Linear Algebra

An MIT 18.06 Companion

After the lectures of **Prof. Gilbert Strang**
Massachusetts Institute of Technology (18.06)
Notes written and LaTeX-typeset by Rui Zhou

(rewritten from the 18.06 lectures and notes, 2026)

How to Read This Book

This is a companion to a one-semester linear algebra course in the spirit of MIT's **18.06**, taught for decades by Gilbert Strang. It is written to be read straight through, like a short textbook, rather than skimmed like a formula sheet. Its goal is the one Strang sets in his first lecture: not just to *solve* $Ax = b$, but to *understand* it — geometrically, as a statement about combinations of vectors, and structurally, through the spaces hiding inside every matrix.

The book follows the arc of the course. We begin with the two pictures of a linear system — the row picture and the *column* picture — and learn to solve it by elimination, recording the work as $A = LU$. We then step back and ask the questions that organize all of linear algebra: which right-hand sides b can be reached (the column space), which solutions x vanish (the null space), and how the *four fundamental subspaces* of a matrix fit together. From there we add geometry (orthogonality, projection, least squares, $A = QR$), then determinants, then the eigenvalue story ($A = S\Lambda S^{-1}$ and its applications to powers, differential equations, and Markov chains), the special structure of symmetric and positive definite matrices, and finally the singular value decomposition — the climax of the subject — before closing with linear transformations and change of basis.

How Each Chapter Is Organized

Each chapter opens with a short *story* in plain English: what problem are we solving, and why are we about to introduce a particular idea? We then state the definitions and theorems, prove the central ones, and — in Strang's style — lean on small, concrete numerical examples to make the ideas tangible. Definitions and key results are boxed for quick review.

Box Color Code

- **Green: Definition**

  — the objects we build on (subspaces, rank, eigenvalues, ...).

- **Blue: Theorem**

  — the main results, usually with a proof or sketch.

- **Pink: Assumption**

  — the standing hypotheses on a matrix.

- Worked **examples** and **remarks** appear in lightly ruled boxes throughout.

A Note on Notation

A is an $m \times n$ matrix; x and b are column vectors. We write A^{T} for the transpose, A^{-1} for the inverse, $\operatorname{rank} A$ for the rank. The four fundamental subspaces are the column space $C(A)$, the null space $N(A)$, the row space $C(A^{\mathsf{T}})$, and the left null space $N(A^{\mathsf{T}})$. After elimination, U is an upper-triangular form and R its reduced row echelon form; pivots are the nonzero entries that remain. Eigenvalues are λ, and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$.

Contents

Chapter 1

The Geometry of Linear Equations

The fundamental problem of linear algebra is to solve a system of n linear equations in n unknowns. That problem is so familiar from school — two lines, find where they cross — that it is easy to think there is nothing left to understand. There is. The whole course turns on *seeing* a linear system the right way, and the right way is not the one most of us were taught.

Take the smallest interesting example, two equations in two unknowns:

$$\begin{aligned} 2x - y &= 0, \\ -x + 2y &= 3. \end{aligned} \tag{1.1}$$

There are three ways to look at this system, and we will use all three throughout the book. The *row picture* plots one line per equation and asks where they meet — this is the picture you already know. The *column picture* reads the same numbers down the columns instead of across the rows, and asks a different question: how do we combine two fixed vectors to produce a third? This second picture is the one to fall in love with. Almost everything later — column space, rank, independence, the four fundamental subspaces — is the column picture growing up. The *matrix picture* then packages everything as $Ax = b$, and the real content of the chapter is learning to read those three symbols in two different ways at once.

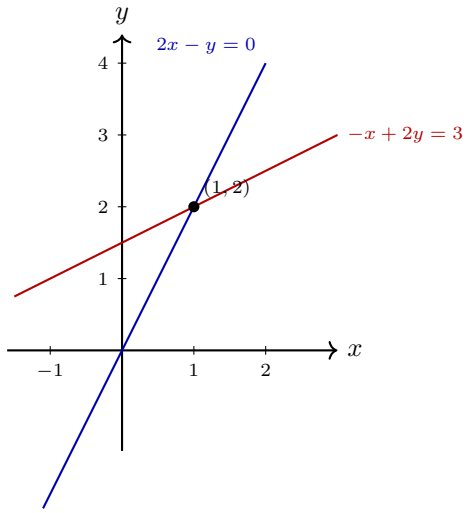
1.1 The Row Picture

In the row picture we take the equations one at a time. Each equation in two unknowns describes a line in the xy-plane, namely all points (x, y) that satisfy it. A solution of the *system* must satisfy *every* equation, so it is a point lying on *all* the lines at once — a point of intersection.

For the system (1.1), the first equation $2x - y = 0$ is the line $y = 2x$ through the origin; the second, $-x + 2y = 3$, is the line $y = \frac{1}{2}x + \frac{3}{2}$. They cross at a single point, and reading it off (or solving the two equations by hand) gives

$$x = 1, \qquad y = 2.$$

We check by substituting back: $2(1) - 2 = 0$ and $-1 + 2(2) = 3$. Both hold, so $(1, 2)$ is the solution.



Nothing here is new. The point of the row picture is that it stops being helpful exactly when we need it most. With three unknowns, each equation becomes a *plane* in $\mathbb{R}^3$, and the solution is where three planes meet — typically in a single point, but two planes meet in a line, and the third plane may or may not pass through it. With ten unknowns we are intersecting ten hyperplanes in $\mathbb{R}^{10}$, and no one can see that. We need a picture that scales. That picture comes from reading the columns.

1.2 The Column Picture

Look at the same system (1.1) again, but now collect the coefficients down the *columns* rather than across the rows. The two numbers multiplying x are 2 and -1; the two multiplying y are -1 and 2; the right-hand sides are 0 and 3. So the system says exactly

$$x \begin{bmatrix} 2 \\ -1 \end{bmatrix} + y \begin{bmatrix} -1 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 3 \end{bmatrix}. \tag{1.2}$$

This is the *same two equations* — read off the top entries and you recover $2x - y = 0$; read off the bottom entries and you recover $-x + 2y = 3$. But the question has changed completely. We are no longer asking "where do two lines cross?" We are asking: *how much of each column vector do we need so that they add up to the right-hand side?*

Definition 1.1: Linear Combination

Given vectors $v_1, \ldots, v_n$ in $\mathbb{R}^m$ and scalars $x_1, \ldots, x_n$, the vector

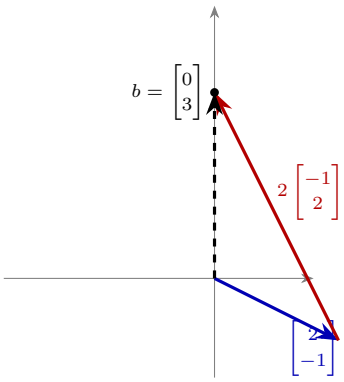
$$x_1 v_1 + x_2 v_2 + \cdots + x_n v_n$$

is called a *linear combination* of $v_1, \ldots, v_n$. To take combinations of vectors — scale them and add them — is the one fundamental operation of the whole subject.

So solving the system means finding the scalars x and y that make the linear combination $x\,(2,-1)+y\,(-1,2)$ land on $(0,3)$. We already know the answer must be $x=1$, $y=2$, and we can watch it work:

$$1\begin{bmatrix}2\\-1\end{bmatrix}+2\begin{bmatrix}-1\\2\end{bmatrix}=\begin{bmatrix}2\\-1\end{bmatrix}+\begin{bmatrix}-2\\4\end{bmatrix}=\begin{bmatrix}0\\3\end{bmatrix}. \quad \checkmark$$

Geometrically, one copy of the first column followed by two copies of the second forms a little path from the origin to the target $b=(0,3)$.



Now ask the deeper question. We solved this for one particular $b=(0,3)$. What if the right-hand side were some *other* vector? Could we still hit it with a combination of the two columns $(2,-1)$ and $(-1,2)$? In two dimensions the answer is yes for *every* b: these two columns point in genuinely different directions, and their combinations $x(2,-1)+y(-1,2)$ sweep out the entire plane. Every vector in $\mathbb{R}^2$ is reachable, so $Ax=b$ is solvable for every b. That "filling the whole space" is the property we will spend the next chapters naming and measuring.

Remark (Why the column picture wins).

> The row picture and the column picture contain identical information — the same coefficients, just read in two directions. But the column picture answers the questions that organize the rest of the course. "Which right-hand sides b can I reach?" becomes "which vectors are combinations of the columns?" — that set will be the *column space* (Chapter 4). "Is the solution unique?" becomes "can a nonzero combination of the columns give zero?" — that is *linear independence*. Keep one eye on the columns from now on.

1.3 The Matrix Picture: $Ax=b$

The two columns and the right-hand side are begging to be packaged. Put the columns of coefficients side by side into a *matrix* A, stack the unknowns into a vector x, and the right-hand sides into a vector b:

$$A=\begin{bmatrix}2&-1\\-1&2\end{bmatrix},\qquad x=\begin{bmatrix}x\\y\end{bmatrix},\qquad b=\begin{bmatrix}0\\3\end{bmatrix}.$$

The matrix A is the *coefficient matrix*; its columns are exactly the vectors from (1.2). The entire system collapses to three symbols:

$$\boxed{Ax = b.}$$

For a general $m \times n$ system — m equations, n unknowns — we write $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$. The matrix A has one row per equation and one column per unknown.

Everything now depends on what the product Ax *means*. There are two ways to compute it, and the difference between them is the difference between the two pictures.

1.3.1 Two ways to read Ax

Consider a concrete product, with the unknowns set to $x = (1, 2)$:

$$\begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = ?$$

Reading 1: rows (the dot-product view). Take each row of A and dot it with x. The first row $(2, 5)$ against $(1, 2)$ gives $2 \cdot 1 + 5 \cdot 2 = 12$; the second row $(1, 3)$ against $(1, 2)$ gives $1 \cdot 1 + 3 \cdot 2 = 7$:

$$\begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \cdot 1 + 5 \cdot 2 \\ 1 \cdot 1 + 3 \cdot 2 \end{bmatrix} = \begin{bmatrix} 12 \\ 7 \end{bmatrix}.$$

This is the mechanical rule everyone learns, and it is the row picture in disguise: each entry of Ax is one equation's left-hand side.

Reading 2: columns (the combination view). Treat the entries of x as the amounts of each column. One copy of the first column plus two copies of the second:

$$\begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = 1 \begin{bmatrix} 2 \\ 1 \end{bmatrix} + 2 \begin{bmatrix} 5 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 10 \\ 6 \end{bmatrix} = \begin{bmatrix} 12 \\ 7 \end{bmatrix}.$$

Same answer, of course. But this reading says something the first one hides: Ax is a *linear combination of the columns of* A, with the entries of x as the coefficients. This is the single most important sentence in the chapter.

The meaning of Ax

For $A \in \mathbb{R}^{m\times n}$ with columns $a_1, \ldots, a_n$ and $x = (x_1, \ldots, x_n)^{\mathsf{T}}$,

$$Ax = x_1 a_1 + x_2 a_2 + \cdots + x_n a_n.$$

The product Ax is the linear combination of the columns of A weighted by the entries of x. Solving $Ax = b$ asks: *which combination of the columns produces b?*

1.3.2 When does $Ax = b$ have a solution?

Reading Ax as a combination of columns answers the existence question immediately. As x ranges over all of $\mathbb{R}^n$, the products Ax range over *all* linear combinations of the columns of

A. So $Ax = b$ has a solution precisely when b is one of those combinations.

Definition 1.2: Column Space

The *column space* of $A \in \mathbb{R}^{m\times n}$, written $C(A)$, is the set of all linear combinations of the columns of A:

$$C(A) = \{\, Ax : x \in \mathbb{R}^n \,\} \subseteq \mathbb{R}^m.$$

It is the set of all right-hand sides b that the columns can reach.

Proposition 1.3: Solvability of $Ax = b$

The system $Ax = b$ has a solution if and only if b lies in the column space $C(A)$.

Proof. If $Ax = b$ for some x, then $b = Ax$ is by definition a linear combination of the columns, so $b \in C(A)$. Conversely, if $b \in C(A)$ then $b = x_1a_1 + \cdots + x_na_n$ for some scalars x_i; collecting those scalars into $x = (x_1, \ldots, x_n)^{\mathsf{T}}$ gives $Ax = b$. □

For the 2×2 matrix $A = \left[\begin{smallmatrix} 2 & -1 \\ -1 & 2 \end{smallmatrix}\right]$, the two columns point in different directions and their combinations fill all of $\mathbb{R}^2$; the column space is the whole plane, so $Ax = b$ is solvable for every b — and, because the columns are independent, the solution is unique. When a square matrix has this property we call it *invertible*, and we may solve in one stroke by writing $x = A^{-1}b$ (we build A^{-1} in Chapter 2). The clean case is the one where the columns fill the space; the rest of this chapter is about what goes wrong when they do not.

1.4 Three Equations, Three Unknowns

The two pictures grow up gracefully. In $\mathbb{R}^3$ each equation is a plane (the row picture), and Ax is a combination of three column vectors in space (the column picture). A good way to feel the difference is Strang's *difference matrix*, where the columns have meaning.

Example (A difference matrix).

Take the three columns

$$u = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \quad v = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}, \quad w = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \qquad A = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix},$$

so A has u, v, w as its columns. Multiplying gives

$$Ax = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 - x_1 \\ x_3 - x_2 \end{bmatrix} = x_1u + x_2v + x_3w.$$

The output records the *differences* of consecutive components of x — that is why A is

called a difference matrix. Now solve $Ax = b$. The equations

$$\begin{aligned} x_1 &= b_1, \\ x_2 - x_1 &= b_2, \\ x_3 - x_2 &= b_3 \end{aligned}$$

unwind from the top: $x_1 = b_1$, then $x_2 = b_1 + b_2$, then $x_3 = b_1 + b_2 + b_3$. The solution *sums* the right-hand side:

$$x = \begin{bmatrix} b_1 \\ b_1 + b_2 \\ b_1 + b_2 + b_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} b = A^{-1}b.$$

A solution exists for *every* $b \in \mathbb{R}^3$, and it is unique — difference and sum undo each other, so A^{-1} is the *sum matrix*. The three columns u, v, w are independent and their combinations fill all of $\mathbb{R}^3$; the column space is the whole space. When that happens for n vectors in $\mathbb{R}^n$, we say they form a *basis* for $\mathbb{R}^n$, and the matrix holding them is invertible.

1.5 Singular Matrices: When the Columns Are Dependent

So far the columns have always filled the space, and life is easy. The interesting — and, for real problems, the usual — case is when they do not. Change just one column of the difference matrix and the whole story changes.

Example (A circular difference matrix).

Replace the last column so that the differences "wrap around":

$$C = \begin{bmatrix} 1 & 0 & -1 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}, \qquad Cx = \begin{bmatrix} x_1 - x_3 \\ x_2 - x_1 \\ x_3 - x_2 \end{bmatrix}.$$

Two things are different now. First, $Cx = 0$ has nonzero solutions: if $x_1 = x_2 = x_3$ then every difference vanishes, so $Cx = 0$ for the whole line of vectors $x = (c, c, c)$. The combination

$$1 \cdot \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} + 1 \cdot \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} + 1 \cdot \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

shows the three columns are *linearly dependent* — a nontrivial combination of them gives zero. Because we cannot reverse this collapse, no inverse C^{-1} exists.

Second, $Cx = b$ is not solvable for every b. Add the three equations

$$x_1 - x_3 = b_1,$$
$$x_2 - x_1 = b_2,$$
$$x_3 - x_2 = b_3$$

and the left side telescopes to 0, forcing

$$b_1 + b_2 + b_3 = 0.$$

A solution exists *only* when the components of b sum to zero. The reachable right-hand sides are not all of $\mathbb{R}^3$; they form the plane $b_1 + b_2 + b_3 = 0$ through the origin. That plane *is* the column space $C(C)$: the three columns, being dependent, lie in a common plane and their combinations never leave it.

This is the general dichotomy, stated for the square case.

Definition 1.4: Singular and Nonsingular Matrices

A square matrix $A \in \mathbb{R}^{n\times n}$ is *nonsingular* (equivalently, *invertible*) if its columns are linearly independent — the only solution of $Ax = 0$ is $x = 0$. Otherwise A is *singular*: its columns are linearly dependent, some nonzero x satisfies $Ax = 0$, and the combinations of the columns fail to fill $\mathbb{R}^n$.

Remark (The geometry of singular).

For a singular A the columns are squashed onto a smaller piece of space through the origin — in $\mathbb{R}^2$ onto a line, in $\mathbb{R}^3$ onto a line or a plane. Then $Ax = b$ has *no* solution when b lies off that line or plane, and *infinitely many* when b lies on it (add any solution of $Ax = 0$ to a particular solution). The clean "exactly one solution for every b" belongs only to the nonsingular case. Telling these apart efficiently is the job of *elimination*, the subject of the next chapter.

The same phenomenon appears already in two dimensions, where it is easy to draw.

Example (A singular 2×2 system).

Consider

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}, \qquad Ax = x_1 \begin{bmatrix} 1 \\ 2 \end{bmatrix} + x_2 \begin{bmatrix} 2 \\ 4 \end{bmatrix}.$$

The second column is exactly twice the first, so both columns lie on the single line through $(1, 2)$. Every combination Ax stays on that line: the column space is a line, not the plane. Thus $Ax = b$ has a solution only when b already lies on the line $\{t(1, 2)\}$ — for instance $b = (3, 6)$ is reachable (take $x = (3, 0)$ or $x = (1, 1)$, infinitely many ways), while $b = (1, 0)$ is unreachable. In the row picture the two equations

$$x_1 + 2x_2 = b_1,$$
$$2x_1 + 4x_2 = b_2$$

are parallel lines: they coincide when $b_2 = 2b_1$ (a whole line of solutions) and never meet otherwise (no solution). Dependent columns and parallel rows are the same defect seen from two sides.

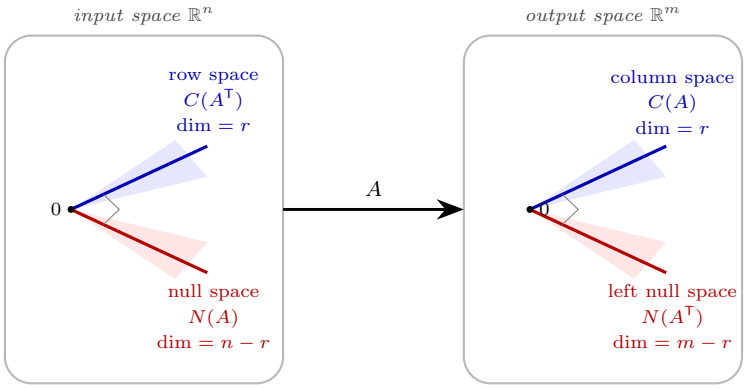
1.6 What to Carry Forward

We met one system written three ways. The *row picture* intersects lines or planes and is the picture from school; it stops scaling past three dimensions. The *column picture* reads $Ax = b$ as a search for the right combination of the columns of A, and it is the picture that organizes everything ahead. The bridge between them is the product Ax, which we can compute by rows (dot products) or — more revealingly — by columns ($Ax = x_1a_1 + \cdots + x_na_n$).

From the column reading, two questions fall out that will drive the next four chapters. *Existence:* $Ax = b$ is solvable exactly when b lies in the column space $C(A)$ — the set of all reachable combinations. *Uniqueness:* the solution is one-of-a-kind exactly when the columns are independent, so that the only combination giving zero is the trivial one. When both hold for a square matrix, it is nonsingular and $x = A^{-1}b$; when the columns collapse onto a lower-dimensional line or plane through the origin, the matrix is singular, and $Ax = b$ has either no solution or infinitely many.

Remark (Looking ahead).

Strang's advice for every matrix you meet: ask *"what is it doing?"* The difference matrix takes differences; its inverse sums. Rectangular matrices — seven equations in three unknowns — are never invertible, yet the symmetric square matrix $A^{\mathsf{T}}A$ that grows out of them often is, and it is the key to least squares (Chapter 5). The list of questions raised here — which b are reachable, which x vanish, how the spaces inside A fit together — becomes the theory of the *four fundamental subspaces*, the backbone of this book.



Chapter 2

Solving Linear Equations by Elimination

In Chapter 1 we learned to *read* a system $Ax = b$ — as intersecting planes, or, better, as a search for the right combination of the columns of A. Now we learn to *solve* it. The method is the one a computer uses millions of times a second, the one you half-remember from school: systematically subtract one equation from another until the unknowns peel off one at a time. It is called *elimination*, and Gauss organized it two hundred years ago.

What is new here is not the arithmetic but the bookkeeping. Every elimination step — "subtract 3 times row 1 from row 2" — *is itself a matrix*, multiplying A from the left. Stringing those steps together turns the messy process of elimination into a single clean statement about matrices: $A = LU$, the original matrix written as a lower triangular L times an upper triangular U. The factorization $A = LU$ is elimination, remembered. It tells us how to solve $Ax = b$ fast, it exposes exactly when a matrix is invertible, and — because L and U are triangular — it is the form in which real software stores a matrix. Along the way we will pin down matrix multiplication itself (there are several illuminating ways to see it), build the inverse A^{-1} by Gauss–Jordan, and handle the one thing that can go wrong — a zero in a pivot position — with a row exchange, the *permutation* matrix P. We close with the transpose and the symmetric matrices $A = A^{\mathsf{T}}$ that will run through the rest of the book.

2.1 Elimination: Forward and Back

Take three equations in three unknowns. Strang's running example for this chapter is

$$\begin{aligned} x + 2y + \ \ z &= 2, \\ 3x + 8y + \ \ z &= 12, \\ 4y + \ \ z &= 2, \end{aligned} \qquad A = \begin{bmatrix} 1 & 2 & 1 \\ 3 & 8 & 1 \\ 0 & 4 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 2 \\ 12 \\ 2 \end{bmatrix}. \tag{2.1}$$

Elimination comes in two passes. The *forward* pass uses the equations from the top down to clear out the unknowns below the diagonal, turning A into an upper triangular matrix U. The *back* pass — back substitution — then solves the triangular system from the bottom

up.

2.1.1 The Forward Pass: Producing the Pivots

The first *pivot* is the entry in the top-left corner, the 1 in row 1, column 1. Its job is to clear the rest of column 1. The entry below it in row 2 is 3, so we subtract 3 times row 1 from row 2; the entry in row 3 is already 0, so that row needs nothing. The multiplier we used, 3, is worth remembering — it will reappear as an entry of L.

$$\begin{bmatrix} 1 & 2 & 1 \\ 3 & 8 & 1 \\ 0 & 4 & 1 \end{bmatrix} \xrightarrow{\text{row}_2 - 3\,\text{row}_1} \begin{bmatrix} 1 & 2 & 1 \\ 0 & 2 & -2 \\ 0 & 4 & 1 \end{bmatrix}.$$

Column 1 is now clean. The second pivot is the entry that has surfaced in row 2, column 2: the 2. Use it to clear column 2 below: the entry in row 3 is 4, so subtract 2 times row 2 from row 3 (multiplier 2):

$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 2 & -2 \\ 0 & 4 & 1 \end{bmatrix} \xrightarrow{\text{row}_3 - 2\,\text{row}_2} \begin{bmatrix} 1 & 2 & 1 \\ 0 & 2 & -2 \\ 0 & 0 & 5 \end{bmatrix} = U.$$

We have arrived at U, upper triangular, with the three *pivots* $1, 2, 5$ sitting on its diagonal. Below the diagonal is all zeros — that is the whole point of the forward pass.

Definition 2.1: Pivots, and the Triangular Form U

In elimination, the **pivot** in a given step is the first nonzero entry in the row doing the eliminating; it occupies a diagonal position and is used to clear the entries directly below it. When elimination runs to completion with no zero in any pivot position, the result is an upper triangular matrix U whose diagonal entries are the pivots. The number we multiply the pivot row by before subtracting — the **multiplier** ℓ_{ij} for clearing row i using pivot row j — records exactly how much of row j was removed from row i.

Remark (Pivots may not be zero).

The pivot is a divisor: each multiplier is (entry to clear) / (pivot), and the back substitution will divide by the pivots too. So a pivot is never allowed to be 0. If a 0 turns up in a pivot position, we try to rescue it by swapping in a lower row that has a nonzero entry there — a *row exchange*, Section 2.5. If no row below has a nonzero entry in that column, elimination has discovered that A is *singular*: there is no full set of pivots, and $Ax = b$ has no unique solution. The pivots are where invertibility lives.

2.1.2 Carrying the Right-Hand Side; Back Substitution

The same operations must hit b. The cleanest way to do it by hand is to append b as an extra column — the *augmented matrix* $[A \mid b]$ — and eliminate the whole thing at once, so

the right-hand side rides along automatically:

$$\left[\begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 3 & 8 & 1 & 12 \\ 0 & 4 & 1 & 2 \end{array}\right] \longrightarrow \left[\begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & 2 & -2 & 6 \\ 0 & 4 & 1 & 2 \end{array}\right] \longrightarrow \left[\begin{array}{ccc|c} 1 & 2 & 1 & 2 \\ 0 & 2 & -2 & 6 \\ 0 & 0 & 5 & -10 \end{array}\right].$$

(In the right-hand column: $12 - 3 \cdot 2 = 6$, then $2 - 2 \cdot 6 = -10$.) Elimination has turned the original $Ax = b$ into the triangular system $Ux = c$ with

$$U = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 2 & -2 \\ 0 & 0 & 5 \end{bmatrix}, \qquad c = \begin{bmatrix} 2 \\ 6 \\ -10 \end{bmatrix}.$$

Because the steps are reversible, $Ux = c$ and $Ax = b$ have exactly the same solutions. And a triangular system is solved instantly from the bottom up — this is *back substitution*. The last equation involves only z:

$$5z = -10 \;\Rightarrow\; z = -2; \qquad 2y - 2z = 6 \;\Rightarrow\; 2y = 6 - 4 \;\Rightarrow\; y = 1; \qquad x + 2y + z = 2 \;\Rightarrow\; x = 2.$$

So $x = (2, 1, -2)$. Substituting back into (2.1) checks all three equations. The forward pass made the matrix triangular; the back pass read off the answer.

Elimination in one breath

To solve $Ax = b$: (1) *forward eliminate* — use each pivot to clear the column below it, carrying b along, until A becomes upper triangular U and b becomes c; (2) *back substitute* — solve $Ux = c$ from the bottom equation up. The pivots are the diagonal entries of U; with no row exchanges the product of the pivots is the determinant of A (Chapter 6; each row exchange flips the sign), and it is nonzero exactly when A is invertible.

2.2 Elimination as Matrix Multiplication

Here is the idea that converts elimination from a procedure into algebra. Each elimination step — "subtract 3 times row 1 from row 2" — can be performed by *multiplying A on the left* by a particular matrix. To see why, we first need to be comfortable with how a matrix multiplies from the left, which is really a statement about combining *rows*.

2.2.1 Rows, Columns, and the Product Ax

We already know from Chapter 1 that A times a *column* vector x is a combination of the *columns* of A. The mirror image is just as useful: a *row* vector times A is a combination of the *rows* of A.

$$\begin{bmatrix} c_1 & c_2 & c_3 \end{bmatrix} \begin{bmatrix} — \,\text{row}_1\, — \\ — \,\text{row}_2\, — \\ — \,\text{row}_3\, — \end{bmatrix} = c_1(\text{row}_1) + c_2(\text{row}_2) + c_3(\text{row}_3).$$

A row vector on the left scoops up a combination of the rows; a column vector on the right scoops up a combination of the columns. *Left-multiply to combine rows; right-multiply to combine columns.* Keep that slogan — it is the whole secret of elimination matrices.

2.2.2 The Elimination Matrices E_{ij}

To subtract 3 times row 1 from row 2 and leave the other rows alone, we want a matrix E whose rows pick out exactly those combinations. Start from the identity I (whose rows reproduce A unchanged) and adjust the row that should change. Row 2 of the output must be $(\text{row}_2 - 3\,\text{row}_1)$, so we put -3 in the $(2,1)$ slot:

$$E_{21} = \begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad E_{21}A = \begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ 3 & 8 & 1 \\ 0 & 4 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 2 & -2 \\ 0 & 4 & 1 \end{bmatrix}.$$

Read the middle row of E_{21}: it says "new row $2 = -3 \cdot \text{row}_1 + 1 \cdot \text{row}_2$," precisely the step we did by hand. The matrix E_{ij} that subtracts ℓ_{ij} times row j from row i is the identity with the single extra entry $-\ell_{ij}$ in position (i,j).

The second forward step — subtract 2 times row 2 from row 3 — is E_{32} with a -2 in position $(3,2)$. (The step that would clear position $(3,1)$ is $E_{31} = I$, since that entry was already 0.) Applying them in order builds U:

$$E_{32}\,E_{21}\,A = U, \qquad E_{32} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2 & 1 \end{bmatrix}.$$

By the associativity of matrix multiplication we may bundle the steps into one matrix $E = E_{32}E_{21}$, so that $EA = U$. The single matrix E records the entire forward elimination.

The meaning of $EA = U$

Forward elimination on A (with no row exchanges) is left-multiplication by a product of elementary matrices $E = \cdots E_{32}E_{31}E_{21}$, each of which is the identity with one off-diagonal entry. The result is the upper triangular U:

$$EA = U.$$

Every step is reversible (you can always add back what you subtracted), so E is invertible and $A = E^{-1}U$. That inverse is the matrix L, and it is the subject of Section 2.4.

2.2.3 Four Ways to Read a Matrix Product

Before cashing in $EA = U$, it pays to slow down on matrix multiplication itself, because the product AB is where every idea in this book is computed. Let A be $m \times n$ and B be $n \times p$, so the product $C = AB$ is $m \times p$. There are four equivalent ways to see it, and each is the right one for some purpose.

1. Entry by entry (row times column). The standard rule: the (i, j) entry of C is row i of A dotted with column j of B,

$$c_{ij} = \sum_{k=1}^{n} a_{ik}\, b_{kj}.$$

This is how you compute by hand, but it hides the structure.

2. Column by column. The product A times column j of B *is* column j of C. So

$$\text{column } j \text{ of } AB = A\,(\text{column } j \text{ of } B)\,:$$

each column of C is a combination of the columns of A. The *columns of C live in the column space of A.*

3. Row by row. Symmetrically, row i of A times the whole matrix B is row i of C. Each row of C is a combination of the rows of B — the *rows of C live in the row space of B.* This is the reading that powers elimination: EA takes combinations of the rows of A.

4. Columns times rows (the outer-product sum). Here is the one Strang loves and the one that returns in the SVD. A column of A (size $m \times 1$) times a row of B (size $1 \times p$) is a full $m \times p$ matrix — an *outer product.* For example

$$\begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} \begin{bmatrix} 1 & 6 \end{bmatrix} = \begin{bmatrix} 2 & 12 \\ 3 & 18 \\ 4 & 24 \end{bmatrix}.$$

Every column of this product is a multiple of $(2, 3, 4)$ and every row is a multiple of $(1, 6)$, so its column space and its row space are each a single line — it is a *rank-one* matrix. The full product AB is the *sum* of these outer products, one per shared index k:

$$AB = \sum_{k=1}^{n} (\text{column } k \text{ of } A)(\text{row } k \text{ of } B).$$

A big matrix product is built from rank-one pieces. Hold onto this; it is the seed of the singular value decomposition in Chapter 9.

Remark (A fifth way: blocks).

> If the matrices are cut into blocks that line up, the product follows the same row-times-column rule with the *blocks* as entries:
>
> $$\begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix} \begin{bmatrix} B_1 & B_2 \\ B_3 & B_4 \end{bmatrix} = \begin{bmatrix} A_1B_1 + A_2B_3 & A_1B_2 + A_2B_4 \\ A_3B_1 + A_4B_3 & A_3B_2 + A_4B_4 \end{bmatrix}.$$
>
> Block multiplication is what lets us carry the identity I alongside A in Gauss–Jordan, and it is how numerical libraries chop a huge matrix into cache-sized pieces.

2.3 The Inverse Matrix and Gauss–Jordan

For a square matrix, the most important question is whether it can be undone.

Definition 2.2: Inverse Matrix

A square matrix $A \in \mathbb{R}^{n\times n}$ is **invertible** (or **nonsingular**) if there is a matrix A^{-1} with

$$A^{-1}A = I = AA^{-1}.$$

The inverse, when it exists, is unique. If no such matrix exists, A is **singular**.

Invertibility is exactly the nonsingular case of Chapter 1: A^{-1} exists if and only if the columns of A are independent, equivalently the only solution of $Ax = 0$ is $x = 0$. If instead some nonzero x has $Ax = 0$, no inverse can exist — multiplying $Ax = 0$ by a would-be A^{-1} would give $x = 0$, a contradiction. The matrix

$$\begin{bmatrix} 1 & 3 \\ 2 & 6 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

is singular: 3 times column 1 minus column 2 is zero, the columns lie on one line, and there is no inverse.

2.3.1 Solving for the Inverse Is Solving $Ax = b$, n Times

What *is* A^{-1}? Reading $AA^{-1} = I$ one column at a time (the column-by-column view of Section 2.2.3) says: column j of A^{-1} is the vector x_j solving $Ax_j = e_j$, where e_j is column j of the identity. Finding the inverse is nothing but solving $Ax = b$ for the n special right-hand sides $e_1, \ldots, e_n$. For

$$A = \begin{bmatrix} 1 & 3 \\ 2 & 7 \end{bmatrix}, \qquad \text{find } A^{-1} = \begin{bmatrix} a & c \\ b & d \end{bmatrix} \text{ with } A\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix},\ A\begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

2.3.2 Gauss–Jordan: Solve Them All at Once

The trick is to handle both right-hand sides simultaneously. Augment A with the *whole* identity, then eliminate. Gauss's forward pass clears below the pivots; Jordan's extra back pass clears *above* them too, and finally scales each pivot to 1. When the left block has become I, the right block has become A^{-1}:

$$\left[\begin{array}{cc|cc} 1 & 3 & 1 & 0 \\ 2 & 7 & 0 & 1 \end{array}\right] \xrightarrow{\text{row}_2 - 2\,\text{row}_1} \left[\begin{array}{cc|cc} 1 & 3 & 1 & 0 \\ 0 & 1 & -2 & 1 \end{array}\right] \xrightarrow{\text{row}_1 - 3\,\text{row}_2} \left[\begin{array}{cc|cc} 1 & 0 & 7 & -3 \\ 0 & 1 & -2 & 1 \end{array}\right].$$

So

$$A^{-1} = \begin{bmatrix} 7 & -3 \\ -2 & 1 \end{bmatrix}, \qquad \text{check} \qquad \begin{bmatrix} 1 & 3 \\ 2 & 7 \end{bmatrix}\begin{bmatrix} 7 & -3 \\ -2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \checkmark$$

Theorem 2.3: Why Gauss–Jordan Produces A^{-1}

Row-reducing $[A \mid I]$ to $[I \mid X]$ shows that $X = A^{-1}$. The whole reduction is left-multiplication by one invertible matrix E (the product of the elementary steps). Applied to the left block, $EA = I$, so $E = A^{-1}$. Applied to the right block, $EI = E = X$. Hence $X = A^{-1}$. If elimination cannot turn the left block into I — a pivot is missing — then A is singular and has no inverse.

Remark (Two facts about inverses we will reuse).

