boldsymbol{e} - \boldsymbol{E}) \end{matrix}$$

When those equations are shown by nearly parallel long sticks, a small shake gives a big jump in the crossing point (x, y). Errors $\boldsymbol{e}$ and $\boldsymbol{E}$ are amplified by 10000.

11.2 Norms and Condition Numbers

How do we measure the size of a matrix? For a vector, the length is $\|\boldsymbol{x}\|$. For a matrix, ***the norm is*** $\|A\|$. This word "*norm*" is sometimes used for vectors, instead of length. It is always used for matrices, and there are many ways to measure $\|A\|$. We look at the requirements on all "matrix norms" and then choose one.

Frobenius squared all the $|a_{ij}|^2$ and added; his norm $\|A\|_{\mathrm{F}}$ is the square root. This treats A like a long vector with n^2 components: sometimes useful, but not the choice here.

I prefer to start with a vector norm. The triangle inequality says that $\|\boldsymbol{x}+\boldsymbol{y}\|$ is not greater than $\|\boldsymbol{x}\|+\|\boldsymbol{y}\|$. The length of $2\boldsymbol{x}$ or $-2\boldsymbol{x}$ is doubled to $2\|\boldsymbol{x}\|$. The same rules will apply to matrix norms:

$$\|A+B\| \le \|A\| + \|B\| \quad \text{and} \quad \|cA\| = |c|\,\|A\|. \tag{1}$$

The second requirements for a matrix norm are new, because matrices multiply. The norm $\|A\|$ controls the growth from $\boldsymbol{x}$ to $A\boldsymbol{x}$, and from B to AB:

Growth factor $\|A\|$
$$\|A\boldsymbol{x}\| \le \|A\|\,\|\boldsymbol{x}\| \quad \text{and} \quad \|AB\| \le \|A\|\,\|B\|. \tag{2}$$

This leads to a natural way to define $\|A\|$, the norm of a matrix:

The norm of A is the largest ratio $\|A\boldsymbol{x}\|/\|\boldsymbol{x}\|$:
$$\|A\| = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|}. \tag{3}$$

$\|A\boldsymbol{x}\|/\|\boldsymbol{x}\|$ is never larger than $\|A\|$ (its maximum). This says that $\|A\boldsymbol{x}\| \le \|A\|\,\|\boldsymbol{x}\|$.

Example 1 If A is the identity matrix I, the ratios are $\|\boldsymbol{x}\|/\|\boldsymbol{x}\|$. Therefore $\|I\| = 1$. If A is an orthogonal matrix Q, lengths are again preserved: $\|Q\boldsymbol{x}\| = \|\boldsymbol{x}\|$. The ratios still give $\|Q\| = 1$. An orthogonal Q is good to compute with: errors don't grow.

Example 2 The norm of a diagonal matrix is its largest entry (using absolute values):

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \quad \text{has norm} \quad \|A\| = 3. \quad \text{The eigenvector} \quad \boldsymbol{x} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \text{has} \quad A\boldsymbol{x} = 3\boldsymbol{x}.$$

The eigenvalue is 3. For this A (but not all A), the largest eigenvalue equals the norm.

For a positive definite symmetric matrix the norm is $\|A\| = \lambda_{\max}(A)$.

Choose $\boldsymbol{x}$ to be the eigenvector with maximum eigenvalue. Then $\|A\boldsymbol{x}\|/\|\boldsymbol{x}\|$ equals $\lambda_{\max}$. The point is that no other $\boldsymbol{x}$ can make the ratio larger. The matrix is $A = Q\Lambda Q^{\mathrm{T}}$, and the orthogonal matrices Q and Q^{T} leave lengths unchanged. So the ratio to maximize is really $\|\Lambda\boldsymbol{x}\|/\|\boldsymbol{x}\|$. The norm is the largest eigenvalue in the diagonal Λ.

Symmetric matrices Suppose A is symmetric but not positive definite. $A = Q\Lambda Q^{\mathrm{T}}$ is still true. Then the norm is the largest of $|\lambda_1|, |\lambda_2|, \ldots, |\lambda_n|$. We take absolute values, because the norm is only concerned with length. For an eigenvector $\|A\boldsymbol{x}\| = \|\lambda\boldsymbol{x}\| = |\lambda|$ times $\|\boldsymbol{x}\|$. The $\boldsymbol{x}$ that gives the maximum ratio is the eigenvector for the maximum $|\lambda|$.

Unsymmetric matrices If A is not symmetric, its eigenvalues may not measure its true size. *The norm can be larger than any eigenvalue.* A very unsymmetric example has $\lambda_1 = \lambda_2 = 0$ but its norm is not zero:

$$\|\boldsymbol{A}\| > \boldsymbol{\lambda_{\max}} \quad A = \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix} \qquad \text{has norm} \qquad \|A\| = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = 2.$$

The vector $\boldsymbol{x} = (0, 1)$ gives $A\boldsymbol{x} = (2, 0)$. The ratio of lengths is $2/1$. This is the maximum ratio $\|A\|$, even though $\boldsymbol{x}$ is not an eigenvector.

It is the ***symmetric matrix*** $A^{\mathrm{T}}A$, not the unsymmetric A, that has eigenvector $\boldsymbol{x} = (0, 1)$. The norm is really decided by ***the largest eigenvalue of*** $A^{\mathrm{T}}A$:

The norm of A (symmetric or not) ***is the square root of*** $\boldsymbol{\lambda_{\max}}(A^{\mathrm{T}}A)$**:**

$$\|A\|^2 = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|A\boldsymbol{x}\|^2}{\|\boldsymbol{x}\|^2} = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\boldsymbol{x}^{\mathrm{T}}A^{\mathrm{T}}A\boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}} = \lambda_{\max}(A^{\mathrm{T}}A). \tag{4}$$

The unsymmetric example with $\lambda_{\max}(A) = 0$ has $\lambda_{\max}(A^{\mathrm{T}}A) = 4$:

$$A = \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix} \text{ leads to } A^{\mathrm{T}}A = \begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix} \text{ with } \lambda_{\max} = 4. \text{ So the norm is } \|A\| = \sqrt{4}.$$

For any A Choose $\boldsymbol{x}$ to be the eigenvector of $A^{\mathrm{T}}A$ with largest eigenvalue $\lambda_{\max}$. The ratio in equation (4) is $\boldsymbol{x}^{\mathrm{T}}A^{\mathrm{T}}A\boldsymbol{x} = \boldsymbol{x}^{\mathrm{T}}(\lambda_{\max})\boldsymbol{x}$ divided by $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$. This is $\lambda_{\max}$.

No $\boldsymbol{x}$ can give a larger ratio. The symmetric matrix $A^{\mathrm{T}}A$ has eigenvalues $\lambda_1, \ldots, \lambda_n$ and orthonormal eigenvectors $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n$. Every $\boldsymbol{x}$ is a combination of those vectors. Try this combination in the ratio and remember that $\boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{q}_j = 0$:

$$\frac{\boldsymbol{x}^{\mathrm{T}}A^{\mathrm{T}}A\boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}} = \frac{(c_1\boldsymbol{q}_1 + \cdots + c_n\boldsymbol{q}_n)^{\mathrm{T}}(c_1\lambda_1\boldsymbol{q}_1 + \cdots + c_n\lambda_n\boldsymbol{q}_n)}{(c_1\boldsymbol{q}_1 + \cdots + c_n\boldsymbol{q}_n)^{\mathrm{T}}(c_1\boldsymbol{q}_1 + \cdots + c_n\boldsymbol{q}_n)} = \frac{c_1^2\lambda_1 + \cdots + c_n^2\lambda_n}{c_1^2 + \cdots + c_n^2}.$$

The maximum ratio $\lambda_{\max}$ is when all c's are zero, except the one that multiplies $\lambda_{\max}$.

Note 1 The ratio in equation (4) is the ***Rayleigh quotient*** for the symmetric matrix $A^{\mathrm{T}}A$. Its maximum is the largest eigenvalue $\lambda_{\max}(A^{\mathrm{T}}A)$. The minimum ratio is $\lambda_{\min}(A^{\mathrm{T}}A)$. If you substitute any vector $\boldsymbol{x}$ into the Rayleigh quotient $\boldsymbol{x}^{\mathrm{T}}A^{\mathrm{T}}A\boldsymbol{x}/\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$, you are guaranteed to get a number between $\lambda_{\min}(A^{\mathrm{T}}A)$ and $\lambda_{\max}(A^{\mathrm{T}}A)$.

Note 2 The norm $\|A\|$ equals the ***largest singular value*** $\sigma_{\max}$ of A. The singular values $\sigma_1, \ldots, \sigma_r$ are the square roots of the positive eigenvalues of $A^{\mathrm{T}}A$. So certainly $\sigma_{\max} = (\lambda_{\max})^{1/2}$. Since U and V are orthogonal in $A = U\Sigma V^{\mathrm{T}}$, the norm is $\|\boldsymbol{A}\| = \boldsymbol{\sigma_{\max}}$.

The Condition Number of A

Section 9.1 showed that roundoff error can be serious. Some systems are sensitive, others are not so sensitive. The sensitivity to error is measured by the ***condition number***. This is the first chapter in the book which intentionally introduces errors. We want to estimate how much they change $\boldsymbol{x}$.

The original equation is $A\boldsymbol{x} = \boldsymbol{b}$. Suppose the right side is changed to $\boldsymbol{b} + \Delta\boldsymbol{b}$ because of roundoff or measurement error. The solution is then changed to $\boldsymbol{x} + \Delta\boldsymbol{x}$. Our goal is to estimate the change $\Delta\boldsymbol{x}$ in the solution from the change $\Delta\boldsymbol{b}$ in the equation. Subtraction gives the *error equation* $A(\Delta\boldsymbol{x}) = \Delta\boldsymbol{b}$:

$$\text{Subtract } A\boldsymbol{x} = \boldsymbol{b} \text{ from } A(\boldsymbol{x} + \Delta\boldsymbol{x}) = \boldsymbol{b} + \Delta\boldsymbol{b} \text{ to find } \boxed{A(\Delta\boldsymbol{x}) = \Delta\boldsymbol{b}.} \tag{5}$$

The error is $\Delta\boldsymbol{x} = A^{-1}\Delta\boldsymbol{b}$. It is large when A^{-1} is large (then A is nearly singular). The error $\Delta\boldsymbol{x}$ is especially large when $\Delta\boldsymbol{b}$ points in the worst direction—which is amplified most by A^{-1}. ***The worst error has*** $\|\Delta\boldsymbol{x}\| = \|A^{-1}\|\,\|\Delta\boldsymbol{b}\|$.

This error bound $\|A^{-1}\|$ has one serious drawback. If we multiply A by 1000, then A^{-1} is divided by 1000. The matrix looks a thousand times better. But a simple rescaling cannot change the reality of the problem. It is true that $\Delta\boldsymbol{x}$ will be divided by 1000, but so will the exact solution $\boldsymbol{x} = A^{-1}\boldsymbol{b}$. The ***relative error*** $\|\Delta\boldsymbol{x}\|/\|\boldsymbol{x}\|$ will stay the same. It is this relative change in $\boldsymbol{x}$ that should be compared to the relative change in $\boldsymbol{b}$.

Comparing relative errors will now lead to the "condition number" $c = \|A\|\,\|A^{-1}\|$. Multiplying A by 1000 does not change this number, because A^{-1} is divided by 1000 and the condition number c stays the same. It measures the sensitivity of $A\boldsymbol{x} = \boldsymbol{b}$.

The solution error is less than $c = \|A\|\,\|A^{-1}\|$ ***times the problem error:***

Condition number c
$$\boxed{\frac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \le c\frac{\|\Delta\boldsymbol{b}\|}{\|\boldsymbol{b}\|}.} \tag{6}$$

If the problem error is ΔA (error in A instead of $\boldsymbol{b}$), still c controls $\Delta\boldsymbol{x}$:

Error ΔA in A
$$\boxed{\frac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x} + \Delta\boldsymbol{x}\|} \le c\frac{\|\Delta A\|}{\|A\|}.} \tag{7}$$

Proof The original equation is $\boldsymbol{b} = A\boldsymbol{x}$. The error equation (5) is $\Delta\boldsymbol{x} = A^{-1}\Delta\boldsymbol{b}$. Apply the key property $\|A\boldsymbol{x}\| \le \|A\|\|\boldsymbol{x}\|$ of matrix norms:

$$\|\boldsymbol{b}\| \le \|A\|\,\|\boldsymbol{x}\| \qquad \text{and} \qquad \|\Delta\boldsymbol{x}\| \le \|A^{-1}\|\,\|\Delta\boldsymbol{b}\|.$$

Multiply the left sides to get $\|\boldsymbol{b}\|\,\|\Delta\boldsymbol{x}\|$, and multiply the right sides to get $c\|\boldsymbol{x}\|\,\|\Delta\boldsymbol{b}\|$. Divide both sides by $\|\boldsymbol{b}\|\,\|\boldsymbol{x}\|$. The left side is now the relative error $\|\Delta\boldsymbol{x}\|/\|\boldsymbol{x}\|$. The right side is now the upper bound in equation (6).

The same condition number $c = \|A\|\,\|A^{-1}\|$ appears when the error is in the matrix. We have ΔA instead of $\Delta\boldsymbol{b}$ in the error equation:

Subtract $A\boldsymbol{x} = \boldsymbol{b}$ from $(A + \Delta A)(\boldsymbol{x} + \Delta\boldsymbol{x}) = \boldsymbol{b}$ to find $A(\Delta\boldsymbol{x}) = -(\Delta A)(\boldsymbol{x} + \Delta\boldsymbol{x})$.

Multiply the last equation by A^{-1} and take norms to reach equation (7):

$$\|\Delta\boldsymbol{x}\| \le \|A^{-1}\|\,\|\Delta A\|\,\|\boldsymbol{x} + \Delta\boldsymbol{x}\| \quad \text{or} \quad \frac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x} + \Delta\boldsymbol{x}\|} \le \|A\|\,\|A^{-1}\|\frac{\|\Delta A\|}{\|A\|}.$$

Conclusion Errors enter in two ways. They begin with an error ΔA or $\Delta\boldsymbol{b}$—a wrong matrix or a wrong $\boldsymbol{b}$. This problem error is amplified (a lot or a little) into the solution error $\Delta\boldsymbol{x}$. That error is bounded, relative to $\boldsymbol{x}$ itself, by the condition number c.

The error $\Delta\boldsymbol{b}$ depends on computer roundoff and on the original measurements of $\boldsymbol{b}$. The error ΔA also depends on the elimination steps. Small pivots tend to produce large errors in L and U. Then $L+\Delta L$ times $U+\Delta U$ equals $A+\Delta A$. When ΔA or the condition number is very large, the error $\Delta\boldsymbol{x}$ can be unacceptable.

Example 3 When A is symmetric, $c = \|A\|\,\|A^{-1}\|$ comes from the eigenvalues:

$$A = \begin{bmatrix} 6 & 0 \\ 0 & 2 \end{bmatrix} \text{ has norm } 6. \qquad A^{-1} = \begin{bmatrix} \frac{1}{6} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \text{ has norm } \tfrac{1}{2}.$$

This A is symmetric positive definite. Its norm is $\lambda_{\max} = 6$. The norm of A^{-1} is $1/\lambda_{\min} = \frac{1}{2}$. Multiplying norms gives the *condition number* $\|A\|\,\|A^{-1}\| = \lambda_{\max}/\lambda_{\min}$:

Condition number for positive definite A $$c = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{6}{2} = 3.$$

Example 4 Keep the same A, with eigenvalues 6 and 2. To make $\boldsymbol{x}$ small, choose $\boldsymbol{b}$ along the first eigenvector $(1, 0)$. To make $\Delta\boldsymbol{x}$ large, choose $\Delta\boldsymbol{b}$ along the second eigenvector $(0, 1)$. Then $\boldsymbol{x} = \frac{1}{6}\boldsymbol{b}$ and $\Delta\boldsymbol{x} = \frac{1}{2}\Delta\boldsymbol{b}$. The ratio $\|\Delta\boldsymbol{x}\|/\|\boldsymbol{x}\|$ is exactly $c = 3$ times the ratio $\|\Delta\boldsymbol{b}\|/\|\boldsymbol{b}\|$.

This shows that the worst error allowed by the condition number $\|A\|\,\|A^{-1}\|$ can actually happen. Here is a useful rule of thumb, experimentally verified for Gaussian elimination: *The computer can lose* $\log c$ *decimal places to roundoff error.*

Problem Set 11.2

1 Find the norms $\|A\| = \lambda_{\max}$ and condition numbers $c = \lambda_{\max}/\lambda_{\min}$ of these positive definite matrices:

$$\begin{bmatrix} .5 & 0 \\ 0 & 2 \end{bmatrix} \qquad \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \qquad \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}.$$

2 Find the norms and condition numbers from the square roots of $\lambda_{\max}(A^{\mathrm{T}}A)$ and $\lambda_{\min}(A^{\mathrm{T}}A)$. Without positive definiteness in A, we go to $A^{\mathrm{T}}A$!

$$\begin{bmatrix} -2 & 0 \\ 0 & 2 \end{bmatrix} \qquad \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \qquad \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$

3 Explain these two inequalities from the definitions (3) of $\|A\|$ and $\|B\|$:

$$\|AB\boldsymbol{x}\| \le \|A\|\,\|B\boldsymbol{x}\| \le \|A\|\,\|B\|\,\|\boldsymbol{x}\|.$$

From the ratio of $\|AB\boldsymbol{x}\|$ to $\|\boldsymbol{x}\|$, deduce that $\|AB\| \le \|A\|\,\|B\|$. This is the key to using matrix norms. The norm of A^n is never larger than $\|A\|^n$.

4 Use $\|AA^{-1}\| \le \|A\|\,\|A^{-1}\|$ to prove that the condition number is at least 1.

5 Why is I the only symmetric positive definite matrix that has $\lambda_{\max} = \lambda_{\min} = 1$? Then the only other matrices with $\|A\| = 1$ and $\|A^{-1}\| = 1$ must have $A^{\mathrm{T}}A = I$. Those are _____ matrices: perfectly conditioned.

6 Orthogonal matrices have norm $\|Q\| = 1$. If $A = QR$ show that $\|A\| \le \|R\|$ and also $\|R\| \le \|A\|$. Then $\|A\| = \|Q\|\,\|R\|$. Find an example of $A = LU$ with $\|A\| < \|L\|\|U\|$.

7 (a) Which famous inequality gives $\|(A+B)\boldsymbol{x}\| \le \|A\boldsymbol{x}\| + \|B\boldsymbol{x}\|$ for every $\boldsymbol{x}$?

(b) Why does the definition (3) of matrix norms lead to $\|A+B\| \le \|A\| + \|B\|$?

8 Show that if λ is any eigenvalue of A, then $|\lambda| \le \|A\|$. Start from $A\boldsymbol{x} = \lambda\boldsymbol{x}$.

9 The "*spectral radius*" $\rho(A) = |\lambda_{\max}|$ is the largest absolute value of the eigenvalues. Show with 2 by 2 examples that $\rho(A+B) \le \rho(A) + \rho(B)$ and $\rho(AB) \le \rho(A)\rho(B)$ can both be *false*. The spectral radius is not acceptable as a norm.

10 (a) Explain why A and A^{-1} have the same condition number.

(b) Explain why A and A^{T} have the same norm, based on $\lambda(A^{\mathrm{T}}A)$ and $\lambda(AA^{\mathrm{T}})$.

11 Estimate the condition number of the ill-conditioned matrix $A = \begin{bmatrix} 1 & 1 \\ 1 & 1.0001 \end{bmatrix}$.

12 Why is the determinant of A no good as a norm? Why is it no good as a condition number?

13 (Suggested by C. Moler and C. Van Loan.) Compute $\boldsymbol{b} - A\boldsymbol{y}$ and $\boldsymbol{b} - A\boldsymbol{z}$ when

$$\boldsymbol{b} = \begin{bmatrix} .217 \\ .254 \end{bmatrix} \quad A = \begin{bmatrix} .780 & .563 \\ .913 & .659 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} .341 \\ -.087 \end{bmatrix} \quad \boldsymbol{z} = \begin{bmatrix} .999 \\ -1.0 \end{bmatrix}.$$

Is $\boldsymbol{y}$ closer than $\boldsymbol{z}$ to solving $A\boldsymbol{x} = \boldsymbol{b}$? Answer in two ways: Compare the *residual* $\boldsymbol{b} - A\boldsymbol{y}$ to $\boldsymbol{b} - A\boldsymbol{z}$. Then compare $\boldsymbol{y}$ and $\boldsymbol{z}$ to the true $\boldsymbol{x} = (1, -1)$. Both answers can be right. Sometimes we want a small residual, sometimes a small $\Delta\boldsymbol{x}$.

14 (a) Compute the determinant of A in Problem 13. Compute A^{-1}.

(b) If possible compute $\|A\|$ and $\|A^{-1}\|$ and show that $c > 10^6$.

Problems 15–19 are about vector norms other than the usual $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x} \cdot \boldsymbol{x}}$.

15 The "ℓ^1 norm" and the "ℓ^∞ norm" of $\boldsymbol{x} = (x_1, \ldots, x_n)$ are

$$\|\boldsymbol{x}\|_1 = |x_1| + \cdots + |x_n| \quad \text{and} \quad \|\boldsymbol{x}\|_\infty = \max_{1 \le i \le n} |x_i|.$$

Compute the norms $\|\boldsymbol{x}\|$ and $\|\boldsymbol{x}\|_1$ and $\|\boldsymbol{x}\|_\infty$ of these two vectors in $\mathbf{R}^5$:

$$\boldsymbol{x} = (1, 1, 1, 1, 1) \qquad \boldsymbol{x} = (.1, .7, .3, .4, .5).$$

16 Prove that $\|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\| \le \|\boldsymbol{x}\|_1$. Show from the Schwarz inequality that the ratios $\|\boldsymbol{x}\|/\|\boldsymbol{x}\|_\infty$ and $\|\boldsymbol{x}\|_1/\|\boldsymbol{x}\|$ are never larger than $\sqrt{n}$. Which vector $(x_1, \ldots, x_n)$ gives ratios equal to $\sqrt{n}$?

17 All vector norms must satisfy the *triangle inequality*. Prove that

$$\|\boldsymbol{x} + \boldsymbol{y}\|_\infty \le \|\boldsymbol{x}\|_\infty + \|\boldsymbol{y}\|_\infty \qquad \text{and} \qquad \|\boldsymbol{x} + \boldsymbol{y}\|_1 \le \|\boldsymbol{x}\|_1 + \|\boldsymbol{y}\|_1.$$

18 Vector norms must also satisfy $\|c\boldsymbol{x}\| = |c|\,\|\boldsymbol{x}\|$. The norm must be positive except when $\boldsymbol{x} = \boldsymbol{0}$. Which of these are norms for vectors (x_1, x_2) in $\mathbf{R}^2$?

$$\|\boldsymbol{x}\|_A = |x_1| + 2|x_2| \qquad \|\boldsymbol{x}\|_B = \min(|x_1|, |x_2|)$$

$$\|\boldsymbol{x}\|_C = \|\boldsymbol{x}\| + \|\boldsymbol{x}\|_\infty \qquad \|\boldsymbol{x}\|_D = \|A\boldsymbol{x}\| \quad \text{(this answer depends on } A\text{)}.$$

Challenge Problems

19 Show that $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} \le \|\boldsymbol{x}\|_1\,\|\boldsymbol{y}\|_\infty$ by choosing components $y_i = \pm 1$ to make $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}$ as large as possible.

20 The eigenvalues of the $-1, 2, -1$ difference matrix K are $\lambda = 2 - 2\cos(j\pi/n{+}1)$. Estimate $\lambda_{\min}$ and $\lambda_{\max}$ and $c = \mathbf{cond}(K) = \lambda_{\max}/\lambda_{\min}$ as n increases: $c \approx Cn^2$ with what constant C?

Test this estimate with $\mathbf{eig}(K)$ and $\mathbf{cond}(K)$ for $n = 10, 100, 1000$.

11.3 Iterative Methods and Preconditioners

Up to now, our approach to $A\boldsymbol{x} = \boldsymbol{b}$ has been direct. We accepted $\boldsymbol{A}$ as it came. We attacked it by elimination with row exchanges. We now look at **iterative methods, which replace $\boldsymbol{A}$ by a simpler matrix $\boldsymbol{S}$**. The difference $T = S - A$ is moved over to the right side of the equation. The problem becomes easier to solve, with S instead of A. But there is a price—*the simpler system has to be solved over and over.*

An iterative method is easy to invent. Just split A (carefully) into $S - T$.

$$\textbf{Rewrite } A\boldsymbol{x} = \boldsymbol{b} \qquad S\boldsymbol{x} = T\boldsymbol{x} + \boldsymbol{b}. \tag{1}$$

The novelty is to solve (1) iteratively. Each guess $\boldsymbol{x}_k$ leads to the next $\boldsymbol{x}_{k+1}$:

$$\textbf{Pure iteration} \qquad S\boldsymbol{x}_{k+1} = T\boldsymbol{x}_k + \boldsymbol{b}. \tag{2}$$

Start with any $\boldsymbol{x}_0$. Then solve $S\boldsymbol{x}_1 = T\boldsymbol{x}_0 + \boldsymbol{b}$. Continue to $S\boldsymbol{x}_2 = T\boldsymbol{x}_1 + \boldsymbol{b}$. A hundred iterations are very common—often more. Stop when (and if!) $\boldsymbol{x}_{k+1}$ is sufficiently close to $\boldsymbol{x}_k$—or when the **residual** $\boldsymbol{r}_k = \boldsymbol{b} - A\boldsymbol{x}_k$ is near zero. Our hope is to get near the true solution, more quickly than by elimination. When the $\boldsymbol{x}_k$ converge, their limit x_∞ does solve equation (1): $S\boldsymbol{x}_\infty = T\boldsymbol{x}_\infty + \boldsymbol{b}$ means $A\boldsymbol{x}_\infty = \boldsymbol{b}$.

The two goals of the splitting $A = S - T$ are ***speed per step*** and ***fast convergence***. The speed of each step depends on S and the speed of convergence depends on $S^{-1}T$:

1 Equation (2) should be easy to solve for $\boldsymbol{x}_{k+1}$. The "***preconditioner***" S could be the diagonal or triangular part of A. A fast way uses $S = L_0U_0$, where those factors have many zeros compared to the exact $A = LU$. This is "*incomplete LU*".

2 The difference $\boldsymbol{x} - \boldsymbol{x}_k$ (this is the error $\boldsymbol{e}_k$) should go quickly to zero. Subtracting equation (2) from (1) cancels $\boldsymbol{b}$, and it leaves the ***equation for the error*** $\boldsymbol{e}_k$:

$$\textbf{Error equation} \qquad S\boldsymbol{e}_{k+1} = T\boldsymbol{e}_k \quad \text{which means} \quad \boldsymbol{e}_{k+1} = S^{-1}T\boldsymbol{e}_k. \tag{3}$$

At every step the error is multiplied by $S^{-1}T$. If $S^{-1}T$ is small, its powers go quickly to zero. But what is "small"?

The extreme splitting is $S = A$ and $T = 0$. Then the first step of the iteration is the original $A\boldsymbol{x} = \boldsymbol{b}$. Convergence is perfect and $S^{-1}T$ is zero. But the cost of that step is what we wanted to avoid. The choice of S is a battle between speed per step (a simple S) and fast convergence (S close to A). Here are some choices of S:

J $S =$ diagonal part of A (the iteration is called *Jacobi's method*)

GS $S =$ lower triangular part of A including the diagonal (*Gauss-Seidel method*)

ILU $S =$ approximate L times approximate U (*incomplete $L\,U$ method*).

Our first question is pure linear algebra: ***When do the x_k's converge to x?*** The answer uncovers the number $|\lambda|_{\max}$ that controls convergence. In examples of Jacobi and Gauss-Seidel, we will compute this "*spectral radius*" $|\lambda|_{\max}$. It is the largest eigenvalue of the ***iteration matrix*** $B = S^{-1}T$.

The Spectral Radius $\rho(B)$ Controls Convergence

Equation (3) is $e_{k+1} = S^{-1}Te_k$. Every iteration step multiplies the error by the same matrix $B = S^{-1}T$. The error after k steps is $e_k = B^k e_0$. ***The error approaches zero if the powers of $B = S^{-1}T$ approach zero.*** It is beautiful to see how the eigenvalues of B—the largest eigenvalue in particular—control the matrix powers B^k.

The powers B^k approach zero if and only if every eigenvalue of B has $|\lambda| < 1$.
The rate of convergence is controlled by the spectral radius of B: $\rho = \max |\lambda(B)|$.

The test for convergence is $|\lambda|_{\max} < 1$. Real eigenvalues must lie between -1 and 1. Complex eigenvalues $\lambda = a + ib$ must have $|\lambda|^2 = a^2 + b^2 < 1$. The spectral radius "*rho*" is the largest distance from 0 to the eigenvalues of $B = S^{-1}T$. This is $\rho = |\lambda|_{\max}$.

To see why $|\lambda|_{\max} < 1$ is necessary, suppose the starting error e_0 happens to be an eigenvector of B. After one step the error is $Be_0 = \lambda e_0$. After k steps the error is $B^k e_0 = \lambda^k e_0$. If we start with an eigenvector, we continue with that eigenvector—and the *factor* λ^k *only goes to zero when* $|\lambda| < 1$. This condition is required of every eigenvalue.

To see why $|\lambda|_{\max} < 1$ is sufficient for the error to approach zero, suppose e_0 is a combination of eigenvectors:

$$e_0 = c_1 x_1 + \cdots + c_n x_n \quad \text{leads to} \quad e_k = c_1(\lambda_1)^k x_1 + \cdots + c_n(\lambda_n)^k x_n. \tag{4}$$

This is the point of eigenvectors! When we multiply by B, each eigenvector x_i is multiplied by λ_i. If all $|\lambda_i| < 1$ then equation (4) ensures that e_k goes to zero.

Example 1 $B = \begin{bmatrix} .6 & .5 \\ .6 & .5 \end{bmatrix}$ has $\lambda_{\max} = 1.1$ $\qquad B' = \begin{bmatrix} .6 & 1.1 \\ 0 & .5 \end{bmatrix}$ has $\lambda_{\max} = .6$

B^2 is 1.1 times B. Then B^3 is $(1.1)^2$ times B. The powers of B will blow up. Contrast with the powers of B'. The matrix $(B')^k$ has $(.6)^k$ and $(.5)^k$ on its diagonal. The off-diagonal entries also involve $\rho^k = (.6)^k$, which sets the speed of convergence.

Note When there are too few eigenvectors, equation (4) is not correct. We turn to the *Jordan form* when eigenvectors are missing and the matrix B can't be diagonalized:

Jordan form J $\qquad B = MJM^{-1}$ $\qquad$ and $\qquad B^k = MJ^kM^{-1}$. (5)

Section 8.3 shows how J and J^k are made of "blocks" with one repeated eigenvalue:

The powers of a 2 by 2 block in J are $\begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}^k = \begin{bmatrix} \lambda^k & k\lambda^{k-1} \\ 0 & \lambda^k \end{bmatrix}$.

If $|\lambda| < 1$ then these powers approach zero. The extra factor k from a double eigenvalue is overwhelmed by the decreasing factor λ^{k-1}. This applies to every block:

Diagonalizable or not: Convergence $B^k \to 0$ and its speed depend on $\rho = |\lambda|_{\max} < 1$.

Jacobi versus Gauss-Seidel

We now solve a specific 2 by 2 problem by splitting A. Watch for that number $|\lambda|_{\max}$.

$$A\boldsymbol{x} = \boldsymbol{b} \qquad \begin{matrix} 2u - v = 4 \\ -u + 2v = -2 \end{matrix} \quad \text{has the solution} \quad \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}. \tag{6}$$

The first splitting is ***Jacobi's method***. Keep the *diagonal* of A on the left side (this is S). Move the off-diagonal part of A to the right side (this is T). Then iterate:

Jacobi iteration $S\boldsymbol{x}_{k+1} = T\boldsymbol{x}_k + \boldsymbol{b}$ $\qquad \begin{matrix} 2u_{k+1} = v_k + 4 \\ 2v_{k+1} = u_k - 2. \end{matrix}$

Start from $u_0 = v_0 = 0$. The first step finds $u_1 = 2$ and $v_1 = -1$. Keep going:

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 2 \\ -1 \end{bmatrix} \quad \begin{bmatrix} 3/2 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 2 \\ -1/4 \end{bmatrix} \quad \begin{bmatrix} 15/8 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 2 \\ -1/16 \end{bmatrix} \quad \text{approaches} \quad \begin{bmatrix} 2 \\ 0 \end{bmatrix}.$$

This shows convergence. At steps 1, 3, 5 the second component is -1, $-1/4$, $-1/16$. Those drop by 4 in each two steps. *The error equation is* $S\boldsymbol{e}_{k+1} = T\boldsymbol{e}_k$:

$$\textbf{Error equation} \qquad \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \boldsymbol{e}_{k+1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \boldsymbol{e}_k \quad \text{or} \quad \boldsymbol{e}_{k+1} = \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix} \boldsymbol{e}_k. \tag{7}$$

That last matrix $S^{-1}T$ has eigenvalues $\frac{1}{2}$ and $-\frac{1}{2}$. So its spectral radius is $\rho(B) = \frac{1}{2}$:

$$B = S^{-1}T = \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix} \quad \text{has } |\lambda|_{\max} = \tfrac{1}{2} \quad \text{and} \quad \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix}^2 = \begin{bmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{4} \end{bmatrix}.$$

Two steps multiply the error by $\frac{1}{4}$ exactly, in this special example. The important message is this: Jacobi's method works well when the main diagonal of A is large compared to the off-diagonal part. The diagonal part is S, the rest is $-T$. We want the diagonal to dominate.

The eigenvalue $\lambda = \frac{1}{2}$ is unusually small. Ten iterations reduce the error by $2^{10} = 1024$. More typical and more expensive is $|\lambda|_{\max} = .99$ or $.999$.

The ***Gauss-Seidel method*** keeps the whole lower triangular part of A as S:

$$\textbf{Gauss-Seidel} \qquad \begin{matrix} 2u_{k+1} = v_k + 4 \\ -u_{k+1} + 2v_{k+1} = -2 \end{matrix} \quad \text{or} \quad \begin{matrix} u_{k+1} = \frac{1}{2}v_k + 2 \\ v_{k+1} = \frac{1}{2}u_{k+1} - 1. \end{matrix} \tag{8}$$

Notice the change. The new u_{k+1} from the first equation is used *immediately* in the second equation. With Jacobi, we saved the old u_k until the whole step was complete. With Gauss-Seidel, the new values enter right away and the old u_k is destroyed. This cuts the storage in half. It also speeds up the iteration (usually). And it costs no more than the Jacobi method.

Test the iteration starting from another start $u_0 = 0$ and $v_0 = -1$:

$$\begin{bmatrix} 0 \\ -1 \end{bmatrix} \quad \begin{bmatrix} 3/2 \\ -1/4 \end{bmatrix} \quad \begin{bmatrix} 15/8 \\ -1/16 \end{bmatrix} \quad \begin{bmatrix} 63/32 \\ -1/64 \end{bmatrix} \quad \text{approaches} \quad \begin{bmatrix} 2 \\ 0 \end{bmatrix}.$$

The errors in the first component are 2, 1/2, 1/8, 1/32. The errors in the second component are -1, $-1/4$, $-1/16$, $-1/32$. We divide by 4 in *one step* not two steps. ***Gauss-Seidel is twice as fast as Jacobi***. We have $\rho_{\mathbf{GS}} = (\rho_{\mathbf{J}})^2$ when A is positive definite tridiagonal:

$$S = \begin{bmatrix} 2 & 0 \\ -1 & 2 \end{bmatrix} \quad \text{and} \quad T = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad S^{-1}T = \begin{bmatrix} \mathbf{0} & \frac{1}{2} \\ 0 & \frac{1}{4} \end{bmatrix}.$$

The Gauss-Seidel eigenvalues are 0 and $\frac{1}{4}$. Compare with $\frac{1}{2}$ and $-\frac{1}{2}$ for Jacobi.

With a small push we can describe the ***successive overrelaxation method*** (**SOR**). The new idea is to introduce a parameter ω (omega) into the iteration. Then choose this number ω to make the spectral radius of $S^{-1}T$ as small as possible.

Rewrite $A\boldsymbol{x} = \boldsymbol{b}$ as $\omega A\boldsymbol{x} = \omega \boldsymbol{b}$. The matrix S in **SOR** has the diagonal of the original A, but below the diagonal we use ωA. On the right side T is $S - \omega A$:

$$\textbf{SOR} \qquad \begin{aligned} 2u_{k+1} &= (2-2\omega)u_k + \omega v_k + 4\omega \\ -\omega u_{k+1} + 2v_{k+1} &= (2-2\omega)v_k - 2\omega. \end{aligned} \tag{9}$$

This looks more complicated to us, but the computer goes as fast as ever. SOR is like Gauss-Seidel, with an adjustable number ω. The best ω makes it faster.

I will put on record the most valuable test matrix of order n. It is our favorite -1, 2, -1 tridiagonal matrix K. The diagonal is $2I$. Below and above are -1's. Our example had $n = 2$, which leads to $\cos\frac{\pi}{3} = \frac{1}{2}$ as the Jacobi eigenvalue found above. Notice especially that this $|\lambda|_{\max}$ is squared for Gauss-Seidel:

The splittings of the $-1, 2, -1$ matrix K of order n yield these eigenvalues of B:

Jacobi ($S = 0, 2, 0$ matrix): $S^{-1}T$ has $|\lambda|_{\max} = \cos\dfrac{\pi}{n+1}$

Gauss-Seidel ($S = -1, 2, 0$ matrix): $S^{-1}T$ has $|\lambda|_{\max} = \left(\cos\dfrac{\pi}{n+1}\right)^2$

SOR (with the best ω): $S^{-1}T$ has $|\lambda|_{\max} = \left(\cos\dfrac{\pi}{n+1}\right)^2 \Big/ \left(1 + \sin\dfrac{\pi}{n+1}\right)^2$.

Let me be clear: For the -1, 2, -1 matrix you should not use any of these iterations! Elimination on a tridiagonal matrix is very fast (exact LU). Iterations are intended for a large sparse matrix that has nonzeros far from the central diagonal. Those create many more nonzeros in the exact L and U. This **fill-in** is why elimination becomes expensive.

We mention one more splitting. The idea of "***incomplete*** $L\,U$" is to set the small nonzeros in L and U *back to zero*. This leaves triangular matrices L_0 and U_0 which are again sparse. The splitting has $S = L_0U_0$ on the left side. Each step is quick:

Incomplete LU $$L_0U_0\boldsymbol{x}_{k+1} = (L_0U_0 - A)\boldsymbol{x}_k + \boldsymbol{b}.$$

On the right side we do sparse matrix-vector multiplications. Don't multiply L_0 times U_0, those are matrices. Multiply $\boldsymbol{x}_k$ by U_0 and then multiply that vector by L_0. On the left side we do forward and back substitutions. If L_0U_0 is close to A, then $|\lambda|_{\max}$ is small. A few iterations will give a close answer.

Multigrid and Conjugate Gradients

I cannot leave the impression that Jacobi and Gauss-Seidel are great methods. Generally the "low-frequency" part of the error decays very slowly, and many iterations are needed. Here are two important ideas that bring tremendous improvement. **Multigrid** can solve problems of size n in $O(n)$ steps. With a good preconditioner, **conjugate gradients** becomes one of the most popular and powerful algorithms in numerical linear algebra.

Multigrid Solve smaller problems with coarser grids. Each iteration will be cheaper and faster. Then interpolate between the coarse grid values to get a quick headstart on the full-size problem. Multigrid might go 4 levels down and back.

Conjugate gradients An ordinary iteration like $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - A\boldsymbol{x}_k + \boldsymbol{b}$ involves multiplication by A at each step. If A is sparse, this is not too expensive: $A\boldsymbol{x}_k$ is what we are willing to do. It adds one more basis vector to the growing "Krylov spaces" that contain our approximations. But $\boldsymbol{x}_{k+1}$ is **not the best combination** of $\boldsymbol{x}_0, A\boldsymbol{x}_0, \ldots, A^k\boldsymbol{x}_0$. The ordinary iterations are simple but far from optimal.

The conjugate gradient method chooses **the best combination** $\boldsymbol{x}_k$ at every step. The extra cost (beyond one multiplication by A) is not great. We will give the CG iteration, emphasizing that this method was created for a *symmetric positive definite matrix*. When A is not symmetric, one good choice is GMRES. When $A = A^{\mathrm{T}}$ is not positive definite, there is MINRES. A world of high-powered iterative methods has been created around the idea of making optimal choices of each successive $\boldsymbol{x}_k$.

My textbook *Computational Science and Engineering* describes multigrid and CG in much more detail. Among books on numerical linear algebra, Trefethen-Bau is deservedly popular (others are terrific too). Golub-Van Loan is a level up.

The Problem Set reproduces the five steps in each conjugate gradient cycle from $\boldsymbol{x}_{k-1}$ to $\boldsymbol{x}_k$. We compute that new approximation $\boldsymbol{x}_k$, the new residual $\boldsymbol{r}_k = \boldsymbol{b} - A\boldsymbol{x}_k$, and the new search direction $\boldsymbol{d}_k$ to look for the next $\boldsymbol{x}_{k+1}$.

I wrote those steps for the original matrix A. But a **preconditioner** S can make convergence much faster. Our original equation is $A\boldsymbol{x} = \boldsymbol{b}$. The preconditioned equation is $S^{-1}A\boldsymbol{x} = S^{-1}\boldsymbol{b}$. Small changes in the code give the *preconditioned conjugate gradient method*—the leading iterative method to solve positive definite systems.

The biggest competition is direct elimination, with the equations reordered to take maximum advantage of the zeros in A. It is not easy to outperform Gauss.

Iterative Methods for Eigenvalues

We move from $A\boldsymbol{x} = \boldsymbol{b}$ to $A\boldsymbol{x} = \lambda\boldsymbol{x}$. Iterations are an option for linear equations. They are a necessity for eigenvalue problems. The eigenvalues of an n by n matrix are the roots of an nth degree polynomial. The determinant of $A - \lambda I$ starts with $(-\lambda)^n$. This book must not leave the impression that eigenvalues should be computed that way! Working from $\det(A - \lambda I) = 0$ is a *very poor approach*—except when n is small.