For invertible A, B of the same size, the inverse of a *product* reverses the order, like socks and shoes:

$$(AB)^{-1} = B^{-1}A^{-1},$$

because $(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = AA^{-1} = I$. And the inverse and transpose commute: $(A^{\mathsf{T}})^{-1} = (A^{-1})^{\mathsf{T}}$, sometimes written $A^{-\mathsf{T}}$.

2.4 The Factorization $A = LU$

Now we cash in. Forward elimination gave $EA = U$ with $E = E_{32}E_{31}E_{21}$. Instead of carrying the awkward product E, undo it: multiply by E^{-1} on the left,

$$EA = U \quad \Longrightarrow \quad A = E^{-1}U = LU, \qquad L := E^{-1}.$$

This is the most important factorization in computational linear algebra. The matrix U is the upper triangular form we already produced. The new matrix $L = E^{-1}$ is *lower* triangular with 1's on its diagonal — and, remarkably, building it requires no work at all.

2.4.1 Why L Is So Simple

Watch the two-by-two case first, because everything important is already visible there. Take

$$A = \begin{bmatrix} 2 & 1 \\ 8 & 7 \end{bmatrix}.$$

One step clears the 8: subtract 4 times row 1 from row 2 (multiplier $\ell_{21} = 4$). The elimination matrix and its inverse are

$$E_{21} = \begin{bmatrix} 1 & 0 \\ -4 & 1 \end{bmatrix}, \qquad L = E_{21}^{-1} = \begin{bmatrix} 1 & 0 \\ 4 & 1 \end{bmatrix}.$$

To *invert* an elementary step, just flip the sign of its one off-diagonal entry: subtracting 4 times row 1 is undone by *adding* 4 times row 1. So L carries $+4$ where E carried -4. Then

$$\begin{bmatrix} 2 & 1 \\ 8 & 7 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 \\ 4 & 1 \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix}}_{U}.$$

The lone multiplier 4 sits in L exactly where it was eliminated, in position $(2,1)$.

The three-by-three case shows the real payoff. We had $A = E_{21}^{-1}E_{31}^{-1}E_{32}^{-1}U$, so $L = E_{21}^{-1}E_{31}^{-1}E_{32}^{-1}$. One might fear that multiplying these inverses together scrambles the entries. It does not — *when there are no row exchanges, the multipliers drop straight into L with no interaction.* For the running example (2.1) the multipliers were $\ell_{21} = 3$, $\ell_{31} = 0$, $\ell_{32} = 2$, and indeed

$$A = \begin{bmatrix} 1 & 2 & 1 \\ 3 & 8 & 1 \\ 0 & 4 & 1 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 0 & 2 & 1 \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} 1 & 2 & 1 \\ 0 & 2 & -2 \\ 0 & 0 & 5 \end{bmatrix}}_{U}.$$

Each multiplier ℓ_{ij} goes verbatim into position (i,j) of L: the 3, the 0, the 2. You can read the entire elimination history off L.

Theorem 2.4: $A = LU$ and the Multipliers

If A is reduced to the upper triangular U by forward elimination *without row exchanges*, then

$$A = LU,$$

where U has the pivots on its diagonal and L is lower triangular with 1's on its diagonal and the multiplier ℓ_{ij} in each below-diagonal position (i,j). The factorization is unique once we fix the 1's on the diagonal of L.

Proof. Forward elimination is a product of elementary matrices $E_{ij}A = U$ (the factors applied left to right, each E_{ij} subtracting a multiple of a higher row from the strictly lower row i, clearing position (i,j)). Let L be the inverses applied in reverse order, so $A = LU$. Each inverse E_{ij}^{-1} is the identity with the single entry $+\ell_{ij}$ in position (i,j). The crucial fact is that these particular inverses multiply with *no carrying*: when we always subtract a higher row from a strictly lower one, the entry $+\ell_{ij}$ in E_{ij}^{-1} never gets multiplied by an off-diagonal entry of another factor, so the product simply collects every ℓ_{ij} into its own slot. Hence L is lower triangular with 1's on the diagonal and ℓ_{ij} below, and $A = LU$. □

Remark (Why $L = E^{-1}$ is nicer than E).

The product $E = E_{32}E_{31}E_{21}$ that takes A *to* U can develop "cross terms" — combining two subtractions can plant a stray number where neither step put one. (Subtracting twice row 1 from row 2, then five times the new row 2 from row 3, leaves a 10 in the corner of E.) But the *inverse* order, $L = E_{21}^{-1}E_{31}^{-1}E_{32}^{-1}$, applies the undo-steps from the bottom up and never collides: the multipliers land untouched. This is exactly why we factor $A = LU$ rather than carry $EA = U$ — L is free, E is not.

2.4.2 Solving $Ax = b$ with LU

Once $A = LU$ is in hand, solving $Ax = b$ for *any* right-hand side is two quick triangular sweeps:

$$Ax = b \iff L(Ux) = b \implies \underbrace{Lc = b}_{\text{forward, solve for } c}, \quad \underbrace{Ux = c}_{\text{back, solve for } x}.$$

Solving $Lc = b$ from the top down is the forward elimination of b (it recreates the c we found by augmenting); solving $Ux = c$ from the bottom up is back substitution. The expensive part — producing L and U — is done *once*; after that each new b costs only the two cheap triangular solves. That is why software stores a matrix as its LU factors, not as A^{-1}.

Remark (Splitting off the pivots: $A = LDU$).

Sometimes we pull the pivots out of U into a diagonal matrix D, leaving 1's on the diagonal of what remains. For the 2×2 example,

$$\begin{bmatrix} 2 & 1 \\ 8 & 7 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 1 & \frac{1}{2} \\ 0 & 1 \end{bmatrix} = LDU.$$

Now L and U both have unit diagonals and $D = \text{diag}(\text{pivots})$ holds the pivots $2, 3$. The symmetric form $A = LDL^{\mathsf{T}}$ is what we will want for symmetric matrices below.

2.4.3 The Cost of Elimination

Why does this matter for big matrices? Count the operations. Clearing column 1 of an $n \times n$ matrix touches roughly n^2 entries; column 2 touches an $(n-1) \times (n-1)$ block, and so on. The total is on the order of

$$n^2 + (n-1)^2 + \cdots + 1^2 \approx \tfrac{1}{3}n^3$$

multiply-subtract operations to produce L and U, plus a cheap n^2 for each right-hand side. Building A^{-1} explicitly and multiplying $A^{-1}b$ would cost more and lose accuracy. Elimination — stored as $A = LU$ — is the workhorse precisely because it is the cheap, stable way to solve $Ax = b$.

2.5 Row Exchanges and Permutations: $PA = LU$

We assumed all along that no pivot came out zero. When one does, $A = LU$ needs a small repair. The fix is a *permutation matrix*.

Definition 2.5: Permutation Matrix

A **permutation matrix** P is obtained by reordering the rows of the identity I. Multiplying PA reorders the rows of A in the same way; multiplying AP reorders the columns. Every P is invertible, and its inverse is its transpose:

$$P^{-1} = P^{\mathsf{T}}, \qquad \text{equivalently} \qquad P^{\mathsf{T}}P = I.$$

For instance, the matrix that swaps the first two rows is

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad PA \text{ has rows } 2, 1, 3 \text{ of } A.$$

That $P^{\mathsf{T}}P = I$ is the column picture again: the columns of P are the standard basis vectors

$e_1, \ldots, e_n$ in some order, so $e_i^{\mathsf{T}} e_j$ is 1 when $i = j$ and 0 otherwise — exactly I. (Note $PA \neq AP$ in general: multiplication is not commutative.)

2.5.1 When Elimination Needs a Swap

Suppose a 0 appears in a pivot position but a row below has a nonzero entry in that column. Swap the two rows — multiply by a permutation P — and elimination continues. If we knew in advance which swaps elimination would need, we could do them all up front, on PA, and then run clean elimination with no exchanges. That gives the general factorization.

Theorem 2.6: $PA = LU$ for Any Invertible A

For every invertible square matrix A there is a permutation matrix P (recording the row exchanges) such that PA factors with no further exchanges:

$$PA = LU,$$

with L lower triangular (1's on the diagonal) and U upper triangular (nonzero pivots on the diagonal). When no exchanges are needed, $P = I$ and this is the earlier $A = LU$.

Example (A swap forces a permutation).

The matrix

$$A = \begin{bmatrix} 0 & 2 \\ 3 & 4 \end{bmatrix}$$

has a 0 in the first pivot position, so elimination stalls immediately. Swap the rows with $P = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$:

$$PA = \begin{bmatrix} 3 & 4 \\ 0 & 2 \end{bmatrix},$$

which is already upper triangular. So $PA = LU$ with $L = I$ and $U = \left[\begin{smallmatrix} 3 & 4 \\ 0 & 2 \end{smallmatrix}\right]$. The pivots are 3 and 2; the original A is invertible, but only after the exchange does it reveal its pivots.

Remark (How many permutations are there?).

There are $n!$ permutation matrices of size n — one for each reordering of n rows. They are closed under multiplication and inverse (the inverse of a reordering is the reverse reordering, which is also the transpose), so they form a group. In practice good software always permutes to put the *largest* available entry in the pivot position ("partial pivoting"), even when a nonzero pivot is available, because dividing by a large pivot keeps round-off small.

2.6 Transposes and Symmetric Matrices

The last ingredient of the chapter is an operation we have used informally and now name. The *transpose* flips a matrix across its diagonal, turning rows into columns.

Definition 2.7: Transpose

The **transpose** A^{T} of an $m \times n$ matrix A is the $n \times m$ matrix whose entries are

$$(A^{\mathsf{T}})_{ij} = A_{ji}.$$

Row i of A becomes column i of A^{T}, and vice versa.

For example,

$$\begin{bmatrix} 1 & 3 \\ 2 & 3 \\ 4 & 1 \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} 1 & 2 & 4 \\ 3 & 3 & 1 \end{bmatrix}.$$

The transpose obeys two rules we will lean on constantly. It distributes over sums, $(A + B)^{\mathsf{T}} = A^{\mathsf{T}} + B^{\mathsf{T}}$, and — like the inverse — it *reverses the order* of a product:

$$(AB)^{\mathsf{T}} = B^{\mathsf{T}} A^{\mathsf{T}}.$$

(Entry (i, j) of $(AB)^{\mathsf{T}}$ is entry (j, i) of AB, which is row j of A dotted with column i of B — and that is exactly row i of B^{T} dotted with column j of A^{T}.) Combining the transpose with the inverse gives $(A^{-1})^{\mathsf{T}} = (A^{\mathsf{T}})^{-1}$, since transposing $A^{-1}A = I$ yields $A^{\mathsf{T}}(A^{-1})^{\mathsf{T}} = I$.

2.6.1 Symmetric Matrices

The matrices that equal their own transpose are special enough to earn a name, and they will dominate the second half of the book (eigenvalues, positive definiteness, the SVD).

Definition 2.8: Symmetric Matrix

A square matrix A is **symmetric** if

$$A^{\mathsf{T}} = A, \qquad \text{equivalently} \qquad A_{ij} = A_{ji} \text{ for all } i, j.$$

The matrix is mirror-symmetric across its main diagonal.

Theorem 2.9: $A^{\mathsf{T}}A$ Is Always Symmetric

For *any* matrix A (square or not), the products $A^{\mathsf{T}}A$ and AA^{T} are symmetric.

Proof. Apply the order-reversing rule to $A^{\mathsf{T}}A$, remembering $(A^{\mathsf{T}})^{\mathsf{T}} = A$:

$$(A^{\mathsf{T}}A)^{\mathsf{T}} = A^{\mathsf{T}}(A^{\mathsf{T}})^{\mathsf{T}} = A^{\mathsf{T}}A.$$

So $A^{\mathsf{T}}A$ equals its own transpose. The same computation gives $(AA^{\mathsf{T}})^{\mathsf{T}} = AA^{\mathsf{T}}$. □

This little fact is the hinge of least squares (Chapter 5): a rectangular A is never invertible, but the symmetric square matrix $A^{\mathsf{T}}A$ it spawns usually is, and it is the matrix

that actually gets solved. Symmetry also makes the LU story cleaner: when A is symmetric (and no row exchanges are needed), elimination produces $U = DL^{\mathsf{T}}$, so

$$A = LDL^{\mathsf{T}},$$

the same L appearing on both sides. Half the work, and the symmetry is visible in the factors. We will meet symmetric matrices again and again — they have real eigenvalues and orthogonal eigenvectors, the cleanest behavior any matrix can have.

2.7 What to Carry Forward

We turned the school-room procedure of elimination into clean linear algebra. The *forward pass* drives A to an upper triangular U with the pivots on its diagonal; *back substitution* then solves the triangular system. Every step is a left-multiplication by an elementary matrix, and bundling the steps gives $EA = U$ — whose inverse is the *whole content* of this chapter:

$$A = LU \quad \text{(or } PA = LU \text{ if rows must be swapped).}$$

The triangular L stores the multipliers, U stores the pivots, and together they let us solve $Ax = b$ for any b by two cheap triangular sweeps. We learned to read a matrix product four ways — entry, column, row, and the outer-product sum of rank-one pieces — and used the column reading to build the inverse A^{-1} by Gauss–Jordan, which is just $Ax = e_j$ solved n times. The transpose and the symmetric matrices $A = A^{\mathsf{T}}$ round out the toolkit.

The throughline

Elimination decides everything. It produces the pivots, and the pivots decide invertibility: a full set of nonzero pivots means A is nonsingular and $A = LU$ (after permutations $PA = LU$); a missing pivot means A is singular. **Counting the pivots is the same as counting the independent columns** — and that count, the *rank* r, is where we go next. In Chapter 3 we name the spaces ($C(A)$, $N(A)$) that elimination has secretly been computing, and in Chapter 4 we read all four fundamental subspaces straight off the reduced form.

Chapter 3

Vector Spaces and the Solution of $Ax = b$

Elimination told us *how* to solve a square system with one solution. Now we ask the harder and more honest questions. When does $Ax = b$ have a solution at all? When it does, is the solution unique — and if not, what does the whole family of solutions look like? A rectangular matrix — more equations than unknowns, or fewer — almost never has the clean "one solution for every b" behavior of an invertible square matrix, and yet rectangular systems are exactly the ones that show up in real problems. To handle them we need a new language: the language of *vector spaces*.

The plan is to attach two spaces to every matrix A. The first, the *column space* $C(A)$, lives on the output side and answers existence: $Ax = b$ is solvable exactly when b lies in $C(A)$. The second, the *null space* $N(A)$, lives on the input side and answers uniqueness: it collects every solution of $Ax = 0$, and it controls how many solutions $Ax = b$ has once it has one. Elimination, run all the way down to the *reduced row echelon form* R, reads both spaces off the matrix mechanically. By the end of the chapter we will be able to write down the complete solution of *any* $Ax = b$ — a particular solution plus the whole null space — and to predict its shape from a single number, the rank r.

3.1 Spaces of Vectors

Before we can talk about the spaces hidden inside a matrix, we need to know what a space *is*. The idea is to isolate the one operation we have used since the first page — taking linear combinations — and to call "space" any collection of vectors closed under it.

Definition 3.1: Vector Space

A **vector space** is a nonempty collection V of vectors that is *closed under linear combinations*: whenever v and w are in V and c, d are real numbers, the combination $cv+dw$ is again in V. (In full, addition and scalar multiplication must obey the usual eight rules — commutativity and associativity of addition, distributive laws, a zero vector, and so on — but closure is the property we check in practice.)

The model example is $\mathbb{R}^n$, the space of all column vectors with n real components. Add two such vectors or scale one by a number and you stay in $\mathbb{R}^n$, so it is closed; $\mathbb{R}^2$ is the xy-plane and $\mathbb{R}^3$ is space. These are the spaces we draw. But the definition is deliberately blind to coordinates, and that is its power: the *solutions* of $Ax = 0$ will form a vector space, and so will spaces of matrices and functions, all governed by the same words.

One demand is built into closure and worth saying out loud. Taking $c = d = 0$ shows $0 \cdot v + 0 \cdot w = 0$ must lie in V: **every vector space contains the zero vector**. A collection that misses the origin cannot be a space.

Example (A collection that fails).

The set of vectors in $\mathbb{R}^2$ with both components positive — the open first quadrant — is *not* a vector space. It is closed under addition (two positive vectors add to a positive vector), but not under scalar multiplication: multiply $(1, 1)$ by -5 and you get $(-5, -5)$, which has left the quadrant. It also fails the zero-vector test, since $(0, 0)$ is not in the open quadrant. Closure under *all* linear combinations, positive and negative, is the real requirement.

3.1.1 Subspaces

Most spaces we care about sit inside a bigger one. A line through the origin lives inside the plane; the column space lives inside $\mathbb{R}^m$. We give the relationship a name.

Definition 3.2: Subspace

A **subspace** of a vector space V is a subset $S \subseteq V$ that is itself a vector space — that is, S is closed under linear combinations: if $v, w \in S$ and $c, d \in \mathbb{R}$, then $cv + dw \in S$. In particular every subspace contains the zero vector.

To test whether a subset is a subspace you do *not* recheck all eight axioms; they are inherited from V. You check two things: that S contains 0, and that S is closed under combinations. That is the whole job.

Example (Lines through the origin, and one that is not).

Fix a nonzero vector $v \in \mathbb{R}^2$. The set of all multiples $\{cv : c \in \mathbb{R}\}$ is a line through the origin, and it is a subspace: $c_1 v$ and $c_2 v$ combine to $(d_1 c_1 + d_2 c_2)v$, still a multiple of v. Now take a line that does *not* pass through the origin — say all points (x, y) with $y = x + 1$. Multiplying a point on it by 0 gives $(0, 0)$, which is off the line, so the line is not closed under scalar multiplication: not a subspace. Every subspace must run through the origin.

Cataloguing the subspaces of a small space makes the idea concrete.

The subspaces of $\mathbb{R}^2$ and $\mathbb{R}^3$

The subspaces of $\mathbb{R}^2$ are exactly:

(1) all of $\mathbb{R}^2$;

(2) any line through the origin;

(3) the zero subspace $Z = \{0\}$.

The subspaces of $\mathbb{R}^3$ are: all of $\mathbb{R}^3$; any plane through the origin; any line through the origin; and $Z = \{0\}$. In every case the subspace is a "flat" through the origin, of some dimension from 0 up to the whole space.

3.1.2 New subspaces from old

Two ways of combining subspaces will matter. The first behaves; the second does not.

Proposition 3.3: Intersection of Subspaces

If S and T are subspaces of V, then their intersection $S \cap T$ — the vectors lying in *both* — is again a subspace.

Proof. The zero vector is in S and in T, so it is in $S \cap T$. If $v, w \in S \cap T$, then $v, w \in S$ gives $cv + dw \in S$ (since S is closed), and likewise $cv + dw \in T$. So $cv + dw$ lies in both, hence in $S \cap T$. The intersection is closed under combinations. □

The *union* $S \cup T$, by contrast, is usually *not* a subspace. Take a plane P and a line L through the origin in $\mathbb{R}^3$. A vector from P added to a vector from L generally lands in neither, so $P \cup L$ is not closed. Two lines through the origin in $\mathbb{R}^2$ make the failure visible: their union is an "X", but adding a vector from one arm to a vector from the other lands off the X. The right way to combine subspaces additively is the *sum* $S + T$ of all $v + w$ with $v \in S$, $w \in T$, which *is* a subspace — but the union is not.

3.2 The Column Space and Solvability of $Ax = b$

Now we attach the first subspace to a matrix. Recall from Chapter 1 the one sentence that organizes everything: the product Ax is a *linear combination of the columns of* A, weighted by the entries of x. As x ranges over all of $\mathbb{R}^n$, the vector Ax sweeps out every combination of the columns. That swept-out set is a subspace, and it is the answer to "which b are reachable?"

Definition 3.4: Column Space

The **column space** of a matrix $A \in \mathbb{R}^{m \times n}$, written $C(A)$, is the set of all linear combinations of its columns:

$$C(A) = \{\, Ax : x \in \mathbb{R}^n \,\} \subseteq \mathbb{R}^m.$$

It is a subspace of $\mathbb{R}^m$ (the output space): a combination of combinations of the columns is again a combination of the columns, so $C(A)$ is closed.

The column space is built to answer the existence question, and it does so by its very definition.

Theorem 3.5: Solvability of $Ax = b$

The system $Ax = b$ has a solution if and only if b lies in the column space $C(A)$.

Proof. If $Ax = b$ for some x, then $b = Ax$ is a combination of the columns of A, so $b \in C(A)$. Conversely, if $b \in C(A)$ then $b = x_1a_1 + \cdots + x_na_n$ for some scalars x_i, where $a_1, \ldots, a_n$ are the columns; collecting those scalars into the vector $x = (x_1, \ldots, x_n)^\mathsf{T}$ gives $Ax = b$. So solvability of $Ax = b$ is exactly membership of b in $C(A)$. □

3.2.1 A rectangular example: when most b are out of reach

The theorem earns its keep on a tall matrix, where the columns cannot possibly fill the output space.

Example (Four equations, three columns).

Consider

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 3 \\ 3 & 1 & 4 \\ 4 & 1 & 5 \end{bmatrix}, \qquad m = 4,\ n = 3.$$

Here $Ax = b$ is four equations in three unknowns. The columns live in $\mathbb{R}^4$, and three vectors can never fill four-dimensional space, so $C(A)$ is a *proper* subspace of $\mathbb{R}^4$ — most right-hand sides b are unreachable, and for those $Ax = b$ has no solution. How big is $C(A)$? Look at the columns: the third is the sum of the first two,

$$\begin{bmatrix} 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix},$$

so the third column contributes nothing new. The first two columns point in genuinely different directions, and their combinations sweep out a *plane* (a two-dimensional subspace) through the origin in $\mathbb{R}^4$. That plane is $C(A)$. A vector b is reachable precisely when it lies on this plane, and then $Ax = b$ is solvable; off the plane there is no solution at all.

This is the typical situation for a rectangular matrix: the column space is a thin slice of the output space, and existence is a real constraint, not a formality.

3.3 The Null Space

The second subspace lives on the input side. Where the column space asks "which b can I hit?", the null space asks "which x get crushed to zero?" — and it is the key to uniqueness.

Definition 3.6: Null Space

The **null space** of $A \in \mathbb{R}^{m\times n}$, written $N(A)$, is the set of all solutions of the homogeneous equation $Ax = 0$:

$$N(A) = \{\, x \in \mathbb{R}^n : Ax = 0 \,\} \subseteq \mathbb{R}^n.$$

The null space is a subspace — and here the check is illuminating, because it shows directly that solution sets of homogeneous equations are spaces.

Proposition 3.7: $N(A)$ Is a Subspace of $\mathbb{R}^n$

The null space $N(A)$ is closed under linear combinations, hence a subspace of the input space $\mathbb{R}^n$.

Proof. First $0 \in N(A)$, since $A0 = 0$. Now suppose $x_1, x_2 \in N(A)$, so $Ax_1 = 0$ and $Ax_2 = 0$. For any scalars c, d,

$$A(cx_1 + dx_2) = c\,Ax_1 + d\,Ax_2 = c \cdot 0 + d \cdot 0 = 0,$$

so $cx_1 + dx_2 \in N(A)$. The null space is closed under combinations. □

Notice the contrast with the column and null spaces of the *same* rectangular matrix. For the 4×3 example above, $C(A)$ sits in $\mathbb{R}^4$ (one coordinate per row), while $N(A)$ sits in $\mathbb{R}^3$ (one coordinate per column). The two spaces live in different worlds. Its null space is the line of multiples of $(1, 1, -1)$, since column 1 plus column 2 minus column 3 is zero:

$$1\begin{bmatrix}1\\2\\3\\4\end{bmatrix} + 1\begin{bmatrix}1\\1\\1\\1\end{bmatrix} - 1\begin{bmatrix}2\\3\\4\\5\end{bmatrix} = \begin{bmatrix}0\\0\\0\\0\end{bmatrix}, \qquad \text{so } A\begin{bmatrix}1\\1\\-1\end{bmatrix} = 0.$$

Remark (Why the solution set of $Ax = b$ is usually *not* a subspace).

> The set of solutions of $Ax = b$ with $b \neq 0$ is *not* a vector space: it misses the origin, because $A0 = 0 \neq b$. Only the homogeneous equation $Ax = 0$ has a solution set ($N(A)$) that is a subspace. The inhomogeneous solution set, as we will see, is a *shifted* copy of $N(A)$ — a plane that has been pushed off the origin by one particular solution.

3.4 Solving $Ax = 0$: Pivots, Free Variables, and Special Solutions

We have defined $N(A)$, but a definition is not a computation. How do we actually find every solution of $Ax = 0$? The answer is elimination — the same row operations as before, now carrying A all the way to its cleanest form. Row operations do not change the solutions of

$Ax = 0$ (adding a multiple of one equation to another leaves the solution set alone), so they do not change the null space. We may reduce freely.

Example (The running matrix).

For the rest of this chapter our example is the 3×4 matrix

$$A = \begin{bmatrix} 1 & 2 & 2 & 2 \\ 2 & 4 & 6 & 8 \\ 3 & 6 & 8 & 10 \end{bmatrix}, \qquad m = 3,\ n = 4.$$

The columns are dependent (column 2 is twice column 1), so we expect a nontrivial null space. Elimination subtracts $2\times$ row 1 from row 2 and $3\times$ row 1 from row 3, then clears below the next pivot, reaching the echelon form

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

The third row collapsed to zero: row 3 of A was a combination of rows 1 and 2, and elimination exposed it. The pivots sit in columns 1 and 3, so there are two of them.

Definition 3.8: Rank, Pivot Columns, Free Columns

The **rank** r of A is the number of pivots produced by elimination. The columns holding pivots are the **pivot columns**; the others are the **free columns**. The variables multiplying free columns are the **free variables**: they can be set to any values, and the pivot variables are then determined.

In the running example $r = 2$. Columns 1 and 3 are pivot columns; columns 2 and 4 are free; so x_2 and x_4 are free variables. The count is already meaningful: $n - r = 4 - 2 = 2$ free variables, and we will get one independent solution of $Ax = 0$ from each.

3.4.1 Reduced row echelon form

We can push elimination one step further. Scale each pivot to 1 and clear *above* each pivot as well as below. The result is the **reduced row echelon form** R — the unique cleanest form of A, with pivots equal to 1 and zeros everywhere else in the pivot columns.

Example (From U to R).

Continuing the running example, divide row 2 by 2 and subtract $2\times$(new row 2) from row 1:

$$U = \begin{bmatrix} 1 & 2 & 2 & 2 \\ 0 & 0 & 2 & 4 \\ 0 & 0 & 0 & 0 \end{bmatrix} \longrightarrow \begin{bmatrix} 1 & 2 & 0 & -2 \\ 0 & 0 & 2 & 4 \\ 0 & 0 & 0 & 0 \end{bmatrix} \longrightarrow R = \begin{bmatrix} 1 & 2 & 0 & -2 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Now the pivot columns (1 and 3) hold the columns of the identity, $(1, 0, 0)$ and $(0, 1, 0)$. The free columns (2 and 4) carry the numbers that tell us exactly how to solve $Ax = 0$.

Why is R worth the extra work? Because we can read the solutions of $Rx = 0$ off the rows with no back-substitution. The pivot rows of R say

$$\begin{aligned} x_1 + 2x_2 \qquad - 2x_4 &= 0, \\ x_3 + 2x_4 &= 0, \end{aligned} \quad \text{i.e.} \quad \begin{aligned} x_1 &= -2x_2 + 2x_4, \\ x_3 &= \qquad - 2x_4. \end{aligned}$$

Each pivot variable (x_1, x_3) is expressed directly in terms of the free variables (x_2, x_4). Choose the free variables, and the rest follows.

3.4.2 Special solutions

The systematic way to produce a basis of solutions is to turn on one free variable at a time.

Definition 3.9: Special Solutions

For each free variable, the **special solution** is the solution of $Ax = 0$ obtained by setting that free variable to 1, all other free variables to 0, and solving for the pivot variables. There is one special solution per free variable, hence $n - r$ of them.

Example (The two special solutions).

For the running matrix the free variables are x_2, x_4.

- Set $(x_2, x_4) = (1, 0)$: then $x_1 = -2(1) + 2(0) = -2$ and $x_3 = -2(0) = 0$, giving

$$s_1 = \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix}.$$

  This reflects that column 2 is twice column 1: $-2(\text{col } 1) + 1(\text{col } 2) = 0$.

- Set $(x_2, x_4) = (0, 1)$: then $x_1 = 2$ and $x_3 = -2$, giving

$$s_2 = \begin{bmatrix} 2 \\ 0 \\ -2 \\ 1 \end{bmatrix}.$$

A direct check confirms $As_1 = 0$ and $As_2 = 0$. Every solution of $Ax = 0$ is a combination $c_1 s_1 + c_2 s_2$, so

$$N(A) = \{\, c_1 s_1 + c_2 s_2 : c_1, c_2 \in \mathbb{R} \,\}$$

is a two-dimensional plane through the origin in $\mathbb{R}^4$ — one dimension for each free variable.

Theorem 3.10: Special Solutions Describe the Null Space

The special solutions of $Ax = 0$ — one for each free variable — span the null space $N(A)$, and they are independent, so they form a basis. Consequently the null space has $n - r$ independent solutions:

$$\dim N(A) = n - r = (\text{number of free variables}).$$

Proof. Independence. Look at the free-variable slots. Special solution s_j carries a 1 in the slot of its own free variable and 0 in the slots of the other free variables. So in a combination $\sum c_j s_j$, the entry in free slot j equals exactly c_j; for the combination to vanish, every c_j must be 0. The special solutions are independent. *Spanning.* Take any $x \in N(A)$, let c_j be its values in the free slots, and form $y = x - \sum c_j s_j$. Then $y \in N(A)$ (a combination of null-space vectors), and by construction y has 0 in every free slot. But the pivot equations express each pivot variable in terms of the free ones, so a null-space vector with all free variables zero has all pivot variables zero too: $y = 0$. Hence $x = \sum c_j s_j$. The special solutions span $N(A)$, and there are $n - r$ of them. □

Remark (The I, F pattern and the null-space matrix).

When the pivot columns happen to come first, the reduced form has the block shape

$$R = \begin{bmatrix} I & F \\ 0 & 0 \end{bmatrix},$$

with an $r \times r$ identity I in the pivot columns and an $r \times (n - r)$ block F in the free columns. Stacking the special solutions side by side gives the **null-space matrix**

$$N = \begin{bmatrix} -F \\ I \end{bmatrix}, \qquad \text{and indeed} \qquad RN = \begin{bmatrix} I & F \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -F \\ I \end{bmatrix} = \begin{bmatrix} -F + F \\ 0 \end{bmatrix} = 0.$$

The bottom I (size $n - r$) puts a 1 in each free slot in turn; the top $-F$ records the resulting pivot values. In general the free columns may be interleaved with the pivot columns, but the recipe is the same: $-F$ on the pivot rows, the identity on the free rows.

3.5 Solving $Ax = b$: The Complete Solution

With $N(A)$ in hand we can solve the inhomogeneous system $Ax = b$ completely. The strategy has three steps: check solvability, find *one* particular solution, then add the entire null space.

3.5.1 Solvability conditions on b

We already know the abstract test: $Ax = b$ is solvable iff $b \in C(A)$. Elimination turns this into concrete conditions on the entries of b. Whenever a combination of the rows of A produces a zero row, the *same* combination of the entries of b must produce zero — otherwise the equation $0 =$(nonzero) appears and there is no solution.

Example (The solvability condition for the running matrix).

Row 3 of A equals row 1 plus row 2:

$$(3, 6, 8, 10) = (1, 2, 2, 2) + (2, 4, 6, 8).$$

So if $Ax = b$ holds, the third component of b must equal the sum of the first two: $b_3 = b_1 + b_2$. We can see this by eliminating on the augmented matrix $[A \;\; b]$, carrying b along:

$$\left[\begin{array}{cccc|c} 1 & 2 & 2 & 2 & b_1 \\ 2 & 4 & 6 & 8 & b_2 \\ 3 & 6 & 8 & 10 & b_3 \end{array}\right] \longrightarrow \left[\begin{array}{cccc|c} 1 & 2 & 2 & 2 & b_1 \\ 0 & 0 & 2 & 4 & b_2 - 2b_1 \\ 0 & 0 & 0 & 0 & b_3 - b_2 - b_1 \end{array}\right].$$

The last row reads $0 = b_3 - b_2 - b_1$. The system is solvable exactly when

$$b_3 - b_1 - b_2 = 0,$$

which is precisely the condition that b lie in the column space $C(A)$. The two conditions — "$b \in C(A)$" and "$b_3 = b_1 + b_2$" — are one and the same. A right-hand side that satisfies it is, for instance, $b = (1, 5, 6)$, since $6 = 1 + 5$.

3.5.2 A particular solution

Once solvability holds, one solution is easy to find: set every free variable to zero and solve for the pivot variables.

Example (The particular solution for $b = (1, 5, 6)$).

Put $x_2 = x_4 = 0$. The two pivot equations (from R, with the reduced right-hand side) become

$$x_1 + 2x_3 = 1,$$
$$x_3 = 3/2,$$

where the reduced right side is $b_1 = 1$ in the first pivot row, and the second pivot row of R has right-hand side $(b_2 - 2b_1)/2 = (5 - 2)/2 = 3/2$ (the pivot 2 has already been scaled to 1), so $x_3 = 3/2$ and then $x_1 = 1 - 2(3/2) = -2$. The particular solution is

$$x_p = \begin{bmatrix} -2 \\ 0 \\ 3/2 \\ 0 \end{bmatrix}, \qquad \text{and indeed} \qquad Ax_p = \begin{bmatrix} 1 \\ 5 \\ 6 \end{bmatrix} = b.$$

Setting the free variables to zero is what makes this one solution stand out from the infinitely many to come.

3.5.3 Adding the null space: the complete solution

The particular solution is one point; the full solution set is that point plus the entire null space. The reason is a one-line algebraic fact.

Theorem 3.11: Complete Solution of $Ax = b$

Suppose $Ax = b$ is solvable, and let x_p be any particular solution. Then the set of *all* solutions is

$$x_{\text{complete}} = x_p + x_n, \qquad x_n \in N(A),$$

that is, x_p plus every vector in the null space. Equivalently, the solution set is the null space shifted by x_p.

Proof. If $x_n \in N(A)$ then $A(x_p + x_n) = Ax_p + Ax_n = b + 0 = b$, so every such vector solves $Ax = b$. Conversely, if x is any solution, then $A(x - x_p) = b - b = 0$, so $x - x_p \in N(A)$; writing $x_n = x - x_p$ gives $x = x_p + x_n$ with x_n in the null space. Thus the solutions are exactly $x_p + N(A)$. □

Example (The complete solution for the running matrix).

Combining the particular solution with the two special solutions found earlier, the complete solution of $Ax = b$ with $b = (1, 5, 6)$ is

$$x = \underbrace{\begin{bmatrix} -2 \\ 0 \\ 3/2 \\ 0 \end{bmatrix}}_{x_p} + c_1 \underbrace{\begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix}}_{s_1} + c_2 \underbrace{\begin{bmatrix} 2 \\ 0 \\ -2 \\ 1 \end{bmatrix}}_{s_2}, \qquad c_1, c_2 \in \mathbb{R}.$$

Geometrically the solutions form a two-dimensional plane in $\mathbb{R}^4$ — a shifted copy of the null-space plane, pushed off the origin by x_p. The particular solution names *one* point on the plane; the special solutions span the directions you may slide within it. There are infinitely many solutions, all of this one form.