For $n > 4$ there is no formula to solve $\det(A - \lambda I) = 0$. Worse than that, the λ's can be very unstable and sensitive. It is much better to work with A itself, gradually making it diagonal or triangular. (Then the eigenvalues appear on the diagonal.) Good computer codes are available in the LAPACK library—individual routines are free on **www.netlib.org/lapack**. This library combines the earlier LINPACK and EISPACK, with many improvements (to use matrix-matrix operations in the Level 3 BLAS). It is a collection of Fortran 77 programs for linear algebra on high-performance computers. For your computer and mine, a high quality matrix package is all we need. For supercomputers with parallel processing, move to ScaLAPACK and block elimination.

We will briefly discuss the power method and the QR method (chosen by LAPACK) for computing eigenvalues. It makes no sense to give full details of the codes.

1 Power methods and inverse power methods. Start with any vector $\boldsymbol{u}_0$. Multiply by A to find $\boldsymbol{u}_1$. Multiply by A again to find $\boldsymbol{u}_2$. If $\boldsymbol{u}_0$ is a combination of the eigenvectors, then A multiplies each eigenvector $\boldsymbol{x}_i$ by λ_i. After k steps we have $(\lambda_i)^k$:

$$\boldsymbol{u}_k = A^k\boldsymbol{u}_0 = c_1(\lambda_1)^k\boldsymbol{x}_1 + \cdots + c_n(\lambda_n)^k\boldsymbol{x}_n. \tag{10}$$

As the power method continues, ***the largest eigenvalue begins to dominate***. The vectors $\boldsymbol{u}_k$ point toward that dominant eigenvector $\boldsymbol{x}_1$. We saw this for Markov matrices:

$$A = \begin{bmatrix} .9 & .3 \\ .1 & .7 \end{bmatrix} \quad \text{has} \quad \lambda_{\max} = 1 \quad \text{with eigenvector} \quad \begin{bmatrix} .75 \\ .25 \end{bmatrix}.$$

Start with $\boldsymbol{u}_0$ and multiply at every step by A:

$$\boldsymbol{u}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \ \boldsymbol{u}_1 = \begin{bmatrix} .9 \\ .1 \end{bmatrix}, \ \boldsymbol{u}_2 = \begin{bmatrix} .84 \\ .16 \end{bmatrix} \quad \text{is approaching} \quad \boldsymbol{u}_\infty = \begin{bmatrix} .75 \\ .25 \end{bmatrix}.$$

The speed of convergence depends on the *ratio* of the second largest eigenvalue λ_2 to the largest λ_1. We don't want λ_1 to be small, we want λ_2/λ_1 to be small. Here $\lambda_2 = .6$ and $\lambda_1 = 1$, giving good speed. For large matrices it often happens that $|\lambda_2/\lambda_1|$ is very close to 1. Then the power method is too slow.

Is there a way to find the *smallest* eigenvalue—which is often the most important in applications? Yes, by the *inverse* power method: Multiply $\boldsymbol{u}_0$ by A^{-1} instead of A. Since we never want to compute A^{-1}, we actually solve $A\boldsymbol{u}_1 = \boldsymbol{u}_0$. By saving the LU factors, the next step $A\boldsymbol{u}_2 = \boldsymbol{u}_1$ is fast. Step k has $Au_k = u_{k-1}$:

Inverse power method $$u_k = A^{-k}u_0 = \frac{c_1 x_1}{(\lambda_1)^k} + \cdots + \frac{c_n x_n}{(\lambda_n)^k}. \tag{11}$$

Now the *smallest* eigenvalue $\lambda_{\min}$ is in control. When it is very small, the factor $1/\lambda_{\min}^k$ is large. For high speed, we make $\boldsymbol{\lambda}_{\min}$ even smaller by shifting the matrix to $A - \lambda^* I$.

That shift doesn't change the eigenvectors. (λ^* might come from the diagonal of A, even better is a Rayleigh quotient $x^{\mathrm{T}}Ax/x^{\mathrm{T}}x$). If λ^* is close to $\lambda_{\min}$ then $(A - \lambda^* I)^{-1}$ has the very large eigenvalue $(\lambda_{\min} - \lambda^*)^{-1}$. Each ***shifted inverse power step*** multiplies the eigenvector by this big number, and that eigenvector quickly dominates.

2 The QR Method This is a major achievement in numerical linear algebra. Sixty years ago, eigenvalue computations were slow and inaccurate. We didn't even realize that solving $\det(A - \lambda I) = 0$ was a terrible method. Jacobi had suggested earlier that A should gradually be made triangular—then the eigenvalues appear automatically on the diagonal. He used 2 by 2 rotations to produce off-diagonal zeros. (Unfortunately the previous zeros can become nonzero again. But Jacobi's method made a partial comeback with parallel computers.) The ***QR method*** is now a leader in eigenvalue computations.

The basic step is to factor A, whose eigenvalues we want, into QR. Remember from Gram-Schmidt (Section 4.4) that Q has orthonormal columns and R is triangular. For eigenvalues the key idea is: ***Reverse Q and R.*** The new matrix (same λ's) is $A_1 = RQ$. *The eigenvalues are not changed* in RQ *because* $A = QR$ *is similar to* $A_1 = Q^{-1}AQ$:

$A_1 = RQ$ has the same λ $$QRx = \lambda x \quad \text{gives} \quad RQ(Q^{-1}x) = \lambda(Q^{-1}x). \tag{12}$$

This process continues. Factor the new matrix A_1 into Q_1R_1. Then reverse the factors to R_1Q_1. This is the similar matrix A_2 and again no change in the eigenvalues. Amazingly, those eigenvalues begin to show up on the diagonal. Soon the last entry of A_4 holds an accurate eigenvalue. In that case we remove the last row and column and continue with a smaller matrix to find the next eigenvalue.

Two extra ideas make this method a success. One is to shift the matrix by a multiple of I, before factoring into QR. Then RQ is shifted back to give A_{k+1}:

$$\text{Factor } A_k - c_k I \text{ into } Q_k R_k. \text{ The next matrix is } A_{k+1} = R_k Q_k + c_k I.$$

A_{k+1} has the same eigenvalues as A_k, and the same as the original $A_0 = A$. A good shift chooses c near an (unknown) eigenvalue. That eigenvalue appears more accurately on the diagonal of A_{k+1}—which tells us a better c for the next step to A_{k+2}.

The second idea is to obtain off-diagonal zeros before the QR method starts. An elimination step E will do it, or a Givens rotation, but don't forget E^{-1} (or λ will change):

$$EAE^{-1} = \begin{bmatrix} 1 & & \\ & 1 & \\ & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ \mathbf{1} & 4 & 5 \\ \mathbf{1} & \mathbf{6} & 7 \end{bmatrix} \begin{bmatrix} 1 & & \\ & 1 & \\ & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 5 & 3 \\ \mathbf{1} & 9 & 5 \\ \mathbf{0} & \mathbf{4} & 2 \end{bmatrix}. \textbf{ Same } \boldsymbol{\lambda}\textbf{'s.}$$

We must leave those nonzeros 1 and 4 along *one subdiagonal*. More E's could remove them, but E^{-1} would fill them in again. This is a "***Hessenberg matrix***" (one nonzero

subdiagonal). The zeros in the lower left corner will stay zero through the QR method. The operation count for each QR factorization drops from $O(n^3)$ to $O(n^2)$.

Golub and Van Loan give this example of one shifted QR step on a Hessenberg matrix. The shift is $7I$, taking 7 from all diagonal entries of A (then shifting back for A_1):

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 0 & .001 & 7 \end{bmatrix} \quad \text{leads to} \quad A_1 = \begin{bmatrix} -.54 & 1.69 & 0.835 \\ .31 & 6.53 & -6.656 \\ 0 & .00002 & 7.012 \end{bmatrix}.$$

Factoring $A - 7I$ into QR produced $A_1 = RQ + 7I$. Notice the very small number .00002. The diagonal entry 7.012 is almost an exact eigenvalue of A_1, and therefore of A. Another QR step on A_1 with shift by $7.012I$ would give terrific accuracy.

For a few eigenvalues of a large sparse matrix I would look to **ARPACK**. Problems 25–27 describe the Arnoldi iteration that orthogonalizes the basis—each step has only three terms when A is symmetric. The matrix becomes *tridiagonal*: a wonderful start for computing eigenvalues.

Problem Set 11.3

Problems 1–12 are about iterative methods for $Ax = b$.

1 Change $Ax = b$ to $x = (I - A)x + b$. What are S and T for this splitting? What matrix $S^{-1}T$ controls the convergence of $x_{k+1} = (I - A)x_k + b$?

2 If λ is an eigenvalue of A, then _____ is an eigenvalue of $B = I - A$. The real eigenvalues of B have absolute value less than 1 if the real eigenvalues of A lie between _____ and _____.

3 Show why the iteration $x_{k+1} = (I - A)x_k + b$ does not converge for $A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$.

4 Why is the norm of B^k never larger than $\|B\|^k$? Then $\|B\| < 1$ guarantees that the powers B^k approach zero (convergence). No surprise since $|\lambda|_{\max}$ is below $\|B\|$.

5 If A is singular then all splittings $A = S - T$ must fail. From $Ax = 0$ show that $S^{-1}Tx = x$. So this matrix $B = S^{-1}T$ has $\lambda = 1$ and fails.

6 Change the 2's to 3's and find the eigenvalues of $S^{-1}T$ for Jacobi's method:

$$Sx_{k+1} = Tx_k + b \quad \text{is} \quad \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix} x_{k+1} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} x_k + b.$$

7 Find the eigenvalues of $S^{-1}T$ for the Gauss-Seidel method applied to Problem 6:

$$\begin{bmatrix} 3 & 0 \\ -1 & 3 \end{bmatrix} x_{k+1} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x_k + b.$$

Does $|\lambda|_{\max}$ for Gauss-Seidel equal $|\lambda|^2_{\max}$ for Jacobi?

8 For any 2 by 2 matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ show that $|\lambda|_{\max}$ equals $|bc/ad|$ for Gauss-Seidel and $|bc/ad|^{1/2}$ for Jacobi. We need $ad \neq 0$ for the matrix S to be invertible.

9 Write a computer code (MATLAB or other) for the Gauss-Seidel method. You can define S and T from A, or set up the iteration loop directly from the entries a_{ij}. Test it on the $-1, 2, -1$ matrices A of order 10, 20, 50 with $\boldsymbol{b} = (1, 0, \ldots, 0)$.

10 The Gauss-Seidel iteration at component i uses earlier parts of $\boldsymbol{x}^{\text{new}}$:

$$\textbf{Gauss-Seidel} \qquad x_i^{\text{new}} = x_i^{\text{old}} + \frac{1}{a_{ii}}\Big(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{\text{new}} - \sum_{j=i}^{n} a_{ij}x_j^{\text{old}}\Big).$$

If every $x_i^{\text{new}} = x_i^{\text{old}}$ how does this show that the solution $\boldsymbol{x}$ is correct? How does the formula change for Jacobi's method? For **SOR** insert ω outside the parentheses.

11 Divide equation (10) by λ_1^k and explain why $|\lambda_2/\lambda_1|$ controls the convergence of the power method. Construct a matrix A for which this method *does not converge*.

12 The Markov matrix $A = \begin{bmatrix} .9 & .3 \\ .1 & .7 \end{bmatrix}$ has $\lambda = 1$ and .6, and the power method $\boldsymbol{u}_k = A^k\boldsymbol{u}_0$ converges to $\begin{bmatrix} .75 \\ .25 \end{bmatrix}$. Find the eigenvectors of A^{-1}. What does the inverse power method $\boldsymbol{u}_{-k} = A^{-k}\boldsymbol{u}_0$ converge to (after you multiply by $.6^k$)?

13 The tridiagonal matrix of size $n - 1$ with diagonals $-1, 2, -1$ has eigenvalues $\lambda_j = 2 - 2\cos(j\pi/n)$. Why are the smallest eigenvalues approximately $(j\pi/n)^2$? The inverse power method converges at the speed $\lambda_1/\lambda_2 \approx 1/4$.

14 For $A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$ apply the power method $\boldsymbol{u}_{k+1} = A\boldsymbol{u}_k$ three times starting with $\boldsymbol{u}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. What eigenvector is the power method converging to?

15 For $A = -1,\ 2, -1$ matrix, apply the *inverse* power method $\boldsymbol{u}_{k+1} = A^{-1}\boldsymbol{u}_k$ three times with the same $\boldsymbol{u}_0$. What eigenvector are the $\boldsymbol{u}_k$'s approaching?

16 In the QR method for eigenvalues when A is shifted to make $A_{22} = 0$, show that the $2, 1$ entry drops from $\sin\theta$ in $A = QR$ to $-\sin^3\theta$ in RQ. (*Compute R and RQ.*) This "cubic convergence" makes the method a success:

$$A = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & 0 \end{bmatrix} = QR = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} 1 & ? \\ 0 & ? \end{bmatrix}.$$

17 If A is an orthogonal matrix, its QR factorization has $Q =$ _____ and $R =$ _____. Therefore $RQ =$ _____. These are among the rare examples when the QR method goes nowhere.

18 The shifted QR method factors $A - cI$ into QR. Show that the next matrix $A_1 = RQ + cI$ equals $Q^{-1}AQ$. Therefore A_1 has the _____ eigenvalues as A (but A_1 is closer to triangular).

19 When $A = A^{\mathrm{T}}$, the "*Lanczos method*" finds a's and b's and orthonormal $\boldsymbol{q}$'s so that $A\boldsymbol{q}_j = b_{j-1}\boldsymbol{q}_{j-1} + a_j\boldsymbol{q}_j + b_j\boldsymbol{q}_{j+1}$ (with $\boldsymbol{q}_0 = \boldsymbol{0}$). Multiply by $\boldsymbol{q}_j^{\mathrm{T}}$ to find a formula for a_j. The equation says that $AQ = QT$ where T is a tridiagonal matrix.

20 The equation in Problem 19 develops from this loop with $b_0 = 1$ and $\boldsymbol{r}_0 =$ any $\boldsymbol{q}_1$:

$$\boldsymbol{q}_{j+1} = \boldsymbol{r}_j/b_j;\ j = j+1;\ a_j = \boldsymbol{q}_j^{\mathrm{T}}A\boldsymbol{q}_j;\ \boldsymbol{r}_j = A\boldsymbol{q}_j - b_{j-1}\boldsymbol{q}_{j-1} - a_j\boldsymbol{q}_j;\ b_j = \|\boldsymbol{r}_j\|.$$

Write a code and test it on the $-1, 2, -1$ matrix A. $Q^{\mathrm{T}}Q$ should be I.

21 Suppose A is *tridiagonal and symmetric in the QR method.* From $A_1 = Q^{-1}AQ$ show that A_1 is symmetric. Write $A_1 = RAR^{-1}$ to show that A_1 is also tridiagonal. (If the lower part of A_1 is proved tridiagonal then by symmetry the upper part is too.) Symmetric tridiagonal matrices are the best way to start in the QR method.

Problems 22–25 present two fundamental iterations. Each step involves $A\boldsymbol{q}$ or $A\boldsymbol{d}$.

The key point for large matrices is that matrix-vector multiplication is much faster than matrix-matrix multiplication. A crucial construction starts with a vector $\boldsymbol{b}$. Repeated multiplication will produce $A\boldsymbol{b}, A^2\boldsymbol{b}, \ldots$ but those vectors are far from orthogonal. The "**Arnoldi iteration**" creates an orthonormal basis $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots$ for the same space by the Gram-Schmidt idea: *orthogonalize each new $A\boldsymbol{q}_n$ against the previous $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{n-1}$*. The "Krylov space" spanned by $\boldsymbol{b}, A\boldsymbol{b}, \ldots, A^{n-1}\boldsymbol{b}$ then has a much better basis $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$.

Here in pseudocode are two of the most important algorithms in numerical linear algebra: Arnoldi gives a good basis and CG gives a good approximation to $\boldsymbol{x} = A^{-1}\boldsymbol{b}$.

Arnoldi Iteration

$\boldsymbol{q}_1 = \boldsymbol{b}/\|\boldsymbol{b}\|$
for $n = 1$ **to** $N-1$
  $\boldsymbol{v} = A\boldsymbol{q}_n$
  for $j = 1$ **to** n
    $h_{jn} = \boldsymbol{q}_j^{\mathrm{T}}\boldsymbol{v}$
    $\boldsymbol{v} = \boldsymbol{v} - h_{jn}\boldsymbol{q}_j$
  $h_{n+1,n} = \|\boldsymbol{v}\|$
  $\boldsymbol{q}_{n+1} = \boldsymbol{v}/h_{n+1,n}$

Conjugate Gradient Iteration for Positive Definite A

$\boldsymbol{x}_0 = 0, \boldsymbol{r}_0 = \boldsymbol{b}, \boldsymbol{d}_0 = \boldsymbol{r}_0$
for $n = 1$ **to** N

$\alpha_n = (\boldsymbol{r}_{n-1}^{\mathrm{T}}\boldsymbol{r}_{n-1})/(\boldsymbol{d}_{n-1}^{\mathrm{T}}A\boldsymbol{d}_{n-1})$	step length $\boldsymbol{x}_{n-1}$ to $\boldsymbol{x}_n$
$\boldsymbol{x}_n = \boldsymbol{x}_{n-1} + \alpha_n\boldsymbol{d}_{n-1}$	approximate solution
$\boldsymbol{r}_n = \boldsymbol{r}_{n-1} - \alpha_n A\boldsymbol{d}_{n-1}$	new residual $\boldsymbol{b} - A\boldsymbol{x}_n$
$\beta_n = (\boldsymbol{r}_n^{\mathrm{T}}\boldsymbol{r}_n)/(\boldsymbol{r}_{n-1}^{\mathrm{T}}\boldsymbol{r}_{n-1})$	improvement this step
$\boldsymbol{d}_n = \boldsymbol{r}_n + \beta_n\boldsymbol{d}_{n-1}$	next search direction

% *Notice: only* 1 *matrix-vector multiplication $A\boldsymbol{q}$ and $A\boldsymbol{d}$*

For conjugate gradients, the residuals $\boldsymbol{r}_n$ are orthogonal and the search directions are A-orthogonal: all $\boldsymbol{d}_j^{\mathrm{T}}A\boldsymbol{d}_k = 0$. The iteration solves $A\boldsymbol{x} = \boldsymbol{b}$ by minimizing the error $\boldsymbol{e}^{\mathrm{T}}A\boldsymbol{e}$ over all vectors in the *Krylov space* = span of $\boldsymbol{b}, A\boldsymbol{b}, \ldots, A^{n-1}\boldsymbol{b}$. It is a fantastic algorithm.

22 For the diagonal matrix $A = \mathrm{diag}([1\ \ 2\ \ 3\ \ 4])$ and the vector $\boldsymbol{b} = (1,1,1,1)$, go through one Arnoldi step to find the orthonormal vectors $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$.

23 Arnoldi's method is finding Q so that $AQ = QH$ (column by column):

$$AQ = \begin{bmatrix} A\boldsymbol{q}_1 & \cdots & A\boldsymbol{q}_N \end{bmatrix} = \begin{bmatrix} \boldsymbol{q}_1 & \cdots & \boldsymbol{q}_N \end{bmatrix} \begin{bmatrix} h_{11} & h_{12} & \cdot & h_{1N} \\ h_{21} & h_{22} & \cdot & h_{2N} \\ 0 & h_{32} & \cdot & \cdot \\ 0 & 0 & \cdot & h_{NN} \end{bmatrix} = QH$$

H is a "Hessenberg matrix" with one nonzero subdiagonal. Here is the crucial fact when A is symmetric: ***The Hessenberg matrix*** $H = Q^{-1}AQ = Q^{\mathrm{T}}AQ$ ***is symmetric and therefore it is tridiagonal***. Explain that sentence.

24 This tridiagonal H (when A is symmetric) gives the **Lanczos iteration**:

$$\textbf{Three terms only} \qquad \boldsymbol{q}_{j+1} = (A\boldsymbol{q}_j - h_{j,j}\boldsymbol{q}_j - h_{j-1,j}\boldsymbol{q}_{j-1})/h_{j+1,j}$$

From $H = Q^{-1}AQ$, why are the eigenvalues of H the same as the eigenvalues of A? For large matrices, the "Lanczos method" computes the leading eigenvalues by stopping at a smaller tridiagonal matrix H_k. The QR method in the text is applied to compute the eigenvalues of H_k.

25 Apply the conjugate gradient method to solve $A\boldsymbol{x} = \boldsymbol{b} = \textbf{ones}(100, 1)$, where A is the $-1, 2, -1$ second difference matrix $A = \textbf{toeplitz}([2 \ -1 \ \textbf{zeros}(1, 98)])$. Graph $\boldsymbol{x}_{10}$ and $\boldsymbol{x}_{20}$ from CG, along with the exact solution $\boldsymbol{x}$. (Its 100 components are $x_i = (ih - i^2h^2)/2$ with $h = 1/101$. "$\textbf{plot}(i, x(i))$" should produce a parabola.)

26 For unsymmetric matrices, the spectral radius $\rho = \max|\lambda_i|$ is not a norm. But still $\|A^n\|$ grows or decays like ρ^n for large n. Compare those numbers for $A = [1 \quad 1; \quad 0 \quad 1.1]$ using the command **norm**.

$\boldsymbol{A^n \to 0}$ **if and only if** $\boldsymbol{\rho < 1}$. When $A = S^{-1}T$, this is the key to convergence.

Chapter 12

Linear Algebra in Probability & Statistics

12.1 Mean, Variance, and Probability

We are starting with the three fundamental words of this chapter: *mean, variance, and probability*. Let me give a rough explanation of their meaning before I write any formulas:

The **mean** is the *average value* or expected value

The **variance** $\boldsymbol{\sigma^2}$ measures the average *squared distance* from the mean m

The **probabilities** of n different outcomes are positive numbers $p_1, \ldots, p_n$ adding to 1.

Certainly the mean is easy to understand. We will start there. But right away we have two different situations that you have to keep straight. On the one hand, we may have the results (*sample values*) from a completed trial. On the other hand, we may have the expected results (*expected values*) from future trials. Let me give examples:

Sample values Five random freshmen have ages $\mathbf{18, 17, 18, 19, 17}$

Sample mean $\frac{\mathbf{1}}{\mathbf{5}}(18 + 17 + 18 + 19 + 17) = \mathbf{17.8}$

Probabilities The ages in a freshmen class are $17\ (\mathbf{20\%}), 18\ (\mathbf{50\%}), 19\ (\mathbf{30\%})$

A random freshman has **expected age** $\mathbf{E}\,[\mathrm{x}] = (0.2)\,17 + (0.5)\,18 + (0.3)\,19 = \mathbf{18.1}$

Both numbers 17.8 and 18.1 are correct averages. The sample mean starts with N samples $x_1, \ldots, x_N$ from a completed trial. Their mean is the *average* of the N observed samples:

$$\textbf{Sample mean} \qquad m = \mu = \frac{\mathbf{1}}{\boldsymbol{N}}(\boldsymbol{x_1} + \boldsymbol{x_2} + \cdots + \boldsymbol{x_N}) \tag{1}$$

The **expected value of x** starts with the probabilities $p_1, \ldots, p_n$ of the ages $x_1, \ldots, x_n$:

$$\textbf{Expected value} \quad m = \mathbf{E}[x] = p_1x_1 + p_2x_2 + \cdots + p_nx_n. \tag{2}$$

This is $\boldsymbol{p} \cdot \boldsymbol{x}$. Notice that $m = \mathrm{E}[x]$ tells us what to expect, $m = \mu$ tells us what we got.

By taking many samples (large N), the sample results will come close to the probabilities. The "Law of Large Numbers" says that with probability 1, the sample mean will converge to its expected value $\mathrm{E}[x]$ as the sample size N increases. A fair coin has probability $p_0 = \frac{1}{2}$ of tails and $p_1 = \frac{1}{2}$ of heads. Then $\mathrm{E}[x] = \left(\frac{1}{2}\right)0 + \frac{1}{2}(1)$. The fraction of heads in N flips of the coin is the sample mean, expected to approach $\mathrm{E}[x] = \frac{1}{2}$.

This does *not* mean that if we have seen more tails than heads, the next sample is likely to be heads. The odds remain 50-50. The first 100 or 1000 flips do affect the sample mean. *But* 1000 *flips will not affect its limit*—because you are dividing by $N \to \infty$.

Variance (around the mean)

The **variance σ^2** measures expected distance (squared) from the expected mean $\mathrm{E}[x]$. The **sample variance S^2** measures actual distance (squared) from the sample mean. The square root is the **standard deviation σ or S**. After an exam, I email μ and S to the class. I don't know the expected mean and variance because I don't know the probabilities p_1 to p_{100} for each score. (After teaching for 50 years, I still have no idea what to expect.)

The deviation is always deviation *from the mean*—sample or expected. We are looking for the size of the "spread" around the mean value $x = m$. Start with N samples.

$$\textbf{Sample variance} \quad S^2 = \frac{1}{N-1}\left[(x_1 - m)^2 + \cdots + (x_N - m)^2\right] \tag{3}$$

The sample ages $x = 18, 17, 18, 19, 17$ have mean $m = 17.8$. That sample has variance 0.7 :

$$S^2 = \frac{1}{4}\left[(.2)^2 + (-.8)^2 + (.2)^2 + (1.2)^2 + (-.8)^2\right] = \frac{1}{4}(2.8) = \mathbf{0.7}$$

The minus signs disappear when we compute squares. Please notice ! Statisticians divide by $N - 1 = 4$ (and not $N = 5$) so that S^2 is an unbiased estimate of σ^2. One degree of freedom is already accounted for in the sample mean.

An important identity comes from splitting each $(x - m)^2$ into $x^2 - 2mx + m^2$:

$$\begin{aligned} \text{sum of } (x_i - m)^2 &= (\text{sum of } x_i^2) - 2m(\text{sum of } x_i) + (\text{sum of } m^2) \\ &= (\text{sum of } x_i^2) - 2m(Nm) + Nm^2 \\ \textbf{sum of } \boldsymbol{(x_i - m)^2} &= \textbf{(sum of } \boldsymbol{x_i^2}\textbf{)} - \boldsymbol{Nm^2}. \end{aligned} \tag{4}$$

This is an equivalent way to find $(x_1 - m)^2 + \cdots + (x_N - m^2)$ by adding $x_1^2 + \cdots + x_N^2$.

Now start with probabilities p_i (never negative!) instead of samples. We find expected values instead of sample values. The variance σ^2 is the crucial number in statistics.

$$\textbf{Variance} \quad \boldsymbol{\sigma^2 = E\left[(x-m)^2\right] = p_1(x_1-m)^2 + \cdots + p_n(x_n-m)^2.} \qquad (5)$$

We are squaring the distance from the expected value $m = E[x]$. We don't have samples, only expectations. We know probabilities but we don't know experimental outcomes.

Example 1 Find the variance σ^2 of the ages of college freshmen.

Solution The probabilities of ages $x_i = 17, 18, 19$ were $p_i = 0.2$ and 0.5 and 0.3. The expected value was $\boldsymbol{m = \sum p_i x_i = 18.1}$. The variance uses those same probabilities:

$$\begin{aligned}\boldsymbol{\sigma^2} &= (0.2)(\mathbf{17-18.1})^2 + (0.5)(\mathbf{18-18.1})^2 + (0.3)(\mathbf{19-18.1})^2 \\ &= (0.2)(\mathbf{1.21}) + (0.5)(\mathbf{0.01}) + (0.3)(\mathbf{0.81}) = \mathbf{0.49}.\end{aligned}$$

The **standard deviation** is the square root $\boldsymbol{\sigma = 0.7}$.
This measures the spread of $17, 18, 19$ around $E[x]$, weighted by probabilities .2, .5, .3.

Continuous Probability Distributions

Up to now we have allowed for n possible outcomes $x_1, \ldots, x_n$. With ages 17, 18, 19, we only had $n = 3$. If we measure age in days instead of years, there will be a thousand possible ages (too many). Better to allow *every number between* 17 *and* 20—a continuum of possible ages. Then the probabilities p_1, p_2, p_3 for ages x_1, x_2, x_3 have to move to a **probability distribution** $\boldsymbol{p(x)}$ for a whole continuous range of ages $17 \le x \le 20$.

The best way to explain probability distributions is to give you two examples. They will be the **uniform distribution** and the **normal distribution**. The first (uniform) is easy. The normal distribution is all-important.

Uniform distribution Suppose ages are uniformly distributed between 17.0 and 20.0. All ages between those numbers are "equally likely". Of course any one exact age has no chance at all. There is zero probability that you will hit the exact number $x = 17.1$ or $x = 17 + \sqrt{2}$. What you can truthfully provide (assuming our uniform distribution) is **the chance $\boldsymbol{F(x)}$ that a random freshman has age less than $\boldsymbol{x}$**:

The chance of age less than $x = 17$ is $F(17) = \mathbf{0}$	$x \le 17$ won't happen
The chance of age less than $x = 20$ is $F(20) = \mathbf{1}$	$x \le 20$ will happen
The chance of age less than x is $\boldsymbol{F(x) = \frac{1}{3}(x-17)}$	$\boldsymbol{F}$ **goes from 0 to 1**

That formula $F(x) = \frac{1}{3}(x - 17)$ gives $F = 0$ at $x = 17$; then $x \le 17$ won't happen. It gives $F(x) = 1$ at $x = 20$; then $x \le 20$ is sure. Between 17 and 20, the graph of the **cumulative distribution $\boldsymbol{F(x)}$** increases linearly for this uniform model.

Let me draw the graphs of $F(x)$ and its derivative $p(x)$ = "probability density function".

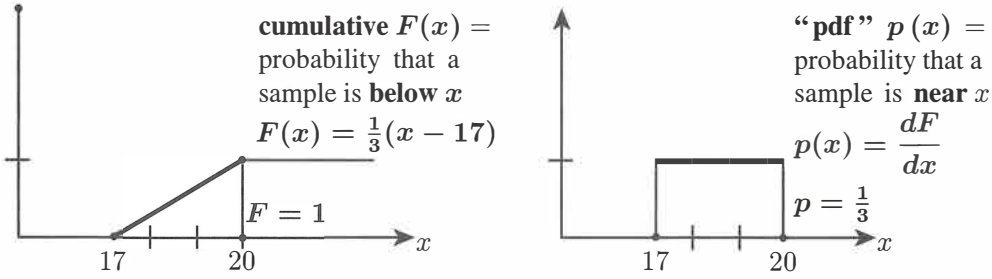


Figure 12.1: $F(x)$ is the cumulative distribution and its derivative $p(x) = dF/dx$ is the **probability density function** (**pdf**). For this uniform distribution, $p(x)$ is constant between 17 and 20. The total area under the graph of $p(x)$ is the total probability $F = 1$.

You could say that $p(x)\,dx$ is the probability of a sample falling in between x and $x + dx$. This is "infinitesimally true": $p(x)\,dx$ is $F(x + dx) - F(x)$. Here is the full truth:

$$\boldsymbol{F = \text{integral of } p} \qquad \textbf{Probability of } a \le x \le b = \int_a^b p(x)\,dx = F(b) - F(a) \quad (6)$$

$F(b)$ is the probability of $x \le b$. I subtract $F(a)$ to keep $x \ge a$. That leaves $a \le x \le b$.

Mean and Variance of $p(x)$

What are the mean m and variance σ^2 for a probability distribution? Previously we added $p_i x_i$ to get the mean (expected value). With a continuous distribution we **integrate** $xp(x)$:

$$\textbf{Mean} \qquad m = \mathbf{E}[x] = \int x\,p(x)\,dx = \int_{x=17}^{20} (x)\left(\frac{1}{3}\right) dx = 18.5$$

For this uniform distribution, the mean m is halfway between 17 and 20. Then the probability of a random value x below this halfway point $m = 18.5$ is $F(m) = \frac{1}{2}$.

In MATLAB, x = rand(1) chooses a random number uniformly between 0 and 1. Then the expected mean is $m = \frac{1}{2}$. The interval from 0 to x has probability $F(x) = x$. The interval below the mean m always has probability $F(m) = \frac{1}{2}$.

The variance is the average squared distance to the mean. With N outcomes, σ^2 is the sum of $p_i(x_i - m)^2$. For a continuous random variable x, the sum changes to an **integral**.

$$\textbf{Variance} \qquad \sigma^2 = \mathbf{E}\left[(x - m)^2\right] = \int p(x)\,(x - m)^2\,dx \quad (7)$$

When ages are uniform between $17 \le x \le 20$, the integral can shift to $0 \le x \le 3$:

$$\sigma^2 = \int_{17}^{20} \frac{1}{3}(x - 18.5)^2\,dx = \int_0^3 \frac{1}{3}(x - 1.5)^2\,dx = \frac{1}{9}(x-1.5)^3\Big]_{x=0}^{x=3} = \frac{2}{9}(1.5)^3 = \frac{3}{4}.$$

That is a typical example, and here is the complete picture for a uniform $p(x)$, 0 to a.

Uniform distribution for $0 \le x \le a$ **Density** $\boldsymbol{p(x) = \frac{1}{a}}$ **Cumulative** $\boldsymbol{F(x) = \frac{x}{a}}$

$$\textbf{Mean } m = \frac{a}{2} \text{ halfway} \qquad \textbf{Variance } \sigma^2 = \int_0^a \frac{1}{a}\left(x - \frac{a}{2}\right)^2 dx = \frac{a^2}{12} \tag{8}$$

The mean is a multiple of a, the variance is a multiple of a^2. For $a = 3$, $\sigma^2 = \frac{9}{12} = \frac{3}{4}$. For one random number between 0 and 1 (mean $\frac{1}{2}$) the variance is $\sigma^2 = \frac{1}{12}$.

Normal Distribution : Bell-shaped Curve

The normal distribution is also called the "Gaussian" distribution. It is the most important of all probability density functions $p(x)$. The reason for its overwhelming importance comes from repeating an experiment and averaging the outcomes. The experiments have their own distribution (like heads and tails). *The average approaches a normal distribution.*

Central Limit Theorem (informal) The average of N samples of "any" probability distribution approaches a normal distribution as $N \to \infty$.

Start with the "standard normal distribution". It is symmetric around $x = 0$, so its mean value is $m = 0$. It is chosen to have a standard variance $\sigma^2 = 1$. It is called $\mathbf{N}(0, 1)$.

$$\textbf{Standard normal distribution} \qquad \boldsymbol{p(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}}. \tag{9}$$

The graph of $p(x)$ is the **bell-shaped curve** in Figure 12.2. The standard facts are

$$\textbf{Total probability} = 1 \qquad \int_{-\infty}^{\infty} p(x)\,dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\,dx = \mathbf{1}$$

$$\textbf{Mean } \mathbf{E}[x] = 0 \qquad m = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-x^2/2}\,dx = \mathbf{0}$$

$$\textbf{Variance } \mathbf{E}[x^2] = 1 \qquad \sigma^2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-0)^2 e^{-x^2/2}\,dx = \mathbf{1}$$

The zero mean was easy because we are integrating an odd function. Changing x to $-x$ shows that "integral $=$ $-$ integral". So that integral must be $m = 0$.

The other two integrals apply the idea in Problem 12 to reach 1. Figure 12.2 shows a graph of $p(x)$ for the normal distribution $\mathbf{N}(0, \sigma)$ and also its cumulative distribution $F(x) =$ integral of $p(x)$. From the symmetry of $p(x)$ you see *mean = zero*. From $F(x)$ you see a very important practical approximation for opinion polling:

The probability that a random sample falls between $-\sigma$ and σ is $\boldsymbol{F(\sigma) - F(-\sigma) \approx \frac{2}{3}}$.

This is because $\int_{-\sigma}^{\sigma} p(x)\,dx$ equals $\int_{-\infty}^{\sigma} p(x)\,dx - \int_{-\infty}^{-\sigma} p(x)\,dx = F(\sigma) - F(-\sigma)$.

Similarly, the probability that a random x lies between -2σ and 2σ ("*less than two standard deviations from the mean*") is $F(2\sigma) - F(-2\sigma) \approx 0.95$. If you have an experimental result further than 2σ from the mean, it is fairly sure to be not accidental: chance $= 0.05$. Drug tests may look for a tighter confirmation, like probability 0.001. Searching for the Higgs boson used a hyper-strict test of 5σ deviation from pure accident.

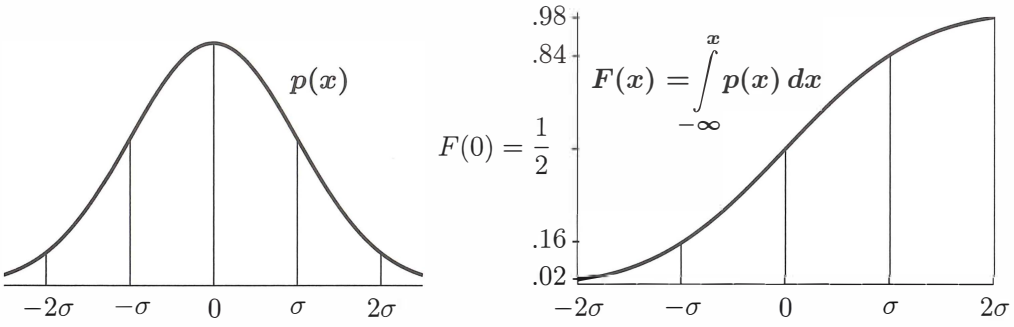


Figure 12.2: The standard normal distribution $p(x)$ has mean $m = 0$ and $\sigma = 1$.

The normal distribution with any mean m and standard deviation σ comes by shifting and stretching the standard $\mathbf{N}(0, 1)$. **Shift x to $x - m$. Stretch $x - m$ to $(x - m)/\sigma$.**

Gaussian density $p(x)$
Normal distribution $\mathbf{N}(m, \sigma)$

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-m)^2/2\sigma^2} \tag{10}$$

The integral of $p(x)$ is $F(x)$—the probability that a random sample will fall below x. The differential $p(x)\,dx = F(x + dx) - F(x)$ is the probability that a random sample will fall between x and $x + dx$. There is no simple formula to integrate $e^{-x^2/2}$, so this cumulative distribution $F(x)$ is computed and tabulated very carefully.

N Coin Flips and $N \to \infty$

Example 2 **Suppose x is 1 or -1 with equal probabilities $p_1 = p_{-1} = \frac{1}{2}$.**

The mean value is $\boldsymbol{m} = \frac{1}{2}(1) + \frac{1}{2}(-1) = \mathbf{0}$. The variance is $\boldsymbol{\sigma^2} = \frac{1}{2}(1)^2 + \frac{1}{2}(-1)^2 = \mathbf{1}$.

The key question is the *average* $A_N = (x_1 + \cdots + x_N)/N$. The independent x_i are ± 1 and we are dividing their sum by N. The expected mean of A_N is still zero. The law of large numbers says that this sample average approaches zero with probability 1. How fast does A_N approach zero? **What is its variance $\boldsymbol{\sigma_N^2}$?**

$$\text{By linearity} \quad \boldsymbol{\sigma_N^2} = \frac{\sigma^2}{N^2} + \frac{\sigma^2}{N^2} + \cdots + \frac{\sigma^2}{N^2} = N\frac{\sigma^2}{N^2} = \frac{\mathbf{1}}{\boldsymbol{N}} \quad \text{since } \sigma^2 = 1. \tag{11}$$

Example 3 **Change outputs from 1 or -1 to $x = 1$ or $x = 0$. Keep $p_1 = p_0 = \frac{1}{2}$.**

The new mean value $\boldsymbol{m} = \frac{1}{2}$ falls halfway between 0 and 1. The variance moves to $\boldsymbol{\sigma^2} = \frac{1}{4}$:

$$\boldsymbol{m} = \frac{1}{2}(1) + \frac{1}{2}(0) = \frac{\mathbf{1}}{\mathbf{2}} \quad \text{and} \quad \boldsymbol{\sigma^2} = \frac{1}{2}\left(1 - \frac{1}{2}\right)^2 + \frac{1}{2}\left(0 - \frac{1}{2}\right)^2 = \frac{\mathbf{1}}{\mathbf{4}}.$$

The average A_N now has mean $\frac{1}{2}$ and variance $\dfrac{1}{4N^2} + \cdots + \dfrac{1}{4N^2} = \dfrac{\mathbf{1}}{\mathbf{4N}} = \boldsymbol{\sigma_N^2}$. (12)

This σ_N is half the size of σ_N in Example 2. This must be correct because the new range 0 to 1 is half as long as -1 to 1. Examples 2-3 are showing a law of linearity.

The new $0 - 1$ variable x_{new} is $\frac{1}{2}x_{\text{old}} + \frac{1}{2}$. So the mean m is increased to $\frac{1}{2}$ and the variance is *multiplied* by $\left(\frac{1}{2}\right)^2$. A shift changes m and the rescaling changes σ^2.

Linearity $x_{\text{new}} = ax_{\text{old}} + b$ has $m_{\text{new}} = am_{\text{old}} + b$ and $\sigma^2{}_{\text{new}} = a^2\sigma^2{}_{\text{old}}$. (13)

Here are the results from three numerical tests: random 0 or 1 averaged over N trials.

[**48** 1's from $N = \mathbf{100}$] [**5035** 1's from $N = \mathbf{10000}$] [**19967** 1's from $N = \mathbf{40000}$].

The standardized X $= (x - m)/\sigma = \left(A_N - \frac{1}{2}\right) / 2\sqrt{N}$ was [**$-.40$**] [**.70**] [**$-.33$**].

The Central Limit Theorem says that the average of many coin flips will approach a normal distribution. Let us begin to see how that happens: **binomial approaches normal**.

For each flip, the probability of heads is $\frac{1}{2}$. For $N = 3$ flips, the probability of heads all three times is $\left(\frac{1}{2}\right)^3 = \frac{1}{8}$. The probability of heads twice and tails once is $\frac{3}{8}$, from three sequences HHT and HTH and THH. These numbers $\frac{1}{8}$ and $\frac{3}{8}$ are pieces of $\left(\frac{1}{2} + \frac{1}{2}\right)^3 = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8} = 1$. *The average number of heads in 3 flips is* 1.5.

Mean $m = (3 \text{ heads})\dfrac{1}{8} + (2 \text{ heads})\dfrac{3}{8} + (1 \text{ head})\dfrac{3}{8} + 0 = \dfrac{3}{8} + \dfrac{6}{8} + \dfrac{3}{8} = \mathbf{1.5 \text{ heads}}$

With N flips, Example 3 (or common sense) gives a mean of $\boldsymbol{m = \Sigma\, x_i p_i = \frac{1}{2}N}$ heads.

The variance σ^2 is based on the *squared distance* from this mean $N/2$. With $N = 3$ the variance is $\sigma^2 = \frac{3}{4}$ (*which is* $N/4$). To find σ^2 we add $(x_i - m)^2 p_i$ with $m = 1.5$:

$$\boldsymbol{\sigma^2} = (3-1.5)^2\frac{1}{8} + (2-1.5)^2\frac{3}{8} + (1-1.5)^2\frac{3}{8} + (0-1.5)^2\frac{1}{8} = \frac{9+3+3+9}{32} = \boldsymbol{\frac{3}{4}}.$$

For any N, the variance is $\boldsymbol{\sigma_N^2 = N/4}$. Then $\sigma_N = \sqrt{N}/2$.