Remark (Particular plus homogeneous).

This $x_p + x_n$ structure is not special to matrices — it is the shape of the solution set of *every* linear equation, from $Ax = b$ to linear differential equations. One particular solution fixes the offset; the homogeneous solutions ($Ax = 0$) fill in all the freedom. Find one, add the null space, and you have them all.

3.6 The Four Cases: Rank and the Shape of the Solution

We can now see the whole landscape at once. Everything about the solutions of $Ax = b$ — how many there are, for which b they exist — is decided by how the rank r compares to the number of rows m and columns n. Since r counts pivots, and a matrix has at most one pivot per row and one per column, we always have $r \le m$ and $r \le n$. Three boundaries matter: whether $r = n$ (every column is a pivot column, so no free variables), and whether

$r = m$ (every row has a pivot, so no zero rows and no solvability conditions). Crossing them gives four cases.

3.6.1 Full column rank: $r = n$

When $r = n$ every column is a pivot column, so there are no free variables and no special solutions. The null space is just $\{0\}$, and the columns are independent.

$r = n$ (full column rank)

The reduced form has the shape $R = \begin{bmatrix} I \\ 0 \end{bmatrix}$ — an $n \times n$ identity stacked above $m - n$ zero rows. Since $N(A) = \{0\}$, a solution of $Ax = b$, *if one exists*, is **unique**. But $r = n \leq m$ allows zero rows in R, which impose solvability conditions on b. So

$$Ax = b \text{ has } \boxed{0 \text{ or } 1} \text{ solution.}$$

There is 0 when b lies off the column space, 1 when $b \in C(A)$. Independent columns are common in real applications, and they are exactly what makes least squares work later (Chapter 5).

3.6.2 Full row rank: $r = m$

When $r = m$ every row has a pivot, so R has no zero rows. Elimination never produces the equation $0 =$(something), so there are *no* solvability conditions: $Ax = b$ is solvable for every $b \in \mathbb{R}^m$. The column space is all of $\mathbb{R}^m$.

$r = m$ (full row rank)

The reduced form has the shape $R = [I \;\; F]$ — an $m \times m$ identity in the pivot columns, followed by the $m \times (n - m)$ free block F. Since there are no zero rows, $C(A) = \mathbb{R}^m$ and $Ax = b$ is solvable for *every* b. But $r = m < n$ leaves $n - r = n - m$ free variables, hence a nontrivial null space. So

$$Ax = b \text{ has } \boxed{\infty} \text{ many solutions for every } b,$$

a full $x_p + N(A)$ family of dimension $n - m$.

3.6.3 Full rank, square: $r = m = n$

When $r = m = n$ the matrix is square with a pivot in every row and column: it is *invertible*. The reduced form is the identity, $R = I$. There are no zero rows (no solvability conditions) and no free variables (no null space, $N(A) = \{0\}$).

$r = m = n$ (invertible)

With $R = I$, the system $Ax = b$ has **exactly one** solution for every $b \in \mathbb{R}^m$:

$$Ax = b \text{ has } \boxed{1} \text{ solution}, \qquad x = A^{-1}b.$$

This is the clean case from Chapter 1 — existence and uniqueness together, for every right-hand side. It is also the *only* case where "solve once and for all with A^{-1}" makes sense.

3.6.4 Deficient rank: $r < m$ and $r < n$

The general case has zero rows in R *and* free variables. There are solvability conditions on b (from the zero rows) and a nontrivial null space (from the free variables). This is our running example: $m = 3$, $n = 4$, $r = 2$, so $r < m$ and $r < n$.

$r < m$ and $r < n$ (the general rectangular case)

The reduced form has the shape $R = \left[\begin{smallmatrix} I & F \\ 0 & 0 \end{smallmatrix}\right]$, with both an F block (free variables present) and zero rows (solvability conditions present). So

$$Ax = b \text{ has } \boxed{0 \text{ or } \infty} \text{ many solutions.}$$

There is no solution when b violates a solvability condition (lies off $C(A)$); when $b \in C(A)$ there are infinitely many, the family $x_p + N(A)$ of dimension $n - r$. Our running matrix sat here: for $b = (1, 5, 6)$ we found the full two-parameter plane of solutions.

3.6.5 The four cases at a glance

Reading the reduced form R tells the whole story. The table collects the four possibilities (with pivot columns ordered first, so the blocks appear cleanly).

	$r = m = n$	$r = n < m$	$r = m < n$	$r < m,\ r < n$
R	I	$\left[\begin{smallmatrix} I \\ 0 \end{smallmatrix}\right]$	$\left[\begin{smallmatrix} I & F \end{smallmatrix}\right]$	$\left[\begin{smallmatrix} I & F \\ 0 & 0 \end{smallmatrix}\right]$
# solutions of $Ax = b$	1	0 or 1	∞	0 or ∞

The logic behind the table

Two yes/no questions decide everything.

- **Are the columns independent?** ($r = n$ vs. $r < n$.) If yes, $N(A) = \{0\}$ and a solution is unique when it exists. If no, there are free variables and a whole null space to add.
- **Do the columns span $\mathbb{R}^m$?** ($r = m$ vs. $r < m$.) If yes, $C(A) = \mathbb{R}^m$ and

$Ax = b$ is always solvable. If no, there are solvability conditions and some b have no solution.

Uniqueness is governed by $N(A)$; existence is governed by $C(A)$. The rank r is the one number that places a matrix into its case.

3.7 What to Carry Forward

Two subspaces, read off one matrix, answer the two questions that drive the subject. The *column space* $C(A) \subseteq \mathbb{R}^m$ collects the reachable right-hand sides: $Ax = b$ has a solution exactly when $b \in C(A)$ — existence. The *null space* $N(A) \subseteq \mathbb{R}^n$ collects the solutions of $Ax = 0$: it is $\{0\}$ precisely when the columns are independent, and otherwise it measures the freedom in the solution — uniqueness.

The computational engine is elimination carried all the way to the reduced row echelon form R. From R we read the rank r (the number of pivots), the pivot and free columns, and the special solutions — one per free variable — that form a basis for $N(A)$. To solve $Ax = b$: check the solvability conditions read off the zero rows of R; find a particular x_p by setting the free variables to zero; then write the complete solution as $x_p + N(A)$, one point plus the whole null-space plane.

Remark (Looking ahead).

We have measured the null space ($\dim N(A) = n - r$) and asserted that the pivot columns of A make a basis for the column space ($\dim C(A) = r$), but we have not yet said carefully what "dimension" and "basis" mean, nor why the row space should also have dimension r. That is the next chapter: independence, basis, dimension, and the *four* fundamental subspaces — column space, null space, row space, and left null space — all tied together by the single number r. The picture that results is the backbone of the whole course.

Chapter 4

Independence, Basis, and the Four Fundamental Subspaces

We have a vector space and we know how to recognize subspaces inside it. The two we care about most — the column space $C(A)$ and the null space $N(A)$ — both came out of a single matrix. But a subspace is an infinite set of vectors, and we cannot understand an infinite set by listing it. We need a small, finite handle on each space: a shortest list of vectors that still reaches everything in it. That list is a *basis*, and the number of vectors in it is the *dimension*. These two words turn vague spaces into objects we can count.

The whole chapter is built on one number. Run elimination on A and count the pivots; call that number the *rank* r. We will see that r is the dimension of the column space, $n - r$ is the dimension of the null space, r is also the dimension of the row space, and $m - r$ is the dimension of the left null space. Four subspaces, two of them sitting in $\mathbb{R}^m$ and two in $\mathbb{R}^n$, and a single number r that ties them together. The chapter ends with the picture that organizes all of linear algebra — Strang's "big picture" of the four fundamental subspaces and how they fit, at right angles, inside $\mathbb{R}^n$ and $\mathbb{R}^m$.

4.1 Linear Independence

Start with the most basic question you can ask about a list of vectors: is any one of them redundant? Could you throw one away and still build everything you could build before? If yes, the list carries dead weight. If no, every vector pulls in a direction the others cannot reach, and we call the list *independent*.

The clean way to say this avoids the words "redundant" and "reach." It asks instead: what combinations of the vectors give the zero vector?

Definition 4.1: Linear Independence

Vectors $v_1, v_2, \ldots, v_n$ are **linearly independent** if the only combination that produces the zero vector is the trivial one:

$$c_1v_1 + c_2v_2 + \cdots + c_nv_n = 0 \quad \Longrightarrow \quad c_1 = c_2 = \cdots = c_n = 0.$$

If some combination with not-all-zero coefficients gives 0, the vectors are **linearly dependent**.

Why is this the right definition? Suppose a nontrivial combination vanishes, say $c_1v_1 + \cdots + c_nv_n = 0$ with $c_1 \neq 0$. Then we can solve for v_1:

$$v_1 = -\frac{c_2}{c_1}\, v_2 - \cdots - \frac{c_n}{c_1}\, v_n,$$

so v_1 is a combination of the others — it *is* redundant. Dependence is exactly the situation where one vector is built from the rest. Independence rules that out.

Remark (Geometry).

Two vectors are independent when they do not lie on the same line through the origin (neither is a multiple of the other). Three vectors are independent when they do not lie in the same plane through the origin. Each new independent vector must point out of the flat span of the ones before it.

4.1.1 Independence and the Null Space

Here is the connection that makes independence computable, and it is pure column picture. Stack the vectors $v_1, \ldots, v_n$ as the columns of a matrix A. Then a combination of the columns is exactly Ax:

$$c_1v_1 + c_2v_2 + \cdots + c_nv_n \;=\; A \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}.$$

Asking "which combinations give zero?" is asking "which x solve $Ax = 0$?" — that is the null space all over again.

Theorem 4.2: Independence Test via $Ax = 0$

Let A have columns $v_1, \ldots, v_n$. The columns are linearly independent if and only if the null space contains only the zero vector, $N(A) = \{0\}$. Equivalently, the only solution of $Ax = 0$ is $x = 0$.

Proof. A nontrivial dependence $c_1v_1 + \cdots + c_nv_n = 0$ is precisely a nonzero vector $x = (c_1, \ldots, c_n)$ with $Ax = 0$, i.e. a nonzero element of $N(A)$. So the columns are dependent exactly when $N(A) \neq \{0\}$, and independent exactly when $N(A) = \{0\}$. □

This hands the whole question to elimination. Reduce A and count pivots. If every column is a pivot column there are no free variables, the only solution of $Ax = 0$ is $x = 0$, and the columns are independent. If even one column is free, set that free variable to 1 and you have produced a nonzero x in the null space — a dependence.

When you can rule out independence on sight

Suppose A is $m \times n$ with $m < n$ — more columns than rows, a wide matrix. Then $Ax = 0$ has more unknowns than equations, so there is at least one free variable, hence a nonzero solution. **Any set of n vectors in $\mathbb{R}^m$ with $n > m$ is automatically dependent.** You cannot have more independent vectors than the dimension of the space they live in.

Example (Two 2×2 matrices: one independent, one not).

The columns of $\left[\begin{smallmatrix} 1 & 2 \\ 3 & 4 \end{smallmatrix}\right]$ are independent: elimination gives the reduced form $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$, two pivots, no free variables, so $N(A) = \{0\}$. The columns of $\left[\begin{smallmatrix} 1 & 3 \\ 2 & 6 \end{smallmatrix}\right]$ are dependent: the second column is 3 times the first, elimination leaves $\left[\begin{smallmatrix} 1 & 3 \\ 0 & 0 \end{smallmatrix}\right]$ with one pivot and a free variable, and $x = (3, -1)$ solves $Ax = 0$.

Remark (A useful summary).

For the columns of A: if they are independent then every column is a pivot column, $\operatorname{rank} A = n$, and there are no free variables. If they are dependent then $\operatorname{rank} A < n$ and there is at least one free variable. Pivots count independent columns.

4.2 Spanning, Basis, and Dimension

Independence is half of what we want; it says a list has no waste. The other half says the list is big enough — that it actually builds the whole space.

Definition 4.3: Spanning a Space

Vectors $v_1, \ldots, v_k$ **span** a space S if every vector in S is a combination of $v_1, \ldots, v_k$. Equivalently, S is the set of all such combinations.

The columns of A span the column space $C(A)$ — that is exactly how $C(A)$ was defined. Spanning vectors are allowed to be wasteful: you can throw in extra, dependent vectors and still span. Independent vectors are allowed to be too few: two independent vectors in $\mathbb{R}^3$ span only a plane. A *basis* asks for both at once — enough to span, but no waste.

Definition 4.4: Basis

A **basis** for a vector space S is a list of vectors $v_1, \ldots, v_d$ that are

(i) linearly independent, and

(ii) span S.

A basis is the perfect handle on a space. Because the vectors span, every $w \in S$ is *some* combination $w = c_1v_1 + \cdots + c_dv_d$. Because they are independent, that combination is *unique* — the coefficients c_i are determined by w. (If w had two representations, their difference would be a nontrivial combination giving 0.) So a basis sets up an exact dictionary between vectors in S and their lists of coefficients.

Example (The standard basis of $\mathbb{R}^3$).

The columns of the identity,

$$\begin{bmatrix}1\\0\\0\end{bmatrix}, \quad \begin{bmatrix}0\\1\\0\end{bmatrix}, \quad \begin{bmatrix}0\\0\\1\end{bmatrix},$$

form a basis for $\mathbb{R}^3$. They are independent because $c_1e_1 + c_2e_2 + c_3e_3 = (c_1, c_2, c_3)$ is zero only when all $c_i = 0$, and they span because $(b_1, b_2, b_3) = b_1e_1 + b_2e_2 + b_3e_3$. This is not the only basis: any three vectors that are the columns of an invertible 3×3 matrix form a basis of $\mathbb{R}^3$, because invertibility means $Ax = 0$ forces $x = 0$ (independence) and $Ax = b$ is always solvable (spanning).

Example (Not a basis).

The vectors $(1, 1, 2)$, $(2, 2, 5)$, $(3, 3, 8)$ do *not* form a basis for $\mathbb{R}^3$. Place them as columns; the matrix has two equal rows (rows 1 and 2 are both $(1, 2, 3)$), so it cannot be invertible. Reducing shows only two pivots: the three vectors are dependent and span just a plane, not all of $\mathbb{R}^3$.

4.2.1 Dimension

A space has many different bases. The next fact — the one that makes "dimension" well defined — is that they all have the *same length*.

Theorem 4.5: Every Basis Has the Same Size

If a space S has a basis of d vectors, then every basis of S has exactly d vectors. This common number d is the **dimension** of S, written $\dim S$.

Proof. Let $v_1, \ldots, v_d$ and $w_1, \ldots, w_p$ both be bases of S, and suppose for contradiction $p > d$. Each w_j lies in S, so it is a combination of the v's: $w_j = \sum_{i=1}^{d} a_{ij}v_i$. Collect the coefficients into a $d \times p$ matrix $A = (a_{ij})$. Since A is $d \times p$ with $p > d$ (wide), the system $Ac = 0$ has a nonzero solution $c \in \mathbb{R}^p$. For that c,

$$\sum_{j=1}^{p} c_j w_j = \sum_{j=1}^{p} c_j \sum_{i=1}^{d} a_{ij}v_i = \sum_{i=1}^{d} \Big(\sum_{j=1}^{p} a_{ij}c_j \Big) v_i = \sum_{i=1}^{d} 0 \cdot v_i = 0,$$

a nontrivial dependence among $w_1, \ldots, w_p$ — contradicting that the w's are independent. So $p \le d$. By symmetry (swap the roles) $d \le p$, hence $p = d$. □

Remark (Rank versus dimension).

> Keep the words straight: a *matrix* has a rank; a *subspace* has a dimension. The rank of A is a property of the array of numbers; the dimension of $C(A)$ is a property of the space those numbers carve out. The whole point of the chapter is that these two numbers agree.

So $\dim \mathbb{R}^n = n$: every basis of $\mathbb{R}^n$ has exactly n vectors, because the standard basis has n. You cannot squeeze $n+1$ independent vectors into $\mathbb{R}^n$, and $n-1$ vectors can never span it.

4.3 Bases for the Column Space and Null Space

Now we put independence, basis, and dimension to work on the two spaces a matrix already gave us. We do it on one running example and keep returning to it for the rest of the chapter.

Example (The running matrix).

> Throughout this chapter we use
>
> $$A = \begin{bmatrix} 1 & 2 & 3 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 2 & 3 & 1 \end{bmatrix}, \qquad m = 3,\ n = 4.$$
>
> Elimination on A ends at the reduced row echelon form
>
> $$R = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$
>
> There are two pivots (in columns 1 and 2), so the rank is $r = 2$. Columns 3 and 4 are free.

Let us read both spaces off this R.

4.3.1 A Basis for the Column Space

The four columns of A span $C(A)$, but they are dependent — look at R. The reduced form tells us the relations among columns, because $Ax = 0$ and $Rx = 0$ have the same solutions. From R, column 3 equals column 1 plus column 2 (read off the entries $(1, 1, 0)$ in column 3 of R); the same relation holds among the columns of A: indeed $(3, 2, 3) = (1, 1, 1) + (2, 1, 2)$. Column 4 equals column 1. The dependent columns add nothing new.

The *pivot columns* — columns 1 and 2 — are independent and they reach everything the others reach. They are a basis for $C(A)$.

Theorem 4.6: Pivot Columns Are a Basis for $C(A)$

The pivot columns of A (the columns that become pivot columns in R) form a basis for the column space $C(A)$. Therefore

$$\dim C(A) = \text{number of pivots} = r = \operatorname{rank} A.$$

Proof. Reduction does not change which columns depend on which: $Ax = 0 \iff Rx = 0$, so any dependence among columns of A is a dependence among columns of R and vice versa. In R the pivot columns are independent (each pivot sits in a column with 1 in its own pivot row and 0 in the other pivot rows), so the corresponding columns of A are independent too. Every free column of R is a combination of the pivot columns of R (that is what the free entries record), so every free column of A is the same combination of the pivot columns of A. Hence the pivot columns span $C(A)$ and are independent: a basis. Their number is r. □

Remark (Take the columns of A, not of R).

A warning Strang stresses: the column space of R is *not* the column space of A. Elimination changes the columns. In our example $C(A)$ lives in $\mathbb{R}^3$ and contains $(1, 1, 1)$, but $C(R)$ contains only vectors with last entry 0. The pivots tell you *which* columns to take, but you must take them from the original A. So a basis for $C(A)$ here is $(1, 1, 1)$ and $(2, 1, 2)$, and $\dim C(A) = 2$.

4.3.2 A Basis for the Null Space

Now the other space. Because there are free variables, $N(A)$ is more than just the zero vector. We build a basis from *special solutions*: one solution of $Ax = 0$ for each free variable, setting that free variable to 1 and the other free variables to 0.

In our example the free columns are 3 and 4, so there are two special solutions. From R, the pivot variables satisfy $x_1 = -x_3 - x_4$ and $x_2 = -x_3$.

- Free variables $(x_3, x_4) = (1, 0)$ give $x_1 = -1$, $x_2 = -1$: the special solution $s_1 = (-1, -1, 1, 0)$.
- Free variables $(x_3, x_4) = (0, 1)$ give $x_1 = -1$, $x_2 = 0$: the special solution $s_2 = (-1, 0, 0, 1)$.

You can check directly that $As_1 = 0$ and $As_2 = 0$.

Theorem 4.7: Special Solutions Are a Basis for $N(A)$

The special solutions — one for each free variable — form a basis for the null space. Therefore

$$\dim N(A) = \text{number of free variables} = n - r.$$

Proof. Independence. Look at the free-variable slots. The special solution s_j has a 1 in the slot of its own free variable and 0 in the slots of the other free variables. So in a combination

$\sum c_j s_j$, the entry in free slot j is exactly c_j; for the combination to be 0 every c_j must be 0. The special solutions are independent.

Spanning. Take any $x \in N(A)$. Let c_j be its values in the free slots, and form $x - \sum c_j s_j$. This vector is again in $N(A)$ (a combination of null-space vectors), and by construction it has 0 in every free slot. But a null-space vector with all free variables zero is forced to have all pivot variables zero as well (the pivot equations express each pivot variable in terms of the free ones), so it is the zero vector. Hence $x = \sum c_j s_j$. The special solutions span $N(A)$.

The number of special solutions is the number of free variables, which is $n - r$. □

In the running example $\dim N(A) = n - r = 4 - 2 = 2$, and $\{s_1, s_2\}$ is a basis. The column space lives in $\mathbb{R}^3$ with dimension 2; the null space lives in $\mathbb{R}^4$ with dimension 2. Already two of the four numbers, r and $n - r$, have appeared.

4.4 The Four Fundamental Subspaces

Every $m \times n$ matrix A comes with *four* subspaces, not two. We have met the column space and the null space; the other two are the same construction applied to the transpose A^T.

Definition 4.8: The Four Fundamental Subspaces

For an $m \times n$ matrix A:

- the **column space** $C(A)$ — all combinations of the columns — a subspace of $\mathbb{R}^m$;
- the **null space** $N(A)$ — all solutions of $Ax = 0$ — a subspace of $\mathbb{R}^n$;
- the **row space** $C(A^\mathsf{T})$ — all combinations of the rows, i.e. the column space of A^T — a subspace of $\mathbb{R}^n$;
- the **left null space** $N(A^\mathsf{T})$ — all solutions y of $A^\mathsf{T} y = 0$ — a subspace of $\mathbb{R}^m$.

The names "row space" and "left null space" deserve a word. The combinations of the rows of A are the same as the combinations of the columns of A^T, so we just write $C(A^\mathsf{T})$. And $A^\mathsf{T} y = 0$ is the same as $y^\mathsf{T} A = 0$ (transpose both sides) — here y^T multiplies A from the *left*, which is why $N(A^\mathsf{T})$ is called the *left* null space.

Two of the spaces sit in $\mathbb{R}^n$ (row space and null space, indexed by columns) and two sit in $\mathbb{R}^m$ (column space and left null space, indexed by rows). We now find a basis and the dimension of each, all from the same elimination.

4.4.1 Column Space and Null Space (Reprise)

We already did these:

$$\dim C(A) = r \quad (\subseteq \mathbb{R}^m), \qquad \dim N(A) = n - r \quad (\subseteq \mathbb{R}^n).$$

A basis for $C(A)$ is the pivot columns of A; a basis for $N(A)$ is the special solutions.

4.4.2 The Row Space

For the row space we could transpose A and run elimination again — but we do not have to. Elimination acts on *rows*, taking combinations of them, so it never leaves the row space. And because every elementary row step is reversible, the rows of A and the rows of R build exactly the same combinations.

Theorem 4.9: The Nonzero Rows of R Are a Basis for the Row Space

The row space of A equals the row space of R. The nonzero rows of R (the r rows containing pivots) form a basis for it, so

$$\dim C(A^{\mathsf{T}}) = r.$$

Proof. Each row of R is a combination of rows of A (elimination only ever adds multiples of rows), so the row space of R sits inside the row space of A. Elimination is reversible, so each row of A is likewise a combination of rows of R; the two row spaces are equal. The nonzero rows of R span that space. They are independent: each contains a pivot 1 in a column where the other rows have 0, so no nontrivial combination of them can vanish. Hence they are a basis, and there are r of them. □

Remark (The headline).

Notice what just happened: $\dim C(A) = r$ and $\dim C(A^{\mathsf{T}}) = r$. **The row space and the column space have the same dimension.** The number of independent columns equals the number of independent rows. That common number is the rank r — the single most important number attached to a matrix.

In the running example the nonzero rows of R are $(1, 0, 1, 1)$ and $(0, 1, 1, 0)$. These two vectors in $\mathbb{R}^4$ are a basis for the row space $C(A^{\mathsf{T}})$, and $\dim C(A^{\mathsf{T}}) = 2 = r$. (Unlike the column space, here we *do* read the basis off R, because row operations preserve the row space.)

4.4.3 The Left Null Space

The fourth space is the null space of A^{T}. Since A^{T} is $n \times m$ with rank r (rows and columns of a matrix have the same rank, and $\operatorname{rank} A^{\mathsf{T}} = \operatorname{rank} A$), the number of free columns of A^{T} is $m - r$:

$$\dim N(A^{\mathsf{T}}) = m - r.$$

A basis comes for free from the same elimination, if we run it on the *augmented* matrix $[A \;\; I]$. Row-reducing A to R is the same as left-multiplying by some invertible matrix E, and carrying I along records that E:

$$E\,[A \;\; I] \longrightarrow [R \;\; E], \qquad \text{so} \qquad EA = R.$$

The bottom $m-r$ rows of R are zero. So the bottom $m-r$ rows of E, when they multiply A, produce zero rows: each such row y^{T} satisfies $y^{\mathsf{T}}A = 0$, i.e. $A^{\mathsf{T}}y = 0$. Those rows are independent (they are rows of an invertible E) and there are $m-r$ of them — a basis for the left null space.

Example (The fourth space on the running matrix).

Carrying I_3 alongside A and reducing gives

$$E = \begin{bmatrix} -1 & 2 & 0 \\ 1 & -1 & 0 \\ -1 & 0 & 1 \end{bmatrix}, \qquad EA = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = R.$$

Here $m-r = 3-2 = 1$: there is one zero row in R, the third. The corresponding bottom row of E is $y^{\mathsf{T}} = (-1, 0, 1)$, and indeed $y^{\mathsf{T}}A = -(\text{row } 1) + (\text{row } 3) = 0$ because rows 1 and 3 of A are equal. So $N(A^{\mathsf{T}})$ is the line through $(-1, 0, 1)$ in $\mathbb{R}^3$, with $\dim N(A^{\mathsf{T}}) = 1$.

4.5 The Rank Theorem: One Number Ties It Together

Collect the four dimensions. They are not four independent facts — they are four faces of one number r.

Theorem 4.10: Dimensions of the Four Subspaces

Let A be $m \times n$ with rank r. Then

$$\begin{aligned} \dim C(A) &= r, & C(A) &\subseteq \mathbb{R}^m, \\ \dim N(A) &= n-r, & N(A) &\subseteq \mathbb{R}^n, \\ \dim C(A^{\mathsf{T}}) &= r, & C(A^{\mathsf{T}}) &\subseteq \mathbb{R}^n, \\ \dim N(A^{\mathsf{T}}) &= m-r, & N(A^{\mathsf{T}}) &\subseteq \mathbb{R}^m. \end{aligned}$$

The most-used consequence is the relation between the two spaces inside $\mathbb{R}^n$. The column count n splits into pivot columns and free columns:

$$\underbrace{r}_{\dim C(A^{\mathsf{T}})} + \underbrace{(n-r)}_{\dim N(A)} = n.$$

Corollary 4.11: Rank–Nullity Theorem

For any $m \times n$ matrix A,

$$\dim C(A^{\mathsf{T}}) + \dim N(A) = r + (n-r) = n,$$

that is, $\operatorname{rank} A + \dim N(A) = n$. The number of pivot columns plus the number of free columns is the total number of columns.

There is a matching split inside $\mathbb{R}^m$, between the column space and the left null space:

$$\dim C(A) + \dim N(A^\mathsf{T}) = r + (m - r) = m.$$

Why this is the whole point

Every dimension question about A reduces to one computation: **run elimination and count the pivots.** That count is r. Then $C(A)$ and $C(A^\mathsf{T})$ have dimension r; $N(A)$ has dimension $n - r$; $N(A^\mathsf{T})$ has dimension $m - r$. Solvability of $Ax = b$ (b must lie in $C(A)$), uniqueness of solutions (controlled by $N(A)$), independence of columns ($N(A) = \{0\}$), independence of rows ($N(A^\mathsf{T}) = \{0\}$) — all of it is governed by where r falls relative to m and n.

4.5.1 Two Extreme Cases: Rank-One and Full-Rank

Example (A rank-one matrix).

The matrix

$$A = \begin{bmatrix} 1 & 4 & 5 \\ 2 & 8 & 10 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} 1 & 4 & 5 \end{bmatrix}$$

has every column a multiple of $(1, 2)$, so $\operatorname{rank} A = 1$. Here $m = 2$, $n = 3$, $r = 1$. The column space is the line through $(1, 2)$ in $\mathbb{R}^2$, dimension 1. The row space is the line through $(1, 4, 5)$ in $\mathbb{R}^3$, dimension 1 — again equal to the column-space dimension, as it must be. The null space has dimension $n - r = 2$ and the left null space has dimension $m - r = 1$. Every rank-one matrix factors as $A = uv^\mathsf{T}$ for column vectors u, v; these are the simplest nonzero matrices, and they are the building blocks of the singular value decomposition later.

Example (Full row rank and a useful null space).

Take the single-row matrix $A = [1\ 1\ 1\ 1]$, so $m = 1$, $n = 4$, $r = 1$. The equation $Ax = 0$ says $x_1 + x_2 + x_3 + x_4 = 0$ — a 3-dimensional subspace of $\mathbb{R}^4$, matching $n - r = 3$. A basis is the three special solutions

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

The column space is all of $\mathbb{R}^1$ (dimension 1), the row space is the line through $(1, 1, 1, 1)$ (dimension 1), and the left null space is $\{0\}$ with dimension $m - r = 0$ — its basis is the empty set.

4.6 The Big Picture

Two of the four subspaces live in $\mathbb{R}^n$: the row space $C(A^\mathsf{T})$ of dimension r, and the null space $N(A)$ of dimension $n-r$. Their dimensions add to n. Two live in $\mathbb{R}^m$: the column space $C(A)$ of dimension r, and the left null space $N(A^\mathsf{T})$ of dimension $m-r$, adding to m. The Fundamental Theorem of Linear Algebra says these pairs do not merely have complementary dimensions — they meet *at right angles*.

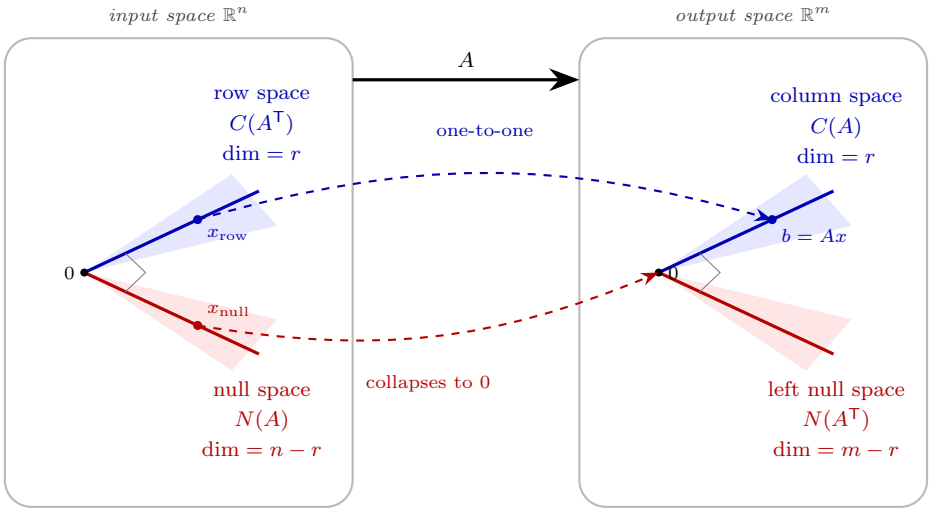
Theorem 4.12: Fundamental Theorem of Linear Algebra (Part 1)

For any $m \times n$ matrix A:

- In $\mathbb{R}^n$, the row space and the null space are **orthogonal complements**: every vector in $C(A^\mathsf{T})$ is perpendicular to every vector in $N(A)$, and together their dimensions fill $\mathbb{R}^n$ $(r + (n-r) = n)$.
- In $\mathbb{R}^m$, the column space and the left null space are orthogonal complements: every vector in $C(A)$ is perpendicular to every vector in $N(A^\mathsf{T})$, and $r + (m-r) = m$.

Proof. Take $x \in N(A)$, so $Ax = 0$. Each entry of the vector Ax is a row of A dotted with x, and all of them are zero. So x is orthogonal to every row of A, hence to every combination of rows — that is, to all of $C(A^\mathsf{T})$. Thus $N(A) \perp C(A^\mathsf{T})$. Their dimensions sum to $(n-r) + r = n$, the dimension of the whole space, so they are orthogonal complements: together they span $\mathbb{R}^n$ and meet only at 0. Applying the same argument to A^T (whose row space is $C(A)$ and whose null space is $N(A^\mathsf{T})$) gives the orthogonal splitting of $\mathbb{R}^m$. □

This is the picture to carry in your head for the rest of the course. On the left, the input space $\mathbb{R}^n$ splits into the row space and the null space at right angles. On the right, the output space $\mathbb{R}^m$ splits into the column space and the left null space at right angles. The matrix A sends $\mathbb{R}^n$ to $\mathbb{R}^m$: it crushes the null space to zero, and it maps the row space one-to-one onto the column space (both of dimension r). Nothing of A reaches the left null space — that is the part of $\mathbb{R}^m$ outside the column space, the directions in which $Ax = b$ has no solution.



Remark (What each space controls).

Read the picture for what it tells you about $Ax = b$. *Existence:* a solution exists exactly when b lies in the column space $C(A)$; the left null space $N(A^\mathsf{T})$ measures the constraints b must satisfy. *Uniqueness:* the solution is unique exactly when the null space $N(A)$ is just $\{0\}$; otherwise any element of $N(A)$ can be added to a solution. The row space $C(A^\mathsf{T})$ holds the one "particular" solution that has no null-space component — the shortest solution. We will return to this when we project onto these spaces and solve least squares.

4.7 Matrices Are Vectors Too

Everything above used vectors in $\mathbb{R}^n$, but the definitions of independence, span, basis, and dimension never mentioned coordinates — they only used addition and scalar multiplication. So they apply to any vector space, and a good test case is a space whose "vectors" are themselves matrices.

Example (The space of 3×3 matrices).

Let M be the set of all 3×3 matrices. We can add two such matrices and scale one by a number, the zero matrix is the additive identity, and all the space axioms hold — so M is a vector space. Choosing an element of M means choosing 9 numbers, so $\dim M = 9$ and M "looks like" $\mathbb{R}^9$. A basis is the nine matrices with a single 1 and the rest 0.

Inside M live natural subspaces. The symmetric matrices S (those with $A = A^\mathsf{T}$) have dimension 6: three free entries on the diagonal and three above it, with the lower triangle forced. The upper triangular matrices U also have dimension 6. Their intersection $D = S \cap U$ is the diagonal matrices, dimension 3. These dimensions obey the counting rule

$$\dim S + \dim U = \dim(S + U) + \dim(S \cap U),$$

which here reads $6 + 6 = 9 + 3$, since the sum $S + U$ (all sums of a symmetric and an upper-triangular matrix) is all of M.

Remark (Subspace, but not the union).

> $S \cup U$ — matrices that are symmetric *or* upper triangular — is *not* a subspace: add a symmetric matrix to an upper-triangular one and you usually land outside both, just as two lines through the origin in $\mathbb{R}^2$ are not a subspace until you fill in the plane between them. The right object is the *sum* $S + U$, which is a subspace.

The lesson is that "vector space" is about structure, not about arrows. Polynomials, functions, and solution sets of linear differential equations are all vector spaces, and the same words — independence, basis, dimension, and the rank that ties the four subspaces together — describe them all.

Chapter 5

Orthogonality, Projection, Least Squares, and $A = QR$

The four fundamental subspaces of Chapter 4 told us *how big* each space is. This chapter tells us how they sit — at right angles. The row space and the null space are not just complementary in dimension; they are perpendicular, every vector in one orthogonal to every vector in the other. The same is true of the column space and the left null space. That is the second half of the Fundamental Theorem, and it turns the abstract "big picture" into a geometric one with right angles you can draw.

Once the angles are right, a new and very practical question opens up. The real world hands us systems $Ax = b$ that have *no* solution — more equations than unknowns, measurements that do not quite agree. We cannot reach b, because b is not in the column space. So we settle for the next best thing: the point p in the column space that is *closest* to b, and we solve $Ax = p$ instead. Finding that closest point is *projection*, and using it to fit a line through scattered data is *least squares* — the most-used idea in all of applied linear algebra. We close with the cleanest possible bases, the orthonormal ones, and the Gram–Schmidt process that manufactures them, packaged as the factorization $A = QR$.

5.1 Orthogonal Vectors and Subspaces

"Orthogonal" is just another word for perpendicular. Two vectors meet at a right angle exactly when a familiar test on their dot product holds, and the dot product is something we can compute for vectors in any dimension — long after we can no longer see the angle.

Definition 5.1: Orthogonal Vectors

Two vectors $x, y \in \mathbb{R}^n$ are **orthogonal** if their dot product is zero:

$$x^{\mathsf{T}} y \;=\; x_1 y_1 + x_2 y_2 + \cdots + x_n y_n \;=\; 0.$$

We write $x \perp y$. The length of x is $\|x\| = \sqrt{x^{\mathsf{T}} x}$, so $\|x\|^2 = x^{\mathsf{T}} x$.

Why is $x^{\mathsf{T}} y = 0$ the right test? Place x and y tail to tail. They are perpendicular

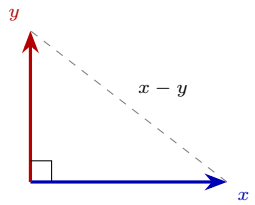
exactly when the triangle with legs x and y and hypotenuse $x - y$ is a right triangle — and Pythagoras says that happens exactly when $\|x\|^2 + \|y\|^2 = \|x - y\|^2$. Expand the right side:

$$\|x - y\|^2 = (x - y)^\mathsf{T}(x - y) = x^\mathsf{T}x - 2x^\mathsf{T}y + y^\mathsf{T}y = \|x\|^2 + \|y\|^2 - 2\,x^\mathsf{T}y.$$

This equals $\|x\|^2 + \|y\|^2$ precisely when the cross term $x^\mathsf{T}y$ vanishes. So the dot-product test and the right-angle picture say the same thing.

Remark (The zero vector).

> Every vector is orthogonal to the zero vector, since $x^\mathsf{T}0 = 0$ always. This is a convenience, not a paradox: the zero vector has no direction to disagree with, so it is perpendicular to everything — and to itself.



A whole *subspace* can be perpendicular to another. We do not check one pair of vectors but all of them at once.

Definition 5.2: Orthogonal Subspaces

Subspaces S and T of $\mathbb{R}^n$ are **orthogonal**, written $S \perp T$, if every vector in S is orthogonal to every vector in T:

$$x^\mathsf{T}y = 0 \qquad \text{for all } x \in S,\ y \in T.$$

The classic warning is the blackboard and the floor. They meet along a line, and you might call them "perpendicular planes," but as *subspaces* they are not orthogonal: a vector running along the bottom edge of the blackboard lies in both the blackboard and the floor, and a nonzero vector is never orthogonal to itself ($x^\mathsf{T}x = \|x\|^2 > 0$). For two subspaces to be orthogonal they may share only the zero vector.

Example (Orthogonal subspaces in the plane).

> In $\mathbb{R}^2$, the only subspaces are $\{0\}$, lines through the origin, and the whole plane. The zero subspace is orthogonal to everything. Two lines through the origin are orthogonal subspaces exactly when they cross at a right angle — the x-axis and the y-axis, for instance. But a line and the whole plane can never be orthogonal: the plane contains the line itself, and the line is not perpendicular to itself.

5.2 The Row Space Is Orthogonal to the Null Space

Now the payoff. The two subspaces inside $\mathbb{R}^n$ that we built from a matrix — the row space $C(A^\mathsf{T})$ and the null space $N(A)$ — are automatically orthogonal. This is not a special property of special matrices. It falls out of the meaning of $Ax = 0$.

Theorem 5.3: Row Space $\perp$ Null Space

For any $m \times n$ matrix A, the null space $N(A)$ and the row space $C(A^\mathsf{T})$ are orthogonal subspaces of $\mathbb{R}^n$:

$$x \in N(A),\ y \in C(A^\mathsf{T}) \quad \Longrightarrow \quad x^\mathsf{T} y = 0.$$

Likewise, in $\mathbb{R}^m$, the left null space $N(A^\mathsf{T})$ is orthogonal to the column space $C(A)$.

Proof. Let $x \in N(A)$, so $Ax = 0$. Reading Ax one row at a time, each entry of Ax is a row of A dotted with x:

$$Ax = \begin{bmatrix} (\text{row } 1) \cdot x \\ \vdots \\ (\text{row } m) \cdot x \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}.$$

So x is orthogonal to *every* row of A. But then x is orthogonal to every combination of the rows — if $y = c_1 r_1 + \cdots + c_m r_m$ is any vector in the row space, then $x^\mathsf{T} y = c_1(x^\mathsf{T} r_1) + \cdots + c_m(x^\mathsf{T} r_m) = 0$. Hence $x \perp C(A^\mathsf{T})$. The column-space statement is the same fact applied to A^T, whose row space is $C(A)$ and whose null space is $N(A^\mathsf{T})$. $\square$

This is the geometric content the dimension count was hiding. In Chapter 4 we found $\dim C(A^\mathsf{T}) = r$ and $\dim N(A) = n - r$, so the two dimensions add up to n. Now we know they are perpendicular as well. Two subspaces whose dimensions fill the space and that meet at right angles deserve a name.

Definition 5.4: Orthogonal Complement

The **orthogonal complement** of a subspace $S \subseteq \mathbb{R}^n$, written $S^\perp$, is the set of *all* vectors orthogonal to everything in S:

$$S^\perp = \left\{ x \in \mathbb{R}^n : x^\mathsf{T} y = 0 \text{ for all } y \in S \right\}.$$

Being orthogonal is not quite enough to be the complement; the complement must catch *every* perpendicular vector. The row space and null space pass that stronger test.

Theorem 5.5: Fundamental Theorem of Linear Algebra, Part 2

For any $m \times n$ matrix A, the row space and null space are orthogonal complements in $\mathbb{R}^n$:

$$N(A) = C(A^\mathsf{T})^\perp, \qquad \dim C(A^\mathsf{T}) + \dim N(A) = r + (n - r) = n.$$

In $\mathbb{R}^m$, the column space and the left null space are orthogonal complements:

$$N(A^\mathsf{T}) = C(A)^\perp, \qquad \dim C(A) + \dim N(A^\mathsf{T}) = r + (m - r) = m.$$

Proof. We saw $N(A) \perp C(A^\mathsf{T})$, so $N(A) \subseteq C(A^\mathsf{T})^\perp$. For equality we count dimensions. The complement $C(A^\mathsf{T})^\perp$ consists of the vectors perpendicular to an r-dimensional subspace

of $\mathbb{R}^n$; such a complement has dimension $n - r$, since a subspace and its perpendicular complement share only 0 and together span $\mathbb{R}^n$, so their dimensions add to n. But $N(A)$ already has dimension $n - r$ and sits inside it, so the two coincide. (A subspace cannot sit strictly inside another of the same dimension.) The column-space case is the same statement for A^T. □

Remark (Orthogonal complements split the space).

When S and $S^\perp$ are orthogonal complements in $\mathbb{R}^n$, every vector $v \in \mathbb{R}^n$ splits *uniquely* as $v = v_S + v_{S^\perp}$ with $v_S \in S$ and $v_{S^\perp} \in S^\perp$. The dimensions add to n and the only shared vector is 0, so the pieces are determined. This splitting is exactly what projection will compute: given b, find its column-space part p and its left-null-space part $e = b - p$.

Example (Reading the orthogonality off a rank-one matrix).

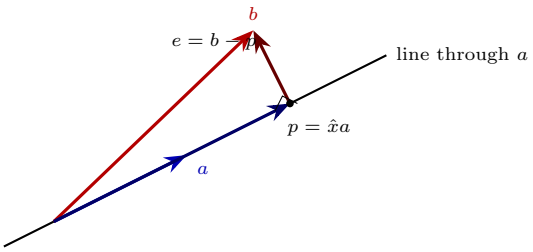
Take

$$A = \begin{bmatrix} 1 & 2 & 5 \\ 2 & 4 & 10 \end{bmatrix},$$

whose second row is twice the first, so $r = 1$. The row space is the line through $(1, 2, 5)$ in $\mathbb{R}^3$, dimension 1. The null space is everything perpendicular to $(1, 2, 5)$ — the solutions of $x_1 + 2x_2 + 5x_3 = 0$ — which is a *plane* through the origin, dimension $n - r = 3 - 1 = 2$. A line and a plane, perpendicular, filling $\mathbb{R}^3$: that is the orthogonal-complement split, drawn for one small matrix.

5.3 Projection onto a Line

We turn to the question that drives the rest of the chapter. Given a vector b and a line through the origin in the direction a, which point p on the line is *closest* to b? Geometry answers instantly: drop a perpendicular from b to the line. The foot of that perpendicular is p, and the leftover $e = b - p$ — the *error* — is orthogonal to the line.



Let us turn the picture into algebra. Since p lies on the line, it is some multiple of a:

$$p = \hat{x}\, a \qquad \text{for an unknown scalar } \hat{x}.$$

The one fact that pins down $\hat{x}$ is that the error $e = b - \hat{x}a$ is perpendicular to a:

$$a^\mathsf{T}(b - \hat{x}a) = 0 \quad \Longrightarrow \quad a^\mathsf{T}b = \hat{x}\, a^\mathsf{T}a \quad \Longrightarrow \quad \boxed{\hat{x} = \frac{a^\mathsf{T}b}{a^\mathsf{T}a}}.$$

The closest point itself is

$$p = \hat{x}\, a = a\, \frac{a^{\mathsf{T}} b}{a^{\mathsf{T}} a}.$$

Two sanity checks Strang likes: doubling b doubles p (the formula is linear in b), while doubling a leaves p unchanged — a appears once on top and twice on bottom, and only the *direction* of a matters, not its length.

5.3.1 The Projection Matrix for a Line

The map $b \mapsto p$ is linear, so it must be a matrix times b. Rearrange, keeping the order of the products straight ($a^{\mathsf{T}} b$ is a number, $a\, a^{\mathsf{T}}$ is a matrix):

$$p = a\, \frac{a^{\mathsf{T}} b}{a^{\mathsf{T}} a} = \frac{a\, a^{\mathsf{T}}}{a^{\mathsf{T}} a}\, b = Pb, \qquad \text{where} \qquad P = \frac{a\, a^{\mathsf{T}}}{a^{\mathsf{T}} a}.$$

Here $a\, a^{\mathsf{T}}$ is an $n \times n$ matrix — a column times a row — not a scalar; only $a^{\mathsf{T}} a$ in the denominator is a number. The matrix P is the **projection matrix** onto the line through a.

Example (Projecting onto a line in $\mathbb{R}^3$).

Let $a = (1, 1, 1)$ and $b = (1, 2, 3)$. Then $a^{\mathsf{T}} a = 3$ and $a^{\mathsf{T}} b = 6$, so

$$\hat{x} = \frac{6}{3} = 2, \qquad p = 2a = (2, 2, 2), \qquad e = b - p = (-1, 0, 1).$$

Check the right angle: $a^{\mathsf{T}} e = -1 + 0 + 1 = 0$. The projection matrix is

$$P = \frac{a\, a^{\mathsf{T}}}{a^{\mathsf{T}} a} = \frac{1}{3} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix},$$

and indeed $Pb = \frac{1}{3}(6, 6, 6) = (2, 2, 2) = p$.

Proposition 5.6: Properties of the Line Projection $P = \frac{a\, a^{\mathsf{T}}}{a^{\mathsf{T}} a}$

The projection matrix onto a line satisfies

$$P^{\mathsf{T}} = P \qquad \text{and} \qquad P^2 = P,$$

and $\operatorname{rank} P = 1$ with column space the line through a.

Proof. For symmetry, $\left(a\, a^{\mathsf{T}}\right)^{\mathsf{T}} = a\, a^{\mathsf{T}}$ and the scalar $a^{\mathsf{T}} a$ is unchanged by transposing, so $P^{\mathsf{T}} = P$. For idempotence,

$$P^2 = \frac{a\, a^{\mathsf{T}}}{a^{\mathsf{T}} a} \cdot \frac{a\, a^{\mathsf{T}}}{a^{\mathsf{T}} a} = \frac{a\, (a^{\mathsf{T}} a)\, a^{\mathsf{T}}}{(a^{\mathsf{T}} a)^2} = \frac{a\, a^{\mathsf{T}}}{a^{\mathsf{T}} a} = P,$$

where the middle $a^{\mathsf{T}} a$ cancels one factor in the denominator. Geometrically $P^2 = P$ says:

project once and you land on the line; project again and a point already on the line does not move. Finally every output Pb is a multiple of a, so the column space is the single line through a and $\operatorname{rank} P = 1$. □

5.4 Projection onto a Subspace

A line is the case of projecting onto a one-dimensional column space. The real prize is projecting b onto a whole subspace — the column space of an $m \times n$ matrix A with independent columns $a_1, \dots, a_n$. The closest point p is now a combination of *all* the columns:

$$p = \hat{x}_1 a_1 + \cdots + \hat{x}_n a_n = A\hat{x}.$$

We need the coefficient vector $\hat{x}$. The principle is identical to the line case: the error $e = b - A\hat{x}$ must be perpendicular to the subspace — that is, perpendicular to *every* column of A. One equation per column:

$$a_1^{\mathsf{T}}(b - A\hat{x}) = 0, \quad \dots, \quad a_n^{\mathsf{T}}(b - A\hat{x}) = 0.$$

Stack these n equations into one. The rows $a_1^{\mathsf{T}}, \dots, a_n^{\mathsf{T}}$ are exactly the rows of A^{T}, so the whole system is

$$A^{\mathsf{T}}(b - A\hat{x}) = 0,$$

which rearranges into the central equation of the entire chapter.

Theorem 5.7: The Normal Equations

Let A be $m \times n$ with linearly independent columns, and let $b \in \mathbb{R}^m$. The projection $p = A\hat{x}$ of b onto the column space $C(A)$ is found by solving the **normal equations**

$$\boxed{A^{\mathsf{T}} A\,\hat{x} = A^{\mathsf{T}} b}.$$

Because the columns of A are independent, $A^{\mathsf{T}}A$ is invertible, and

$$\hat{x} = (A^{\mathsf{T}}A)^{-1} A^{\mathsf{T}} b, \qquad p = A\hat{x} = A(A^{\mathsf{T}}A)^{-1} A^{\mathsf{T}}\, b.$$

Proof. The closest p in $C(A)$ is characterized by $e = b - p \perp C(A)$, i.e. $A^{\mathsf{T}} e = 0$. Writing $p = A\hat{x}$ turns this into $A^{\mathsf{T}}(b - A\hat{x}) = 0$, that is $A^{\mathsf{T}}A\hat{x} = A^{\mathsf{T}}b$. Invertibility of $A^{\mathsf{T}}A$ (proved in Section 5.5.2) makes $\hat{x}$ unique. □

Remark (Where the error lives).

> The condition $A^{\mathsf{T}}e = 0$ says exactly that the error e is in the null space of A^{T} — the *left null space* of A. And we proved that $N(A^{\mathsf{T}}) = C(A)^{\perp}$. So the projection splits b into a piece p in the column space and a piece e in its orthogonal complement, $b = p + e$. The whole machine is the orthogonal-complement decomposition from Part 2 of the Fundamental Theorem, made computational.

5.4.1 The Projection Matrix

Just as on the line, the map $b \mapsto p$ is linear, so $p = Pb$ for a matrix P. Reading it off the formula $p = A(A^\mathsf{T}A)^{-1}A^\mathsf{T}b$:

$$\boxed{P = A(A^\mathsf{T}A)^{-1}A^\mathsf{T}}.$$

Resist the urge to "simplify." If A were square and invertible we could split $(A^\mathsf{T}A)^{-1} = A^{-1}(A^\mathsf{T})^{-1}$ and watch everything collapse to $P = I$ — which is correct (a square invertible A has column space all of $\mathbb{R}^m$, so b is already in it and projects to itself). But for a genuinely rectangular A there is no A^{-1}, and $P = A(A^\mathsf{T}A)^{-1}A^\mathsf{T}$ does not reduce. The formula generalizes the line case perfectly: there we divided by the number $a^\mathsf{T}a$; here we multiply by the matrix $(A^\mathsf{T}A)^{-1}$.

Proposition 5.8: Properties of the Projection Matrix $P = A(A^\mathsf{T}A)^{-1}A^\mathsf{T}$

For A with independent columns, P satisfies

$$P^\mathsf{T} = P \qquad \text{and} \qquad P^2 = P.$$

Its column space is $C(A)$, and $I-P$ is the projection onto the orthogonal complement $N(A^\mathsf{T})$.

Proof. Symmetric: using $(BC)^\mathsf{T} = C^\mathsf{T}B^\mathsf{T}$ and $\left((A^\mathsf{T}A)^{-1}\right)^\mathsf{T} = (A^\mathsf{T}A)^{-1}$ (since $A^\mathsf{T}A$ is symmetric),

$$P^\mathsf{T} = \left(A(A^\mathsf{T}A)^{-1}A^\mathsf{T}\right)^\mathsf{T} = A\,(A^\mathsf{T}A)^{-1}A^\mathsf{T} = P.$$

Idempotent: the middle factors telescope,

$$P^2 = A(A^\mathsf{T}A)^{-1}\underbrace{A^\mathsf{T}A(A^\mathsf{T}A)^{-1}}_{=I}A^\mathsf{T} = A(A^\mathsf{T}A)^{-1}A^\mathsf{T} = P.$$

Every output $Pb = A\hat{x}$ is a combination of columns of A, so $C(P) = C(A)$. Finally $(I-P)b = b - p = e$ lands in $N(A^\mathsf{T})$, and $I-P$ inherits $(I-P)^\mathsf{T} = I-P$ and $(I-P)^2 = I - 2P + P^2 = I - P$, so it too is a projection — onto the complementary space. □

Two extreme inputs

The projection P sorts every b by where it lives. If b is already in the column space, $b = Ax$, then $Pb = b$ — nothing to project. If b is perpendicular to the column space, $b \in N(A^\mathsf{T})$, then $A^\mathsf{T}b = 0$, so $\hat{x} = 0$ and $Pb = 0$ — it projects to the origin. A general b is a mix of the two, $b = p + e$, and P keeps the column-space part p and discards e.

5.5 Least Squares

Here is why projection matters. When $Ax = b$ has *no* solution — the common case with more equations than unknowns — we cannot make Ax equal b, because Ax is trapped in the column space and b is not. The honest goal is to make Ax as *close* to b as possible: minimize the length of the error $Ax - b$. Minimizing the length is the same as minimizing its square, the sum of squared errors — hence *least squares*.

Definition 5.9: Least-Squares Solution

Given $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, a **least-squares solution** $\hat{x}$ is one that minimizes the squared error

$$\|Ax - b\|^2 = (Ax - b)^{\mathsf{T}}(Ax - b) = e_1^2 + e_2^2 + \cdots + e_m^2.$$

The minimizer is exactly the projection problem in disguise. Among all vectors Ax in the column space, the one closest to b is the projection p. So the best $A\hat{x}$ equals p, and $\hat{x}$ solves the very normal equations we already derived:

$$\boxed{A^{\mathsf{T}}A\,\hat{x} = A^{\mathsf{T}}b}.$$

No calculus is needed — geometry hands us the answer — but calculus agrees: setting the partial derivatives of $\|Ax - b\|^2$ to zero gives the same linear system.

5.5.1 Fitting a Straight Line

The canonical use is fitting a line $b = C + Dt$ to data points. With more than two points the line generally cannot pass through all of them, so we fit the closest line in the least-squares sense.

Example (The closest line through three points).

Fit a line $b = C + Dt$ to the points

$$(t, b) = (1, 1),\ (2, 2),\ (3, 2).$$

If the line passed through all three, we would need

$$\begin{aligned} C + D &= 1, \\ C + 2D &= 2, \\ C + 3D &= 2, \end{aligned} \qquad \text{i.e.} \qquad \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} C \\ D \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}}_{b}.$$

This $Ax = b$ has no solution — three points, not collinear. Form the normal equations $A^{\mathsf{T}}A\hat{x} = A^{\mathsf{T}}b$. Compute

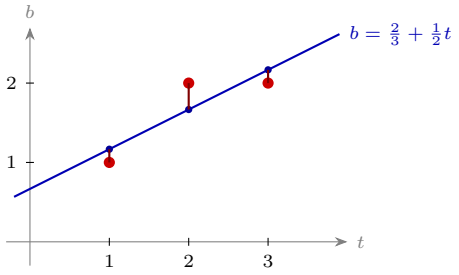
$$A^{\mathsf{T}}A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}, \qquad A^{\mathsf{T}}b = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 5 \\ 11 \end{bmatrix}.$$

The normal equations are

$$3\hat{C} + 6\hat{D} = 5,$$
$$6\hat{C} + 14\hat{D} = 11.$$

Subtract twice the first from the second: $(14 - 12)\hat{D} = 11 - 10$, so $\hat{D} = \frac{1}{2}$, and then $3\hat{C} = 5 - 3 = 2$ gives $\hat{C} = \frac{2}{3}$. The best line is

$$b = \tfrac{2}{3} + \tfrac{1}{2}\,t.$$



Remark (Reading off the projection and the error).

The fitted values are the projection $p = A\hat{x}$ — the heights of the line at $t = 1, 2, 3$:

$$p_1 = \tfrac{2}{3} + \tfrac{1}{2} = \tfrac{7}{6}, \quad p_2 = \tfrac{2}{3} + 1 = \tfrac{5}{3}, \quad p_3 = \tfrac{2}{3} + \tfrac{3}{2} = \tfrac{13}{6}.$$

The errors are the vertical gaps $e = b - p$:

$$e = \left(1 - \tfrac{7}{6},\ 2 - \tfrac{5}{3},\ 2 - \tfrac{13}{6}\right) = \left(-\tfrac{1}{6},\ \tfrac{1}{3},\ -\tfrac{1}{6}\right).$$

Two checks confirm the geometry. The errors sum to zero, $-\frac{1}{6} + \frac{1}{3} - \frac{1}{6} = 0$ (that is e orthogonal to the all-ones first column of A). And e is orthogonal to the second column $(1, 2, 3)$: $-\frac{1}{6} + \frac{2}{3} - \frac{1}{2} = 0$. So $e \in N(A^{\mathsf{T}})$ and $p \in C(A)$, perpendicular, as projection demands.

5.5.2 Why $A^{\mathsf{T}}A$ Is Invertible

Everything above leaned on one assumption: $A^{\mathsf{T}}A$ is invertible. This is where the column-independence hypothesis earns its keep. The key is that $A^{\mathsf{T}}A$ has the *same* null space as A.

Theorem 5.10: $N(A^{\mathsf{T}}A) = N(A)$, and $A^{\mathsf{T}}A$ Is Invertible iff A Has Independent Columns

For any $m \times n$ matrix A,

$$N(A^{\mathsf{T}}A) = N(A) \qquad \text{and} \qquad \operatorname{rank}(A^{\mathsf{T}}A) = \operatorname{rank} A.$$

Consequently the square matrix $A^{\mathsf{T}}A$ is invertible if and only if A has linearly independent columns.

Proof. If $Ax = 0$ then $A^\mathsf{T}Ax = A^\mathsf{T}0 = 0$, so $N(A) \subseteq N(A^\mathsf{T}A)$. Conversely suppose $A^\mathsf{T}Ax = 0$. Multiply on the left by x^T:

$$x^\mathsf{T}A^\mathsf{T}Ax = 0 \quad \Longrightarrow \quad (Ax)^\mathsf{T}(Ax) = 0 \quad \Longrightarrow \quad \|Ax\|^2 = 0 \quad \Longrightarrow \quad Ax = 0,$$

since only the zero vector has zero length. So $N(A^\mathsf{T}A) \subseteq N(A)$, and the two null spaces are equal. Equal null spaces force equal nullity, and since $A^\mathsf{T}A$ has n columns, $\operatorname{rank}(A^\mathsf{T}A) = n - \dim N(A^\mathsf{T}A) = n - \dim N(A) = \operatorname{rank} A$.

Now $A^\mathsf{T}A$ is $n \times n$. It is invertible iff its null space is $\{0\}$, iff $N(A) = \{0\}$, iff the columns of A are independent. □

Remark (Independent columns guarantee a unique fit).

> This theorem is the license for least squares. As long as the columns of A are independent — and for a line fit they are, since $(1, 1, 1)$ and $(1, 2, 3)$ point in different directions — the matrix $A^\mathsf{T}A$ is invertible, the normal equations have exactly one solution $\hat{x}$, and the best line is unique. One easy way to *guarantee* independent columns is to make them orthonormal, which is where we head next.

5.6 Orthonormal Bases and Orthogonal Matrices

Independent columns make least squares work, but *orthonormal* columns make it trivial. Orthonormal means perpendicular and unit length — the best-behaved vectors there are.

Definition 5.11: Orthonormal Vectors

Vectors $q_1, q_2, \ldots, q_n$ are **orthonormal** if

$$q_i^\mathsf{T} q_j = \begin{cases} 0 & \text{if } i \neq j \quad \text{(perpendicular)}, \\ 1 & \text{if } i = j \quad \text{(unit length)}. \end{cases}$$

Orthonormal vectors are automatically independent: if $\sum c_i q_i = 0$, dot both sides with q_j to get $c_j = 0$ for every j.

Collect orthonormal vectors as the columns of a matrix and call it Q. The orthonormality conditions are precisely the statement that $Q^\mathsf{T}Q$ is the identity.

Definition 5.12: Orthogonal Matrix

A matrix Q with orthonormal columns satisfies

$$Q^\mathsf{T}Q = I.$$

The (i, j) entry of $Q^\mathsf{T}Q$ is $q_i^\mathsf{T} q_j$, which is 1 on the diagonal and 0 off it. If Q is *square* with orthonormal columns it is called an **orthogonal matrix**, and then $Q^\mathsf{T}Q = I$ means $Q^\mathsf{T} = Q^{-1}$: the inverse is just the transpose.

Example (Three orthogonal matrices).

A permutation matrix shuffles the standard basis, so its columns are orthonormal:

$$Q = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \qquad Q^{\mathsf{T}} = Q^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

A rotation of the plane is orthogonal:

$$Q = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \qquad Q^{\mathsf{T}}Q = I$$

because $\cos^2\theta + \sin^2\theta = 1$ and the two columns are perpendicular. And rescaling fixes a near-miss: $\left[\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right]$ has perpendicular columns but length $\sqrt{2}$, so divide by $\sqrt{2}$:

$$Q = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \qquad Q^{\mathsf{T}}Q = I.$$

A rectangular Q (more rows than columns) can still have orthonormal columns; then $Q^{\mathsf{T}}Q = I$ holds, but QQ^{T} is generally *not* the identity — it is the projection onto the column space, as we now see.

5.6.1 Projection with Orthonormal Columns

Watch the projection machinery collapse when $A = Q$ has orthonormal columns. Since $Q^{\mathsf{T}}Q = I$, the inverse $(Q^{\mathsf{T}}Q)^{-1} = I$ disappears, and

$$P = Q(Q^{\mathsf{T}}Q)^{-1}Q^{\mathsf{T}} = Q\,Q^{\mathsf{T}}, \qquad \hat{x} = (Q^{\mathsf{T}}Q)^{-1}Q^{\mathsf{T}}b = Q^{\mathsf{T}}b.$$

The normal equations $Q^{\mathsf{T}}Q\hat{x} = Q^{\mathsf{T}}b$ become simply $\hat{x} = Q^{\mathsf{T}}b$: *no system to solve.* Each coefficient is one dot product,

$$\hat{x}_i = q_i^{\mathsf{T}}b,$$

so the projection is $p = \sum_i (q_i^{\mathsf{T}}b)\,q_i$ — add up the components of b along each orthonormal direction. If Q is square these directions span the whole space, every b is already inside, and $P = QQ^{\mathsf{T}} = I$.

Why we want orthonormal columns

With orthonormal columns there is nothing to invert. The coefficient of b along direction q_i is just $q_i^{\mathsf{T}}b$, computed independently of the others. This decoupling — changing one basis vector does not disturb the coefficients of the rest — is the entire payoff of orthonormality, and the reason the next section builds an orthonormal basis from any starting basis.

5.7 Gram–Schmidt and $A = QR$

We can always *make* an orthonormal basis. Start with any independent vectors $a, b, c, \dots$ spanning a subspace; the **Gram–Schmidt** process produces orthonormal $q_1, q_2, q_3, \dots$ spanning the same subspace. The idea is exactly the projection we just learned: to make a new vector perpendicular to the directions already chosen, *subtract off its projections* onto them.

5.7.1 The Process

Take the first vector as the first direction: $A_1 = a$. For the second, subtract from b its projection onto A_1, leaving the part perpendicular to A_1 — which is the error vector e from before:

$$A_2 = b - \frac{A_1^{\mathsf{T}} b}{A_1^{\mathsf{T}} A_1} A_1.$$

By construction $A_1^{\mathsf{T}} A_2 = A_1^{\mathsf{T}} b - A_1^{\mathsf{T}} b = 0$, so $A_2 \perp A_1$. For a third vector c, subtract its projections onto *both* A_1 and A_2:

$$A_3 = c - \frac{A_1^{\mathsf{T}} c}{A_1^{\mathsf{T}} A_1} A_1 - \frac{A_2^{\mathsf{T}} c}{A_2^{\mathsf{T}} A_2} A_2.$$

At the end, normalize each to unit length:

$$q_1 = \frac{A_1}{\|A_1\|}, \qquad q_2 = \frac{A_2}{\|A_2\|}, \qquad q_3 = \frac{A_3}{\|A_3\|}, \ \dots$$

The q's are orthonormal and span the same space as the original vectors at every stage.

Example (Gram–Schmidt on two vectors).

Take

$$a = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}.$$

Set $A_1 = a$. Then $A_1^{\mathsf{T}} A_1 = 3$ and $A_1^{\mathsf{T}} b = 1 + 0 + 2 = 3$, so

$$A_2 = b - \frac{A_1^{\mathsf{T}} b}{A_1^{\mathsf{T}} A_1} A_1 = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} - \frac{3}{3} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}.$$

Check: $A_1^{\mathsf{T}} A_2 = 0 - 1 + 1 = 0$. Normalizing ($\|A_1\| = \sqrt{3}$, $\|A_2\| = \sqrt{2}$):

$$q_1 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \qquad q_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}, \qquad Q = \begin{bmatrix} 1/\sqrt{3} & 0 \\ 1/\sqrt{3} & -1/\sqrt{2} \\ 1/\sqrt{3} & 1/\sqrt{2} \end{bmatrix}.$$

The two columns of Q are orthonormal and span the same plane as a and b.

5.7.2 The Factorization $A = QR$

Elimination wrote $A = LU$; Gram–Schmidt writes $A = QR$. Where L was lower triangular, R is upper triangular, and Q holds the orthonormal vectors we just built.

To see why R is triangular, note how Gram–Schmidt builds the q's: q_1 comes only from a_1; q_2 comes from a_1, a_2; in general q_k uses only $a_1, \dots, a_k$. Turn this around — each original column is a combination of the q's up to its own index:

$$a_1 = (q_1^{\mathsf{T}} a_1)\, q_1, \qquad a_2 = (q_1^{\mathsf{T}} a_2)\, q_1 + (q_2^{\mathsf{T}} a_2)\, q_2, \qquad \dots$$

Because $q_j^{\mathsf{T}} a_k = 0$ whenever $j > k$ (later q's are perpendicular to earlier columns), these coefficients fill an *upper*-triangular matrix. In matrix form, for $A = [a_1\ a_2]$,

$$\underbrace{\begin{bmatrix} a_1 & a_2 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} q_1 & q_2 \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} q_1^{\mathsf{T}} a_1 & q_1^{\mathsf{T}} a_2 \\ 0 & q_2^{\mathsf{T}} a_2 \end{bmatrix}}_{R}.$$

Theorem 5.13: The QR Factorization

Every $m \times n$ matrix A with independent columns factors as

$$A = QR,$$

where Q is $m \times n$ with orthonormal columns ($Q^{\mathsf{T}} Q = I$) and R is $n \times n$ upper triangular and invertible. The entries of R are the dot products

$$R_{jk} = q_j^{\mathsf{T}} a_k \quad (j \le k), \qquad \text{and} \qquad R = Q^{\mathsf{T}} A.$$

Proof. Gram–Schmidt produces the orthonormal columns of Q and expresses each a_k as a combination of $q_1, \dots, q_k$, so $A = QR$ with R upper triangular and $R_{jk} = q_j^{\mathsf{T}} a_k$. Multiplying $A = QR$ on the left by Q^{T} and using $Q^{\mathsf{T}} Q = I$ gives $R = Q^{\mathsf{T}} A$. The diagonal entries $R_{kk} = q_k^{\mathsf{T}} a_k = \|A_k\| > 0$ are nonzero, so R is invertible. □

Remark (Least squares becomes back-substitution).

With $A = QR$ the normal equations simplify beautifully. Substitute into $A^{\mathsf{T}} A \hat{x} = A^{\mathsf{T}} b$:

$$(QR)^{\mathsf{T}} (QR)\, \hat{x} = (QR)^{\mathsf{T}} b \implies R^{\mathsf{T}} \underbrace{Q^{\mathsf{T}} Q}_{I} R\, \hat{x} = R^{\mathsf{T}} Q^{\mathsf{T}} b \implies R\, \hat{x} = Q^{\mathsf{T}} b,$$

canceling the invertible R^{T}. Since R is upper triangular, $\hat{x}$ comes straight from back-substitution — no $A^{\mathsf{T}} A$ to form or invert. This numerically stable route is how least-squares problems are actually solved in practice.

5.8 What to Carry Forward

Orthogonality turned the four-subspace picture from a statement about dimensions into one about right angles: in $\mathbb{R}^n$ the row space and null space are orthogonal complements, and in $\mathbb{R}^m$ so are the column space and left null space. That perpendicular splitting is what makes *projection* possible — every b breaks uniquely into a part p in the column space and a part $e = b - p$ in its complement, computed by the projection matrix $P = A(A^\mathsf{T}A)^{-1}A^\mathsf{T}$ with its signature properties $P^\mathsf{T} = P$ and $P^2 = P$.

Projection solved the practical crisis of an unsolvable $Ax = b$: when b is out of reach, hit the closest point p instead, which is *least squares*, governed by the normal equations $A^\mathsf{T}A\hat{x} = A^\mathsf{T}b$. The matrix $A^\mathsf{T}A$ is invertible exactly when A has independent columns, because $N(A^\mathsf{T}A) = N(A)$. Fitting a line through scattered data is this idea at its most useful.

Finally, *orthonormal* columns make every formula collapse: $Q^\mathsf{T}Q = I$, projection becomes $\hat{x} = Q^\mathsf{T}b$ with no inverse, and Gram–Schmidt manufactures such columns from any basis, recorded as the factorization $A = QR$. With QR in hand, least squares is just $R\hat{x} = Q^\mathsf{T}b$ by back-substitution.