Figure 12.3 shows how the probabilities of 0, 1, 2, 3, 4 heads in $N = 4$ flips come close to a bell-shaped Gaussian. That Gaussian is centered at the mean value $N/2 = 2$. To reach the standard Gaussian (mean 0 and variance 1) we shift and rescale that graph. If x is the number of heads in N flips—the average of N zero-one outcomes—then x is shifted by its mean $m = N/2$ and rescaled by $\sigma = \sqrt{N}/2$ to produce the standard X:

Shifted and scaled $\quad \boldsymbol{X = \dfrac{x - m}{\sigma} = \dfrac{x - \frac{1}{2}N}{\sqrt{N}/2}} \quad (N = 4 \text{ has } X = x - 2)$

Subtracting m is "centering" or "detrending". The mean of X is zero.

Dividing by σ is "normalizing" or "standardizing". The variance of X is 1.

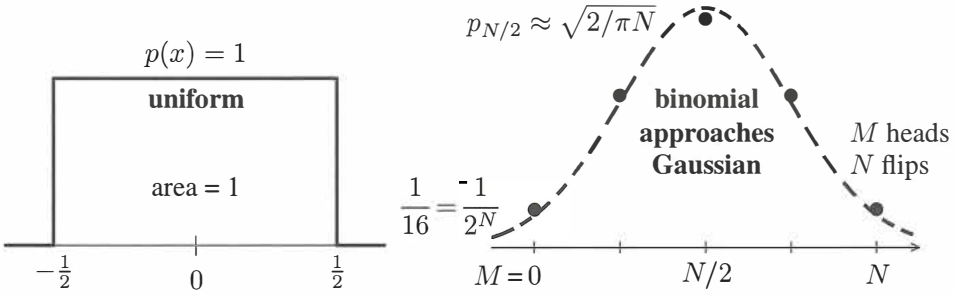


Figure 12.3: The probabilities $p = (1, 4, 6, 4, 1)/16$ for the number of heads in 4 flips. These p_i approach a Gaussian distribution with variance $\sigma^2 = N/4$ centered at $m = N/2$. For X, the Central Limit Theorem gives convergence to the normal distribution $\mathbf{N}(0, 1)$.

It is fun to see the Central Limit Theorem giving the right answer at the center point $X = 0$. At that point, the factor $e^{-X^2/2}$ equals 1. We know that the variance for N coin flips is $\sigma^2 = N/4$. The center of the bell-shaped curve has height $1/\sqrt{2\pi\sigma^2} = \boldsymbol{\sqrt{2/N\pi}}$.

What is the height at the center of the coin-flip distribution p_0 to p_N (the binomial distribution)? For $N = 4$, the probabilities for $0, 1, 2, 3, 4$ heads come from $\left(\frac{1}{2} + \frac{1}{2}\right)^4$.

$$\textbf{Center probability } \boldsymbol{\frac{6}{16}} \qquad \left(\frac{1}{2} + \frac{1}{2}\right)^4 = \frac{1}{16} + \frac{4}{16} + \boldsymbol{\frac{6}{16}} + \frac{4}{16} + \frac{1}{16} = 1.$$

The binomial theorem in Problem 8 tells us the center probability $p_{N/2}$ for any even N:

The center probability $\left(\frac{N}{2} \text{ heads}, \frac{N}{2} \text{ tails}\right)$ is $\dfrac{1}{2^N}\dfrac{N!}{(N/2)!\,(N/2)!}$

For $N = 4$, those factorials produce $4!/2!\,2! = 24/4 = 6$. For large N, Stirling's formula $\sqrt{2\pi N}(N/e)^N$ is a close approximation to $N!$. Use Stirling for N and twice for $N/2$:

Limit of coin-flip Center probability

$$p_{N/2} \approx \frac{1}{2^N}\frac{\sqrt{2\pi N}(N/e)^N}{\pi N(N/2e)^N} = \frac{\sqrt{2}}{\sqrt{\pi N}} = \frac{1}{\sqrt{2\pi\sigma}}. \tag{14}$$

At that last step we used the variance $\sigma^2 = N/4$ for the coin-tossing problem. The result $1/\sqrt{2\pi\sigma}$ matches the center value (above) for the Gaussian. The Central Limit Theorem is true: The "binomial distribution" approaches the normal distribution as $N \to \infty$.

Monte Carlo Estimation Methods

Scientific computing has to work with errors in the data. Financial computing has to work with unsure numbers and uncertain predictions. All of applied mathematics has moved to **accepting uncertainty in the inputs and estimating the variance in the outputs**.

How to estimate that variance? Often probability distributions $p(x)$ are not known. What we can do is to try different inputs $\boldsymbol{b}$ and compute the outputs $\boldsymbol{x}$ and take an average. This is the simplest form of a **Monte Carlo method** (named after the gambling palace on the Riviera, where I once saw a fight about whether the bet was placed in time). Monte Carlo approximates an expected value $\mathrm{E}[x]$ by a sample average $(x_1 + \cdots + x_N)/N$.

Please understand that every x_k can be expensive to compute. We are not just flipping coins. Each sample comes from a set of data b_k. *Monte Carlo randomly chooses this data b_k, it computes the outputs x_k, and then it averages those x's.* Decent accuracy for $\mathrm{E}[x]$ often requires many samples b and huge computing cost. The error in approximating $\mathrm{E}[x]$ by $(x_1 + \cdots + x_N)/N$ is normally of order $1/\sqrt{N}$. *Slow improvement as N increases.*

That $1/\sqrt{N}$ estimate came for coin flips in equation (11). Averaging N independent samples x_k of variance σ^2 reduces the variance to σ^2/N.

"Quasi-Monte Carlo" can sometimes reduce this variance to σ^2/N^2: a big difference! The inputs b_k are selected very carefully—not just randomly. This QMC approach is surveyed in the journal *Acta Numerica* 2013. The newer idea of "Multilevel Monte Carlo" is outlined by Michael Giles in *Acta Numerica* 2015. Here is how it works.

Suppose it is much simpler to simulate another variable $y(b)$ close to $x(b)$. Then use N computations of $y(b_k)$ and only $N^* < N$ computations of $x(b_k)$ to estimate $\mathrm{E}[x]$.

2-level Monte Carlo

$$\mathrm{E}[\boldsymbol{x}] \approx \frac{1}{N}\sum_{1}^{N} \boldsymbol{y}(\boldsymbol{b_k}) + \frac{1}{N^*}\sum_{1}^{N^*} [\boldsymbol{x}(\boldsymbol{b_k}) - \boldsymbol{y}(\boldsymbol{b_k})]\,.$$

The idea is that $x - y$ has a smaller variance σ^* than the original x. Therefore N^* can be smaller than N, with the same accuracy for $\mathrm{E}[x]$. We do N cheap simulations to find the y's. Those cost C each. We only do N^* expensive simulations involving x's. Those cost C^* each. The total computing cost is $NC + N^*C^*$.

Calculus minimizes the overall variance for a fixed total cost. The optimal ratio N^*/N is $\sqrt{C/C^*}\ \sigma^*/\sigma$. Three-level Monte Carlo would simulate x, y, and z:

$$\mathrm{E}[x] \approx \frac{1}{N}\sum_{1}^{N} \boldsymbol{z(b_k)} + \frac{1}{N^*}\sum_{1}^{N^*} [\boldsymbol{y(b_k)} - \boldsymbol{z(b_k)}] + \frac{1}{N^{**}}\sum_{1}^{N^{**}} [\boldsymbol{x(b_k)} - \boldsymbol{y(b_k)}]\,.$$

Giles optimizes $N, N^*, N^{**}, \ldots$ to keep $\mathrm{E}[x] \le$ fixed E_0, and provides a MATLAB code.

Review: Three Formulas for the Mean and the Variance

The formulas for m and σ^2 are the starting point for all of probability and statistics. There are three different cases to keep straight: **sample** values X_i, **expected** values (discrete p_i), and a range of **expected** values (continuous $p(x)$). Here are the mean and the variance:

	$\boldsymbol{m}$	$\boldsymbol{S^2}$ or $\boldsymbol{\sigma^2}$
Samples X_1 to X_N	$\boldsymbol{m} = \dfrac{X_1 + \cdots + X_N}{N}$	$\boldsymbol{S^2} = \dfrac{(X_1 - m)^2 + \cdots + (X_N - m)^2}{N - 1}$
n possible outputs with probabilities p_i	$\boldsymbol{m} = \sum_{1}^{n} p_i x_i$	$\boldsymbol{\sigma^2} = \sum_{1}^{n} p_i (x_i - m)^2$
Range of outputs with probability density	$\boldsymbol{m} = \int x\, p(x) dx$	$\boldsymbol{\sigma^2} = \int (x - m)^2 p(x) dx$

A natural question: Why are there no probabilities p on the first line? How can these formulas be parallel? Answer: *We expect a fraction p_i of the samples to be $X = x_i$.* If this is exactly true, $X = x_i$ is repeated $p_i N$ times. Then lines 1 and 2 give the same m.

When we work with samples, we don't know the p_i. We just include each output X as often as it comes. We get the "empirical" mean instead of the expected mean.

Problem Set 12.1

1 Add 7 to every output x. What happens to the mean and the variance? What are the new sample mean, the new expected mean, and the new variance?

2 We know: $\frac{1}{3}$ of all integers are divisible by 3 and $\frac{1}{7}$ of integers are divisible by 7. What fraction of integers will be divisible by 3 or 7 or both?

3 Suppose you sample from the numbers 1 to 1000 with equal probabilities $1/1000$. What are the probabilities p_0 to p_9 that the last digit of your sample is $0, \ldots, 9$? What is the expected mean m of that last digit? What is its variance σ^2?

4 Sample again from 1 to 1000 but look at the last digit of the sample *squared*. That square could end with $x = 0, 1, 4, 5, 6$, or 9. What are the probabilities $p_0, p_1, p_4, p_5, p_6, p_9$? What are the (expected) mean m and variance σ^2 of that number x?

5 (a little tricky) Sample again from 1 to 1000 with equal probabilities and let x be the *first* digit ($x = 1$ if the number is 15). What are the probabilities p_1 to p_9 (adding to 1) of $x = 1, \dots, 9$? What are the mean and variance of x?

6 Suppose you have $N = 4$ samples $157, 312, 696, 602$ in Problem 5. What are the first digits x_1 to x_4 of the squares? What is the sample mean μ ? What is the sample variance S^2 ? Remember to divide by $N - 1 = 3$ and not $N = 4$.

7 Equation (4) gave a second equivalent form for S^2 (the variance using samples):

$$\boldsymbol{S^2} = \frac{1}{N-1} \text{ sum of } (x_i - m)^2 = \frac{1}{N-1}\left[\left(\text{sum of } x_i^2\right) - Nm^2\right].$$

Verify the matching identity for the expected variance σ^2 (using $m = \Sigma p_i x_i$):

$$\boldsymbol{\sigma^2 = \text{ sum of } p_i\,(x_i - m)^2 = \left(\text{sum of } p_i\, x_i^2\right) - m^2.}$$

8 If all 24 samples from a population produce the same age $x = 20$, what are the sample mean μ and the sample variance S^2 ? What if $x = 20$ or 21, 12 times each ?

9 Computer experiment as on page 541: Find the average $A_{1000000}$ of a million random 0-1 samples ! What is $X = \left(A_N - \frac{1}{2}\right)/2\sqrt{N}$?

10 The probability p_i to get i heads in N coin flips is the *binomial number* $b_i = \binom{N}{i}$ divided by 2^N. The b_i add to $(1+1)^N = 2^N$ so the probabilities p_i add to 1.

$$p_0 + \cdots + p_N = \left(\frac{1}{2} + \frac{1}{2}\right)^N = \frac{1}{2^N}(b_0 + \cdots + b_N) \text{ with } b_i = \frac{N!}{i!\,(N-i)!}$$

$N = 4$ leads to $b_0 = \dfrac{24}{\underline{24}}$, $b_1 = \dfrac{24}{(1)(6)} = 4$, $b_2 = \dfrac{24}{(2)(2)} = 6$, $p_i = \dfrac{1}{\underline{16}}(1, 4, 6, 4, 1)$.

Notice $b_i = b_{N-i}$. *Problem*: Confirm that the mean $m = 0p_0 + \cdots + Np_N$ equals $\frac{N}{2}$.

11 For any function $f(x)$ the expected value is $\mathrm{E}[f] = \sum p_i\, f(x_i)$ or $\int p(x)\, f(x)\, dx$ (discrete probability or continuous probability). Suppose the mean is $\mathrm{E}[x] = m$ and the variance is $\mathrm{E}[(x-m)^2] = \sigma^2$. **What is $\mathrm{E}[\boldsymbol{x^2}]$** ?

12 Show that the standard normal distribution $p(x)$ has total probability $\int p(x)\, dx = 1$ as required. A famous trick multiplies $\int p(x)\, dx$ by $\int p(y)\, dy$ and computes the integral over all x and all y ($-\infty$ to ∞). The trick is to replace $dx\, dy$ in that double integral by $r\, dr\, d\theta$ (polar coordinates with $x^2 + y^2 = r^2$). Explain each step:

$$2\pi \int_{-\infty}^{\infty} p(x)\, dx \int_{-\infty}^{\infty} p(y)\, dy = \iint_{-\infty}^{\infty} e^{-(x^2+y^2)/2} dx\, dy = \int_{\theta=0}^{2\pi} \int_{r=0}^{\infty} e^{-r^2/2}\, r\, dr\, d\theta = \boldsymbol{2\pi}.$$

12.2 Covariance Matrices and Joint Probabilities

Linear algebra enters when we run M different experiments at once. We might measure age and height and weight ($M = 3$ measurements of N people). Each experiment has its own mean value. So we have a vector $\boldsymbol{m} = (m_1, m_2, m_3)$ containing the M mean values. Those could be *sample means* of age and height and weight. Or m_1, m_2, m_3 could be *expected values* of age, height, weight based on known probabilities.

A matrix becomes involved when we look at variances. Each experiment will have a sample variance S_i^2 or an expected $\sigma_i^2 = \mathrm{E}\left[(x_i - m_i)^2\right]$ based on the squared distance from its mean. Those M numbers $\sigma_1^2, \ldots, \sigma_M^2$ will go on the main diagonal of the matrix. So far we have made no connection between the M parallel experiments. They measure M different random variables, but the experiments are not necessarily independent!

If we measure age and height and weight $(\boldsymbol{a}, \boldsymbol{h}, \boldsymbol{w})$ for children, the results will be strongly correlated. Older children are generally taller and heavier. Suppose the means m_a, m_h, m_w are known. Then $\sigma_a^2, \sigma_h^2, \sigma_w^2$ are the separate variances in age, height, weight. **The new numbers are the covariances like $\boldsymbol{\sigma_{ah}}$, where age multiplies height.**

$$\textbf{Covariance} \quad \boldsymbol{\sigma_{ah}} = \mathbf{E}\left[(\textbf{age} - \textbf{mean age})\ (\textbf{height} - \textbf{mean height})\right]. \tag{1}$$

This definition needs a close look. To compute σ_{ah}, it is not enough to know the probability of each age and the probability of each height. We have to know the **joint probability of each pair** (**age and height**). This is because age is connected to height.

$\boldsymbol{p_{ah}}$ = probability that a random child has age $= \boldsymbol{a}$ **and** height $= \boldsymbol{h}$: both at once

$\boldsymbol{p_{ij}}$ = **probability that experiment 1 produces $\boldsymbol{x_i}$ and experiment 2 produces $\boldsymbol{y_j}$**

Suppose experiment 1 (age) has mean m_1. Experiment 2 (height) has mean m_2. The covariance in (1) between experiments 1 and 2 looks at **all pairs** of ages x_i, heights y_j :

$$\textbf{Covariance} \quad \boldsymbol{\sigma_{12}} = \sum_{\textbf{all } i,j} \sum \boldsymbol{p_{ij}}(\boldsymbol{x_i} - \boldsymbol{m_1})(\boldsymbol{y_j} - \boldsymbol{m_2}) \tag{2}$$

To capture this idea of "joint probability p_{ij}" we begin with two small examples.

Example 1 Flip two coins separately. With 1 for heads and 0 for tails, the results can be $(1,1)$ or $(1,0)$ or $(0,1)$ or $(0,0)$. Those four outcomes all have probability $p_{11} = p_{10} = p_{01} = p_{00} = \frac{1}{4}$. **Independent experiments have Prob of (i, j) = (Prob of i) (Prob of j).**

Example 2 *Glue the coins together*, facing the same way. The only possibilities are $(1,1)$ and $(0,0)$. Those have probabilities $\frac{1}{2}$ and $\frac{1}{2}$. The probabilities p_{10} and p_{01} are zero. $(1,0)$ and $(0,1)$ won't happen because the coins stick together: both heads or both tails.

Probability matrices for Examples 1 and 2

$$\boldsymbol{P} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{bmatrix} \qquad \boldsymbol{P} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}.$$

Let me stay longer with P, to show it in good matrix notation. The matrix shows the probability p_{ij} of each pair (x_i, y_j)—starting with (x_1, y_1) = (heads, heads) and (x_1, y_2) = (heads, tails). Notice the row sums p_i and column sums P_j and the total sum = 1.

$$\textbf{Probability matrix} \quad P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \quad \begin{matrix} p_{11}+p_{12} = \boldsymbol{p_1} \\ p_{21}+p_{22} = \boldsymbol{p_2} \end{matrix} \begin{pmatrix} \text{first} \\ \text{coin} \end{pmatrix}$$

$$\text{(second coin) column sums} \quad \boldsymbol{P_1} \quad \boldsymbol{P_2} \qquad 4 \text{ entries add to } 1$$

Those numbers p_1, p_2 and P_1, P_2 are called the **marginals** of the matrix P:

$p_1 = p_{11} + p_{12}$ = chance of heads from **coin 1** (coin 2 can be heads or tails)
$P_1 = p_{11} + p_{21}$ = chance of heads from **coin 2** (coin 1 can be heads or tails)

Example 1 showed *independent* variables. Every probability p_{ij} equals p_i times p_j ($\frac{1}{2}$ times $\frac{1}{2}$ gave $p_{ij} = \frac{1}{4}$ in that example). In this case **the covariance σ_{12} will be zero**. Heads or tails from the first coin gave no information about the second coin.

Zero covariance σ_{12} for independent trials $\quad V = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$ = **diagonal covariance matrix**.

Independent experiments have $\sigma_{12} = 0$ because every $p_{ij} = (p_i)(p_j)$ in equation (2):

$$\sigma_{12} = \sum_i \sum_j (p_i)(p_j)(x_i - m_1)(y_j - m_2) = \left[\sum_i (p_i)(x_i - m_1)\right]\left[\sum_j (p_j)(y_j - m_2)\right] = [0][0].$$

The glued coins show perfect correlation. Heads on one means heads on the other. The covariance σ_{12} moves from 0 to $\sigma_1\sigma_2 = \frac{1}{4}$—this is the largest possible value of σ_{12}:

$$\textbf{Means} = \frac{1}{2} \qquad \boldsymbol{\sigma_{12}} = \frac{1}{2}\left(1-\frac{1}{2}\right)\left(1-\frac{1}{2}\right) + 0 + 0 + \frac{1}{2}\left(0-\frac{1}{2}\right)\left(0-\frac{1}{2}\right) = \frac{1}{4}$$

Heads or tails from coin 1 gives complete information about heads or tails from coin 2:

Glued coins give largest possible covariances
Singular covariance matrix: determinant = 0 $\quad V_{\text{glue}} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2 \\ \sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$

Always $\sigma_1^2\sigma_2^2 \geq \sigma_{12}^2$. Thus σ_{12} is *between* $-\sigma_1\sigma_2$ *and* $\sigma_1\sigma_2$. The covariance matrix V is **positive definite** (or in this singular case of glued coins, V is **positive semidefinite**). That is an important fact about M by M covariance matrices for M experiments.

Note that the **sample covariance matrix S** from N trials is certainly semidefinite. Every new sample X = (age, height, weight) contributes to the **sample mean $\overline{X}$** and to S. Each term $(X_i - \overline{X})(X_i - \overline{X})^{\mathrm{T}}$ is positive semidefinite and we just add to reach S:

$$\overline{X} = \frac{X_1 + \cdots + X_N}{N} \qquad S = \frac{(X_1 - \overline{X})(X_1 - \overline{X})^{\mathrm{T}} + \cdots + (X_N - \overline{X})(X_N - \overline{X})^{\mathrm{T}}}{N-1} \quad (3)$$

The Covariance Matrix V is Positive Semidefinite

Come back to the *expected* covariance σ_{12} between two experiments 1 and 2 (two coins):

$$\boxed{\begin{aligned}\sigma_{12} &= \text{expected value of } [(\textit{output}\,1 - \textit{mean}\,1) \text{ times } (\textit{output}\,2 - \textit{mean}\,2)] \\ &= \sum_{\text{all } i,j} \sum p_{ij}\,(x_i - m_1)\,(y_j - m_2).\end{aligned}} \tag{4}$$

$p_{ij} \geq 0$ is the probability of seeing output x_i in experiment 1 **and** y_j in experiment 2. Some pair of outputs must appear. Therefore the N^2 probabilities p_{ij} add to 1.