Remark (Looking ahead).

> We have now met two of the three great factorizations in outline: $A = LU$ from elimination (Chapter 2) and $A = QR$ from Gram–Schmidt here. The third, built from eigenvectors, diagonalizes a matrix and is the subject of the chapters ahead; when the eigenvectors can be chosen *orthonormal* — which happens precisely for symmetric matrices like the $A^\mathsf{T}A$ of this chapter — that factorization becomes $A = Q\Lambda Q^\mathsf{T}$, and the whole story of orthogonality returns in full force.

Chapter 6

Determinants

We are halfway through the course, and we leave rectangular matrices behind. From here on the matrices are square, and the two big topics ahead are *determinants* and *eigenvalues*. The determinant comes first because the eigenvalue equation $\det(A - \lambda I) = 0$ will lean on it directly.

The determinant is a single number $\det A$ — also written $|A|$ — attached to every square matrix A. It is a remarkable amount of information squeezed into one number. The headline fact is the cleanest test for invertibility we will ever have: A is invertible exactly when $\det A \neq 0$, and singular exactly when $\det A = 0$. You already know one determinant from school,

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc,$$

and the whole chapter is about understanding where that formula comes from, what it generalizes to, and what it is good for.

Strang's strategy — and it is the right one — is to resist writing down a big formula first. The determinant of an $n \times n$ matrix is a sum of $n!$ terms, and meeting that sum cold teaches you nothing. Instead we pin the determinant down by *three properties* (think of them as axioms), and then *derive* everything else: seven more properties, the big formula, the cofactor expansion, the inverse, Cramer's rule, and finally the geometric punchline that $|\det A|$ is the volume of a box. Three rules in, a whole theory out.

6.1 The Three Defining Properties

Here are the three properties that *define* the determinant. Every other fact in the chapter is a consequence of these.

Definition 6.1: The Determinant

The *determinant* is the unique function det that assigns a number to each square matrix and satisfies three properties:

(P1) The identity has determinant 1: $\det I = 1$.

(P2) Row exchange reverses the sign: swapping any two rows of a matrix multiplies its determinant by -1.

(P3) The determinant is linear in each row separately. Holding all other rows fixed, the determinant is a linear function of the one row that changes. This has two parts. Scaling one row by t scales the determinant by t:

$$\det \begin{bmatrix} ta & tb \\ c & d \end{bmatrix} = t \det \begin{bmatrix} a & b \\ c & d \end{bmatrix};$$

and a sum in one row splits the determinant into a sum:

$$\det \begin{bmatrix} a+a' & b+b' \\ c & d \end{bmatrix} = \det \begin{bmatrix} a & b \\ c & d \end{bmatrix} + \det \begin{bmatrix} a' & b' \\ c & d \end{bmatrix}.$$

A word of warning about (P3), because it is the property people misread. Linearity is *one row at a time*. It does *not* say $\det(A+B) = \det A + \det B$ — that is false (try two 2×2 matrices). It says that if you change only the first row, leaving the second row alone, the determinant responds linearly to that one row. The other rows are frozen.

Two quick consequences of (P1) and (P2) are worth seeing right away. From (P1), $\det\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right] = 1$. Exchange the two rows and (P2) says

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

So a permutation matrix — an identity with its rows shuffled — has determinant $+1$ or -1 depending on whether it took an *even* or *odd* number of row exchanges to build it. That sign is the seed of the whole theory.

6.2 Seven Properties Derived from the Three

Now the payoff. Properties **(P4)** through **(P10)** are not new assumptions — each one follows from (P1), (P2), (P3). We prove them in order, because later ones use earlier ones.

Proposition 6.2: (P4) Equal rows give determinant zero

If two rows of A are identical, then $\det A = 0$.

Proof. Suppose rows i and j are equal. Exchanging them changes nothing about the matrix, so the determinant is unchanged. But (P2) says a row exchange flips the sign, so the determinant equals its own negative: $\det A = -\det A$. Therefore $\det A = 0$. □

Proposition 6.3: (P5) Subtracting a multiple of one row from another doesn't change the determinant

For $i \neq j$, subtracting t times row i from row j leaves $\det A$ unchanged. These are exactly the row operations of elimination.

Proof. Apply linearity (P3) to row j, which becomes (row j) $- t$ (row i):

$$\det \begin{bmatrix} \vdots \\ \text{row } i \\ \vdots \\ \text{row } j - t\,\text{row } i \\ \vdots \end{bmatrix} = \det \begin{bmatrix} \vdots \\ \text{row } i \\ \vdots \\ \text{row } j \\ \vdots \end{bmatrix} - t \det \begin{bmatrix} \vdots \\ \text{row } i \\ \vdots \\ \text{row } i \\ \vdots \end{bmatrix}.$$

The first term is $\det A$. The second matrix has row i appearing twice (once in its own slot, once in slot j), so by (P4) its determinant is 0. Hence the whole thing is $\det A$. □

This is the property that makes determinants *computable*: the elimination steps that simplify a matrix do not change its determinant at all. In two dimensions the argument reads

$$\det \begin{bmatrix} a & b \\ c - ta & d - tb \end{bmatrix} = \det \begin{bmatrix} a & b \\ c & d \end{bmatrix} - t \det \begin{bmatrix} a & b \\ a & b \end{bmatrix} = \det \begin{bmatrix} a & b \\ c & d \end{bmatrix} - 0.$$

Proposition 6.4: (P6) A zero row gives determinant zero

If any row of A is entirely zero, then $\det A = 0$.

Proof. The zero row is 0 times itself, so by the scaling part of (P3) with $t = 0$ we may pull out the factor 0, giving $\det A = 0 \cdot \det(\cdots) = 0$. □

Proposition 6.5: (P7) The determinant of a triangular matrix is the product of its diagonal entries

If A is upper triangular (or lower triangular) with diagonal entries $d_1, d_2, \ldots, d_n$, then

$$\det A = d_1 d_2 \cdots d_n.$$

Proof. First suppose every $d_i \neq 0$. Because A is triangular, we can use the elimination operations of (P5) — which do not change the determinant — to clear out every off-diagonal entry, arriving at the diagonal matrix $\operatorname{diag}(d_1, \ldots, d_n)$. Now factor d_i out of row i using the scaling part of (P3), one row at a time:

$$\det \operatorname{diag}(d_1, \ldots, d_n) = d_1 d_2 \cdots d_n \det I = d_1 d_2 \cdots d_n$$

by (P1). If instead some $d_i = 0$, then elimination produces a row of zeros before we finish, and (P6) gives $\det A = 0$, which is also the product $d_1 \cdots d_n$ since one factor is zero. Either way the formula holds. □

This is the single most useful computational fact in the chapter. To find a determinant, run elimination down to a triangular form (the pivots sit on the diagonal), keep track of how many row exchanges you used, and multiply the pivots:

$$\det A = (-1)^{(\#\text{ exchanges})} \,(\text{product of the pivots}).$$

That is exactly how a computer evaluates the determinant of a large matrix. It costs about $n^3/3$ operations — nothing like the $n!$ terms of the big formula coming later.

Example (The 2×2 formula falls out of elimination).

With $a \neq 0$, one elimination step clears the lower-left corner:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \longrightarrow \begin{bmatrix} a & b \\ 0 & d - \frac{c}{a}b \end{bmatrix}.$$

By (P5) the determinant is unchanged, and now the matrix is triangular, so by (P7)

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = a\Big(d - \tfrac{c}{a}b\Big) = ad - bc. \quad \checkmark$$

The pivots are a and $d - \frac{c}{a}b$, and their product is the determinant.

Proposition 6.6: (P8) The determinant detects singularity

$\det A = 0$ if and only if A is singular (not invertible). Equivalently, $\det A \neq 0$ exactly when A is invertible.

Proof. Run elimination on A. If A is singular, elimination produces a row of zeros (a missing pivot), so by (P6) the determinant is 0. If A is nonsingular, elimination produces a full set of n nonzero pivots $d_1, \ldots, d_n$, and by (P7) the determinant is $\pm d_1 d_2 \cdots d_n \neq 0$ (the sign records the row exchanges). So $\det A = 0$ exactly in the singular case. □

This is the property that makes the determinant matter. It ties together every thread of the first half of the book: $\det A \neq 0$ is the same as "the columns are independent," the same as "$N(A) = \{0\}$," the same as "$Ax = b$ has exactly one solution for every b," the same as "$\operatorname{rank} A = n$" (see Chapter 4).

Proposition 6.7: (P9) The determinant of a product is the product of determinants

For square matrices A and B of the same size,

$$\det(AB) = (\det A)(\det B).$$

We will not prove (P9) from scratch here — the cleanest proof uses the $A = LU$ factorization, which we sketch for (P10) below — but its consequences are immediate and constantly used.

Remark (Consequences of the product rule).

A handful of corollaries come straight out of (P9):

- **Inverse:** from $A^{-1}A = I$ and (P1), $(\det A^{-1})(\det A) = 1$, so
$$\det(A^{-1}) = \frac{1}{\det A} \qquad \text{(when } A \text{ is invertible).}$$
- **Powers:** $\det(A^2) = (\det A)^2$, and in general $\det(A^k) = (\det A)^k$.
- **Scaling the whole matrix:** multiplying *every* row by t scales the determinant once per row, so for an $n \times n$ matrix
$$\det(tA) = t^n \det A.$$

That last one is a volume statement in disguise: doubling all three edges of a box in $\mathbb{R}^3$ multiplies its volume by $2^3 = 8$, which is exactly $\det(2A) = 2^3 \det A$.

Proposition 6.8: (P10) The determinant of the transpose equals the determinant

For every square matrix, $\det(A^\mathsf{T}) = \det A$.

Proof. Use the factorization $A = LU$, where L is lower triangular with 1's on its diagonal and U is upper triangular with the pivots on its diagonal. By the product rule (P9), the claim $\det(A^\mathsf{T}) = \det A$ becomes $\det(U^\mathsf{T}L^\mathsf{T}) = \det(LU)$, i.e. $\det(U^\mathsf{T})\det(L^\mathsf{T}) = \det(L)\det(U)$. Now L is triangular with diagonal 1's, so $\det L = \det(L^\mathsf{T}) = 1$ by (P7); and U is triangular, so $\det(U^\mathsf{T}) = \det U$ (transposing a triangular matrix keeps the same diagonal). Both sides are therefore equal to $\det U$. (When A needs row exchanges, the same argument runs through $PA = LU$, and the permutation contributes ± 1 on both sides.) □

The point of (P10) is leverage. The three defining properties were stated for *rows*. Because transposing swaps rows and columns without changing the determinant, every row property automatically becomes a *column* property. So: exchanging two columns flips the sign; the determinant is linear in each column; a repeated column or a zero column makes the determinant 0; and so on. Rows and columns play identical roles.

The ten properties at a glance

From the three axioms (P1)–(P3) we have derived everything:

(P4) equal rows ⇒ determinant 0;

(P5) elimination steps don't change it;

(P6) a zero row $\Rightarrow$ determinant 0;

(P7) triangular $\Rightarrow$ product of the diagonal;

(P8) $\det A = 0 \iff A$ singular;

(P9) $\det(AB) = \det A \det B$;

(P10) $\det(A^{\mathsf{T}}) = \det A$ (so columns behave like rows).

Two of these are the workhorses: (P7) lets you *compute* a determinant by elimination, and (P8) tells you what the answer *means*.

Remark (The one loose end).

There is a gap we have quietly stepped over. Property (P2) says a row exchange flips the sign, so the determinant of a permutation is $(-1)^{\#\text{exchanges}}$. But a given permutation can be reached by different numbers of exchanges — could one route use seven swaps and another ten, giving both -1 and $+1$? If so the determinant would equal its own negative and the whole definition would collapse. It is a genuine theorem (the *parity* of a permutation is well defined: an even permutation can never be reached by an odd number of swaps) that closes this gap. We take it as known; with it, the three properties really do define det uniquely.

6.3 The Big Formula: a Sum of $n!$ Terms

The three properties pin down the determinant, but so far our only recipe is "run elimination." There is also an explicit closed formula, and deriving it is a beautiful application of linearity in the rows. We build it for 2×2, watch the pattern in 3×3, and then state it in general.

6.3.1 Building the 2×2 formula from scratch

Take a general 2×2 matrix and split each row into its two "one-entry" pieces using linearity (P3). The first row (a, b) is $(a, 0) + (0, b)$; the second row (c, d) is $(c, 0) + (0, d)$. Linearity in each row separately breaks the determinant into $2 \times 2 = 4$ pieces:

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \det \begin{bmatrix} a & 0 \\ c & 0 \end{bmatrix} + \det \begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix} + \det \begin{bmatrix} 0 & b \\ c & 0 \end{bmatrix} + \det \begin{bmatrix} 0 & b \\ 0 & d \end{bmatrix}.$$

Now look at the four pieces. The first and last each have a zero *column*, so by (P6)-for-columns they vanish. Only the two pieces that pick *one entry from each row and each column* survive:

$$= \det \begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix} + \det \begin{bmatrix} 0 & b \\ c & 0 \end{bmatrix} = ad \det \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + cb \det \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = ad - bc.$$

The surviving determinants are determinants of permutation matrices: the diagonal one is $+1$, the off-diagonal one is -1. The number of survivors is exactly $2! = 2$.

6.3.2 The same idea in 3×3

Splitting each of the three rows into its three one-entry pieces gives $3 \times 3 \times 3 = 27$ little determinants. A piece survives only if it has no repeated column — i.e. it picks one entry from each row *and* each column. The number of such choices is $3! = 6$, and these are the six permutations of $(1, 2, 3)$. Each survivor is the chosen product times the sign (± 1) of its permutation:

$$\det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{array}{l} a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} \\ \quad + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}. \end{array}$$

A handy mnemonic for 3×3 (and *only* 3×3): the three products running "down-and-to-the-right" get a plus sign, the three running "down-and-to-the-left" get a minus sign. Do not try to extend this trick to 4×4 — it simply fails there.

6.3.3 The general formula

The pattern is now clear. For an $n \times n$ matrix there are $n!$ surviving terms, one for each permutation $(\alpha, \beta, \gamma, \ldots, \omega)$ of $(1, 2, \ldots, n)$, each term taking one entry from every row and every column.

Theorem 6.9: The Big Formula

For an $n \times n$ matrix $A = (a_{ij})$,

$$\det A = \sum_{\text{all } n! \text{ permutations}} (\pm)\, a_{1\alpha}\, a_{2\beta}\, a_{3\gamma} \cdots a_{n\omega},$$

where $(\alpha, \beta, \gamma, \ldots, \omega)$ runs over all permutations of the column indices $(1, 2, \ldots, n)$, and the sign is $+$ for an even permutation and $-$ for an odd one (the determinant of the corresponding permutation matrix).

You can sanity-check the formula on the identity: every term contains a zero except the one from the trivial permutation $\alpha = 1, \beta = 2, \ldots, \omega = n$, which contributes $a_{11}a_{22} \cdots a_{nn} = 1$. So $\det I = 1$, consistent with (P1). With more work one can verify (P2) and (P3) from the formula too; we take that on faith. The count $n!$ grows ferociously — $2, 6, 24, 120, \ldots$ — which is exactly why the elimination method (P7) is the one to use in practice.

Example (Mostly zeros: spotting the surviving permutations).

Compute

$$\det \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

With so many zeros, almost every one of the $4! = 24$ terms dies. A term survives only if it picks a nonzero entry from each row and column. Hunting them down:

- The choice $a_{14}a_{23}a_{32}a_{41} = 1 \cdot 1 \cdot 1 \cdot 1$ uses columns $(4, 3, 2, 1)$. Turning $(4, 3, 2, 1)$ back

into $(1,2,3,4)$ takes *two* exchanges (swap the outer pair, then the inner pair), so the sign is $+$: this term is $+1$.

- The choice $a_{13}a_{22}a_{31}a_{44} = 1\cdot 1\cdot 1\cdot 1$ uses columns $(3,2,1,4)$. That is *one* exchange (swap positions 1 and 3) away from the identity, so the sign is $-$: this term is -1.

No other permutation avoids all the zeros. So $\det = (+1) + (-1) = 0$. The matrix is singular, and indeed (row 1) $-$ (row 2) $+$ (row 3) $-$ (row 4) $= 0$ is a dependence among the rows.

6.4 Cofactors: Expanding Along a Row

The big formula is correct but unwieldy. The *cofactor expansion* reorganizes its $n!$ terms by factoring out the entries of one row, turning an $n\times n$ determinant into a combination of $(n-1)\times(n-1)$ determinants. It is the recursion that makes hand computation manageable, and it is the form we need for the inverse and for the eigenvalue polynomial later.

6.4.1 The idea: group the big formula by the first row

Go back to the 3×3 big formula and gather all the terms containing a_{11}, then all containing a_{12}, then all containing a_{13}:

$$\det A = a_{11}\underbrace{(a_{22}a_{33}-a_{23}a_{32})}_{C_{11}} + a_{12}\underbrace{(-a_{21}a_{33}+a_{23}a_{31})}_{C_{12}} + a_{13}\underbrace{(a_{21}a_{32}-a_{22}a_{31})}_{C_{13}}.$$

Each parenthesis is (up to a sign) a 2×2 determinant — and not just any one. The factor multiplying a_{11} uses only the entries *not* in row 1 or column 1:

$$C_{11} = +\det\begin{bmatrix} a_{22} & a_{23}\\ a_{32} & a_{33}\end{bmatrix},\quad C_{12} = -\det\begin{bmatrix} a_{21} & a_{23}\\ a_{31} & a_{33}\end{bmatrix},\quad C_{13} = +\det\begin{bmatrix} a_{21} & a_{22}\\ a_{31} & a_{32}\end{bmatrix}.$$

The sign in front alternates $+,-,+$. This is the structure that holds in every dimension.

Definition 6.10: Minor and Cofactor

Let A be $n\times n$. The *minor* M_{ij} is the determinant of the $(n-1)\times(n-1)$ matrix obtained by deleting row i and column j of A. The *cofactor* of a_{ij} is the signed minor

$$C_{ij} = (-1)^{i+j}\,M_{ij}.$$

The sign $(-1)^{i+j}$ follows the checkerboard pattern

$$\begin{bmatrix} + & - & + & \cdots\\ - & + & - & \cdots\\ + & - & + & \cdots\\ \vdots & & & \ddots\end{bmatrix},$$

positive when $i+j$ is even, negative when $i+j$ is odd.

Theorem 6.11: Cofactor Expansion (Laplace Expansion)

For any $n \times n$ matrix A, the determinant may be expanded along any row i:

$$\det A = a_{i1}C_{i1} + a_{i2}C_{i2} + \cdots + a_{in}C_{in} = \sum_{j=1}^{n} a_{ij}C_{ij},$$

or, because $\det A^{\mathsf{T}} = \det A$, along any column j:

$$\det A = a_{1j}C_{1j} + a_{2j}C_{2j} + \cdots + a_{nj}C_{nj} = \sum_{i=1}^{n} a_{ij}C_{ij}.$$

Proof. The cofactor C_{ij} collects exactly the terms of the big formula that contain a_{ij}, with a_{ij} divided out. Every term of the big formula contains exactly one entry from row i, so grouping all $n!$ terms by which row-i entry they use partitions the sum into the n groups $a_{ij}C_{ij}$. That the leftover factor in each group is $(-1)^{i+j}$ times the determinant of the matrix with row i and column j deleted is a bookkeeping check on the signs of the permutations; it works out to exactly the checkerboard sign. The column version follows by applying the row version to A^{T}. □

Remark (The two-by-two cofactor sanity check).

Expanding $\det\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ along the first row: the cofactor of a is $+\det[d] = d$, and the cofactor of b is $-\det[c] = -c$. So $\det = a \cdot d + b \cdot (-c) = ad - bc$, as it must be. The minus sign on the b term is the $(-1)^{1+2}$ from the checkerboard.

6.4.2 Choose a good row or column

The freedom to expand along *any* row or column is the practical gift here: pick the one with the most zeros, so that most cofactors are never computed.

Example (A tridiagonal recursion).

The $n \times n$ tridiagonal matrix of 1's — with 1's on the diagonal and on the two adjacent diagonals, 0's elsewhere — has a determinant that satisfies a tidy recursion. Write $D_n = \det A_n$. Expanding along the first row of

$$A_4 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

gives two surviving cofactors. The $a_{11} = 1$ cofactor is D_3, the determinant of the lower-right 3×3 tridiagonal block. The $a_{12} = 1$ cofactor, after deleting row 1 and column 2 and applying the checkerboard minus sign, expands once more to leave $-1 \cdot D_2$. So

$$D_n = D_{n-1} - D_{n-2}.$$

Starting from $D_1 = 1$ and $D_2 = \det\left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right] = 0$, the sequence is

$$1,\ 0,\ -1,\ -1,\ 0,\ 1,\ 1,\ 0,\ -1, \ldots$$

which repeats with period 6. The determinant never grows — a hint that these matrices hover near singularity.

6.5 The Inverse Formula

Cofactors do more than compute a single number. Assembled correctly, they build the entire inverse A^{-1}. This is the explicit formula that the 2×2 rule

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

is secretly an instance of. Notice the ingredients on the right: the leading factor $1/\det A$, and a matrix of cofactors. The general pattern keeps both.

Definition 6.12: Cofactor Matrix

The *cofactor matrix* of A is the $n \times n$ matrix C whose (i, j) entry is the cofactor $C_{ij} = (-1)^{i+j} M_{ij}$.

Theorem 6.13: The Cofactor Formula for the Inverse

If A is invertible, then

$$A^{-1} = \frac{1}{\det A}\, C^{\mathsf{T}},$$

where C is the cofactor matrix. *Mind the transpose:* the cofactors of row i of A go into column i of A^{-1}.

Proof. We show $A\,C^{\mathsf{T}} = (\det A)\, I$; dividing by $\det A$ then gives the formula. Look at the (i, k) entry of $A\,C^{\mathsf{T}}$. It is row i of A dotted with column k of C^{T} — and column k of C^{T} is row k of C, namely the cofactors $C_{k1}, \ldots, C_{kn}$:

$$(A\,C^{\mathsf{T}})_{ik} = \sum_{j=1}^{n} a_{ij}\, C_{kj}.$$

Diagonal entries ($i = k$). The sum $\sum_j a_{ij} C_{ij}$ is exactly the cofactor expansion of $\det A$ along row i. So every diagonal entry equals $\det A$.

Off-diagonal entries ($i \neq k$). The sum $\sum_j a_{ij} C_{kj}$ uses the *entries* of row i but the *cofactors* of row k. That is precisely the cofactor expansion (along row k) of a modified matrix in which row k has been replaced by a copy of row i. This modified matrix has two equal rows (rows i and k are both row i of A), so by (P4) its determinant is 0. Hence every off-diagonal entry is 0.

Therefore $A\,C^{\mathsf{T}} = (\det A)\, I$, and $A^{-1} = \frac{1}{\det A} C^{\mathsf{T}}$. □

Example (Back to the 2×2 case).

For $A = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ the cofactors are single entries with signs: $C_{11} = +d$, $C_{12} = -c$, $C_{21} = -b$, $C_{22} = +a$. So

$$C = \begin{bmatrix} d & -c \\ -b & a \end{bmatrix}, \qquad C^{\mathsf{T}} = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}, \qquad A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

The transpose is what swaps the off-diagonal cofactors into place. This is the familiar rule, now seen as the $n = 2$ case of the cofactor formula.

Remark (Why the cancellation works).

Strang's intuition: $\det A$ is a sum of products of n entries, while each cofactor is a sum of products of $n - 1$ entries. When A multiplies $\frac{1}{\det A} C^{\mathsf{T}}$, the missing entry on each side fills in so that the numerator becomes $\det A$ on the diagonal (giving 1 after dividing) and a determinant-with-a-repeated-row off the diagonal (giving 0). It is far easier to see this from the cofactor structure than from Gauss–Jordan elimination.

6.6 Cramer's Rule

The inverse formula gives a startlingly explicit answer for the solution of $Ax = b$. When A is invertible, $x = A^{-1}b = \frac{1}{\det A} C^{\mathsf{T}} b$. Reading off the components of this product turns each unknown x_j into a ratio of two determinants.

Theorem 6.14: Cramer's Rule

Let A be invertible and $Ax = b$. Then for each j,

$$x_j = \frac{\det B_j}{\det A},$$

where B_j is the matrix formed by replacing the j-th column of A with the right-hand side b (and leaving every other column of A alone).

Proof. From $x = \frac{1}{\det A} C^{\mathsf{T}} b$, the j-th component is $x_j = \frac{1}{\det A} \sum_{i=1}^{n} C_{ij}\, b_i$ (the j-th row of C^{T} is the j-th column of C, namely $C_{1j}, \ldots, C_{nj}$). Now the sum $\sum_i b_i C_{ij}$ is exactly the cofactor expansion, along column j, of the matrix whose column j is b and whose other columns are those of A — because the cofactors C_{ij} depend only on the *other* columns of A, which are untouched. That matrix is B_j, so $\sum_i b_i C_{ij} = \det B_j$, and $x_j = \det B_j / \det A$. $\square$

Example (Cramer on a 2×2 system).

Solve

$$\begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 5 \end{bmatrix}.$$

Here $\det A = 2 \cdot 3 - 1 \cdot 1 = 5$. Replace column 1 by b to get B_1, and column 2 by b to

get B_2:

$$\det B_1 = \det \begin{bmatrix} 3 & 1 \\ 5 & 3 \end{bmatrix} = 9 - 5 = 4, \qquad \det B_2 = \det \begin{bmatrix} 2 & 3 \\ 1 & 5 \end{bmatrix} = 10 - 3 = 7.$$

So $x_1 = \frac{4}{5}$ and $x_2 = \frac{7}{5}$. A quick check: $2 \cdot \frac{4}{5} + \frac{7}{5} = \frac{8+7}{5} = 3$ and $\frac{4}{5} + 3 \cdot \frac{7}{5} = \frac{4+21}{5} = 5$. ✓

Remark (Beautiful, but not for computing).

Cramer's rule is a gem of theory — it writes each unknown as a clean ratio of determinants, and that explicit form is invaluable when you want to see *how* the solution depends on the data. But as a numerical method it is terrible: evaluating $n + 1$ determinants of size n is wildly more work than the single elimination ($A = LU$, then back-substitute) that solves $Ax = b$ directly. Use Cramer's rule to understand, not to compute.

6.7 Determinant as Volume

We close with the geometric meaning that justifies why this one number is worth all the trouble. The determinant is not just an algebraic test for invertibility — it is, in absolute value, a *volume*.

Theorem 6.15: The Determinant Is a Volume

Let A be an $n \times n$ matrix. Then $|\det A|$ is the volume of the n-dimensional box (parallelepiped) whose edges are the rows of A — equivalently the columns, since $\det A^{\mathsf{T}} = \det A$. In $\mathbb{R}^2$ this is the area of the parallelogram spanned by the two rows; in $\mathbb{R}^3$, the volume of the parallelepiped spanned by the three rows.

The clean way to believe this is to check that "volume of the box" obeys the same three defining properties (P1), (P2), (P3) that define the determinant — and if two functions of a matrix obey the same three defining rules, they are the same function (up to the sign that the absolute value erases).

- **(P1) for volume.** When $A = I$ the box is the unit cube, volume 1. So volume agrees with $\det I = 1$.
- **(P2) for volume.** Swapping two edges does not change the box, so it does not change the volume. The determinant flips sign, but $|\det A|$ does not — matched once we take absolute value. (More: for an orthogonal matrix Q the box is a unit cube turned in space, volume 1, and indeed $\det Q = \pm 1$.)
- **(P3a) for volume.** Doubling one edge of the box doubles its volume; scaling that edge by t scales the volume by $|t|$. This matches the scaling part of (P3).
- **(P3b) for volume.** The additive part is the only subtle one. It says that if one edge is split as a sum $v + v'$, the box on $v + v'$ has volume equal to the sum of the volumes of the boxes on v and on v' (the other edges shared). For a parallelogram this is a cut-and-paste fact: sliding along the shared base preserves area, and the two pieces reassemble correctly.

Since volume satisfies the same three rules, $|\det A|$ *is* the volume. Property (P4) is visible too: if two edges of the box coincide, the box is flat and has zero volume — matching $\det A = 0$ for a repeated row.

Remark (A geometric reading of singularity).

Now (P8) has a picture. $\det A = 0$ means the box has collapsed to zero volume: the edge vectors (rows, or columns) lie in a lower-dimensional flat through the origin — in $\mathbb{R}^3$, squashed into a plane or a line. That is exactly the singular case from Chapter 1: dependent columns, a nontrivial null space, no inverse. Invertibility is the box having genuine n-dimensional volume.

Example (Areas of a parallelogram and a triangle).

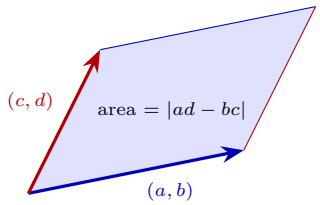
The parallelogram with edge vectors (a, b) and (c, d) has area

$$\left|\det \begin{bmatrix} a & b \\ c & d \end{bmatrix}\right| = |ad - bc| .$$

The triangle formed by those same two edges is half the parallelogram, area $\frac{1}{2}|ad - bc|$. And a triangle with arbitrary vertices (x_1, y_1), (x_2, y_2), (x_3, y_3) has area

$$\frac{1}{2}\left|\det \begin{bmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{bmatrix}\right|,$$

the extra column of 1's accounting for the fact that the vertices need not sit at the origin. Coordinates of corners in, area out — the determinant is doing geometry.



The parallelogram spanned by the two rows of A; its area is $|\det A|$.

6.8 What to Carry Forward

We built the determinant from three rules — $\det I = 1$, sign flip on row exchange, linearity in each row — and out of them came the whole machine. Two derived properties carry the weight. Property (P7), the triangular product rule, makes the determinant *computable*: eliminate, count exchanges, multiply the pivots. Property (P8), $\det A = 0 \iff A$ singular, makes it *meaningful*: one number decides invertibility, independence, and solvability all at once.

From the same three rules we derived the big formula (a signed sum of $n!$ products, one per permutation), the cofactor expansion (the recursion down to smaller determinants),

and from cofactors both the inverse $A^{-1} = \frac{1}{\det A} C^{\mathsf{T}}$ and Cramer's rule $x_j = \det B_j / \det A$. These explicit formulas are theoretical treasures and computational dead ends — when you actually need to solve a system, eliminate. Finally, the determinant is a volume: $|\det A|$ measures the box on the rows (or columns) of A, and singular matrices are exactly the ones whose box has collapsed to zero volume.

Remark (Looking ahead).

> The reason we needed determinants now is the next chapter. Eigenvalues are the numbers λ for which $A - \lambda I$ is singular — which by (P8) means
>
> $$\det(A - \lambda I) = 0.$$
>
> Expanding that determinant (cofactors will do it) produces the *characteristic polynomial*, whose roots are the eigenvalues. Everything in Chapter 7 starts here.

Chapter 7

Eigenvalues, Diagonalization, and Applications

Until now a matrix has been a verb: A acts on a vector x and produces a new vector Ax, usually pointing in a brand-new direction. This chapter hunts for the rare inputs on which A does almost nothing interesting — the vectors x that A merely *stretches*, leaving their direction alone. For such an x we have $Ax = \lambda x$ for some number λ: the matrix collapses, on that one line, to multiplication by a scalar. These special directions are the *eigenvectors*, the stretching factors are the *eigenvalues*, and finding them is the second great problem of the subject, standing beside the first ($Ax = b$).

Why care about a handful of special directions? Because if we can find *enough* of them — a full set of n independent eigenvectors — we can change to coordinates in which A is diagonal. In those coordinates the hard things become easy. Powers A^k, which would take k matrix multiplications, become powers of single numbers λ^k. The matrix exponential e^{At}, which solves $du/dt = Au$, becomes ordinary scalar exponentials $e^{\lambda t}$. The long-run behavior of a Markov chain becomes "keep the eigenvalue $\lambda = 1$, throw the rest away." Eigenvalues turn coupled, tangled systems into a bundle of independent one-dimensional problems. That is the whole story, and we tell it on small 2×2 examples you can check by hand.

7.1 Eigenvalues and Eigenvectors

A matrix A takes a vector x to Ax. For most x the output points somewhere new. But certain special vectors come back parallel to themselves — only their length (and possibly their sign) changes. Those are the directions worth knowing.

Definition 7.1: Eigenvalue and Eigenvector

Let A be an $n \times n$ matrix. A nonzero vector $x \in \mathbb{R}^n$ is an *eigenvector* of A, with *eigenvalue* λ, if

$$Ax = \lambda x.$$

The output Ax lies on the same line as x; the matrix acts on that line as multiplication by the single number λ. We require $x \neq 0$ (the zero vector satisfies $A0 = \lambda 0$ for *every* λ and tells us nothing), but $\lambda = 0$ is allowed.

The eigenvalue λ can be zero: $Ax = 0$ with $x \neq 0$ says x is in the null space, so a singular matrix has $\lambda = 0$ as an eigenvalue, with the null space as its eigenvectors. The eigenvalue is never forced to be real — a rotation has no real eigenvectors at all — but the eigenvector is never zero.

Remark (Read $Ax = \lambda x$ as "no direction change").

Geometrically, A usually rotates and stretches. An eigenvector is a direction that does not rotate: A only stretches it by λ. If $\lambda > 1$ the vector grows, if $0 < \lambda < 1$ it shrinks, if $\lambda < 0$ it flips and stretches, if $\lambda = 0$ it is annihilated. The eigenvectors are the "axes" along which the action of A is purely a scaling.

7.1.1 Eigenvalues you can see without computing

Some matrices wear their eigenvectors on the outside.

Example (Projection: eigenvalues 1 and 0).

Let P be the matrix that projects $\mathbb{R}^3$ orthogonally onto a plane through the origin (Chapter 5). If x already lies *in* the plane, projecting changes nothing: $Px = x$, so x is an eigenvector with $\lambda = 1$. If x is *perpendicular* to the plane, its projection is the zero vector: $Px = 0 = 0 \cdot x$, an eigenvector with $\lambda = 0$. The eigenvalues of any projection are just 1 (the directions kept) and 0 (the directions killed). Here the eigenvectors span all of $\mathbb{R}^3$ — two independent directions in the plane and one perpendicular to it.

Example (A swap matrix: eigenvalues 1 and -1).

The matrix

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

swaps the two components of a vector. The vector $(1, 1)$ is unchanged by the swap, $B(1, 1) = (1, 1)$, so $\lambda = 1$. The vector $(1, -1)$ becomes $(-1, 1) = -(1, -1)$, so $\lambda = -1$. These two eigenvectors are perpendicular — no accident, since $B = B^{\mathsf{T}}$ is symmetric, and symmetric matrices always have orthogonal eigenvectors (Chapter 8). They span $\mathbb{R}^2$.

These examples are lucky: the geometry handed us the eigenvectors. For a general matrix we need a method.

7.1.2 The characteristic equation $\det(A - \lambda I) = 0$

The equation $Ax = \lambda x$ has *two* unknowns wrapped together: the number λ and the vector x. The trick is to move everything to one side and refuse to divide:

$$Ax - \lambda x = 0 \quad \Longleftrightarrow \quad (A - \lambda I)\, x = 0.$$

We are looking for a *nonzero* x in the null space of $A - \lambda I$. But a matrix has a nonzero null vector exactly when it is singular — when its determinant vanishes. This converts the eigenvalue problem into a determinant problem, with λ as the only unknown.