Total probability (all pairs) is 1 $$\sum_{\text{all } i,j} \sum p_{ij} = 1. \tag{5}$$

Here is another fact we need. *Fix on one particular output* x_i in experiment 1. Allow *all outputs* y_j in experiment 2. Add the probabilities of $(x_i, y_1), (x_i, y_2), \ldots, (x_i, y_n)$:

$$\textbf{Row sum } p_i \textbf{ of } P \quad \sum_{j=1}^{n} p_{ij} = \textbf{probability } p_i \textbf{ of } x_i \textbf{ in experiment 1.} \tag{6}$$

Some y_j must happen in experiment 2! Whether the two coins are completely separate or glued together, we still get $\frac{1}{2}$ for the probability $p_H = p_{HH} + p_{HT}$ that coin 1 is heads:

$$\text{(separate) } P_{HH} + P_{HT} = \frac{1}{4} + \frac{1}{4} = \mathbf{\frac{1}{2}} \qquad \text{(glued) } P_{HH} + P_{HT} = \frac{1}{2} + 0 = \mathbf{\frac{1}{2}}.$$

That basic reasoning allows us to write one matrix formula that includes the covariance σ_{12} along with the separate variances σ_1^2 and σ_2^2 for experiment 1 and experiment 2. We get the whole covariance matrix V by adding the matrices V_{ij} for each pair (i, j):

$$\boxed{\begin{aligned}&\textbf{Covariance matrix} \\ &\boldsymbol{V = \Sigma\Sigma V_{ij}}\end{aligned} \qquad V = \sum_{\text{all } i,j} \sum p_{ij} \begin{bmatrix} (x_i - m_1)^2 & (x_i - m_1)(y_j - m_2) \\ (x_i - m_1)(y_j - m_2) & (y_j - m_2)^2 \end{bmatrix}} \tag{7}$$

Off the diagonal, this is equation (2) for the covariance σ_{12}. On the diagonal, we are getting the ordinary variances σ_1^2 and σ_2^2. I will show in detail how we get $V_{11} = \sigma_1^2$ by using equation (6). Allowing all j just leaves the probability p_i of x_i in experiment 1:

$$V_{11} = \sum_{\text{all } i,j} \sum p_{ij}(x_i - m_1)^2 = \sum_{\text{all } i} (\text{probability of } x_i)\,(x_i - m_1)^2 = \boldsymbol{\sigma_1^2}. \tag{8}$$

Please look at that twice. It is the key to producing the whole covariance matrix by one formula (7). The beauty of that formula is that it combines 2 by 2 matrices V_{ij}. And the matrix V_{ij} in (7) for each pair of outcomes i, j is **positive semidefinite**:

V_{ij} has diagonal entries $p_{ij}(x_i - m_1)^2 \geq 0$ and $p_{ij}(y_j - m_2)^2 \geq 0$ and $\det(V_{ij}) = 0$.

That matrix V_{ij} has rank 1. Equation (7) multiplies p_{ij} *times column* $\boldsymbol{U}$ *times row* $\boldsymbol{U}^{\mathrm{T}}$:

$$\begin{bmatrix} (x_i - m_1)^2 & (x_i - m_1)(y_j - m_2) \\ (x_i - m_1)(y_j - m_2) & (y_j - m_2)^2 \end{bmatrix} = \begin{bmatrix} x_i - m_1 \\ y_j - m_2 \end{bmatrix} \begin{bmatrix} x_i - m_1 & y_j - m_2 \end{bmatrix} \tag{9}$$

Every matrix $\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}$ *is positive semidefinite.* So the whole matrix V (combining these matrices $\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}$ with weights $p_{ij} \geq 0$) is **at least semidefinite**—and probably V is definite.

The covariance matrix V is positive definite unless the experiments are dependent.

Now we move from two variables x and y to M variables like age-height-weight. The output from each trial is a vector X with M components. (Each child has an age-height-weight vector with 3 components.) The covariance matrix V is now M by M. V is created from the output vectors $\boldsymbol{X}$ and their average $\overline{\boldsymbol{X}} = \mathbf{E}[\boldsymbol{X}]$:

$$\textbf{Covariance matrix} \qquad \boldsymbol{V} = \mathbf{E}\left[(\boldsymbol{X} - \overline{\boldsymbol{X}})(\boldsymbol{X} - \overline{\boldsymbol{X}})^{\mathrm{T}}\right] \tag{10}$$

Remember that $\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ and $\overline{\boldsymbol{X}}\,\overline{\boldsymbol{X}}^{\mathrm{T}}$ = (column)(row) are M by M matrices.

For $M = 1$ (one variable) you see that $\overline{X}$ is the mean m and V is σ^2 (Section 12.1). For $M = 2$ (two coins) you see that $\overline{X}$ is (m_1, m_2) and V matches equation (10). The expectation E always adds up outputs times their probabilities. For age-height-weight the output could be $X = (5 \text{ years}, 31 \text{ inches}, 48 \text{ pounds})$ and its probability is $\boldsymbol{p_{5,31,48}}$.

Now comes a new idea. *Take any linear combination* $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X} = c_1X_1 + \cdots + c_MX_M$. With $c = (6, 2, 5)$ this would be $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X} = 6\,(\text{age}) + 2\,(\text{height}) + 5\,(\text{weight})$. By linearity we know that its expected value $\mathrm{E}[\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X}]$ is $\boldsymbol{c}^{\mathrm{T}}\mathrm{E}[\boldsymbol{X}] = \boldsymbol{c}^{\mathrm{T}}\overline{\boldsymbol{X}}$:

$$\mathbf{E}[\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X}] = \boldsymbol{c}^{\mathrm{T}}\mathbf{E}[\boldsymbol{X}] = 6\,(\text{expected age}) + 2\,(\text{expected height}) + 5\,(\text{expected weight}).$$

More than that, we also know the *variance* σ^2 of that number $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X}$:

$$\begin{aligned} \text{Variance of } \boldsymbol{c}^{\mathrm{T}}\boldsymbol{X} &= \mathrm{E}\left[(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X} - \boldsymbol{c}^{\mathrm{T}}\overline{\boldsymbol{X}})(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X} - \boldsymbol{c}^{\mathrm{T}}\overline{\boldsymbol{X}})^{\mathrm{T}}\right] \\ &= \boldsymbol{c}^{\mathrm{T}}\mathrm{E}\left[(\boldsymbol{X} - \overline{\boldsymbol{X}})(\boldsymbol{X} - \overline{\boldsymbol{X}})^{\mathrm{T}}\right]\boldsymbol{c} = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{c}! \end{aligned} \tag{11}$$

Now the key point: *The variance of* $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X}$ *can never be negative.* So $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{c} \geq \boldsymbol{0}$. *The covariance matrix* V *is therefore positive semidefinite by the energy test* $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{c} \geq \boldsymbol{0}$.

Covariance matrices V open up the link between probability and linear algebra: V equals $Q\Lambda Q^{\mathrm{T}}$ with eigenvalues $\lambda_i \geq 0$ and orthonormal eigenvectors $\boldsymbol{q}_1$ to $\boldsymbol{q}_M$.

Diagonalizing the covariance matrix means finding M *independent* experiments as combinations of the original M experiments.

Confession I am not entirely happy with that proof based on $\boldsymbol{c}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{c} \geq \boldsymbol{0}$. The expectation symbol **E** is burying the key idea of **joint probability**. Allow me to show directly that V is positive semidefinite (at least for the age-height-weight example). The proof is simply that V **is the sum of the joint probability p_{ahw} of each combination (age, height, weight) times the positive semidefinite matrix $\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}$.** Here U is $X - \overline{X}$:

$$\boldsymbol{V} = \sum_{\text{all } a,h,w} p_{ahw}\, \boldsymbol{U}\,\boldsymbol{U}^{\mathrm{T}} \quad \text{with} \quad \boldsymbol{U} = \begin{bmatrix} \text{age} \\ \text{height} \\ \text{weight} \end{bmatrix} - \begin{bmatrix} \text{mean age} \\ \text{mean height} \\ \text{mean weight} \end{bmatrix}. \tag{12}$$

This is exactly like the 2 by 2 coin flip matrix V in equation (7). Now $M = 3$.

The value of the expectation symbol E is that it also allows *pdf*'s (probability density functions like $p(x, y, z)$ for continuous random variables x and y and z). If we allow all numbers as ages and heights and weights, instead of age $i = 0, 1, 2, 3 \ldots$, then we need $p(x, y, z)$ instead of p_{ijk}. The sums in this section of the book would all change to integrals. But we still have $V = \mathrm{E}\,[\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}]$:

$$\textbf{Covariance matrix } V = \iiint p(x,y,z)\, \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\, dx\, dy\, dz \quad \text{with} \quad \boldsymbol{U} = \begin{bmatrix} x - \overline{x} \\ y - \overline{y} \\ z - \overline{z} \end{bmatrix}. \tag{13}$$

Always $\iiint p = 1$. Examples 1–2 emphasized how p can give diagonal V or singular V:

Independent variables x, y, z $\quad p(x, y, z) = \boldsymbol{p_1(x)\, p_2(y)\, p_3(z)}$.

Dependent variables x, y, z $\quad p(x, y, z) = \boldsymbol{0}$ except when $cx + dy + ez = 0$.

The Mean and Variance of $z = x + y$

Start with the sample mean. We have N samples of x. Their mean (= average) is m_x. We also have N samples of y and their mean is m_y. **The sample mean of $\boldsymbol{z = x + y}$ is clearly $\boldsymbol{m_z = m_x + m_y}$:**

$$\textbf{Mean of sum = Sum of means} \quad \frac{1}{N}\sum_{1}^{N}(x_i + y_i) = \frac{1}{N}\sum_{1}^{N} x_i + \frac{1}{N}\sum_{1}^{N} y_i. \tag{14}$$

Nice to see something that simple. The *expected* mean of $z = x + y$ doesn't look so simple, but it must come out as $\mathbf{E}[z] = \mathbf{E}[x] + \mathbf{E}[y]$. Here is one way to see this.

The joint probability of the pair (x_i, y_j) is p_{ij}. Its value depends on whether the experiments are independent, which we don't know. But for the mean of the sum $z = x + y$,

dependence or independence of x and y doesn't matter. *Expected values still add*:

$$\mathbf{E}[\boldsymbol{x}+\boldsymbol{y}] = \sum_i \sum_j p_{ij}(x_i + y_j) = \sum_i \sum_j p_{ij}x_i + \sum_i \sum_j p_{ij}y_j. \tag{15}$$

All the sums go from 1 to N. We can add in any order. For the first term on the right side, add the p_{ij} along row i of the probability matrix P to get p_i. That double sum gives $\mathrm{E}[x]$:

$$\sum_i \sum_j p_{ij}x_i = \sum_i (p_{i1} + \cdots + p_{iN})x_i = \sum_i p_i x_i = \mathrm{E}[x].$$

For the last term, add p_{ij} down column j of the matrix to get the probability P_j of y_j. Those pairs (x_1, y_j) and (x_2, y_j) and ... and (x_N, y_j) are all the ways to produce y_j:

$$\sum_i \sum_j p_{ij}y_j = \sum_j (p_{1j} + \cdots + p_{Nj})y_j = \sum_j P_j y_j = \mathrm{E}[y].$$

Now equation (15) says that $\mathbf{E}[\boldsymbol{x}+\boldsymbol{y}] = \mathrm{E}[\boldsymbol{x}] + \mathrm{E}[\boldsymbol{y}]$.

What about the variance of $z = x + y$? The joint probabilities p_{ij} and the covariance σ_{xy} will be involved. Let me separate the variance of $x + y$ into three simple pieces:

$$\begin{aligned}\boldsymbol{\sigma_z^2} &= \sum\sum p_{ij}(x_i + y_j - m_x - m_y)^2 \\ &= \sum\sum p_{ij}(x_i - m_x)^2 + \sum\sum p_{ij}(y_j - m_y)^2 + 2\sum\sum p_{ij}(x_i - m_x)(y_j - m_y)\end{aligned}$$

The first piece is $\boldsymbol{\sigma_x^2}$. The second piece is $\boldsymbol{\sigma_y^2}$. The last piece is $\boldsymbol{2\sigma_{xy}}$.

$$\textbf{The variance of } \boldsymbol{z = x + y} \quad \textbf{is} \quad \boldsymbol{\sigma_z^2 = \sigma_x^2 + \sigma_y^2 + 2\sigma_{xy}}. \tag{16}$$

The Covariance Matrix for $Z = AX$

Here is a good way to see σ_z^2 when $z = x + y$. Think of (x, y) as a column vector $\boldsymbol{X}$. Think of the 1 by 2 matrix $A = \begin{bmatrix} 1 & 1 \end{bmatrix}$ multiplying that vector $\boldsymbol{X}$. Then $A\boldsymbol{X}$ is the sum $z = x + y$. The variance $\boldsymbol{\sigma_z^2}$ in equation (16) goes into matrix notation as

$$\boldsymbol{\sigma_z^2} = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{which is} \quad \boldsymbol{\sigma_z^2 = AVA^{\mathrm{T}}}. \tag{17}$$

You can see that $\sigma_z^2 = AVA^{\mathrm{T}}$ in (17) agrees with $\sigma_x^2 + \sigma_y^2 + 2\sigma_{xy}$ in (16).

Now for the main point. The vector $\boldsymbol{X}$ could have M components coming from M experiments (instead of only 2). Those experiments will have an M by M covariance matrix $\boldsymbol{V_X}$. The matrix A could be K by M. Then $A\boldsymbol{X}$ is a vector with K combinations of the M outputs (instead of 1 combination $x + y$ of 2 outputs).

That vector $\boldsymbol{Z} = A\boldsymbol{X}$ of length K has a K by K covariance matrix $V_{\boldsymbol{Z}}$. Then the great rule for covariance matrices—of which equation (17) was only a 1 by 2 example—is this beautiful formula: Covariance matrix of $A\boldsymbol{X}$ is A (covariance matrix of $\boldsymbol{X}$) A^{T}:

$$\boxed{\textbf{The covariance matrix of } \boldsymbol{Z = AX} \textbf{ is } \boldsymbol{V_Z = AV_X A^{\mathrm{T}}}} \tag{18}$$

To me, this neat formula shows the beauty of matrix multiplication. I won't prove this formula, just admire it. It is constantly used in applications—coming in Section 12.3.

The Correlation ρ

Correlation ρ_{xy} is closely related to covariance σ_{xy}. They both measure dependence or independence. Start by rescaling or "standardizing" the random variables x and y **The new $X = x/\sigma_x$ and $Y = y/\sigma_y$ have variance $\sigma_X^2 = \sigma_Y^2 = 1$.** This is just like dividing a vector v by its length to produce a unit vector $v/||v||$ of length 1.

The correlation of x and y is the covariance of X and Y. If the original covariance of x and y was σ_{xy}, then rescaling to X and Y will divide by σ_x and σ_y :

$$\textbf{Correlation } \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \textbf{covariance of } \frac{x}{\sigma_x} \textbf{ and } \frac{y}{\sigma_y} \qquad \textbf{Always } -1 \le \rho_{xy} \le 1$$

Zero covariance gives zero correlation. *Independent random variables* produce $\rho_{xy} = 0$.

We know that always $\sigma_{xy}^2 \le \sigma_x^2 \sigma_y^2$ (the covariance matrix V is at least positive semidefinite). Then $\rho_{xy}^2 \le 1$. Correlation near $\rho = +1$ means strong dependence in the same direction: often voting the same. Negative correlation means that y tends to be below its mean when x is above its mean: Voting in opposite directions.

Example 3 *Suppose that y is just $-x$.* A coin flip has outputs $x = 0$ or 1. The same flip has outputs $y = 0$ or -1. The mean m_x is $\frac{1}{2}$ for a fair coin, and m_y is $-\frac{1}{2}$. The covariance is $\sigma_{xy} = -\sigma_x \sigma_y$. The correlation divides by $\sigma_x \sigma_y$ to get $\rho_{xy} = -1$. In this case the correlation matrix R has determinant zero (singular and only semidefinite):

$$\textbf{Correlation matrix } R = \begin{bmatrix} 1 & \rho_{xy} \\ \rho_{xy} & 1 \end{bmatrix} \qquad R = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \text{ when } y = -x$$

R always has 1's on the diagonal because we normalized to $\sigma_X = \sigma_Y = 1$. R is the correlation matrix for x and y, and the covariance matrix for $X = x/\sigma_x$ and $Y = y/\sigma_y$.

That number ρ_{xy} is also called the Pearson coefficient.

Example 4 Suppose the random variables x, y, z are *independent. What matrix is R?*

Answer *R is the identity matrix.* All three correlations $\rho_{xx}, \rho_{yy}, \rho_{zz}$ are 1 by definition. All three cross-correlations $\rho_{xy}, \rho_{xz}, \rho_{yz}$ are zero by independence.

The correlation matrix R comes from the covariance matrix V, when we rescale every row and every column. Divide each row i and column i by the ith standard deviation σ_i.

(a) $R = DVD$ for the diagonal matrix $D = \text{diag}\,[1/\sigma_1, \ldots, 1/\sigma_M]$.

(b) If covariance V is positive definite, correlation $R = DVD$ is also positive definite.

■ WORKED EXAMPLES ■

12.2 A Suppose x and y are independent random variables with mean 0 and variance 1. Then the covariance matrix $\boldsymbol{V_X}$ for $\boldsymbol{X} = (x, y)$ is the 2 by 2 identity matrix. What are the mean $\boldsymbol{m_Z}$ and the covariance matrix $\boldsymbol{V_Z}$ for the 3-component vector $\boldsymbol{Z} = (x, y, ax + by)$?

Solution

Z is connected to X by A $\quad \boldsymbol{Z} = \begin{bmatrix} x \\ y \\ ax+by \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ a & b \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = A\boldsymbol{X}.$

The vector $\boldsymbol{m_X}$ contains the means of the M components of $\boldsymbol{X}$. The vector $\boldsymbol{m_Z}$ contains the means of the K components of $\boldsymbol{Z} = A\boldsymbol{X}$. The matrix connection between the means of $\boldsymbol{X}$ and $\boldsymbol{Z}$ has to be linear: $\boldsymbol{m_Z} = A\boldsymbol{m_X}$. The mean of $ax + by$ is $am_x + bm_y$.

The covariance matrix for $\boldsymbol{Z}$ is $\boldsymbol{V_Z} = AA^{\mathrm{T}}$, when $\boldsymbol{V_X}$ is the 2 by 2 identity matrix:

$$\boldsymbol{V_Z} = \begin{array}{l}\textbf{covariance matrix for} \\ \boldsymbol{Z = (x, y, ax + by)}\end{array} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ a & b \end{bmatrix} \begin{bmatrix} 1 & 0 & a \\ 0 & 1 & b \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \boldsymbol{a} \\ \mathbf{0} & \mathbf{1} & \boldsymbol{b} \\ \boldsymbol{a} & \boldsymbol{b} & \boldsymbol{a^2 + b^2} \end{bmatrix}.$$

Interpretation: x and y are independent so $\sigma_{xy} = 0$. Then the covariance of x with $ax + by$ is a and the covariance of y with $ax + by$ is b. Those just come from the two independent parts of $ax + by$. Finally, equation (18) gives the variance of $ax + by$:

$$\textbf{Use } \boldsymbol{V_Z = AV_X A^{\mathrm{T}}} \qquad \sigma^2_{ax+by} = \sigma^2_{ax} + \sigma^2_{by} + 2\sigma_{ax,by} = a^2 + b^2 + 0.$$

The 3 by 3 matrix $\boldsymbol{V_Z}$ is *singular*. Its determinant is $a^2 + b^2 - a^2 - b^2 = 0$. The third component $z = ax + by$ is completely dependent on x and y. The rank of $\boldsymbol{V_Z}$ is only 2.

GPS Example The signal from a GPS satellite includes its departure time. The receiver clock gives the arrival time. The receiver multiplies the travel time by the speed of light. Then it knows the distance from that satellite. Distances from four or more satellites pinpoint the receiver position (using least squares !).

One problem: The speed of light changes in the ionosphere. But the correction will be almost the same for all nearby receivers. If one receiver stays in a known position, we can take differences from that position. **Differential GPS** reduces the error variance:

Difference matrix $\boldsymbol{A} = [\,1 \quad -1\,]$ **Covariance matrix** $\boldsymbol{V_Z = AV_X A^{\mathrm{T}}}$

$$\boldsymbol{V_Z} = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \boldsymbol{\sigma_1^2 - 2\sigma_{12} + \sigma_2^2}$$

Errors in the speed of light are gone. Then centimeter positioning accuracy is achievable. (The key ideas are on page 320 of *Algorithms for Global Positioning* by Borre and Strang.) The GPS world is all about time and space and amazing accuracy.

Problem Set 12.2

1 (a) Compute the variance σ^2 when the coin flip probabilities are p and $1-p$ (tails $= 0$, heads $= 1$).

(b) The sum of N independent flips (0 or 1) is the count of heads after N tries. The rule (16-17-18) for the variance of a sum gives $\sigma^2 =$ _____.

2 What is the covariance σ_{kl} between the results $x_1, \ldots, x_n$ of Experiment 3 and the results $y_1, \ldots, y_n$ of Experiment 5 ? Your formula will look like σ_{12} in equation (2). Then the $(3,5)$ and $(5,3)$ entries of the covariance matrix V are $\sigma_{35} = \sigma_{53}$.