Theorem 7.2: The Characteristic Equation

The number λ is an eigenvalue of the $n \times n$ matrix A if and only if $A - \lambda I$ is singular, that is,

$$\det(A - \lambda I) = 0.$$

This is the *characteristic equation*. Its left side is a polynomial of degree n in λ (the *characteristic polynomial*), so A has n eigenvalues, counted with multiplicity. For each eigenvalue λ, the eigenvectors are the nonzero vectors in the null space $N(A - \lambda I)$.

Proof. A nonzero x with $(A - \lambda I)x = 0$ exists if and only if $A - \lambda I$ has a nontrivial null space, which happens exactly when $A - \lambda I$ is not invertible, i.e. $\det(A - \lambda I) = 0$ (Chapter 6). Expanding the determinant of the $n \times n$ matrix $A - \lambda I$ gives a polynomial in λ whose leading term is $(-\lambda)^n$, hence degree n; by the fundamental theorem of algebra it has n roots in $\mathbb{C}$. □

The procedure is now mechanical:

How to find eigenvalues and eigenvectors

Step 1. Form $A - \lambda I$ and compute the characteristic polynomial $\det(A - \lambda I)$.
Step 2. Solve $\det(A - \lambda I) = 0$ for the n eigenvalues $\lambda_1, \ldots, \lambda_n$.
Step 3. For each λ_i, find the null space of $A - \lambda_i I$ by elimination; its nonzero vectors are the eigenvectors for λ_i.

7.1.3 Trace and determinant: two free checks

Before computing, notice two shortcuts that come straight from the characteristic polynomial. Write it with its roots:

$$\det(A - \lambda I) = (\lambda_1 - \lambda)(\lambda_2 - \lambda) \cdots (\lambda_n - \lambda).$$

Setting $\lambda = 0$ gives $\det A = \lambda_1 \lambda_2 \cdots \lambda_n$. And matching the coefficient of λ^{n-1} on both sides ties the sum of the roots to the sum of the diagonal entries.

Theorem 7.3: Trace and Determinant from Eigenvalues

For any $n \times n$ matrix A with eigenvalues $\lambda_1, \dots, \lambda_n$ (with multiplicity),

$$\lambda_1 + \lambda_2 + \cdots + \lambda_n = \operatorname{trace} A \qquad \text{and} \qquad \lambda_1 \lambda_2 \cdots \lambda_n = \det A.$$

The eigenvalues *sum* to the trace (the sum of the diagonal entries) and *multiply* to the determinant.

Proof. Both sides of $\det(A - \lambda I) = \prod_i (\lambda_i - \lambda)$ are polynomials in λ. Put $\lambda = 0$: the right side is $\prod_i \lambda_i$ and the left side is $\det A$, giving the product formula. For the sum, expand the determinant: the only term producing λ^{n-1} comes from the product of the diagonal entries $(a_{11} - \lambda) \cdots (a_{nn} - \lambda)$, whose λ^{n-1} coefficient is $(-1)^{n-1}(a_{11} + \cdots + a_{nn}) = (-1)^{n-1} \operatorname{trace} A$. On the right, the λ^{n-1} coefficient of $\prod_i (\lambda_i - \lambda)$ is $(-1)^{n-1}(\lambda_1 + \cdots + \lambda_n)$. Matching gives $\operatorname{trace} A = \sum_i \lambda_i$. □

For a 2×2 matrix these two facts are especially handy: the characteristic equation collapses to

$$\lambda^2 - (\operatorname{trace} A)\, \lambda + \det A = 0.$$

If you spot one eigenvalue, the trace hands you the other for free.

Example (A symmetric 2×2: $A = \left[\begin{smallmatrix} 3 & 1 \\ 1 & 3 \end{smallmatrix}\right]$).

Compute the characteristic polynomial:

$$\det(A - \lambda I) = \det \begin{bmatrix} 3 - \lambda & 1 \\ 1 & 3 - \lambda \end{bmatrix} = (3 - \lambda)^2 - 1 = \lambda^2 - 6\lambda + 8.$$

The coefficient 6 is the trace $(3+3)$ and the constant 8 is the determinant $(9-1)$, exactly as the formula $\lambda^2 - (\operatorname{trace} A)\lambda + \det A$ predicts. Factoring,

$$\lambda^2 - 6\lambda + 8 = (\lambda - 4)(\lambda - 2) = 0, \qquad \lambda_1 = 4,\ \lambda_2 = 2.$$

Check: $\lambda_1 + \lambda_2 = 6 = \operatorname{trace} A$ and $\lambda_1 \lambda_2 = 8 = \det A$. Now the eigenvectors.

For $\lambda_1 = 4$, solve $(A - 4I)x = 0$:

$$A - 4I = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}, \qquad x_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

For $\lambda_2 = 2$, solve $(A - 2I)x = 0$:

$$A - 2I = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \qquad x_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

The two eigenvectors are perpendicular (again because $A = A^{\mathsf{T}}$), and they are exactly the eigenvectors of the swap matrix B above. That is no coincidence: $A = B + 3I$, and adding $3I$ adds 3 to every eigenvalue while leaving the eigenvectors fixed.

Remark (Shifting by cI shifts the eigenvalues).

If $Ax = \lambda x$ then $(A + cI)x = (\lambda + c)x$: the same eigenvector, with eigenvalue $\lambda + c$. Shifting A by a multiple of the identity slides all eigenvalues by c and changes nothing else. But beware the general warning: for unrelated A and B, the eigenvalues of $A + B$ are *not* the sums $\lambda(A) + \lambda(B)$, and the eigenvalues of AB are not the products — those clean rules hold only when A and B share the same eigenvectors. The shift $A + cI$ works precisely because A and cI trivially share all eigenvectors.

7.1.4 Two warnings: complex and repeated eigenvalues

Not every matrix has a tidy set of real, distinct eigenvalues. Two things can go wrong, and both matter later.

Example (A rotation: complex eigenvalues).

The matrix

$$Q = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

rotates every vector in the plane by 90°. No real vector keeps its direction under a 90° turn, so we expect trouble. Indeed

$$\det(Q - \lambda I) = \det \begin{bmatrix} -\lambda & -1 \\ 1 & -\lambda \end{bmatrix} = \lambda^2 + 1 = 0, \qquad \lambda = \pm i.$$

The eigenvalues are the imaginary numbers i and $-i$. Note $\lambda_1 + \lambda_2 = 0 = \operatorname{trace} Q$ and $\lambda_1 \lambda_2 = 1 = \det Q$ still hold. Complex eigenvalues of a real matrix always come in conjugate pairs $a \pm bi$. Antisymmetric matrices like Q (with $Q^{\mathsf{T}} = -Q$) have purely imaginary eigenvalues; symmetric matrices, at the other extreme, have purely real ones.

Example (A triangular matrix: a repeated eigenvalue with too few eigenvectors).

For a triangular matrix the eigenvalues sit right on the diagonal, because $\det(A - \lambda I)$ is the product of the diagonal entries of $A - \lambda I$. Take

$$A = \begin{bmatrix} 3 & 1 \\ 0 & 3 \end{bmatrix}, \qquad \det(A - \lambda I) = (3 - \lambda)(3 - \lambda) = 0, \qquad \lambda = 3,\, 3.$$

The eigenvalue 3 is repeated. Its eigenvectors solve $(A - 3I)x = 0$:

$$A - 3I = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x = 0 \quad \Longrightarrow \quad x = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

There is only *one* independent eigenvector, not two. This matrix is *defective*: a repeated eigenvalue did not supply enough eigenvectors to span $\mathbb{R}^2$. We will see in a moment that exactly this shortage is what blocks diagonalization.

7.2 Diagonalization: $A = S\Lambda S^{-1}$

Here is the payoff. Suppose A has n independent eigenvectors. Put them, as columns, into a matrix S. Watch what A does to all of them at once.

Since $Ax_i = \lambda_i x_i$ for each eigenvector,

$$AS = A\begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} = \begin{bmatrix} \lambda_1 x_1 & \lambda_2 x_2 & \cdots & \lambda_n x_n \end{bmatrix}$$
$$= \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} \lambda_1 & & & \\ & \ddots & & \\ & & & \lambda_n \end{bmatrix} = S\Lambda,$$

where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ is the diagonal matrix of eigenvalues. Each column of AS is $Ax_i = \lambda_i x_i$, which is the ith column of S scaled by λ_i — and scaling the ith column is exactly what multiplying S on the right by $\operatorname{diag}(\lambda_i)$ does. So $AS = S\Lambda$.

Because the eigenvectors are independent, S is invertible. Multiply by S^{-1} from the left or right:

$$S^{-1}AS = \Lambda \qquad \text{and equivalently} \qquad A = S\Lambda S^{-1}.$$

Theorem 7.4: Diagonalization

An $n \times n$ matrix A is *diagonalizable* if and only if it has n linearly independent eigenvectors. In that case, with S the matrix whose columns are the eigenvectors and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ the matrix of corresponding eigenvalues,

$$A = S\Lambda S^{-1}, \qquad S^{-1}AS = \Lambda.$$

The matrix S is called the *eigenvector matrix*; Λ is the *eigenvalue matrix*.

Proof. If A has n independent eigenvectors, the computation above gives $AS = S\Lambda$; independence makes S invertible, so $A = S\Lambda S^{-1}$. Conversely, if $A = S\Lambda S^{-1}$ with Λ diagonal, then $AS = S\Lambda$, and reading this column by column gives $A(\text{column } i \text{ of } S) = \lambda_i(\text{column } i \text{ of } S)$: the columns of S are eigenvectors. They are independent because S is invertible. So A has n independent eigenvectors. □

Remark (Diagonalization is a change of coordinates).

Read $A = S\Lambda S^{-1}$ from right to left as a three-step recipe for A. First S^{-1} rewrites the input in the eigenvector coordinate system (how much of each eigenvector is present). Then Λ scales each eigen-coordinate by its λ_i — the easy diagonal action. Then S translates back to the standard coordinates. In the eigenvector basis, A *is* the diagonal matrix Λ; the matrices S and S^{-1} are just the dictionaries in and out of that basis. This is the same change-of-basis idea that organizes Chapter 10.

7.2.1 When is a matrix diagonalizable?

The theorem says everything hinges on having enough independent eigenvectors. When are we guaranteed them?

Theorem 7.5: Distinct Eigenvalues Give Independent Eigenvectors

Eigenvectors that belong to distinct eigenvalues are linearly independent. Consequently, if an $n \times n$ matrix has n distinct eigenvalues, it has n independent eigenvectors and is diagonalizable.

Proof. Suppose, for a contradiction, that some eigenvectors with distinct eigenvalues are dependent, and take a shortest dependent relation $c_1x_1 + \cdots + c_kx_k = 0$ with all $c_i \neq 0$ and the λ_i distinct. Apply A: $\sum_i c_i\lambda_i x_i = 0$. Now subtract λ_k times the original relation: $\sum_{i=1}^{k-1} c_i(\lambda_i - \lambda_k)x_i = 0$. The coefficients $c_i(\lambda_i - \lambda_k)$ are nonzero (distinct eigenvalues), giving a *shorter* dependent relation — contradiction. So the eigenvectors are independent. □

So distinct eigenvalues are a sufficient (not necessary) condition, and "most" matrices have distinct eigenvalues and are diagonalizable. The danger lurks only at *repeated* eigenvalues.

Repeated eigenvalues: diagonalizable or not

A repeated eigenvalue *may or may not* have enough eigenvectors.

• The identity I has eigenvalue 1 repeated n times, yet *every* vector is an eigenvector — a full set of n of them. I is diagonalizable (it already is diagonal).

• The matrix $\left[\begin{smallmatrix} 2 & 1 \\ 0 & 2 \end{smallmatrix}\right]$ has eigenvalue 2 repeated twice but only *one* independent eigenvector $(1, 0)$. It is *defective* and *cannot* be diagonalized.

A repeated eigenvalue is diagonalizable exactly when its eigenspace $N(A - \lambda I)$ has dimension equal to the multiplicity of the root. When it falls short, no eigenvector matrix S exists.

Remark (The two failures are different).

Be careful to separate two things. *Complex eigenvalues* (like the rotation's $\pm i$) do not block diagonalization — they just force us to work over $\mathbb{C}$, where the rotation *is* diagonalizable. What truly blocks diagonalization is a *shortage of eigenvectors* at a repeated eigenvalue, as in $\left[\begin{smallmatrix} 2 & 1 \\ 0 & 2 \end{smallmatrix}\right]$. Invertibility of A is a separate matter again: a singular matrix ($\lambda = 0$) can be perfectly diagonalizable. Diagonalizability is about S; invertibility is about Λ.

7.3 Powers of a Matrix: $A^k = S\Lambda^k S^{-1}$ and Difference Equations

The first reward of diagonalization is that powers become trivial. Multiplying $A = S\Lambda S^{-1}$ by itself, the inner $S^{-1}S$ collapses to I:

$$A^2 = (S\Lambda S^{-1})(S\Lambda S^{-1}) = S\Lambda(S^{-1}S)\Lambda S^{-1} = S\Lambda^2 S^{-1}.$$

The same telescoping repeats k times.

Theorem 7.6: Powers of a Diagonalizable Matrix

If $A = S\Lambda S^{-1}$ then for every positive integer k,

$$A^k = S\Lambda^k S^{-1}, \qquad \Lambda^k = \operatorname{diag}(\lambda_1^k, \dots, \lambda_n^k).$$

The eigenvectors of A^k are the same as those of A; the eigenvalues are raised to the kth power. (Directly: $Ax = \lambda x$ gives $A^2x = \lambda Ax = \lambda^2 x$, and so on.)

Raising a full matrix to the 100th power would be 99 matrix multiplications; with diagonalization it is one cube-easy step, λ_i^{100}, sandwiched between S and S^{-1}. And the eigenvalues immediately reveal the long-run behavior.

Corollary 7.7: Long-Run Behavior of A^k

If A has n independent eigenvectors, then $A^k \to 0$ as $k \to \infty$ if and only if every eigenvalue satisfies $|\lambda_i| < 1$. Eigenvalues with $|\lambda| > 1$ make A^k blow up; the eigenvalue of largest magnitude dominates.

7.3.1 Difference equations $u_{k+1} = Au_k$

A great many processes step forward by multiplying the current state by a fixed matrix: $u_{k+1} = Au_k$. The solution is $u_k = A^k u_0$, and diagonalization turns that into pure scalar powers. The trick is to expand the starting vector u_0 in the eigenvector basis.

Write $u_0 = c_1x_1 + \cdots + c_nx_n = Sc$ (find the coefficients by $c = S^{-1}u_0$). Then applying A scales each piece by its eigenvalue, and applying A a total of k times scales by λ_i^k:

$$u_k = A^k u_0 = c_1\lambda_1^k\, x_1 + c_2\lambda_2^k\, x_2 + \cdots + c_n\lambda_n^k\, x_n.$$

Each eigenvector evolves *independently*, growing or shrinking by its own factor λ_i^k. The coupled system has been uncoupled into n separate one-dimensional ones.

Solving $u_{k+1} = Au_k$

(1) Find the eigenvalues λ_i and eigenvectors x_i of A.
(2) Write the start as a combination of eigenvectors: $u_0 = \sum_i c_i x_i$, i.e. $c = S^{-1}u_0$.
(3) Multiply each coefficient by λ_i^k: $u_k = \sum_i c_i\lambda_i^k\, x_i$.

The largest $|\lambda_i|$ controls the long-run growth; eigenvalues with $|\lambda_i| < 1$ fade away.

Example (The Fibonacci numbers).

The Fibonacci sequence $0, 1, 1, 2, 3, 5, 8, 13, \ldots$ obeys $F_{k+2} = F_{k+1} + F_k$. This is one scalar equation of *second* order, but the standard trick turns it into a first-order *system*. Track two numbers at once, $u_k = \begin{bmatrix} F_{k+1} \\ F_k \end{bmatrix}$. Then

$$\begin{aligned} F_{k+2} &= F_{k+1} + F_k, \\ F_{k+1} &= F_{k+1}, \end{aligned} \qquad \Longrightarrow \qquad u_{k+1} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} u_k = Au_k.$$

The matrix $A = \left[\begin{smallmatrix} 1 & 1 \\ 1 & 0 \end{smallmatrix}\right]$ is symmetric, so its eigenvalues are real. Its trace is 1 and its determinant is -1, so by $\lambda^2 - (\operatorname{trace} A)\lambda + \det A = 0$,

$$\lambda^2 - \lambda - 1 = 0, \qquad \lambda_1 = \frac{1+\sqrt{5}}{2} \approx 1.618, \qquad \lambda_2 = \frac{1-\sqrt{5}}{2} \approx -0.618.$$

Only λ_1 exceeds 1 in magnitude, so it governs the growth: F_k grows like λ_1^k, multiplying by the golden ratio each step, while the λ_2^k term decays to 0. Solving $(A - \lambda I)x = 0$ gives eigenvectors $x_i = \left[\begin{smallmatrix} \lambda_i \\ 1 \end{smallmatrix}\right]$. Expanding $u_0 = \left[\begin{smallmatrix} F_1 \\ F_0 \end{smallmatrix}\right] = \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$ in this basis gives $c_1 = -c_2 = 1/\sqrt{5}$, and reading off the second component of $u_k = c_1\lambda_1^k x_1 + c_2\lambda_2^k x_2$ yields the closed form

$$F_k = \frac{1}{\sqrt{5}} \left[\left(\frac{1+\sqrt{5}}{2} \right)^k - \left(\frac{1-\sqrt{5}}{2} \right)^k \right].$$

An exact integer formula for every Fibonacci number, built entirely from eigenvalues.

7.4 Differential Equations $du/dt = Au$ and the Matrix Exponential

Difference equations step in discrete jumps; differential equations flow continuously. The continuous analogue of $u_{k+1} = Au_k$ is

$$\frac{du}{dt} = Au, \qquad u(0) \text{ given.}$$

This couples the components of u together: each derivative du_i/dt depends on all the u_j. The same idea unties them — expand in eigenvectors, where the matrix acts as a scalar.

7.4.1 Pure exponential solutions

Guess that one eigenvector evolves on its own as $u(t) = e^{\lambda t}x$, where $Ax = \lambda x$. Check it:

$$\frac{du}{dt} = \lambda e^{\lambda t}x \qquad \text{and} \qquad Au = e^{\lambda t}Ax = \lambda e^{\lambda t}x.$$

They agree, so $e^{\lambda t}x$ really solves $du/dt = Au$. These "pure" exponential solutions are the continuous cousins of the terms $\lambda_i^k x_i$ from the difference equation: there λ^k, here $e^{\lambda t}$.

Because the equation is linear, any combination of them is again a solution.

Theorem 7.8: General Solution of $du/dt = Au$

If A has independent eigenvectors $x_1, \dots, x_n$ with eigenvalues $\lambda_1, \dots, \lambda_n$, the general solution of $du/dt = Au$ is

$$u(t) = c_1 e^{\lambda_1 t} x_1 + c_2 e^{\lambda_2 t} x_2 + \cdots + c_n e^{\lambda_n t} x_n.$$

The constants c_i are fixed by the initial condition: $u(0) = \sum_i c_i x_i$, i.e. $c = S^{-1} u(0)$.

The eigenvalues now read off the *stability* of the flow. Each term carries $e^{\lambda t}$; for real λ it decays when $\lambda < 0$ and grows when $\lambda > 0$, and in general $e^{\lambda t} = e^{(\operatorname{Re}\lambda) t}(\cos + i \sin)$ decays exactly when $\operatorname{Re}\lambda < 0$.

Stability from the eigenvalues

For $du/dt = Au$:

- **Stable** ($u(t) \to 0$) when *every* eigenvalue has $\operatorname{Re}\lambda < 0$.
- **Steady state** when one eigenvalue is $\lambda = 0$ and all others have $\operatorname{Re}\lambda < 0$: the $\lambda = 0$ eigenvector survives as $t \to \infty$, everything else decays.
- **Blow-up** when some eigenvalue has $\operatorname{Re}\lambda > 0$.

Contrast with difference equations, where the boundary is $|\lambda| = 1$, not $\operatorname{Re}\lambda = 0$: decay needs $|\lambda| < 1$ for A^k, but $\operatorname{Re}\lambda < 0$ for e^{At}.

Example (A 2×2 flow with a steady state).

Take the system

$$\frac{du_1}{dt} = -u_1 + 2u_2, \qquad \frac{du_2}{dt} = u_1 - 2u_2, \qquad A = \begin{bmatrix} -1 & 2 \\ 1 & -2 \end{bmatrix}, \qquad u(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

The columns of A each sum to 0, so A is singular and $\lambda = 0$ is one eigenvalue; the trace is -3, so the other is $\lambda_2 = -3$. The eigenvectors:

$$\lambda_1 = 0: \ Ax = 0 \Rightarrow x_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \qquad \lambda_2 = -3: \ (A + 3I)x = 0 \Rightarrow x_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

The general solution is

$$u(t) = c_1 e^{0 \cdot t} \begin{bmatrix} 2 \\ 1 \end{bmatrix} + c_2 e^{-3t} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = c_1 \begin{bmatrix} 2 \\ 1 \end{bmatrix} + c_2 e^{-3t} \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Fitting $u(0) = (1, 0)$ gives $c_1 = c_2 = \frac{1}{3}$. As $t \to \infty$ the e^{-3t} term dies, leaving the *steady state*

$$u(\infty) = \frac{1}{3} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 2/3 \\ 1/3 \end{bmatrix}.$$

The eigenvalue $\lambda = 0$ pins down where the system settles; the eigenvalue $\lambda = -3$ sets

how fast it gets there. Notice the total $u_1 + u_2 = 1$ is conserved (the columns of A sum to zero) — "stuff" flows from u_1 to u_2 until balance.

7.4.2 The matrix exponential e^{At}

We can package the whole solution as $u(t) = e^{At}u(0)$, in exact analogy with the scalar $u(t) = e^{at}u(0)$ for $du/dt = au$. But what does it mean to exponentiate a *matrix*? Copy the power series for e^x, replacing x by the matrix At.

Definition 7.9: Matrix Exponential

For a square matrix M, define

$$e^M = I + M + \frac{M^2}{2!} + \frac{M^3}{3!} + \cdots = \sum_{n=0}^{\infty} \frac{M^n}{n!}.$$

In particular $e^{At} = I + At + \dfrac{(At)^2}{2!} + \dfrac{(At)^3}{3!} + \cdots$. The series converges for every square matrix.

This series is the right object: differentiating term by term gives $\frac{d}{dt}e^{At} = Ae^{At}$, so $u(t) = e^{At}u(0)$ solves $du/dt = Au$ with the correct start. Summing infinitely many matrices is impractical, but diagonalization gives a closed form. Substitute $A = S\Lambda S^{-1}$, note $(S\Lambda S^{-1})^n = S\Lambda^n S^{-1}$, and factor S and S^{-1} out of every term:

$$e^{At} = \sum_{n=0}^{\infty} \frac{(S\Lambda S^{-1}t)^n}{n!} = S\left(\sum_{n=0}^{\infty} \frac{(\Lambda t)^n}{n!}\right) S^{-1} = S\, e^{\Lambda t}\, S^{-1}.$$

Theorem 7.10: Computing e^{At} by Diagonalization

If $A = S\Lambda S^{-1}$ is diagonalizable, then

$$e^{At} = S\, e^{\Lambda t}\, S^{-1}, \qquad e^{\Lambda t} = \operatorname{diag}\left(e^{\lambda_1 t}, e^{\lambda_2 t}, \ldots, e^{\lambda_n t}\right).$$

The exponential of a diagonal matrix is just the diagonal of scalar exponentials. Thus $u(t) = e^{At}u(0) = Se^{\Lambda t}S^{-1}u(0)$, which expands to exactly $\sum_i c_i e^{\lambda_i t}x_i$ with $c = S^{-1}u(0)$.

Proof. For diagonal Λ, $\Lambda^n = \operatorname{diag}(\lambda_i^n)$, so the series $\sum_n (\Lambda t)^n/n!$ acts entry by entry on the diagonal and equals $\operatorname{diag}(e^{\lambda_i t})$. Inserting $A^n = S\Lambda^n S^{-1}$ into the definition and pulling the constant S, S^{-1} through the (convergent) sum gives $e^{At} = S\, e^{\Lambda t} S^{-1}$. □

Remark (Same diagonalization, two flavors of time).

The pattern is identical to powers. There, $A^k = S\Lambda^k S^{-1}$ replaced the matrix power by scalar powers λ_i^k. Here, $e^{At} = Se^{\Lambda t}S^{-1}$ replaces the matrix exponential by scalar exponentials $e^{\lambda_i t}$. Diagonalization is the single tool that linearizes both discrete and

continuous dynamics: move to the eigenvector basis, solve n scalar problems, move back. (Higher-order scalar ODEs $y'' + by' + ky = 0$ fold into this same machine by the companion-matrix trick that gave us Fibonacci.)

7.5 Markov Matrices: the Eigenvalue $\lambda = 1$ and Steady States

A particularly important difference equation $u_{k+1} = Au_k$ comes from *Markov matrices* — the matrices of probability and of populations flowing between states.

Definition 7.11: Markov Matrix

A square matrix A is a *Markov matrix* if every entry is nonnegative and every column sums to 1. The columns are probability vectors: column j says where the contents of state j go in one step. Powers A^k of a Markov matrix are again Markov matrices.

Because each column sums to 1, the total amount is conserved: $u_{k+1} = Au_k$ redistributes the entries of u_k without changing their sum. So if u_0 holds a population of 1000, so does every u_k. The question is how that population settles, and the answer is one special eigenvalue.

Theorem 7.12: Markov Matrices Have $\lambda = 1$

Every Markov matrix A has $\lambda = 1$ as an eigenvalue, and all its eigenvalues satisfy $|\lambda| \le 1$. The eigenvector for $\lambda = 1$ has nonnegative entries; scaled to sum to 1, it is the *steady-state* distribution.

Proof. To show $\lambda = 1$ is an eigenvalue we show $A - I$ is singular. Subtracting 1 from each diagonal entry makes every column of $A - I$ sum to 0 (each column of A summed to 1). But "all columns sum to zero" says the rows of $A - I$ add up to the zero row: the row vector $(1, 1, \dots, 1)$ satisfies $(1, \dots, 1)(A - I) = 0$. So $A - I$ has dependent rows, hence is singular, and $\det(A - I) = 0$: $\lambda = 1$ is an eigenvalue. (Equivalently, A and A^{T} share eigenvalues since $\det(A - \lambda I) = \det(A^{\mathsf{T}} - \lambda I)$, and A^{T} has row sums 1, so $A^{\mathsf{T}}(1, \dots, 1)^{\mathsf{T}} = (1, \dots, 1)^{\mathsf{T}}$.) The bound $|\lambda| \le 1$ follows because larger eigenvalues would make A^k blow up, impossible when every A^k is again Markov with entries in $[0, 1]$. □

Now run the difference equation. If A has independent eigenvectors with $\lambda_1 = 1$ and $|\lambda_i| < 1$ for $i \ge 2$, then

$$u_k = c_1 \cdot 1^k\, x_1 + c_2 \lambda_2^k\, x_2 + \cdots + c_n \lambda_n^k\, x_n \xrightarrow{k \to \infty} c_1 x_1,$$

because every other λ_i^k decays to 0. The system forgets its start (except for the single number c_1) and converges to the $\lambda = 1$ eigenvector. That is the steady state.

Example (Migration between two states).

Each year 10% of California's people move to Massachusetts and 20% of Massachusetts'

people move to California; the rest stay. With the state vector $u = \left[\begin{smallmatrix} u_{\text{Cal}} \\ u_{\text{Mass}} \end{smallmatrix}\right]$,

$$u_{k+1} = \begin{bmatrix} 0.9 & 0.2 \\ 0.1 & 0.8 \end{bmatrix} u_k = Au_k.$$

Columns sum to 1 — a Markov matrix. One eigenvalue is $\lambda_1 = 1$; the trace $0.9+0.8 = 1.7$ then forces $\lambda_2 = 0.7$. The eigenvectors:

$$\lambda_1 = 1 : \ (A - I)x = 0 \Rightarrow x_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \qquad \lambda_2 = 0.7 : \ (A - 0.7I)x = 0 \Rightarrow x_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Starting from $u_0 = \left[\begin{smallmatrix} 0 \\ 1000 \end{smallmatrix}\right]$ (everyone in Massachusetts), expand $u_0 = c_1x_1 + c_2x_2$ to get $c_1 = \frac{1000}{3}$, $c_2 = -\frac{2000}{3}$, so

$$u_k = \frac{1000}{3}\, 1^k \begin{bmatrix} 2 \\ 1 \end{bmatrix} - \frac{2000}{3}\, (0.7)^k \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

As $k \to \infty$ the $(0.7)^k$ term vanishes and

$$u_\infty = \frac{1000}{3} \begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 2000/3 \\ 1000/3 \end{bmatrix} :$$

two-thirds end up in California, one-third in Massachusetts — the steady state x_1 (normalized), reached from *any* starting distribution. The eigenvalue 1 gives the destination; the eigenvalue 0.7 gives the rate of approach.

Remark (Steady state: $\lambda = 1$ versus $\lambda = 0$).

Watch the two settings side by side. For *differential* equations $du/dt = Au$ the steady state lives in the eigenvector with $\lambda = 0$ (it is the part with $e^{0\cdot t} = 1$ frozen while $e^{\lambda t} \to 0$). For *Markov* difference equations $u_{k+1} = Au_k$ the steady state lives in the eigenvector with $\lambda = 1$ (the part with $1^k = 1$ frozen while $\lambda^k \to 0$). Same idea, shifted by the difference between $e^{\lambda t}$ and λ^k as the time-evolution factor.

7.6 A Short Bridge: Fourier Series and Projection

The chapter has leaned on one move again and again: *expand the input in a good basis, where each basis vector evolves on its own.* For eigenvectors that basis diagonalizes A. When the basis is also *orthonormal*, finding the coefficients becomes effortless — and that effortlessness is exactly what powers Fourier series.

If $q_1, \ldots, q_n$ are an orthonormal basis ($q_i^{\mathsf{T}} q_j = 0$ for $i \neq j$, and $q_i^{\mathsf{T}} q_i = 1$), then any vector $v = x_1q_1 + \cdots + x_nq_n$ has its coefficients given by a single dot product. Hit both sides with q_i^{T} and all cross terms vanish:

$$q_i^{\mathsf{T}} v = x_1\, q_i^{\mathsf{T}} q_1 + \cdots + x_n\, q_i^{\mathsf{T}} q_n = x_i.$$

So $x_i = q_i^{\mathsf{T}} v$: each coordinate is a projection of v onto one basis direction (Chapter 5).

In matrix form, with $Q = [\, q_1 \ \cdots \ q_n \,]$, this is $Qx = v$ with $x = Q^{-1}v = Q^{\mathsf{T}}v$, since an orthonormal Q has $Q^{-1} = Q^{\mathsf{T}}$.

Remark (Fourier series is this idea, infinitely).

Fourier series writes a periodic function as

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

The "vectors" are now functions, the "basis" is $1, \cos x, \sin x, \cos 2x, \ldots$, and the inner product is an integral instead of a sum:

$$\langle f, g \rangle = \int_0^{2\pi} f(x)\, g(x)\, dx.$$

The basic functions are orthogonal under this inner product (for instance $\sin x$ and $\cos x$ integrate to zero against each other), so each Fourier coefficient is found by the same projection formula. For instance,

$$a_1 = \frac{1}{\pi} \int_0^{2\pi} f(x) \cos x \, dx$$

is the component of f along $\cos x$. An infinite-dimensional orthonormal basis, but the eigenvector and projection logic of this chapter is unchanged.

7.7 What to Carry Forward

A square matrix A has special directions, its *eigenvectors* x, on which it acts as plain scaling: $Ax = \lambda x$. We find the eigenvalues by solving $\det(A - \lambda I) = 0$ and the eigenvectors as null spaces of $A - \lambda I$. Two cheap checks — $\sum \lambda_i = \operatorname{trace} A$ and $\prod \lambda_i = \det A$ — guard the arithmetic, and for 2×2 matrices they nearly do the work themselves through $\lambda^2 - (\operatorname{trace} A)\lambda + \det A = 0$.

When A has n independent eigenvectors we can *diagonalize*: $A = S\Lambda S^{-1}$, with eigenvectors in S and eigenvalues in Λ. Distinct eigenvalues guarantee this; a repeated eigenvalue may fall short of eigenvectors and leave A defective. Diagonalization is the master key. It makes powers easy, $A^k = S\Lambda^k S^{-1}$, solving difference equations $u_{k+1} = Au_k$ via $u_k = \sum c_i \lambda_i^k x_i$. It makes the matrix exponential easy, $e^{At} = Se^{\Lambda t}S^{-1}$, solving differential equations $du/dt = Au$ via $u(t) = \sum c_i e^{\lambda_i t} x_i$. The eigenvalues read off stability ($|\lambda| < 1$ for $A^k \to 0$; $\operatorname{Re}\lambda < 0$ for $e^{At} \to 0$) and steady states ($\lambda = 1$ for Markov chains, $\lambda = 0$ for continuous flows).

Remark (Looking ahead).

One question dangles: *which* matrices are sure to have a full set of independent eigenvectors? The cleanest answer is symmetry. A symmetric matrix $A = A^{\mathsf{T}}$ always has real eigenvalues and a complete, *orthonormal* set of eigenvectors — so S can be chosen orthogonal and $A = Q\Lambda Q^{\mathsf{T}}$. That is the *spectral theorem* of Chapter 8, and dropping the requirement that A be square leads to the singular value decomposition of Chapter 9, where eigenvalues of $A^{\mathsf{T}}A$ take over the role λ played here.

Chapter 8

Symmetric and Positive Definite Matrices

Up to now an eigenvalue could be anything — real, complex, repeated, defective. The matrix $A = S\Lambda S^{-1}$ might or might not have enough eigenvectors to diagonalize, and the eigenvector matrix S was just some invertible matrix with no special structure. This chapter is about the family of matrices where everything goes right. A *symmetric* matrix $A = A^{\mathsf{T}}$ always has real eigenvalues, always has a full set of eigenvectors, and — this is the gift — those eigenvectors can be chosen *orthonormal*. So S becomes an *orthogonal* matrix Q, the inverse S^{-1} becomes the cheap transpose Q^{T}, and the diagonalization reads

$$A = Q\Lambda Q^{\mathsf{T}}.$$

This is the **spectral theorem**, the cleanest factorization in the subject, and most of the chapter unfolds from it.

The second half asks a sharper question. Once we know a symmetric matrix has real eigenvalues, we want to know their *signs*. A symmetric matrix whose eigenvalues are all positive is called *positive definite*, and these matrices are everywhere: they are the matrices $A^{\mathsf{T}}A$ of least squares, the covariance matrices of statistics, the stiffness matrices of mechanics, and — the connection that closes the chapter — the second-derivative (Hessian) matrices that detect a genuine minimum. We will collect four different-looking tests for positive definiteness (eigenvalues, determinants, pivots, and the energy $x^{\mathsf{T}}Ax$) and watch them turn out to be the same test in four costumes. Strang likes to say that positive definite matrices bring the whole course together: pivots, determinants, eigenvalues, and stability all meet here.