3 For $M = 3$ experiments, the variance-covariance matrix V will be 3 by 3. There will be a probability p_{ijk} that the three outputs are x_i and y_j and z_k. Write down a formula like equation (7) for the matrix V.

4 What is the covariance matrix V for $M = 3$ independent experiments with means m_1, m_2, m_3 and variances $\sigma_1^2, \sigma_2^2, \sigma_3^2$?

Problems 5–9 are about the conditional probability that $Y = y_j$ when we know $X = x_i$.
Notation: **Prob $(Y = y_j | X = x_i)$** = probability of the outcome y_j given that $X = x_i$.

Example 1 *Coin* 1 *is glued to coin* 2. Then Prob $(Y =$ heads when $X =$ heads$)$ is **1**.

Example 2 *Independent coin flips*: X gives no information about Y. Useless to know X. Then Prob $(Y =$ heads $| X =$ heads$)$ is the same as Prob $(Y =$ heads$)$.

5 Explain the **sum rule** of conditional probability:

$$\text{Prob}\,(Y = y_j) = \text{ sum over all outputs } x_i \text{ of Prob}\,(Y = y_j | X = x_i).$$

6 The n by n matrix P contains **joint probabilities p_{ij}** = Prob $(X = x_i$ **and** $Y = y_j)$.

Explain why the conditional Prob $(Y = y_j | X = x_i)$ equals $\dfrac{p_{ij}}{p_{i1} + \cdots + p_{in}} = \dfrac{p_{ij}}{p_i}$.

7 For this joint probability matrix with Prob $(x_1, y_2) = 0.3$, find Prob $(y_2 | x_1)$ and Prob (x_1).

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} 0.1 & 0.3 \\ 0.2 & 0.4 \end{bmatrix}$$

The entries p_{ij} add to 1.
Some i, j must happen.

8 Explain the **product rule** of conditional probability:

p_{ij} = Prob $(X = x_i$ **and** $Y = y_j)$ equals Prob $(Y = y_j | X = x_i)$ times Prob $(X = x_i)$.

9 Derive this **Bayes Theorem** for p_{ij} from the product rule in Problem 8:

$$\text{Prob}\,(Y = y_j \text{ and } X = x_i) = \frac{\text{Prob}\,(X = x_i | Y = y_j)\ \text{Prob}\,(Y = y_j)}{\text{Prob}\,(X = x_i)}$$

"Bayesians" use prior information. "Frequentists" only use sampling information.

12.3 Multivariate Gaussian and Weighted Least Squares

The normal probability density $p(x)$ (the Gaussian) depends on only two numbers:

$$\textbf{Mean } m \textbf{ and variance } \sigma^2 \qquad p(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-(x-m)^2/2\sigma^2}. \tag{1}$$

The graph of $p(x)$ is a bell-shaped curve centered at $x = m$. The continuous variable x can be anywhere between $-\infty$ and ∞. With probability close to $\frac{2}{3}$, that random x will lie between $m - \sigma$ and $m + \sigma$ (less than one standard deviation σ from its mean value m).

$$\int_{-\infty}^{\infty} p(x)\,dx = 1 \quad \text{and} \quad \int_{m-\sigma}^{m+\sigma} p(x)\,dx = \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} e^{-X^2/2}\,dX \approx \frac{2}{3}. \tag{2}$$

That integral has a change of variables from x to $X = (x - m)/\sigma$. This simplifies the exponent to $-X^2/2$ and it simplifies the limits of integration to -1 and 1. Even the $1/\sigma$ from p disappears outside the integral because dX equals dx/σ. Every Gaussian turns into a **standard Gaussian** $p(X)$ with mean $m = 0$ and variance $\sigma^2 = 1$. Just call it $p(x)$:

$$\textbf{The standard normal distribution } N(0,1) \textbf{ has } \quad p(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}. \tag{3}$$

Integrating $p(x)$ from $-\infty$ to x gives the cumulative distribution $F(x)$: the probability that a random sample is below x. That probability will be $F = \frac{1}{2}$ at $x = 0$ (the mean).

Two-dimensional Gaussians

Now we have $M = 2$ Gaussian random variables x and y. They have means m_1 and m_2. They have variances σ_1^2 and σ_2^2. If they are *independent*, then their probability density $p(x, y)$ is just $\boldsymbol{p_1(x)}$ **times** $\boldsymbol{p_2(y)}$. Multiply probabilities when variables are independent:

$$\textbf{Independent } x \textbf{ and } y \quad p(x,y) = \frac{1}{2\pi\sigma_1\sigma_2}\, e^{-(x-m_1)^2/2\sigma_1^2}\, e^{-(y-m_2)^2/2\sigma_2^2} \tag{4}$$

The covariance of x and y will be $\boldsymbol{\sigma_{12} = 0}$. The covariance matrix V will be *diagonal*. The variances σ_1^2 and σ_2^2 are always on the main diagonal of V. The exponent in $p(x, y)$ is just the sum of the x-exponent and the y-exponent. Good to notice that the two exponents can be combined into $-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m})^{\mathrm{T}} V^{-1} (\boldsymbol{x} - \boldsymbol{m})$ with V^{-1} in the middle:

$$-\frac{(x-m_1)^2}{2\sigma_1^2} - \frac{(y-m_2)^2}{2\sigma_2^2} = -\frac{1}{2} \begin{bmatrix} x - m_1 & y - m_2 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} x - m_1 \\ y - m_2 \end{bmatrix} \tag{5}$$

Non-independent x and y

We are ready to give up independence. The exponent (5) with V^{-1} is still correct when V is no longer a diagonal matrix. **Now the Gaussian depends on a vector m and a matrix V.**

When $M = 2$, the first variable x may give partial information about the second variable y (and vice versa). Maybe part of y is decided by x and part is truly independent. It is the M by M covariance matrix V that accounts for dependencies between the M variables $\boldsymbol{x} = x_1, \ldots, x_M$. Its inverse V^{-1} goes into $p(\boldsymbol{x})$:

Multivariate Gaussian probability distribution

$$\boxed{p(\boldsymbol{x}) = \frac{1}{(\sqrt{2\pi})^M \sqrt{\det V}}\, e^{-(\boldsymbol{x}-\boldsymbol{m})^{\mathrm{T}} V^{-1}(\boldsymbol{x}-\boldsymbol{m})/2}} \quad (6)$$

The vectors $\boldsymbol{x} = (x_1, \ldots, x_M)$ and $\boldsymbol{m} = (m_1, \ldots, m_M)$ contain the random variables and their means. The M square roots of 2π and the determinant of V are included to make the total probability equal to 1. Let me check that by linear algebra. I use the eigenvalues λ and orthonormal eigenvectors $\boldsymbol{q}$ of the symmetric matrix $V = Q\Lambda Q^{\mathrm{T}}$. So $\boldsymbol{V^{-1} = Q\Lambda^{-1}Q^{\mathrm{T}}}$:

$$\boldsymbol{X = x - m} \qquad (x-m)^{\mathrm{T}} V^{-1}(x-m) = \boldsymbol{X^{\mathrm{T}} Q \Lambda^{-1} Q^{\mathrm{T}} X = Y^{\mathrm{T}} \Lambda^{-1} Y}$$

Notice! The combinations $\boldsymbol{Y = Q^{\mathrm{T}} X = Q^{\mathrm{T}}(x - m)}$ are statistically independent. *Their covariance matrix Λ is diagonal.*

This step of diagonalizing V by its eigenvector matrix Q is the same as "uncorrelating" the random variables. Covariances are zero for the new variables $X_1, \ldots X_m$. This is the point where linear algebra helps calculus to compute multidimensional integrals.

The integral of $p(\boldsymbol{x})$ is not changed when we center the variable $\boldsymbol{x}$ by subtracting $\boldsymbol{m}$ to reach $\boldsymbol{X}$, and rotate that variable to reach $\boldsymbol{Y = Q^{\mathrm{T}} X}$. The matrix $\boldsymbol{\Lambda}$ is diagonal! So the integral we want splits into M separate one-dimensional integrals that we know:

$$\int \cdots \int e^{-\boldsymbol{Y^{\mathrm{T}} \Lambda^{-1} Y/2}}\, \boldsymbol{dY} = \left(\int_{-\infty}^{\infty} e^{-y_1^2/2\lambda_1}\, dy_1 \right) \cdots \left(\int_{-\infty}^{\infty} e^{-y_M^2/2\lambda_M}\, dy_M \right)$$

$$= \left(\sqrt{2\pi\lambda_1}\right) \cdots \left(\sqrt{2\pi\lambda_M}\right) = \left(\sqrt{2\pi}\right)^M \sqrt{\det V}. \quad (7)$$

The determinant of V (also the determinant of Λ) is the product $(\lambda_1) \ldots (\lambda_M)$ of the eigenvalues. Then (7) gives the correct number to divide by so that $p(x_1, \ldots, x_M)$ in equation (6) has integral $= 1$ as desired.

The mean and variance of $p(\boldsymbol{x})$ are also M-dimensional integrals. The same idea of diagonalizing V by its eigenvectors and introducing $\boldsymbol{Y = Q^{\mathrm{T}} X}$ will find those integrals:

Vector $\boldsymbol{m}$ of means
$$\int \cdots \int \boldsymbol{x}\, p(\boldsymbol{x})\, \boldsymbol{dx} = (m_1, m_2, \ldots) = \boldsymbol{m} \quad (8)$$

Covariance matrix $\boldsymbol{V}$
$$\int \cdots \int (\boldsymbol{x} - \boldsymbol{m})\, p(\boldsymbol{x}) (\boldsymbol{x} - \boldsymbol{m})^{\mathrm{T}}\, \boldsymbol{dx} = \boldsymbol{V}. \quad (9)$$

Conclusion: Formula (6) for the probability density $p(\boldsymbol{x})$ has all the properties we want.

Weighted Least Squares

In Chapter 4, least squares started from an unsolvable system $A\boldsymbol{x} = \boldsymbol{b}$. We chose $\widehat{\boldsymbol{x}}$ to minimize the error $||\boldsymbol{b} - A\boldsymbol{x}||^2$. That led us to the least squares equation $A^{\mathrm{T}}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}\boldsymbol{b}$. The best $A\widehat{\boldsymbol{x}}$ is the projection of $\boldsymbol{b}$ onto the column space of A. But is this squared distance $E = ||\boldsymbol{b} - A\boldsymbol{x}||^2$ the right error measure to minimize?

If the measurement errors in $\boldsymbol{b}$ are independent random variables, with mean $m = 0$ and variance $\sigma^2 = 1$ and a normal distribution, Gauss would say **yes**: *Use least squares*. If the errors are not independent or their variances are not equal. Gauss would say **no**: *Use **weighted** least squares*. This section will show that the good measure of error is $\boldsymbol{E} = (\boldsymbol{b} - A\boldsymbol{x})^{\mathrm{T}}\boldsymbol{V^{-1}}(\boldsymbol{b} - A\boldsymbol{x})$. The equation for the best $\widehat{\boldsymbol{x}}$ uses the covariance matrix V:

Weighted least squares $$A^{\mathrm{T}}V^{-1}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}V^{-1}\boldsymbol{b}. \tag{10}$$

The most important examples have m *independent* errors in $\boldsymbol{b}$. Those errors have variances $\sigma_1^2, \ldots, \sigma_m^2$. By independence, V is a diagonal matrix. The good weights $1/\sigma_1^2, \ldots, 1/\sigma_m^2$ come from V^{-1}. *We are weighting the errors in $\boldsymbol{b}$ to have* **variance = 1**:

Weighted least squares
Independent errors in b

$$\text{Minimize} \quad E = \sum_{i=1}^{m} \frac{(\boldsymbol{b} - A\boldsymbol{x})_i^2}{\sigma_i^2} \tag{11}$$

By weighting the errors, we are "whitening" the noise. **White noise** is a quick description of independent errors based on the standard Gaussian $\mathbf{N}(0, 1)$ with mean zero and $\sigma^2 = 1$.

Let me write down the steps to equations (10) and (11) for the best $\widehat{\boldsymbol{x}}$:

Start with $A\boldsymbol{x} = \boldsymbol{b}$ (m equations, n unknowns, $m > n$, no solution)

Each right side b_i has mean zero and variance σ_i^2. The b_i are independent.

Divide the ith equation by σ_i to have variance $= 1$ for every b_i/σ_i

That division turns $A\boldsymbol{x} = \boldsymbol{b}$ into $V^{-1/2}A\boldsymbol{x} = V^{-1/2}\boldsymbol{b}$ with $V^{-1/2} = \text{diag}\,(1/\sigma_1, \ldots, 1/\sigma_m)$

Ordinary least squares on those weighted equations has $A \to V^{-1/2}A$ and $\boldsymbol{b} \to V^{-1/2}\boldsymbol{b}$

$$(V^{-1/2}A)^{\mathrm{T}}(V^{-1/2}A)\widehat{\boldsymbol{x}} = (V^{-1/2}A)^{\mathrm{T}}V^{-1/2}\,\boldsymbol{b} \;\text{ is }\; A^{\mathrm{T}}V^{-1}A\widehat{\boldsymbol{x}} = A^{\mathrm{T}}V^{-1}\boldsymbol{b}. \tag{12}$$

Because of $1/\sigma^2$ in V^{-1}, more reliable equations (*smaller* σ) get heavier weights. This is the main point of weighted least squares.

Those diagonal weightings (uncoupled equations) are the most frequent and the simplest. They apply to *independent errors in the* b_i. When these measurement errors are not independent, V is no longer diagonal—but (12) is still the correct weighted equation.

In practice, finding all the covariances can be serious work. Diagonal V is simpler.

The Variance in the Estimated $\widehat{x}$

One more point: Often the important question is not the best $\widehat{x}$ for one particular set of measurements $\boldsymbol{b}$. This is only one sample! The real goal is to know the reliability of the whole experiment. That is measured (as reliability always is) by the **variance in the estimate** $\widehat{x}$. First, zero mean in $\boldsymbol{b}$ gives zero mean in $\widehat{x}$. Then the formula connecting variance $\boldsymbol{V}$ in the inputs $\boldsymbol{b}$ to variance $\boldsymbol{W}$ in the outputs $\widehat{x}$ turns out to be beautiful:

Variance-covariance matrix $\boldsymbol{W}$ for $\widehat{x}$ $\quad \mathrm{E}[(\widehat{x}-x)(\widehat{x}-x)^{\mathrm{T}}] = (\boldsymbol{A^{\mathrm{T}}V^{-1}A})^{-1}$. (13)

That smallest possible variance comes from the best possible weighting, which is V^{-1}.

This key formula is a perfect application of Section 12.2. **If $\boldsymbol{b}$ has covariance matrix V, then $\widehat{x} = \boldsymbol{Lb}$ has covariance matrix $\boldsymbol{LVL^{\mathrm{T}}}$.** Equation (12) above tells us that L is $(A^{\mathrm{T}}V^{-1}A)^{-1}A^{\mathrm{T}}V^{-1}$. Now substitute this into LVL^{T} and watch equation (13) appear:

$$\boldsymbol{LVL^{\mathrm{T}}} = (A^{\mathrm{T}}V^{-1}A)^{-1}A^{\mathrm{T}}V^{-1} \quad V \quad V^{-1}A(A^{\mathrm{T}}V^{-1}A)^{-1} = \boldsymbol{(A^{\mathrm{T}}V^{-1}A)^{-1}}.$$

This is the covariance W of the output, our best estimate $\widehat{x}$. It is time for examples.

Example 1 Suppose a doctor measures your heart rate x three times $(m = 3, n = 1)$:

$$\begin{matrix} x = b_1 \\ x = b_2 \\ x = b_3 \end{matrix} \quad \text{is} \quad A\boldsymbol{x} = \boldsymbol{b} \quad \text{with} \quad A = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \quad \text{and} \quad V = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}$$

The variances could be $\sigma_1^2 \cdot \quad /9$ and $\sigma_2^2 = 1/4$ and $\sigma_3^2 = \mathbf{1}$. You are getting more nervous as measurements are taken: $\quad ;$ is less reliable than b_2 and b_1. All three measurements contain some information, so they all go into the best (weighted) estimate $\widehat{x}$:

$$V^{-1/2}A\widehat{x} = V^{-1/2}\boldsymbol{b} \quad \text{is} \quad \begin{matrix} 3x = 3b_1 \\ 2x = 2b_2 \\ 1x = 1b_3 \end{matrix} \quad \text{leading to} \quad \boldsymbol{A^{\mathrm{T}}V^{-1}A\widehat{x} = A^{\mathrm{T}}V^{-1}b}$$

$$\begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 9 & & \\ & 4 & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \widehat{x} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 9 & & \\ & 4 & \\ & & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

$$\boxed{\widehat{x} = \frac{9b_1 + 4b_2 + b_3}{14} \quad \textbf{is a weighted average of } b_1, b_2, b_3}$$

Most weight is on b_1 since its variance σ_1 is smallest. The variance of $\widehat{\boldsymbol{x}}$ has the beautiful formula $W = (A^{\mathrm{T}}V^{-1}A)^{-1} = 1/14$:

$$\textbf{Variance of } \widehat{x} \qquad \left(\begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 9 & & \\ & 4 & \\ & & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right)^{-1} = \frac{1}{14} \quad \text{is smaller than} \quad \frac{1}{9}$$

The BLUE theorem of Gauss (proved on the website) says that our $\widehat{\boldsymbol{x}} = L\boldsymbol{b}$ is the best linear unbiased estimate of the solution to $A\boldsymbol{x} = \boldsymbol{b}$. Any other unbiased choice $\boldsymbol{x}^* = L^*\boldsymbol{b}$ has greater variance than $\widehat{\boldsymbol{x}}$. All unbiased choices have $L^*A = I$ so that an exact $A\boldsymbol{x} = \boldsymbol{b}$ will produce the right answer $\boldsymbol{x} = L^*\boldsymbol{b} = L^*A\boldsymbol{x}$.

Note. I must add that there are reasons not to minimize squared errors in the first place. One reason: This $\widehat{\boldsymbol{x}}$ often has many small components. The squares of small numbers are very small, and they appear when we minimize. It is easier to make sense of *sparse* vectors—only a few nonzeros. Statisticians often prefer to minimize **unsquared errors**: **the sum of** $|(\boldsymbol{b} - A\boldsymbol{x})_i|$. *This error measure is* L^1 *instead of* L^2. Because of the absolute values, the equation for $\widehat{\boldsymbol{x}}$ becomes nonlinear (it is actually piecewise linear).

Fast new algorithms are computing a sparse $\widehat{\boldsymbol{x}}$ quickly and the future may belong to L^1.

The Kalman Filter

The "Kalman filter" is the great algorithm in dynamic least squares. That word *dynamic* means that new measurements $\boldsymbol{b}_k$ keep coming. So the best estimate $\widehat{\boldsymbol{x}}_k$ keeps changing (based on all of $\boldsymbol{b}_0, \ldots, \boldsymbol{b}_k$). More than that, the matrix A is also changing. So $\widehat{\boldsymbol{x}}_2$ will be our best least squares estimate of the latest solution $\boldsymbol{x}_k$ to the **whole history of observation equations and update equations (state equations) up to time 2**:

$$A_0\boldsymbol{x}_0 = \boldsymbol{b}_0 \qquad \boldsymbol{x}_1 = F_0\boldsymbol{x}_0 \qquad A_1\boldsymbol{x}_1 = \boldsymbol{b}_1 \qquad \boldsymbol{x}_2 = F_1\boldsymbol{x}_1 \qquad A_2\boldsymbol{x}_2 = \boldsymbol{b}_2 \tag{14}$$

The Kalman idea is to introduce one equation at a time. There will be errors in each equation. With every new equation, we update the best estimate $\widehat{\boldsymbol{x}}_k$ for the current $\boldsymbol{x}_k$. But history is not forgotten! This new estimate $\widehat{\boldsymbol{x}}_k$ uses all the past observations $\boldsymbol{b}_0$ to $\boldsymbol{b}_{k-1}$ and all the state equations $\boldsymbol{x}_{\text{new}} = F_{\text{old}}\,\boldsymbol{x}_{\text{old}}$. A large and growing least squares problem.

One more important point. Each least squares equation is **weighted** using the covariance matrix V_k for the error in b_k. There is even a covariance matrix C_k for errors in the update equations $\boldsymbol{x}_{k+1} = F_k x_k$. The best $\widehat{\boldsymbol{x}}_2$ then depends on $\boldsymbol{b}_0, \boldsymbol{b}_1, \boldsymbol{b}_2$ and V_0, V_1, V_2 and C_1, C_2. The good way to write $\widehat{\boldsymbol{x}}_k$ is as an update to the previous $\widehat{\boldsymbol{x}}_{k-1}$.

Let me concentrate on a simplified problem, without the matrices F_k and the covariances C_k. We are estimating the same true $\boldsymbol{x}$ at every step. How do we get $\widehat{\boldsymbol{x}}_1$ from $\widehat{\boldsymbol{x}}_0$?