8.1 Symmetric Matrices Have Real Eigenvalues

A symmetric matrix is one that equals its own transpose, $A = A^{\mathsf{T}}$, so the entry in row i, column j matches the entry in row j, column i. The matrix is a mirror image across its main diagonal. This is a strong constraint with strong rewards, and the first reward concerns the eigenvalues.

For a general real matrix the eigenvalues can be complex — a rotation matrix has eigenvalues $e^{\pm i\theta}$ with no real part at all (Chapter 7). One might expect symmetry to do something nice, and it does the nicest possible thing: it forces every eigenvalue back onto the real line.

Theorem 8.1: Real Eigenvalues of a Symmetric Matrix

If A is a real symmetric matrix, $A = A^{\mathsf{T}}$, then every eigenvalue of A is real.

Proof. Let $Ax = \lambda x$ with $x \neq 0$, where a priori λ and x may be complex. Write $\bar{z}$ for the complex conjugate. Conjugating the equation entrywise, and using that A has real entries so $\bar{A} = A$,

$$A\bar{x} = \bar{\lambda}\,\bar{x}.$$

Now transpose this and use $A^{\mathsf{T}} = A$:

$$\bar{x}^{\mathsf{T}}A = \bar{\lambda}\,\bar{x}^{\mathsf{T}}.$$

Multiply on the right by x:

$$\bar{x}^{\mathsf{T}}Ax = \bar{\lambda}\,\bar{x}^{\mathsf{T}}x.$$

On the other hand, multiply the original $Ax = \lambda x$ on the left by $\bar{x}^{\mathsf{T}}$:

$$\bar{x}^{\mathsf{T}}Ax = \lambda\,\bar{x}^{\mathsf{T}}x.$$

The left sides are identical, so $\bar{\lambda}\,\bar{x}^{\mathsf{T}}x = \lambda\,\bar{x}^{\mathsf{T}}x$. The number $\bar{x}^{\mathsf{T}}x$ is

$$\bar{x}^{\mathsf{T}}x = |x_1|^2 + |x_2|^2 + \cdots + |x_n|^2,$$

a sum of squared magnitudes, which is strictly positive because $x \neq 0$. Dividing it out leaves $\bar{\lambda} = \lambda$, and a number equal to its own conjugate is real. □

The quantity $\bar{x}^{\mathsf{T}}x$ is worth a second look — it is the right way to measure the squared length of a complex vector, and it reappears when we meet Hermitian matrices in Section 8.7. For a *real* eigenvector it is just the ordinary $x^{\mathsf{T}}x = \|x\|^2$.

Remark (A bonus from the same computation).

> The proof conjugated $Ax = \lambda x$ to get $A\bar{x} = \bar{\lambda}\,\bar{x}$. Read that line by itself: if λ is an eigenvalue of a real matrix with eigenvector x, then $\bar{\lambda}$ is an eigenvalue with eigenvector $\bar{x}$. **For any real matrix, complex eigenvalues come in conjugate pairs.** Symmetry is what then collapses each pair onto the real axis.

8.2 Orthogonal Eigenvectors and the Spectral Theorem

Real eigenvalues are only the first gift. The second is that eigenvectors belonging to *different* eigenvalues are automatically perpendicular — no work required, it falls out of symmetry.

Theorem 8.2: Eigenvectors of Distinct Eigenvalues Are Orthogonal

Let $A = A^{\mathsf{T}}$ be real symmetric, and suppose $Ax = \lambda x$ and $Ay = \mu y$ with $\lambda \neq \mu$. Then $x^{\mathsf{T}}y = 0$: the eigenvectors are orthogonal.

Proof. Compute $x^{\mathsf{T}}Ay$ two ways. Using $Ay = \mu y$,

$$x^{\mathsf{T}}Ay = x^{\mathsf{T}}(\mu y) = \mu\, x^{\mathsf{T}}y.$$

Using $A = A^{\mathsf{T}}$ and $Ax = \lambda x$, transpose $Ax = \lambda x$ to get $x^{\mathsf{T}}A = \lambda x^{\mathsf{T}}$, so

$$x^{\mathsf{T}}Ay = (\lambda x^{\mathsf{T}})y = \lambda\, x^{\mathsf{T}}y.$$

Subtracting, $(\lambda - \mu)\, x^{\mathsf{T}}y = 0$. Since $\lambda \neq \mu$, the factor $\lambda - \mu$ is nonzero, forcing $x^{\mathsf{T}}y = 0$. □

So eigenvectors for distinct eigenvalues are orthogonal on the nose. When an eigenvalue is repeated, its eigenspace has dimension equal to the multiplicity (this is the part that needs proof, and it is exactly where symmetry rescues us from the defective matrices of Chapter 7), and *within* that eigenspace we are free to pick an orthonormal basis by Gram–Schmidt (Chapter 5). Stitching the pieces together gives a full orthonormal set of n eigenvectors. Normalizing each to unit length, collect them as the columns of a matrix Q. Orthonormal columns mean $Q^{\mathsf{T}}Q = I$, i.e. Q is an *orthogonal* matrix with $Q^{-1} = Q^{\mathsf{T}}$.

Theorem 8.3: Spectral Theorem

Every real symmetric matrix $A = A^{\mathsf{T}}$ can be factored as

$$A = Q\Lambda Q^{\mathsf{T}},$$

where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ holds the (real) eigenvalues and Q is an orthogonal matrix ($Q^{\mathsf{T}}Q = I$) whose columns are orthonormal eigenvectors of A. Conversely, any matrix of the form $Q\Lambda Q^{\mathsf{T}}$ with real Λ is symmetric.

Proof. With n orthonormal eigenvectors $q_1, \ldots, q_n$ as the columns of Q and eigenvalues on the diagonal of Λ, the relations $Aq_k = \lambda_k q_k$ assemble columnwise into $AQ = Q\Lambda$. Right-multiplying by $Q^{-1} = Q^{\mathsf{T}}$ gives $A = Q\Lambda Q^{\mathsf{T}}$. For the converse, if $A = Q\Lambda Q^{\mathsf{T}}$ then

$$A^{\mathsf{T}} = (Q\Lambda Q^{\mathsf{T}})^{\mathsf{T}} = Q\,\Lambda^{\mathsf{T}}Q^{\mathsf{T}} = Q\Lambda Q^{\mathsf{T}} = A,$$

because a diagonal matrix is its own transpose. So A is symmetric. □

This is the same diagonalization $A = S\Lambda S^{-1}$ from Chapter 7, but with the eigenvector matrix upgraded from a generic invertible S to an orthogonal Q. The payoff is that S^{-1}, normally a separate computation, is handed to us for free as Q^{T}. Mathematicians call the set of eigenvalues the *spectrum* of A, which is where the theorem gets its name; engineers

call the same result the *principal axis theorem*, for a reason we will see when we draw the ellipse of $x^{\mathsf{T}}Ax$.

Example (A symmetric 2×2 diagonalized).

Take

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

The eigenvalues solve $\det(A - \lambda I) = (1-\lambda)^2 - 4 = 0$, so $\lambda = 3$ and $\lambda = -1$ (real, as promised). For $\lambda = 3$, $(A - 3I)x = 0$ gives the eigenvector $(1, 1)$; for $\lambda = -1$, the eigenvector is $(1, -1)$. These two are perpendicular — $(1,1)\cdot(1,-1) = 0$ — exactly as the theorem guarantees, with no checking of the matrix beyond its symmetry. Normalizing each by $1/\sqrt{2}$,

$$Q = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix}, \qquad A = Q\Lambda Q^{\mathsf{T}}.$$

Here Q is symmetric and orthogonal at once, so $Q^{\mathsf{T}} = Q$; you can multiply $Q\Lambda Q$ out and recover A.

8.2.1 Every Symmetric Matrix Is a Sum of Projections

The factorization $A = Q\Lambda Q^{\mathsf{T}}$ has a reading that is easy to miss inside the matrix notation. Expand the product the way we expand any $Q\Lambda Q^{\mathsf{T}}$ — as a sum of rank-one pieces, one per eigenvalue:

$$A = Q\Lambda Q^{\mathsf{T}} = \begin{bmatrix} q_1 & \cdots & q_n \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \begin{bmatrix} q_1^{\mathsf{T}} \\ \vdots \\ q_n^{\mathsf{T}} \end{bmatrix} = \lambda_1 q_1 q_1^{\mathsf{T}} + \lambda_2 q_2 q_2^{\mathsf{T}} + \cdots + \lambda_n q_n q_n^{\mathsf{T}}.$$

Each term $q_k q_k^{\mathsf{T}}$ is exactly the projection matrix onto the line through the unit vector q_k (recall from Chapter 5 that the projection onto a unit vector q is qq^{T}). So the spectral theorem says something concrete and physical.

Spectral decomposition as projections

Every symmetric matrix breaks into a weighted sum of perpendicular projections:

$$A = \lambda_1 P_1 + \lambda_2 P_2 + \cdots + \lambda_n P_n, \qquad P_k = q_k q_k^{\mathsf{T}}.$$

Each P_k projects onto one eigenvector direction, the P_k are mutually orthogonal projections ($P_j P_k = 0$ for $j \neq k$, $P_k^2 = P_k$), and they sum to the identity ($P_1 + \cdots + P_n = I$). Applying A to any vector means: split the vector into its components along the eigen-directions, then stretch the k-th component by λ_k.

That last sentence is the whole geometric content of a symmetric matrix. It does nothing but stretch along n perpendicular axes. There is no rotation, no shear — the orthonormal eigenvectors are the axes, and the eigenvalues are the stretch factors.

8.3 Signs of the Eigenvalues: Pivots and Inertia

Knowing the eigenvalues are real, the next question is their *signs* — how many are positive, how many negative, how many zero. For differential equations the signs decide stability; for the minimum problem ahead they decide whether a critical point is a true minimum. But computing eigenvalues means solving $\det(A - \lambda I) = 0$, an n-th degree equation, which is hopeless by hand for large n. Elimination, by contrast, is cheap. The remarkable fact is that elimination already knows the signs.

Theorem 8.4: Pivots and Eigenvalues Share Signs (Sylvester's Law of Inertia)

Let $A = A^{\mathsf{T}}$ be symmetric and invertible, factored by symmetric elimination into pivots $d_1, \dots, d_n$ (no row exchanges). Then the number of positive pivots equals the number of positive eigenvalues, and the number of negative pivots equals the number of negative eigenvalues:

$$\#\{\text{positive pivots}\} = \#\{\text{positive } \lambda\}, \qquad \#\{\text{negative pivots}\} = \#\{\text{negative } \lambda\}.$$

This count — the triple (positives, negatives, zeros) — is the *inertia* of A.

We will not prove the law in full, but the mechanism is visible in the LDL^{T} factorization. Symmetric elimination writes $A = LDL^{\mathsf{T}}$ with L lower triangular (ones on the diagonal) and $D = \operatorname{diag}(d_1, \dots, d_n)$ the pivots. The eigenvalues of A and of D are generally different numbers, yet they have the same *signs*: passing from A to D via the invertible L never moves an eigenvalue across zero. The pivots are the cheap shadow of the eigenvalues, accurate in sign even though wrong in value.

This is more powerful than it looks, because of a shift trick. The eigenvalues of $A + bI$ are exactly b more than the eigenvalues of A (if $Ax = \lambda x$ then $(A + bI)x = (\lambda + b)x$). So to find out how many eigenvalues of A exceed a chosen number b, count the positive pivots of $A - bI$. Sweeping b across the line locates every eigenvalue between cheap elimination steps, without ever solving the characteristic equation.

Example (Counting signs without eigenvalues).

The symmetric matrix

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

has eigenvalues $+1$ and -1 (one of each sign). Elimination with a row exchange is awkward here, but the determinant test agrees: $\det A = -1 < 0$ means the two eigenvalues have opposite signs, so exactly one is negative. For a clean pivot example, take

$$B = \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix}, \qquad B = LDL^{\mathsf{T}}, \quad D = \begin{bmatrix} 1 & 0 \\ 0 & -8 \end{bmatrix},$$

since subtracting $3\times$row 1 from row 2 gives pivots 1 and $1 - 9 = -8$. One positive pivot, one negative pivot — so one positive eigenvalue and one negative eigenvalue. Indeed the

eigenvalues of B are 4 and -2. The pivots got the *values* wrong but the *signs* exactly right.

8.4 Positive Definite Matrices

Now specialize to the best-behaved symmetric matrices of all: those whose eigenvalues are not merely real but strictly positive.

Definition 8.5: Positive Definite Matrix

A symmetric matrix $A = A^{\mathsf{T}}$ is **positive definite** if all of its eigenvalues are strictly positive:

$$\lambda_1 > 0,\ \lambda_2 > 0,\ \ldots,\ \lambda_n > 0.$$

If the eigenvalues are merely nonnegative ($\lambda_k \geq 0$, allowing zeros), A is **positive semidefinite**.

Checking every eigenvalue is exactly the test we just said is too expensive. The wonderful thing about positive definiteness is that it can be detected several other ways, each cheaper or more meaningful than computing eigenvalues. We collect them as a single theorem; the rest of the chapter is really the proof that these tests agree and the story of why the last one matters.

Theorem 8.6: Four Tests for Positive Definiteness

For a symmetric matrix $A = A^{\mathsf{T}}$, the following are equivalent. Each is a complete test for positive definiteness.

(1) **Eigenvalue test.** All eigenvalues are positive: $\lambda_k > 0$ for every k.

(2) **Determinant test.** All leading principal minors are positive: the determinant of every upper-left $k \times k$ block of A is positive, for $k = 1, \ldots, n$.

(3) **Pivot test.** All n pivots are positive (symmetric elimination, no row exchanges).

(4) **Energy test.** The quadratic form is positive away from the origin: $x^{\mathsf{T}}Ax > 0$ for every $x \neq 0$.

A few words on why the four hang together. Tests (2) and (3) are linked by the fact that each pivot is a ratio of consecutive leading minors,

$$d_k = \frac{\det A_k}{\det A_{k-1}},$$

where A_k is the upper-left $k \times k$ block (and $\det A_0 = 1$). So the pivots are positive exactly when the leading minors are positive — tests (2) and (3) are the same statement, read through the bookkeeping of elimination. Test (1) connects to (3) through inertia: positive pivots mean positive eigenvalues by Sylvester's law. And test (4), the energy, is the deepest — it is the one that does not even mention computing anything about A, and it is the one

that connects to minima. We prove the equivalence of (1) and (4) below, and treat the energy test as the real definition.

Remark (Why the determinant alone is not enough).

> A natural wrong guess is "positive definite means positive determinant." But $\det A = \lambda_1\lambda_2\cdots\lambda_n$ is positive whenever an *even* number of eigenvalues are negative. The matrix
>
> $$\begin{bmatrix} -1 & 0 \\ 0 & -3 \end{bmatrix}$$
>
> has determinant $+3 > 0$, yet both eigenvalues are negative — it is negative definite, the opposite of what we want. Test (2) repairs this by demanding that *every* leading minor be positive, not just the last one. The first minor ($-1 < 0$) catches the imposter immediately.

8.4.1 The energy $x^\mathsf{T}Ax$

The fourth test introduces the quantity that ties the chapter to calculus. For a symmetric A, the scalar

$$x^\mathsf{T}Ax$$

is called the *quadratic form* or the *energy* of A. Writing it out for a symmetric 2×2 matrix shows its shape. With

$$A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \qquad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

we compute

$$x^\mathsf{T}Ax = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} ax_1 + bx_2 \\ bx_1 + cx_2 \end{bmatrix} = ax_1^2 + 2b\,x_1x_2 + cx_2^2.$$

This is a pure second-degree polynomial in x_1, x_2 — no linear or constant term — so its graph $z = x^\mathsf{T}Ax$ passes through the origin and is tangent to the flat plane there. Positive definiteness asks: is the origin the bottom of a bowl (every other point higher), or could the surface dip below zero in some direction? That is precisely the question of a minimum, which is why test (4) is the bridge to calculus.

Theorem 8.7: Energy Test Equals Eigenvalue Test

For a symmetric matrix A, the energy $x^\mathsf{T}Ax > 0$ for all $x \neq 0$ if and only if every eigenvalue of A is positive.

Proof. Diagonalize by the spectral theorem, $A = Q\Lambda Q^\mathsf{T}$. Substitute and change variables to $y = Q^\mathsf{T}x$:

$$x^\mathsf{T}Ax = x^\mathsf{T}Q\Lambda Q^\mathsf{T}x = (Q^\mathsf{T}x)^\mathsf{T}\Lambda(Q^\mathsf{T}x) = y^\mathsf{T}\Lambda y = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2.$$

Because Q is invertible, x ranges over all nonzero vectors exactly when y does. If every $\lambda_k > 0$, this weighted sum of squares is positive whenever some $y_k \neq 0$, i.e. whenever $x \neq 0$ — so A is positive definite. Conversely, if some $\lambda_j \leq 0$, choose $x = q_j$ (the j-th eigenvector),

giving $y = e_j$ and $x^\mathsf{T}Ax = \lambda_j \le 0$ at a nonzero vector — so the energy test fails. The two conditions are equivalent. □

The change of variables in this proof is the principal axis theorem at work: in the eigenvector coordinates y, the energy is a clean sum $\sum \lambda_k y_k^2$ with no cross terms. The eigenvectors are the axes along which the quadratic form decouples; the eigenvalues are the curvatures along those axes.

8.4.2 Completing the square: energy as a sum of squares with pivots

There is a second, hands-on way to see why positive pivots guarantee positive energy, and it is just elimination in disguise. Take the positive definite matrix

$$A = \begin{bmatrix} 2 & 6 \\ 6 & 20 \end{bmatrix}, \qquad x^\mathsf{T}Ax = 2x^2 + 12xy + 20y^2.$$

Complete the square on the x terms:

$$2x^2 + 12xy + 20y^2 = 2\big(x^2 + 6xy\big) + 20y^2 = 2\big(x + 3y\big)^2 - 18y^2 + 20y^2 = 2\,(x + 3y)^2 + 2\,y^2.$$

Both squares carry positive coefficients, so the energy is a sum of two genuine squares and can never be negative; it is zero only when $x + 3y = 0$ and $y = 0$, i.e. only at $x = y = 0$. That is positive definiteness, proved by hand.

Now look at where the numbers came from. Symmetric elimination on A subtracts 3×row 1 from row 2:

$$\begin{bmatrix} 2 & 6 \\ 6 & 20 \end{bmatrix} \xrightarrow{R_2 - 3R_1} \begin{bmatrix} 2 & 6 \\ 0 & 2 \end{bmatrix} = U, \qquad L = \begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix}.$$

The pivots are 2 and 2 — exactly the coefficients in front of the two squares. And the multiplier 3 is exactly the number inside the first square $(x + 3y)$. This is no accident.

Completing the square is elimination

For any symmetric $A = LDL^\mathsf{T}$, the energy splits as

$$x^\mathsf{T}Ax = \sum_{k=1}^{n} d_k\,(k\text{-th row of } L^\mathsf{T}x)^2,$$

a weighted sum of squares whose weights are the pivots d_k and whose squared linear forms come from L. So **positive pivots $\iff$ a sum of *positive* multiples of squares $\iff$ positive energy.** Completing the square by hand and running elimination are literally the same computation. When even one pivot is negative, that square contributes a downward direction, and the energy goes negative.

This makes the link between tests (3) and (4) tangible. The matrix

$$\begin{bmatrix} 2 & 6 \\ 6 & 7 \end{bmatrix}$$

has $\det = 14 - 36 = -22 < 0$, so its eigenvalues have opposite signs and it is *not* positive definite. Completing the square gives $2x^2+12xy+7y^2 = 2(x+3y)^2-11y^2$, which is negative at $x = -3$, $y = 1$. The second pivot turned negative ($7-18 = -11$), and the negative square is the downhill direction.

8.4.3 The boundary: positive semidefinite

Between the positive definite matrices and the indefinite ones sits the boundary case. Tune the corner entry of the last matrix until the determinant just hits zero:

$$A = \begin{bmatrix} 2 & 6 \\ 6 & 18 \end{bmatrix}, \qquad \det A = 2 \cdot 18 - 36 = 0.$$

This A is singular: it has eigenvalues 0 and 20, a single positive pivot, and rank 1. Its energy

$$2x^2 + 12xy + 18y^2 = 2\,(x + 3y)^2$$

is a single square — it is ≥ 0 always, but it equals zero along the whole line $x + 3y = 0$ (for instance at $x = 3$, $y = -1$), not only at the origin. This is the model of a *positive semidefinite* matrix: energy never negative, but flat in some direction. The eigenvalue 0 is exactly that flat direction. Positive definite is the strict interior ($\lambda > 0$, energy a sum of n full squares); positive semidefinite is the closed boundary ($\lambda \geq 0$, energy short by at least one square).

Example (A 3×3 positive definite matrix).

The second-difference matrix

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

is positive definite, and all four tests agree. *Determinants:* the leading minors are

$$\det[2] = 2, \qquad \det \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} = 3, \qquad \det A = 4,$$

all positive. *Pivots:* since each pivot is the ratio of consecutive leading minors (as noted above), the pivots are $2,\ \frac{3}{2},\ \frac{4}{3}$ — all positive, and their product $2 \cdot \frac{3}{2} \cdot \frac{4}{3} = 4$ matches $\det A$. *Eigenvalues:* they work out to $2 - \sqrt{2},\ 2,\ 2 + \sqrt{2}$, all positive, with product 4. *Energy:*

$$x^{\mathsf{T}} A x = 2x_1^2 + 2x_2^2 + 2x_3^2 - 2x_1x_2 - 2x_2x_3,$$

which a sum-of-squares rearrangement (using the pivots $2, \frac{3}{2}, \frac{4}{3}$) shows is positive except at $x = 0$. Four windows onto one fact.

8.5 The Geometry: Minima, Hessians, and the Energy Bowl

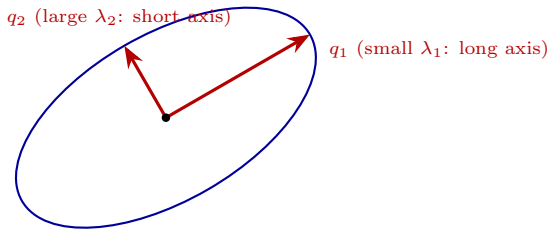
We promised that positive definiteness is the linear-algebra heart of the second-derivative test from calculus. Here is the connection, made precise.

8.5.1 Why $x^{\mathsf{T}}Ax$ is a bowl

The graph $z = x^{\mathsf{T}}Ax$ of a positive definite form is a bowl opening upward, with its single lowest point at the origin. Cutting it horizontally at height $z = k > 0$ gives the curve $x^{\mathsf{T}}Ax = k$, which is an *ellipse* (in $\mathbb{R}^2$) or an *ellipsoid* (in $\mathbb{R}^n$) — a closed, bounded curve precisely because the form is positive in every direction. Run the principal axis theorem on it: writing $A = Q\Lambda Q^{\mathsf{T}}$ and $y = Q^{\mathsf{T}}x$, the level set becomes

$$\lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2 = k.$$

This is an ellipsoid in standard position, axes along the coordinate directions y. Translating back, **the eigenvectors of A point along the principal axes of the ellipsoid, and the eigenvalues set their lengths** — a large λ_k means a steep, short axis (you reach height k quickly), a small λ_k a long, shallow one. This is exactly why engineers call the spectral theorem the principal axis theorem.



level curve $x^{\mathsf{T}}Ax = k$

If instead A were indefinite, the graph would be a *saddle*: rising in the directions of positive eigenvalues, falling in the directions of negative ones, with the origin a critical point that is neither a maximum nor a minimum. The level sets become hyperbolas, which run off to infinity — the signature of a saddle rather than a bowl.

8.5.2 The Hessian and the second-derivative test

In single-variable calculus, a critical point $f'(x) = 0$ is a minimum when $f''(x) > 0$ — the curve bends upward. For a function $f(x_1, \ldots, x_n)$ of several variables, the role of f'' is played by the matrix of second partial derivatives, the *Hessian*:

$$H = \begin{bmatrix} f_{x_1x_1} & f_{x_1x_2} & \cdots & f_{x_1x_n} \\ f_{x_2x_1} & f_{x_2x_2} & \cdots & f_{x_2x_n} \\ \vdots & & \ddots & \vdots \\ f_{x_nx_1} & \cdots & & f_{x_nx_n} \end{bmatrix}.$$

The Hessian is *symmetric*, because mixed partials are equal ($f_{x_ix_j} = f_{x_jx_i}$) for any smooth f — which is exactly why symmetric matrices are the right object for the minimum problem. Near a critical point (where all first derivatives vanish), Taylor's theorem says f looks like its quadratic part,

$$f(x) \approx f(x_0) + \tfrac{1}{2}\,(x - x_0)^{\mathsf{T}} H\,(x - x_0),$$

and the energy $(x - x_0)^{\mathsf{T}} H (x - x_0)$ decides the shape.

The second-derivative test in n variables

At a critical point of f (all first partials zero), let H be the Hessian.

- H positive definite ⇒ the energy is a bowl ⇒ a **local minimum**.
- H negative definite ⇒ a **local maximum**.
- H indefinite (mixed-sign eigenvalues) ⇒ a **saddle point**.

The familiar 2×2 rule from calculus — minimum when $f_{xx} > 0$ and $f_{xx}f_{yy} - f_{xy}^2 > 0$ — is exactly the determinant test (2) applied to H: the first leading minor $f_{xx} > 0$ and the second leading minor $\det H > 0$.

So the calculus test you memorized ($f_{xx} > 0$ and $f_{xx}f_{yy} > f_{xy}^2$) was the determinant test for positive definiteness of the Hessian all along. The two pieces of that rule are precisely the two leading principal minors of H being positive.

Example (Bowl versus saddle).

The form $f(x, y) = 2x^2 + 12xy + 20y^2$ has Hessian

$$H = \begin{bmatrix} 4 & 12 \\ 12 & 40 \end{bmatrix} = 2 \begin{bmatrix} 2 & 6 \\ 6 & 20 \end{bmatrix},$$

positive definite (leading minors $4 > 0$ and $4 \cdot 40 - 144 = 16 > 0$), so the origin is a minimum — a bowl. Change the corner: $g(x, y) = 2x^2 + 12xy + 7y^2$ has $\det\left[\begin{smallmatrix} 4 & 12 \\ 12 & 14 \end{smallmatrix}\right] = 56 - 144 < 0$, so its Hessian is indefinite and the origin is a saddle. The two surfaces differ only in one coefficient, 20 versus 7, and that single number is the difference between a true minimum and no minimum at all.

8.6 The Matrices $A^{\mathsf{T}}A$: Positive Semidefinite for Free

Positive definite matrices might seem like a special breed you would have to go looking for. In fact they manufacture themselves out of *any* matrix, through the symmetric product $A^{\mathsf{T}}A$ that ran the whole least-squares story of Chapter 5.

Start with any $A \in \mathbb{R}^{m \times n}$, rectangular or square. The product $A^{\mathsf{T}}A$ is $n \times n$ and *symmetric*, since $(A^{\mathsf{T}}A)^{\mathsf{T}} = A^{\mathsf{T}}(A^{\mathsf{T}})^{\mathsf{T}} = A^{\mathsf{T}}A$. More is true: its energy is never negative.

Theorem 8.8: $A^\mathsf{T}A$ Is Positive Semidefinite

For any real matrix A, the matrix $A^\mathsf{T}A$ is symmetric positive semidefinite. It is positive *definite* if and only if the columns of A are linearly independent (i.e. A has full column rank, $N(A) = \{0\}$).

Proof. Symmetry was checked above. For the energy, compute for any x:

$$x^\mathsf{T}(A^\mathsf{T}A)x = (Ax)^\mathsf{T}(Ax) = \|Ax\|^2 \geq 0.$$

A squared length is never negative, so $A^\mathsf{T}A$ is positive semidefinite. Equality $\|Ax\|^2 = 0$ holds exactly when $Ax = 0$, i.e. when $x \in N(A)$. If the columns of A are independent then $N(A) = \{0\}$, so $x^\mathsf{T}(A^\mathsf{T}A)x > 0$ for every $x \neq 0$ and $A^\mathsf{T}A$ is positive definite. If the columns are dependent, some $x \neq 0$ has $Ax = 0$, giving zero energy at a nonzero vector — only semidefinite. □

This is why least squares works. The normal equations $A^\mathsf{T}A\hat{x} = A^\mathsf{T}b$ have a unique solution precisely when $A^\mathsf{T}A$ is *invertible*, and we now see that happens exactly when A has independent columns — the same full-column-rank condition that made the projection well defined. Positive definite is the strong form of invertible for symmetric matrices: not just $\det \neq 0$, but every eigenvalue strictly on the positive side.

Remark (Where positive definite matrices come from).

The reading $x^\mathsf{T}(A^\mathsf{T}A)x = \|Ax\|^2$ explains the ubiquity of these matrices. A *covariance matrix* in statistics is $\frac{1}{m}A^\mathsf{T}A$ for a data matrix A — positive semidefinite because variances cannot be negative. A *stiffness* or *Gram matrix* in mechanics and geometry has entries $a_i^\mathsf{T}a_j$, i.e. it is $A^\mathsf{T}A$ for $A = [a_1 \ \cdots \ a_n]$ — positive definite exactly when the a_i are independent. And the next chapter (Chapter 9) builds the singular value decomposition directly on the eigenvalues of $A^\mathsf{T}A$, which are ≥ 0 for exactly the reason above; their square roots are the singular values of A.

8.7 A Glance at Complex Matrices: Hermitian, Unitary, and Fourier

Real symmetric matrices have a complex twin, and meeting it briefly explains why the recipe "$A = A^\mathsf{T}$" has to be adjusted when entries become complex — and it leads to the single most important complex matrix in all of computing.

8.7.1 Length and inner product for complex vectors

For a complex vector $z \in \mathbb{C}^n$ the old formula $z^\mathsf{T}z$ is wrong: it can vanish for nonzero z. For instance $z = (1, i)$ gives $z^\mathsf{T}z = 1 + i^2 = 0$, yet $z \neq 0$. The fix is to conjugate one factor. Write $z^H = \bar{z}^\mathsf{T}$ (conjugate-transpose, the *Hermitian transpose*, named for Hermite). Then

$$z^H z = |z_1|^2 + |z_2|^2 + \cdots + |z_n|^2 \geq 0$$

is a genuine squared length, and the right inner product of two complex vectors is $y^H x$ rather than $y^\mathsf{T} x$. This is the same $\bar{x}^\mathsf{T} x$ that appeared in the proof that symmetric matrices have real eigenvalues — there it quietly guaranteed $\bar{x}^\mathsf{T} x > 0$.

8.7.2 Hermitian and unitary matrices

With H replacing T, the two starring families of the real theory carry straight over.

Definition 8.9: Hermitian and Unitary Matrices

A complex matrix A is **Hermitian** if $A^H = A$ (its conjugate-transpose equals itself); this is the complex version of symmetric. A square complex matrix U is **unitary** if $U^H U = I$ (its columns are orthonormal under the Hermitian inner product); this is the complex version of orthogonal.

A Hermitian matrix must have real diagonal entries (each satisfies $\bar{a} = a$), and its off-diagonal pairs are conjugates, as in

$$A = \begin{bmatrix} 2 & 3+i \\ 3-i & 5 \end{bmatrix}, \qquad A^H = A.$$

Everything good about real symmetric matrices holds for Hermitian ones, by the same proofs with H in place of T: **Hermitian matrices have real eigenvalues and orthonormal eigenvectors**, and they diagonalize as $A = U\Lambda U^H$ with U unitary — the spectral theorem in complex dress. Unitary matrices are the complex isometries: like orthogonal matrices, they preserve length ($\|Ux\| = \|x\|$) and so have eigenvalues on the unit circle.

8.7.3 The Fourier matrix and the FFT

The most important unitary matrix in applications is the *Fourier matrix* F_n, the engine of the discrete Fourier transform. Indexing rows and columns from 0 to $n-1$ (the engineering convention) and setting $w = e^{2\pi i/n}$ — a primitive n-th root of unity, $w^n = 1$ — its entries are

$$(F_n)_{jk} = w^{jk}, \qquad F_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{bmatrix} \quad (w = i).$$

The columns of F_n are orthogonal under the Hermitian inner product (each has length $\sqrt{n}$), so $\frac{1}{\sqrt{n}} F_n$ is unitary and the transform inverts simply: the inverse is $F_n^{-1} = \frac{1}{n} F_n^H = \frac{1}{n}\overline{F_n}$, so multiplying by the conjugate $\overline{F_n}$ and dividing by n undoes it. Thus the discrete Fourier transform is, at heart, a change of basis into an orthonormal basis of complex exponentials — exactly the spectral viewpoint of this chapter, now over $\mathbb{C}$.

The reason F_n runs the digital world is the *fast Fourier transform*. A naive multiply by F_n costs about n^2 operations. But F_{2n} factors into sparse pieces built from F_n,

$$F_{2n} = \begin{bmatrix} I & D \\ I & -D \end{bmatrix} \begin{bmatrix} F_n & 0 \\ 0 & F_n \end{bmatrix} P,$$

where D is diagonal (powers of w) and P is the permutation that sorts even-indexed entries before odd. So a size-$2n$ transform reduces to two size-n transforms plus $O(n)$ cheap work. Recursing all the way down replaces n^2 by about $\frac{1}{2}n\log_2 n$ operations. For $n = 1024$ that is the difference between a million multiplications and roughly five thousand — a couple hundredfold speedup, and the reason Fourier methods are everywhere from audio to image compression. The whole acceleration rides on the special orthogonal structure of F_n — the same structure, orthonormal eigen-directions, that made symmetric matrices so good in the first place.

8.8 What to Carry Forward

Symmetry is the property that makes the eigenvalue story end happily. A real symmetric matrix $A = A^{\mathsf{T}}$ has real eigenvalues and a full set of orthonormal eigenvectors, so it factors as $A = Q\Lambda Q^{\mathsf{T}}$ with Q orthogonal — the spectral theorem. Geometrically that means a symmetric matrix only stretches space along n perpendicular axes, the eigenvectors, by the factors λ_k; and it decomposes into a weighted sum of perpendicular projections $\sum \lambda_k\, q_k q_k^{\mathsf{T}}$.

The signs of those eigenvalues are read off cheaply from the pivots (Sylvester's law of inertia), and when they are all positive we have a *positive definite* matrix. The four tests — positive eigenvalues, positive leading minors, positive pivots, positive energy $x^{\mathsf{T}}Ax$ — are one condition seen four ways, linked by the chain "minors → pivots → sum of squares → energy → eigenvalues." The energy test is the one that matters most: it makes $x^{\mathsf{T}}Ax$ a bowl, so a critical point of a multivariable function is a true minimum exactly when its symmetric Hessian is positive definite. The calculus rule $f_{xx} > 0,\ f_{xx}f_{yy} > f_{xy}^2$ is just the determinant test in disguise.

Remark (Looking ahead).

> Positive (semi)definiteness is not a curiosity but the normal state of the matrices that arise in practice: every $A^{\mathsf{T}}A$ is positive semidefinite, and positive definite exactly when A has independent columns — which is why least squares and covariance and Gram matrices are all built from it. The complex cousins, Hermitian and unitary matrices, repeat the entire theory over $\mathbb{C}$ and hand us the Fourier matrix and the FFT. And the eigenvalues of $A^{\mathsf{T}}A$, nonnegative for the reason in this chapter, become the squared *singular values* of A — the foundation of the singular value decomposition $A = U\Sigma V^{\mathsf{T}}$, the climax of the next chapter (Chapter 9), which extends $A = Q\Lambda Q^{\mathsf{T}}$ from symmetric matrices to *every* matrix, rectangular ones included.