OLD $A_0\,\boldsymbol{x}_0 = \boldsymbol{b}_0$ leads to the weighted equation $A_0^{\mathrm{T}}\,V_0^{-1}\,A_0\,\widehat{\boldsymbol{x}}_0 = A_0^{\mathrm{T}}\,V_0^{-1}\,\boldsymbol{b}_0$. (15)

NEW $\begin{bmatrix} A_0 \\ A_1 \end{bmatrix} \widehat{\boldsymbol{x}}_1 = \begin{bmatrix} \boldsymbol{b}_0 \\ \boldsymbol{b}_1 \end{bmatrix}$ leads to the following weighted equation for $\widehat{\boldsymbol{x}}_1$:

$$\begin{bmatrix} A_0^{\mathrm{T}} & A_1^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} V_0^{-1} & \\ & V_1^{-1} \end{bmatrix} \begin{bmatrix} A_0 \\ A_1 \end{bmatrix} \widehat{\boldsymbol{x}}_1 = \begin{bmatrix} A_0^{\mathrm{T}} & A_1^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} V_0^{-1} & \\ & V_1^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_0 \\ \boldsymbol{b}_1 \end{bmatrix}. \quad (16)$$

Yes, we could just solve that new problem and forget the old one. But the old solution $\widehat{\boldsymbol{x}}_0$ needed work that we hope to reuse in $\widehat{\boldsymbol{x}}_1$. What we look for is **an update to $\widehat{\boldsymbol{x}}_0$**:

Kalman update gives $\widehat{\boldsymbol{x}}_1$ from $\widehat{\boldsymbol{x}}_0$ $\quad \widehat{\boldsymbol{x}}_1 = \widehat{\boldsymbol{x}}_0 + K_1(\boldsymbol{b}_1 - A_1\,\widehat{\boldsymbol{x}}_0).$ (17)

The update correction is the mismatch $\boldsymbol{b}_1 - A_1\widehat{\boldsymbol{x}}_0$ between the old state $\widehat{\boldsymbol{x}}_0$ and the new measurements $\boldsymbol{b}_1$—multiplied by the *Kalman gain matrix* K_1. The formula for K_1 comes from comparing the solutions $\widehat{\boldsymbol{x}}_1$ and $\widehat{\boldsymbol{x}}_0$ to (15) and (16). And when we update $\widehat{\boldsymbol{x}}_0$ to $\widehat{\boldsymbol{x}}_1$ based on new data $\boldsymbol{b}_1$, **we also update the covariance matrix $\boldsymbol{W_0}$ to $\boldsymbol{W_1}$**. Remember $W_0 = (A_0^{\mathrm{T}}\,V_0^{-1}\,A_0)^{-1}$ from equation (13). Update its inverse to W_1^{-1}:

Covariance W_1 of errors in $\widehat{\boldsymbol{x}}_1$ $\quad \boldsymbol{W_1^{-1} = W_0^{-1} + A_1^{\mathrm{T}}\,V_1^{-1}\,A_1}$ (18)

Kalman gain matrix K_1 $\quad \boldsymbol{K_1 = W_1\,A_1^{\mathrm{T}}\,V_1^{-1}}$ (19)

This is the heart of the Kalman filter. Notice the importance of the W_k. Those matrices measure the reliability of the whole process, where the vector $\widehat{\boldsymbol{x}}_k$ estimates the current state based on the particular measurements $\boldsymbol{b}_0$ to $\boldsymbol{b}_k$.

Whole chapters and whole books are written to explain the dynamic Kalman filter, when the states $\boldsymbol{x}_k$ are also changing (based on the matrices F_k). There is a *prediction* of $\boldsymbol{x}_k$ using F, followed by a *correction* using the new data $\boldsymbol{b}$. Perhaps best to stop here.

This page was about **recursive least squares**: adding new data b_k and updating both $\widehat{\boldsymbol{x}}$ and W: the best current estimate based on all the data, and its covariance matrix.

Problem Set 12.3

1 Two measurements of the same variable x give two equations $x = b_1$ and $x = b_2$. Suppose the means are zero and the variances are σ_1^2 and σ_2^2, with independent errors: V is diagonal with entries σ_1^2 and σ_2^2. Write the two equations as $A\boldsymbol{x} = \boldsymbol{b}$ (A is 2 by 1). As in the text Example 1, find this best estimate $\widehat{\boldsymbol{x}}$ based on b_1 and b_2:

$$\widehat{\boldsymbol{x}} = \frac{b_1/\sigma_1^2 + b_2/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2} \qquad \mathrm{E}\left[\widehat{\boldsymbol{x}}\,\widehat{\boldsymbol{x}}^{\mathrm{T}}\right] = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1}.$$

2 (a) In Problem 1, suppose the second measurement b_2 becomes super-exact and its variance $\sigma_2 \to 0$. What is the best estimate $\widehat{\boldsymbol{x}}$ when σ_2 reaches zero?

(b) The opposite case has $\sigma_2 \to \infty$ and no information in b_2. What is now the best estimate $\widehat{\boldsymbol{x}}$ based on b_1 and b_2?

3 If x and y are independent with probabilities $p_1(x)$ and $p_2(y)$, then $p(x,y) = p_1(x)\,p_2(y)$. By separating double integrals into products of single integrals ($-\infty$ to ∞) show that

$$\iint p(x,y)\,dx\,dy = \mathbf{1} \qquad \text{and} \qquad \iint (x+y)\,p(x,y)\,dx\,dy = \boldsymbol{m_1 + m_2}.$$

4 Continue Problem 3 for independent x, y to show that $p(x,y) = p_1(x)\,p_2(y)$ has

$$\iint (x-m_1)^2\,p(x,y)\,dx\,dy = \boldsymbol{\sigma_1^2} \qquad \iint (x-m_1)(y-m_2)\,p(x,y)\,dx\,dy = \mathbf{0}.$$

So the 2 by 2 covariance matrix V is diagonal and its entries are _____.

5 Show that the inverse of a 2 by 2 covariance matrix V is

$$V^{-1} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} 1/\sigma_1^2 & -\rho/\sigma_1\sigma_2 \\ -\rho/\sigma_1\sigma_2 & 1/\sigma_2^2 \end{bmatrix} \quad \begin{matrix} \text{with correlation} \\ \rho = \sigma_{12}/\sigma_1\sigma_2. \end{matrix}$$

This produces the exponent $-(\boldsymbol{x}-\boldsymbol{m})^{\mathrm{T}}\,\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{m})$ in a 2-variable Gaussian.

6 Suppose $\widehat{x}_k$ is the average of $b_1, \ldots, b_k$. A new measurement b_{k+1} arrives and we want the new average $\widehat{x}_{k+1}$. The Kalman update equation (17) is

$$\textbf{New average} \qquad \widehat{x}_{k+1} = \widehat{x}_k + \frac{1}{k+1}\,(b_{k+1} - \widehat{x}_k)\,.$$

Verify that $\widehat{x}_{k+1}$ is the correct average of $b_1 \ldots, b_{k+1}$.

7 Also check the update equation (18) for the variance $W_{k+1} = \sigma^2/(k+1)$ of this average $\widehat{x}$ assuming that $W_k = \sigma^2/k$ and b_{k+1} has variance $V = \sigma^2$.

8 (**Steady model**) Problems 6–7 were *static* least squares. All the sample averages $\widehat{x}_k$ were estimates of the same x. To make the Kalman filter *dynamic*, include also a *state equation $x_{k+1} = Fx_k$ with its own error variance s^2*. The dynamic least squares problem allows x to "drift" as k increases:

$$\begin{bmatrix} 1 & \\ -F & 1 \\ & 1 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} b_0 \\ 0 \\ b_1 \end{bmatrix} \text{ with variances } \begin{bmatrix} \sigma^2 \\ s^2 \\ \sigma^2 \end{bmatrix}.$$

With $F = 1$, divide both sides of those three equations by σ, s, and σ. Find $\widehat{x_0}$ and $\widehat{x_1}$ by least squares, which gives more weight to the recent b_1. The Kalman filter is developed in *Algorithms for Global Positioning* (Borre and Strang, Wellesley-Cambridge Press).

Change in A^{-1} from a Change in A

This final page connects the beginning of the book (inverses and rank one matrices) with the end of the book (dynamic least squares and filters). Begin with this basic formula:

The inverse of $M = I - uv^{\mathrm{T}}$ is $M^{-1} = I + \dfrac{uv^{\mathrm{T}}}{1 - v^{\mathrm{T}}u}$

The quickest proof is $MM^{-1} = I - uv^{\mathrm{T}} + (1 - uv^{\mathrm{T}})\dfrac{uv^{\mathrm{T}}}{1 - v^{\mathrm{T}}u} = I - uv^{\mathrm{T}} + uv^{\mathrm{T}} = I$.

M is not invertible if $v^{\mathrm{T}}u = 1$ (then $Mu = 0$). Here $v^{\mathrm{T}} = u^{\mathrm{T}} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$:

Example The inverse of $M = I - \begin{bmatrix} 1&1&1 \\ 1&1&1 \\ 1&1&1 \end{bmatrix}$ is $M^{-1} = I + \dfrac{1}{1-3}\begin{bmatrix} 1&1&1 \\ 1&1&1 \\ 1&1&1 \end{bmatrix}$

But we don't always start from the identity matrix. Many applications need to invert $M = A - uv^{\mathrm{T}}$. After we solve $Ax = b$ we expect a rank one change to give $My = b$. The division by $1 - v^{\mathrm{T}}u$ above will become a division by $c = 1 - v^{\mathrm{T}}A^{-1}u = 1 - v^{\mathrm{T}}z$.

Step 1 Solve $Az = u$ and compute $c = 1 - v^{\mathrm{T}}z$.

Step 2 If $c \neq 0$ then $M^{-1}b$ is $y = x + \dfrac{v^{\mathrm{T}}x}{c}z$.

Suppose A is easy to work with. A might already be factored into LU by elimination. Then this Sherman-Woodbury-Morrison formula is the fast way to solve $My = b$. Here are three problems to end the book !

9 Take Steps 1– 2to find y when $A = I$ and $u^{\mathrm{T}} = v^{\mathrm{T}} = [1\ \ 2\ \ 3]$ and $b^{\mathrm{T}} = [2\ \ 1\ \ 4]$.

10 Step 2 in this "update formula" claims that $My = \left(A - uv^{\mathrm{T}}\right)\left(x + \dfrac{v^{\mathrm{T}}x}{c}z\right) = b$.

Simplify this to $\dfrac{uv^{\mathrm{T}}x}{c}[1 - c - v^{\mathrm{T}}z] = 0$. This is true since $c = 1 - v^{\mathrm{T}}z$.

11 When A has a new row v^{T}, $A^{\mathrm{T}}A$ in the least squares equation changes to M:

$$M = \begin{bmatrix} A^{\mathrm{T}} & v \end{bmatrix}\begin{bmatrix} A \\ v^{\mathrm{T}} \end{bmatrix} = A^{\mathrm{T}}A + vv^{\mathrm{T}} = \text{ rank one change in } A^{\mathrm{T}}A.$$

Why is that multiplication correct? The updated $\widehat{x}_{\text{new}}$ comes from Steps 1 and 2.

For reference here are four formulas for M^{-1}. The first two were given above, when the change was uv^{T}. Formulas 3 and 4 go beyond rank one to allow matrices U, V, W.

1	$M = I - uv^{\mathrm{T}}$	and	$M^{-1} = I + uv^{\mathrm{T}}/(1 - v^{\mathrm{T}}u)$ (*rank*1 *change*)
2	$M = A - uv^{\mathrm{T}}$	and	$M^{-1} = A^{-1} + A^{-1}uv^{\mathrm{T}}A^{-1}/(1 - v^{\mathrm{T}}A^{-1}u)$
3	$M = I - UV$	and	$M^{-1} = I_n + U(I_m - VU)^{-1}V$
4	$M = A - UW^{-1}V$	and	$M^{-1} = A^{-1} + A^{-1}U(W - VA^{-1}U)^{-1}VA^{-1}$

Formula **4** is the "matrix inversion lemma" in engineering. Not seen until now ! The Kalman filter for solving block tridiagonal systems uses formula **4** at each step.

MATRIX FACTORIZATIONS

1. $\boldsymbol{A = LU} = \begin{pmatrix} \text{lower triangular } L \\ \text{1's on the diagonal} \end{pmatrix} \begin{pmatrix} \text{upper triangular } U \\ \text{pivots on the diagonal} \end{pmatrix}$

   Requirements: No row exchanges as Gaussian elimination reduces square A to U.

2. $\boldsymbol{A = LDU} = \begin{pmatrix} \text{lower triangular } L \\ \text{1's on the diagonal} \end{pmatrix} \begin{pmatrix} \text{pivot matrix} \\ D \text{ is diagonal} \end{pmatrix} \begin{pmatrix} \text{upper triangular } U \\ \text{1's on the diagonal} \end{pmatrix}$

   Requirements: No row exchanges. The pivots in D are divided out to leave 1's on the diagonal of U. If A is symmetric then U is L^{T} and $\boldsymbol{A = LDL^{\mathrm{T}}}$.

3. $\boldsymbol{PA = LU}$ (permutation matrix P to avoid zeros in the pivot positions).

   Requirements: A is invertible. Then P, L, U are invertible. P does all of the row exchanges on A in advance, to allow normal LU. Alternative: $A = L_1 P_1 U_1$.

4. $\boldsymbol{EA = R}$ (m by m invertible E) (any m by n matrix A) = $\mathsf{rref}(A)$.

   Requirements: None ! *The reduced row echelon form* R has r pivot rows and pivot columns, containing the identity matrix. The last $m - r$ rows of E are a basis for the left nullspace of A; they multiply A to give $m - r$ zero rows in R. The first r columns of E^{-1} are a basis for the column space of A.

5. $\boldsymbol{S = C^{\mathrm{T}}C}$ = (lower triangular) (upper triangular) with $\sqrt{D}$ on both diagonals

   Requirements: S is symmetric and positive definite (all n pivots in D are positive). This *Cholesky factorization* $C = \mathsf{chol}(S)$ has $C^{\mathrm{T}} = L\sqrt{D}$, so $S = C^{\mathrm{T}}C = LDL^{\mathrm{T}}$.

6. $\boldsymbol{A = QR}$ = (orthonormal columns in Q) (upper triangular R).

   Requirements: A has independent columns. Those are *orthogonalized* in Q by the Gram-Schmidt or Householder process. If A is square then $Q^{-1} = Q^{\mathrm{T}}$.

7. $\boldsymbol{A = X\Lambda X^{-1}}$ = (eigenvectors in X) (eigenvalues in Λ) (left eigenvectors in X^{-1}).

   Requirements: A must have n linearly independent eigenvectors.

8. $\boldsymbol{S = Q\Lambda Q^{\mathrm{T}}}$ = (orthogonal matrix Q) (real eigenvalue matrix Λ) (Q^{T} is Q^{-1}).

   Requirements: S is *real and symmetric*: $S^{\mathrm{T}} = S$. This is the Spectral Theorem.

9. $\boldsymbol{A = BJB^{-1}}$ = (generalized eigenvectors in B) (Jordan blocks in J) (B^{-1}).

Requirements: A is any square matrix. This *Jordan form* J has a block for each independent eigenvector of A. Every block has only one eigenvalue.

10. $\boldsymbol{A = U\Sigma V^{T}} = \begin{pmatrix} \text{orthogonal} \\ U \text{ is } m \times m \end{pmatrix} \begin{pmatrix} m \times n \text{ singular value matrix} \\ \sigma_1, \ldots, \sigma_r \text{ on its diagonal} \end{pmatrix} \begin{pmatrix} \text{orthogonal} \\ V \text{ is } n \times n \end{pmatrix}$.

Requirements: None. This ***Singular Value Decomposition*** (SVD) has the eigenvectors of AA^{T} in U and eigenvectors of $A^{\mathrm{T}}A$ in V; $\sigma_i = \sqrt{\lambda_i(A^{\mathrm{T}}A)} = \sqrt{\lambda_i(AA^{\mathrm{T}})}$.

Those singular values are $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. By column-row multiplication

$$\boldsymbol{A = U\Sigma V^{T}} = \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^{\mathrm{T}} + \cdots + \sigma_r \boldsymbol{u}_r \boldsymbol{v}_r^{\mathrm{T}}.$$

If S is symmetric positive definite then $U = V = Q$ and $\Sigma = \Lambda$ and $S = Q\Lambda Q^{\mathrm{T}}$.

11. $\boldsymbol{A^{+} = V\Sigma^{+}U^{T}} = \begin{pmatrix} \text{orthogonal} \\ n \times n \end{pmatrix} \begin{pmatrix} n \times m \text{ pseudoinverse of } \Sigma \\ 1/\sigma_1, \ldots, 1/\sigma_r \text{ on diagonal} \end{pmatrix} \begin{pmatrix} \text{orthogonal} \\ m \times m \end{pmatrix}$.

Requirements: None. The *pseudoinverse* A^{+} has $A^{+}A =$ projection onto row space of A and $AA^{+} =$ projection onto column space. $A^{+} = A^{-1}$ if A is invertible. The shortest least-squares solution to $A\boldsymbol{x} = \boldsymbol{b}$ is $\boldsymbol{x}^{+} = A^{+}\boldsymbol{b}$. This solves $A^{\mathrm{T}}A\boldsymbol{x}^{+} = A^{\mathrm{T}}\boldsymbol{b}$.

12. $\boldsymbol{A = QS}$ = (orthogonal matrix Q) (symmetric positive definite matrix S).

Requirements: A is invertible. This *polar decomposition* has $S^2 = A^{\mathrm{T}}A$. The factor S is semidefinite if A is singular. The reverse polar decomposition $A = KQ$ has $K^2 = AA^{\mathrm{T}}$. Both have $Q = UV^{\mathrm{T}}$ from the SVD.

13. $\boldsymbol{A = U\Lambda U^{-1}}$ = (unitary U) (eigenvalue matrix Λ) (U^{-1} which is $U^{\mathrm{H}} = \overline{U}^{\mathrm{T}}$).

Requirements: A is *normal*: $A^{\mathrm{H}}A = AA^{\mathrm{H}}$. Its orthonormal (and possibly complex) eigenvectors are the columns of U. Complex λ's unless $S = S^{\mathrm{H}}$: Hermitian case.

14. $\boldsymbol{A = QTQ^{-1}}$ = (unitary Q) (triangular T with λ's on diagonal) ($Q^{-1} = Q^{\mathrm{H}}$).

Requirements: *Schur triangularization* of any square A. There is a matrix Q with orthonormal columns that makes $Q^{-1}AQ$ triangular: Section 6.4.

15. $\boldsymbol{F_n} = \begin{bmatrix} I & D \\ I & -D \end{bmatrix} \begin{bmatrix} \boldsymbol{F_{n/2}} & \\ & \boldsymbol{F_{n/2}} \end{bmatrix} \begin{bmatrix} \text{even-odd} \\ \text{permutation} \end{bmatrix} =$ one step of the recursive **FFT**.

Requirements: $F_n =$ Fourier matrix with entries w^{jk} where $w^n = 1$: $F_n\overline{F}_n = nI$. D has $1, w, \ldots, w^{n/2-1}$ on its diagonal. For $n = 2^{\ell}$ the *Fast Fourier Transform* will compute $F_n\boldsymbol{x}$ with only $\frac{1}{2}n\ell = \frac{1}{2}n\log_2 n$ multiplications from ℓ stages of D's.

Index

Index of Symbols and Computer Codes

Computer Packages

Code Names

Linear Algebra Websites and Email Address

math.mit.edu/linearalgebra Dedicated to readers and teachers working with this book

ocw.mit.edu MIT's OpenCourseWare site including video lectures in 18.06 and 18.085-6

web.mit.edu/18.06 Current and past exams and homeworks with extra materials

wellesleycambridge.com Ordering information for books by Gilbert Strang

linearalgebrabook@gmail.com Direct email contact about this book

Six Great Theorems of Linear Algebra

Dimension Theorem All bases for a vector space have the same number of vectors.

Counting Theorem Dimension of column space + dimension of nullspace = number of columns.

Rank Theorem Dimension of column space = dimension of row space. This is the rank.

Fundamental Theorem The row space and nullspace of A are orthogonal complements in $\mathbf{R}^n$.

SVD There are orthonormal bases ($\boldsymbol{v}$'s and $\boldsymbol{u}$'s for the row and column spaces) so that $A\boldsymbol{v}_i = \boldsymbol{\sigma}_i \boldsymbol{u}_i$.

Spectral Theorem If $A^{\mathrm{T}} = A$ there are orthonormal $\boldsymbol{q}$'s so that $\boldsymbol{A}\boldsymbol{q}_i = \boldsymbol{\lambda}_i \boldsymbol{q}_i$ and $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\mathrm{T}}$.

LINEAR ALGEBRA IN A NUTSHELL

((***The matrix*** A ***is*** n ***by*** n))

Nonsingular	Singular
A is invertible	A is not invertible
The columns are independent	The columns are dependent
The rows are independent	The rows are dependent
The determinant is not zero	The determinant is zero
$A\boldsymbol{x} = \mathbf{0}$ has one solution $\boldsymbol{x} = \mathbf{0}$	$A\boldsymbol{x} = \mathbf{0}$ has infinitely many solutions
$A\boldsymbol{x} = \boldsymbol{b}$ has one solution $\boldsymbol{x} = A^{-1}\boldsymbol{b}$	$A\boldsymbol{x} = \boldsymbol{b}$ has no solution or infinitely many
A has n (nonzero) pivots	A has $r < n$ pivots
A has full rank $r = n$	A has rank $r < n$
The reduced row echelon form is $R = I$	R has at least one zero row
The column space is all of $\mathbf{R}^n$	The column space has dimension $r < n$
The row space is all of $\mathbf{R}^n$	The row space has dimension $r < n$
All eigenvalues are nonzero	Zero is an eigenvalue of A
$A^{\mathrm{T}}A$ is symmetric positive definite	$A^{\mathrm{T}}A$ is only semidefinite
A has n (positive) singular values	A has $r < n$ singular values