Chapter 9

Similar Matrices, the Jordan Form, and the SVD

We have reached the summit of the course. The eigenvalue chapters taught us to read a square matrix by what it does to its own special directions: a diagonalizable A factors as $A = S\Lambda S^{-1}$, and a symmetric A does it even better, $A = Q\Lambda Q^{\mathsf{T}}$, with perpendicular axes. But two clouds still hang over that story. First, $A = S\Lambda S^{-1}$ secretly says that A and Λ are the "same" matrix written in two coordinate systems — and we never made that idea precise. Second, some matrices have too few eigenvectors to diagonalize at all, and rectangular matrices have no eigenvalues to speak of. This chapter clears both clouds.

We start by naming the relation hiding inside $A = S\Lambda S^{-1}$: *similarity*, $B = M^{-1}AM$. Similar matrices are the same linear map seen in different bases, so they share everything basis-free — above all, their eigenvalues. Diagonalizing is the search for the simplest matrix similar to A. When no diagonal matrix is similar to A, Camille Jordan's theorem tells us the next-best representative, almost diagonal, with a few 1's above the diagonal recording the missing eigenvectors. Then comes the climax. The *singular value decomposition* $A = U\Sigma V^{\mathsf{T}}$ works for *every* matrix, square or rectangular, diagonalizable or not. It takes the eigenvalue idea — orthonormal axes and pure stretching — and makes it succeed universally, by allowing two different orthonormal bases instead of one. The SVD hands us orthonormal bases for all four fundamental subspaces at once, and it says that every matrix, no matter how complicated, does nothing more than *rotate, stretch, and rotate.*

9.1 Similar Matrices

The factorization $A = S\Lambda S^{-1}$ from Chapter 7 has a meaning we have been circling without naming. It says A and the diagonal matrix Λ are linked by a particular kind of sandwich: a matrix on the left, its inverse on the right, the same matrix throughout. That sandwich is worth a definition of its own, because it organizes all square matrices into families.

Definition 9.1: Similar Matrices

Two square matrices A and B of the same size are **similar** if there is an invertible matrix M with

$$B = M^{-1}AM.$$

We write $A \sim B$. The passage from A to $M^{-1}AM$ is a *similarity transformation.*

The first thing to know is that similarity sorts every $n \times n$ matrix into families. It is an *equivalence relation*: $A \sim A$ (take $M = I$); if $A \sim B$ then $B \sim A$ (replace M by M^{-1}, since $A = MBM^{-1} = (M^{-1})^{-1}B(M^{-1})$); and if $A \sim B$ and $B \sim C$ then $A \sim C$ (compose the two M's). So the $n \times n$ matrices break into disjoint *similarity families*, and inside each family we will hunt for the simplest representative — diagonal if we are lucky, "nearly diagonal" (Jordan form) if we are not.

9.1.1 Why similarity is a change of basis

The definition looks algebraic, but its content is geometric. Recall from the eigenvalue chapter that $A = S\Lambda S^{-1}$ means: to apply A, first write the input in the eigenvector basis (multiply by S^{-1}), then stretch each coordinate (multiply by Λ), then translate back to standard coordinates (multiply by S). The same reading works for any M. If M holds a new basis in its columns, then $M^{-1}AM$ is *the very same linear transformation A, written in the basis M*. Similar matrices are one map seen through different coordinate systems.

Similarity is sameness of map, difference of basis

If $M = \begin{bmatrix} m_1 & \cdots & m_n \end{bmatrix}$ holds a basis of $\mathbb{R}^n$ in its columns, then $B = M^{-1}AM$ is the matrix of the *same* linear transformation as A, but with inputs and outputs expressed in the basis $m_1, \ldots, m_n$ rather than the standard basis. Everything that does not depend on the choice of basis must agree for A and B.

Chapter 10 develops change of basis fully. For now, the slogan "same map, new coordinates" is exactly the tool we need to see why similar matrices share their deepest features.

9.1.2 Similar matrices have the same eigenvalues

The headline of the section follows in two lines. If A and B are similar, every eigenvalue of A is an eigenvalue of B, with the eigenvector merely relabeled.

Theorem 9.2: Similar Matrices Share Their Eigenvalues

If $B = M^{-1}AM$, then A and B have the same characteristic polynomial, hence the same eigenvalues with the same algebraic multiplicities. Moreover, if $Ax = \lambda x$ then $M^{-1}x$ is an eigenvector of B for the same λ.

Proof. Take an eigenpair of A: $Ax = \lambda x$ with $x \neq 0$. Insert $MM^{-1} = I$ and multiply on

the left by M^{-1}:

$$A\,(MM^{-1})\,x = \lambda x \quad\Longrightarrow\quad M^{-1}AM\,(M^{-1}x) = \lambda\,(M^{-1}x), \quad \text{i.e.} \quad B\,(M^{-1}x) = \lambda\,(M^{-1}x).$$

Since M^{-1} is invertible and $x \neq 0$, the vector $M^{-1}x$ is nonzero, so it is an eigenvector of B with eigenvalue λ. For the stronger statement about the whole polynomial, use the multiplicativity of determinants:

$$\begin{aligned}\det(B - \lambda I) &= \det(M^{-1}AM - \lambda M^{-1}M) = \det\big(M^{-1}(A - \lambda I)M\big) \\ &= \det(M^{-1})\det(A - \lambda I)\det(M).\end{aligned}$$

Because $\det(M^{-1})\det(M) = 1$, the two characteristic polynomials are equal, so A and B have identical eigenvalues and multiplicities. □

The same computation shows that the *number* of independent eigenvectors is also preserved: M^{-1} carries each eigenspace of A one-to-one onto an eigenspace of B of the same dimension. So similar matrices agree on eigenvalues, their multiplicities, and the count of eigenvectors. They generally do *not* have the same eigenvectors — those rotate with the basis — and they need not have the same entries at all.

Remark (Two free invariants: trace and determinant).

Two consequences of equal eigenvalues are worth keeping at your fingertips, because they are cheap to compute and catch most errors. Similar matrices have the same *determinant* (the product of the eigenvalues) and the same *trace* (their sum). The trace identity also follows directly from $\operatorname{trace}(XY) = \operatorname{trace}(YX)$: $\operatorname{trace}(M^{-1}AM) = \operatorname{trace}(AMM^{-1}) = \operatorname{trace} A$. If two matrices disagree in trace or determinant, they cannot be similar — no need to look further.

Example (A whole family with eigenvalues 3 and 1).

Start from the symmetric matrix

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \qquad \operatorname{trace} A = 4, \quad \det A = 3,$$

whose eigenvalues are 3 and 1. Diagonalizing makes A similar to $\Lambda = \left[\begin{smallmatrix} 3 & 0 \\ 0 & 1 \end{smallmatrix}\right]$. But A is similar to many non-diagonal matrices too. Pick the simple change of basis

$$M = \begin{bmatrix} 1 & 4 \\ 0 & 1 \end{bmatrix}, \qquad M^{-1} = \begin{bmatrix} 1 & -4 \\ 0 & 1 \end{bmatrix},$$

and compute

$$B = M^{-1}AM = \begin{bmatrix} 1 & -4 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} 1 & 4 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} -2 & -15 \\ 1 & 6 \end{bmatrix}.$$

This B looks nothing like A, yet $\operatorname{trace} B = 4$ and $\det B = 3$ — the same eigenvalues 3 and 1. In fact the matrices similar to A are *all* the 2×2 matrices with eigenvalues 3

and 1; other members of the family include $\left[\begin{smallmatrix} 3 & 7 \\ 0 & 1 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 1 & 7 \\ 0 & 3 \end{smallmatrix}\right]$. Diagonalizing is choosing the cleanest representative, Λ, out of this infinite family.

9.2 Repeated Eigenvalues and the Jordan Form

When the eigenvalues are distinct, similarity to a diagonal matrix is automatic — the eigenvectors are independent (Chapter 7), so S is invertible and $S^{-1}AS = \Lambda$. The trouble starts when an eigenvalue repeats and the matrix runs short of eigenvectors. Then no diagonal matrix is similar to A, and we need a substitute.

9.2.1 Two families with the same eigenvalues

The cleanest way to feel the problem is to fix the eigenvalues and watch the families split. Suppose both eigenvalues equal 4. Two very different matrices have characteristic polynomial $(\lambda - 4)^2$:

$$\begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix} 4 & 1 \\ 0 & 4 \end{bmatrix}.$$

They are *not* similar, even though their eigenvalues match. The first is special: the scalar matrix $4I$ commutes with everything, so $M^{-1}(4I)M = 4M^{-1}M = 4I$ for every invertible M. It is similar only to itself — a family of one. It has a full set of eigenvectors (every vector is an eigenvector), so it is diagonal already.

The second matrix is *defective*. Solving $(A - 4I)x = 0$ means $\left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right] x = 0$, which forces $x_2 = 0$: the eigenspace is only the line through $(1, 0)$, one-dimensional. With two eigenvalues but a single eigenvector, no eigenvector matrix S can be invertible, and A cannot be diagonalized. This matrix heads its own family, and every matrix in that family has exactly one eigenvector.

Remark (Diagonalizable is a property of the family).

Whether a matrix can be diagonalized is a property of its whole similarity family, since similar matrices have the same number of eigenvectors. The family of $4I$ is diagonalizable (it *is* the diagonal matrix). The family of $\left[\begin{smallmatrix} 4 & 1 \\ 0 & 4 \end{smallmatrix}\right]$ is not — no member has enough eigenvectors. Counting eigenvectors decides the matter once and for all.

9.2.2 Jordan blocks

Jordan found the "most diagonal" representative of each family. For a single repeated eigenvalue with a shortage of eigenvectors, that representative is a *Jordan block*: the eigenvalue down the diagonal, a row of 1's just above it, zeros elsewhere.

Definition 9.3: Jordan Block

A **Jordan block** of size k for the eigenvalue λ is the $k \times k$ matrix

$$J_k(\lambda) = \begin{bmatrix} \lambda & 1 & & \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix}$$

with λ on the diagonal, 1's on the first superdiagonal, and 0's everywhere else. A block of size k has the single eigenvalue λ (repeated k times) but only *one* eigenvector, e_1; the $k-1$ ones above the diagonal record its $k-1$ missing eigenvectors.

A diagonal matrix is the special case where every Jordan block has size 1. The general theorem assembles blocks down the diagonal of a big matrix.

Theorem 9.4: Jordan's Theorem

Every square matrix A is similar to a **Jordan matrix**

$$J = M^{-1}AM = \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_d \end{bmatrix},$$

a block-diagonal matrix whose diagonal blocks $J_1, \ldots, J_d$ are Jordan blocks. The Jordan matrix is unique up to the order of the blocks. The eigenvalues sit on the diagonal; the number of blocks equals the number of independent eigenvectors of A (one eigenvector per block); and the total number of 1's above the diagonal is $n-d$, the count of *missing* eigenvectors. Two matrices are similar if and only if they have the same Jordan form (same block sizes for the same eigenvalues).

We will not prove existence here — it is the deepest theorem of the chapter, and computing J for a defective matrix is beyond our running examples. What matters is the picture: every matrix is *almost* diagonalizable, and the only obstruction is a handful of 1's above the diagonal, exactly as many as the eigenvectors it lacks.

Remark (The two clean extremes).

> The two ends of Jordan's theorem are the cases we already understand. If A has n distinct eigenvalues it has n eigenvectors, so $d = n$, every block has size 1, and $J = \Lambda$ is diagonal — ordinary diagonalization. If A is symmetric, the spectral theorem (Chapter 8) again gives a full set of eigenvectors, so $J = \Lambda$ with no 1's above the diagonal. The Jordan form earns its keep only for *defective* matrices, the ones missing eigenvectors.

9.2.3 Same eigenvalues need not mean similar

Equal eigenvalues, even the same number of eigenvectors, are not enough for similarity — the block *sizes* must match too. The standard example uses four-fold zero eigenvalues.

Example (Two nilpotent matrices that are not similar).

Consider

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad C = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Both have all four eigenvalues equal to 0 (they are upper triangular with zero diagonal). Both have rank 2, so both have null space of dimension $4 - 2 = 2$: *each has exactly two independent eigenvectors.* On the surface they look identical — same eigenvalues, same number of eigenvectors. Yet they are *not* similar, and the Jordan form shows why. Reading off the blocks,

$$A = \begin{bmatrix} J_3(0) & \\ & J_1(0) \end{bmatrix} \quad \text{(a } 3 \times 3 \text{ block and a } 1 \times 1 \text{ block)},$$

$$C = \begin{bmatrix} J_2(0) & \\ & J_2(0) \end{bmatrix} \quad \text{(two } 2 \times 2 \text{ blocks)}.$$

The block sizes differ, so $A \not\sim C$. A basis-free quantity catches the difference at once: square them. A short computation gives $\operatorname{rank}(A^2) = 1$, while $\operatorname{rank}(C^2) = 0$ (indeed $C^2 = 0$). Now $B = M^{-1}AM$ forces $B^2 = M^{-1}A^2M$, so similar matrices have squares of equal rank. Here the ranks disagree, so A and C cannot be similar. The Jordan form is precisely the bookkeeping that distinguishes families like these.

9.3 The Singular Value Decomposition

Now the climax. Diagonalization $A = S\Lambda S^{-1}$ is wonderful but fragile: it needs A square, it needs enough eigenvectors, and even then S is usually not orthogonal. The spectral theorem $A = Q\Lambda Q^{\mathsf{T}}$ fixes the orthogonality but demands $A = A^{\mathsf{T}}$. The singular value decomposition keeps everything good and drops every restriction.

The goal of the SVD

We want a factorization that works for *every* $A \in \mathbb{R}^{m \times n}$ — rectangular, defective, anything — and that uses *orthonormal* bases throughout. The price is one honest concession: a general matrix maps inputs in $\mathbb{R}^n$ to outputs in $\mathbb{R}^m$, so we will need *two* orthonormal bases, one V for the input space $\mathbb{R}^n$ and one U for the output space $\mathbb{R}^m$, rather than a single eigenvector basis.

9.3.1 The geometric idea: one orthonormal basis into another

Think of A as the linear map that carries the row space of A onto the column space of A (Chapter 4: it sends $\mathbb{R}^n$ to $\mathbb{R}^m$, crushing the null space and mapping the row space one-to-one onto the column space). An orthonormal basis of the row space is easy to get — Gram–Schmidt (Chapter 5) produces one immediately. But there is no reason a generic

matrix should send that orthonormal basis to another *orthogonal* set in the column space; it will usually skew the right angles.

The SVD is the statement that, with the *right* choice of orthonormal basis $v_1, \dots, v_r$ for the row space, the images come out orthogonal after all:

$$Av_i = \sigma_i \, u_i, \qquad i = 1, \dots, r,$$

where the u_i are orthonormal in the column space and the $\sigma_i > 0$ are the stretching factors. Each input axis v_i goes to an output axis u_i, stretched by σ_i and otherwise rigid. The null spaces cause no trouble: the leftover v's span $N(A)$ and get sent to 0, recorded as zeros in Σ.

Definition 9.5: Singular Values and Singular Vectors

For $A \in \mathbb{R}^{m \times n}$ of rank r, the **singular values** $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ are the positive square roots of the nonzero eigenvalues of $A^\mathsf{T}A$ (equivalently of AA^T). The orthonormal vectors $v_1, \dots, v_n$ (eigenvectors of $A^\mathsf{T}A$) are the **right singular vectors**, and the orthonormal vectors $u_1, \dots, u_m$ (eigenvectors of AA^T) are the **left singular vectors**. They satisfy $Av_i = \sigma_i u_i$ for $i \leq r$ and $Av_i = 0$ for $i > r$.

9.3.2 The theorem

Collecting the v's as columns of V and the u's as columns of U, the relations $Av_i = \sigma_i u_i$ assemble columnwise into $AV = U\Sigma$. Since V is orthogonal, $V^{-1} = V^\mathsf{T}$, and multiplying on the right gives the factorization.

Theorem 9.6: Singular Value Decomposition

Every matrix $A \in \mathbb{R}^{m \times n}$ factors as

$$A = U \, \Sigma \, V^\mathsf{T},$$

where U is an $m \times m$ orthogonal matrix ($U^\mathsf{T}U = I_m$), V is an $n \times n$ orthogonal matrix ($V^\mathsf{T}V = I_n$), and Σ is an $m \times n$ "diagonal" matrix with the singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$ on its diagonal and zeros elsewhere. The columns of V are orthonormal eigenvectors of $A^\mathsf{T}A$; the columns of U are orthonormal eigenvectors of AA^T; and the σ_i^2 are the corresponding (nonnegative) eigenvalues.

Proof. The construction *is* the proof, and it rests entirely on the spectral theorem applied to the symmetric matrix $A^\mathsf{T}A$. Suppose the factorization holds, $A = U\Sigma V^\mathsf{T}$. Multiply by $A^\mathsf{T} = V\Sigma^\mathsf{T}U^\mathsf{T}$ on the left:

$$A^\mathsf{T}A = V\Sigma^\mathsf{T}U^\mathsf{T} \, U\Sigma V^\mathsf{T} = V \, (\Sigma^\mathsf{T}\Sigma) \, V^\mathsf{T},$$

because $U^\mathsf{T}U = I$. Now $\Sigma^\mathsf{T}\Sigma$ is the $n \times n$ diagonal matrix $\mathrm{diag}(\sigma_1^2, \dots, \sigma_r^2, 0, \dots, 0)$, so this is exactly the spectral decomposition $A^\mathsf{T}A = V\Lambda V^\mathsf{T}$ of the symmetric matrix $A^\mathsf{T}A$. That tells us how to *build* V and Σ in the first place: $A^\mathsf{T}A$ is symmetric, and by Chapter 8 it

is positive semidefinite (its energy is $x^{\mathsf{T}}A^{\mathsf{T}}Ax = \|Ax\|^2 \geq 0$), so its eigenvalues are real and nonnegative. Diagonalize it, $A^{\mathsf{T}}A = V\Lambda V^{\mathsf{T}}$ with V orthogonal; the eigenvalues $\lambda_i \geq 0$ define the singular values $\sigma_i = \sqrt{\lambda_i}$.

It remains to produce U and verify $Av_i = \sigma_i u_i$. For each i with $\sigma_i > 0$, set

$$u_i = \frac{1}{\sigma_i} Av_i.$$

These u_i are orthonormal: using $A^{\mathsf{T}}Av_j = \sigma_j^2 v_j$,

$$u_i^{\mathsf{T}} u_j = \frac{1}{\sigma_i\sigma_j} (Av_i)^{\mathsf{T}}(Av_j) = \frac{1}{\sigma_i\sigma_j} v_i^{\mathsf{T}}(A^{\mathsf{T}}Av_j) = \frac{\sigma_j^2}{\sigma_i\sigma_j} v_i^{\mathsf{T}} v_j = \frac{\sigma_j}{\sigma_i}\delta_{ij} = \delta_{ij},$$

since the v's are orthonormal. So $u_1, \ldots, u_r$ are an orthonormal set in $\mathbb{R}^m$, and by construction $Av_i = \sigma_i u_i$. Extend $u_1, \ldots, u_r$ to an orthonormal basis $u_1, \ldots, u_m$ of $\mathbb{R}^m$ (Gram–Schmidt), and similarly the v_i for $i > r$ are eigenvectors of $A^{\mathsf{T}}A$ with eigenvalue 0, hence lie in $N(A^{\mathsf{T}}A) = N(A)$ and satisfy $Av_i = 0$. All the relations $Av_i = \sigma_i u_i$ (with $\sigma_i = 0$ for $i > r$) assemble into $AV = U\Sigma$, and right-multiplying by $V^{\mathsf{T}} = V^{-1}$ gives $A = U\Sigma V^{\mathsf{T}}$. □

Remark (Where each piece lives, and why U falls out of AA^{T}).

> The same algebra run on the *other* side explains U directly. From $A = U\Sigma V^{\mathsf{T}}$,
>
> $$AA^{\mathsf{T}} = U\Sigma V^{\mathsf{T}}\, V\Sigma^{\mathsf{T}}U^{\mathsf{T}} = U\,(\Sigma\Sigma^{\mathsf{T}})\,U^{\mathsf{T}},$$
>
> so the columns of U are orthonormal eigenvectors of the $m \times m$ symmetric matrix AA^{T}, with the *same* nonzero eigenvalues σ_i^2. (The nonzero eigenvalues of $A^{\mathsf{T}}A$ and AA^{T} always agree — only the number of extra zeros differs, accounting for the different sizes n and m.) In practice you can find V from $A^{\mathsf{T}}A$ and then get U for free from $u_i = Av_i/\sigma_i$, without diagonalizing AA^{T} separately.

9.4 Computing the SVD

The proof is a recipe. To factor $A = U\Sigma V^{\mathsf{T}}$:

1. Form the symmetric matrix $A^{\mathsf{T}}A$ and diagonalize it: its orthonormal eigenvectors are the columns of V, its eigenvalues are σ_i^2.
2. Take $\sigma_i = \sqrt{\lambda_i}$ (positive square roots), ordered $\sigma_1 \geq \sigma_2 \geq \cdots$, and place them on the diagonal of Σ.
3. Recover each left singular vector from $u_i = Av_i/\sigma_i$ (for $\sigma_i > 0$); fill out U with an orthonormal basis of the left null space if needed.

Two worked examples make the mechanics concrete — first an invertible matrix, then one with a null space.

Example (An invertible 2×2 matrix).

Take

$$A = \begin{bmatrix} 4 & 4 \\ -3 & 3 \end{bmatrix}.$$

Step 1: diagonalize $A^{\mathsf{T}}A$.

$$A^{\mathsf{T}}A = \begin{bmatrix} 4 & -3 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 4 & 4 \\ -3 & 3 \end{bmatrix} = \begin{bmatrix} 25 & 7 \\ 7 & 25 \end{bmatrix}.$$

This symmetric matrix has eigenvalues 25 ± 7, namely $\lambda_1 = 32$ and $\lambda_2 = 18$, with eigenvectors $(1, 1)$ and $(1, -1)$ (orthogonal, as the spectral theorem promises). Normalizing,

$$v_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad v_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Step 2: singular values.

$$\sigma_1 = \sqrt{32} = 4\sqrt{2}, \qquad \sigma_2 = \sqrt{18} = 3\sqrt{2}.$$

Step 3: left singular vectors.

$$Av_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 4+4 \\ -3+3 \end{bmatrix} = \begin{bmatrix} 4\sqrt{2} \\ 0 \end{bmatrix} = \sigma_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

$$Av_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 4-4 \\ -3-3 \end{bmatrix} = \begin{bmatrix} 0 \\ -3\sqrt{2} \end{bmatrix} = \sigma_2 \begin{bmatrix} 0 \\ -1 \end{bmatrix},$$

so $u_1 = (1, 0)$ and $u_2 = (0, -1)$. (Watch the sign: Av_2 points in the $-y$ direction, so $u_2 = (0, -1)$, not $(0, 1)$ — letting $u_i = Av_i/\sigma_i$ fix the signs is the safe habit.) Assembling,

$$\begin{aligned} A &= U\Sigma V^{\mathsf{T}} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 4\sqrt{2} & 0 \\ 0 & 3\sqrt{2} \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}. \end{aligned}$$

Multiplying the three factors back out returns A exactly. The whole geometry is in the diagonal Σ: A stretches one perpendicular direction by $4\sqrt{2}$ and the other by $3\sqrt{2}$, with rigid rotations on either side.

Example (A matrix with a null space).

Now a rank-one example,

$$A = \begin{bmatrix} 4 & 3 \\ 8 & 6 \end{bmatrix},$$

whose second row is twice the first. It has a one-dimensional row space (multiples of $(4, 3)$), a one-dimensional column space (multiples of $(4, 8)$, i.e. $(1, 2)$), and one-

dimensional null and left null spaces.

$$A^\mathsf{T}A = \begin{bmatrix} 4 & 8 \\ 3 & 6 \end{bmatrix}\begin{bmatrix} 4 & 3 \\ 8 & 6 \end{bmatrix} = \begin{bmatrix} 80 & 60 \\ 60 & 45 \end{bmatrix}.$$

This matrix is rank one, so one eigenvalue is 0; the other equals the trace, $\lambda_1 = 80+45 = 125$. Thus $\sigma_1 = \sqrt{125} = 5\sqrt{5}$ and $\sigma_2 = 0$. The unit eigenvector for λ_1 lies along the row space, $v_1 = \frac{1}{5}(4,3) = (0.8, 0.6)$, and $v_2 = (0.6, -0.8)$ spans the null space. On the output side the column space direction is $u_1 = \frac{1}{\sqrt{5}}(1,2)$, and $u_2 = \frac{1}{\sqrt{5}}(2,-1)$ spans the left null space. The SVD is

$$A = \underbrace{\frac{1}{\sqrt{5}}\begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix}}_{U}\underbrace{\begin{bmatrix} 5\sqrt{5} & 0 \\ 0 & 0 \end{bmatrix}}_{\Sigma}\underbrace{\begin{bmatrix} 0.8 & 0.6 \\ 0.6 & -0.8 \end{bmatrix}}_{V^\mathsf{T}}.$$

The single zero on the diagonal of Σ is doing exactly the job advertised: it absorbs the null space, sending v_2 to 0. Notice that the SVD has just handed us orthonormal bases for all four subspaces in one shot.

9.5 The SVD and the Four Fundamental Subspaces

The second example is no accident. The singular value decomposition is the natural climax of the four-subspaces story (Chapter 4), because it produces *orthonormal* bases for all four at once, perfectly matched so that A acts as pure stretching between them.

Orthonormal bases for the four subspaces

Write the SVD of a rank-r matrix $A \in \mathbb{R}^{m\times n}$ as $A = U\Sigma V^\mathsf{T}$ with $V = [v_1 \ \cdots \ v_n]$ and $U = [u_1 \ \cdots \ u_m]$. Then:

$$\begin{aligned} v_1, \ldots, v_r &\text{ is an orthonormal basis for the row space } C(A^\mathsf{T}) \subseteq \mathbb{R}^n, \\ v_{r+1}, \ldots, v_n &\text{ is an orthonormal basis for the null space } N(A) \subseteq \mathbb{R}^n, \\ u_1, \ldots, u_r &\text{ is an orthonormal basis for the column space } C(A) \subseteq \mathbb{R}^m, \\ u_{r+1}, \ldots, u_m &\text{ is an orthonormal basis for the left null space } N(A^\mathsf{T}) \subseteq \mathbb{R}^m, \end{aligned}$$

and A maps each row-space axis to a column-space axis by $Av_i = \sigma_i u_i$.

These are the "right" bases, the ones Chapter 4 could promise existed but not name. The first columns of V ($\sigma_i > 0$) are eigenvectors of $A^\mathsf{T}A$ with positive eigenvalue, so they lie in the row space; the last columns ($\sigma_i = 0$) lie in $N(A^\mathsf{T}A) = N(A)$. On the output side $u_i = Av_i/\sigma_i$ is a combination of columns of A, hence in $C(A)$, while the extra u's fill out $N(A^\mathsf{T})$. The orthogonality of the four subspaces in pairs — row space $\perp$ null space, column space $\perp$ left null space — is built into the orthogonality of V and U.

9.5.1 The sum of rank-one pieces

Just as the spectral theorem wrote a symmetric matrix as $\sum \lambda_k q_k q_k^\mathsf{T}$, the SVD expands any matrix as a sum of rank-one pieces, one per singular value.

Theorem 9.7: The SVD as a Sum of Rank-One Matrices

If $A = U\Sigma V^\mathsf{T}$ has rank r, then

$$A = \sigma_1\, u_1 v_1^\mathsf{T} + \sigma_2\, u_2 v_2^\mathsf{T} + \cdots + \sigma_r\, u_r v_r^\mathsf{T}.$$

Each term $u_i v_i^\mathsf{T}$ is a rank-one matrix (Chapter 4), and the singular values $\sigma_1 \geq \cdots \geq \sigma_r$ order the pieces from most to least important.

Proof. Expand the product $U\Sigma V^\mathsf{T}$ the same way we expanded $Q\Lambda Q^\mathsf{T}$ in Chapter 8: the diagonal Σ picks out one outer product per nonzero entry,

$$U\Sigma V^\mathsf{T} = \begin{bmatrix} u_1 & \cdots & u_m \end{bmatrix} \Sigma \begin{bmatrix} v_1^\mathsf{T} \\ \vdots \\ v_n^\mathsf{T} \end{bmatrix} = \sum_{i=1}^{r} \sigma_i\, u_i v_i^\mathsf{T},$$

since the only nonzero diagonal entries of Σ are $\sigma_1, \ldots, \sigma_r$. □

Remark (Best low-rank approximation: why the SVD runs data science).

Because the pieces are ordered by size, keeping only the largest few gives the best possible approximation of A by a low-rank matrix. Truncating after k terms,

$$A_k = \sigma_1 u_1 v_1^\mathsf{T} + \cdots + \sigma_k u_k v_k^\mathsf{T},$$

is the closest rank-k matrix to A (the Eckart–Young theorem, measured in either the Frobenius or the spectral norm), and the error is governed by the discarded singular values $\sigma_{k+1}, \ldots$. This single fact powers image compression, noise removal, latent-semantic indexing, recommender systems, and *principal component analysis*: the leading singular vectors are the directions of greatest variation in data. The SVD is the most useful matrix factorization in all of applied mathematics, and this is why.

9.6 The Geometry: Rotate, Stretch, Rotate

Stand back and read $A = U\Sigma V^\mathsf{T}$ as a sequence of three motions applied to a vector x, right to left. First V^T acts. Since V is orthogonal, V^T is a *rotation* (or rotation with a reflection): it preserves all lengths and angles, merely realigning the axes so that the right singular vectors v_i become the coordinate directions. Next Σ acts. It is diagonal, so it *stretches* along those coordinate axes, scaling the i-th direction by the singular value σ_i — and, if A is rectangular, dropping or appending zero coordinates to move from $\mathbb{R}^n$ to $\mathbb{R}^m$. Finally U acts: another orthogonal matrix, another *rotation*, aligning the stretched axes with the left singular vectors u_i in the output space.

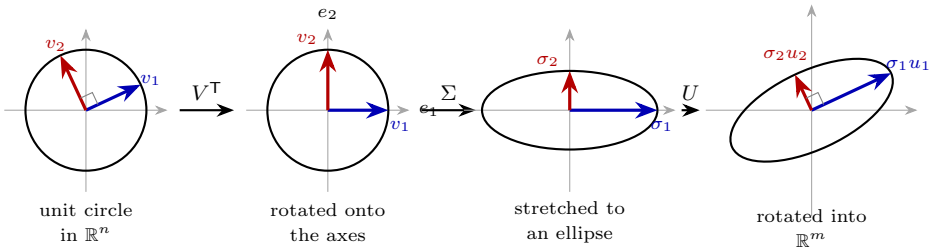
Every matrix is a rotation, a stretch, and a rotation

For any $A = U\Sigma V^{\mathsf{T}}$, applying A to a vector means

$$x \xrightarrow{V^{\mathsf{T}}} \text{(rotate)} \xrightarrow{\Sigma} \text{(stretch along axes)} \xrightarrow{U} \text{(rotate)} \to Ax.$$

No matter how tangled A looks, it does nothing more exotic than this. The rotations U, V contribute no stretching — they only spin space rigidly — and *all* the deformation lives in the singular values on the diagonal of Σ.

The cleanest way to see it is to watch what A does to the unit circle (or the unit sphere in higher dimensions). The right singular vectors v_1, v_2 are a pair of perpendicular radii. The matrix sends them to $\sigma_1 u_1$ and $\sigma_2 u_2$ — still perpendicular, because the u's are orthonormal, but now of lengths σ_1 and σ_2. So A carries the unit circle to an *ellipse* whose semi-axes point along u_1, u_2 with lengths exactly the singular values. The largest singular value σ_1 is how far A can stretch any unit vector; the smallest is how far it can shrink one. This is the picture to keep.



$Av_i = \sigma_i u_i$: perpendicular radii $\mapsto$ perpendicular semi-axes

Remark (How the three factorizations line up).

The course has built three diagonalizing factorizations, each more general than the last. For a matrix with n independent eigenvectors, $A = S\Lambda S^{-1}$ diagonalizes, but S need not be orthogonal and the picture can include shear. For a *symmetric* matrix the eigenvectors become orthonormal and $A = Q\Lambda Q^{\mathsf{T}}$ — pure stretching along perpendicular axes, a single basis Q doing double duty for input and output. The SVD $A = U\Sigma V^{\mathsf{T}}$ keeps the orthonormality but pays for full generality with two bases: it applies to *every* matrix, and when A happens to be symmetric positive definite it collapses back to $A = Q\Lambda Q^{\mathsf{T}}$ with $U = V = Q$ and $\sigma_i = \lambda_i$. The SVD is the spectral theorem set free from symmetry.

9.7 What to Carry Forward

This chapter tied off the eigenvalue story and then transcended it. *Similarity*, $B = M^{-1}AM$, names the relation that diagonalization quietly used: similar matrices are one linear map in two coordinate systems, so they share eigenvalues, multiplicities, eigenvector counts, trace, and determinant — everything basis-free. Diagonalizing is choosing the simplest representative of a similarity family. When a matrix is *defective* — short on eigenvectors because of repeated eigenvalues — no diagonal representative exists, and the *Jordan form* provides the

next best: block-diagonal, with a 1 above the diagonal for each missing eigenvector. Two matrices are similar exactly when their Jordan forms match, block size for block size.

The *singular value decomposition* $A = U\Sigma V^{\mathsf{T}}$ is the summit. It exists for every matrix, square or rectangular, by diagonalizing the symmetric positive semidefinite matrix $A^{\mathsf{T}}A$: its eigenvectors are the columns of V, the square roots of its eigenvalues are the singular values σ_i, and $u_i = Av_i/\sigma_i$ gives the columns of U. The SVD delivers orthonormal bases for all four fundamental subspaces simultaneously, perfectly matched by $Av_i = \sigma_i u_i$; it expands A into ordered rank-one pieces $\sum \sigma_i u_i v_i^{\mathsf{T}}$, whose truncation is the best low-rank approximation and the engine of principal component analysis and data compression; and geometrically it reveals that every matrix is a *rotation, a stretch, and a rotation*, turning the unit sphere into an ellipsoid with semi-axes the singular values.

Remark (Looking ahead).

> The thread running through these last chapters is a single question — "what is the simplest matrix similar (or otherwise equivalent) to A, and in what basis?" Chapter 10 answers it head-on: it develops *change of basis* and *linear transformations*, showing that every matrix in this book has been the coordinate expression of an underlying map, and that choosing coordinates wisely is what all our factorizations $A = LU$, $A = QR$, $A = S\Lambda S^{-1}$, $A = Q\Lambda Q^{\mathsf{T}}$, and $A = U\Sigma V^{\mathsf{T}}$ have really been doing. The SVD is the most democratic of these: it asks nothing of A and gives back orthonormal axes and clean stretching for free.

Chapter 10

Linear Transformations and Change of Basis

For nine chapters a matrix has been an array of numbers and Ax has been a combination of its columns. That is the right working picture, but it hides where matrices come from. A matrix is the *shadow* of something more basic and coordinate-free: a *linear transformation*, a rule T that takes vectors to vectors and respects addition and scaling. The transformation knows nothing about rows, columns, or numbers. It just says "project onto this line" or "rotate by 45°" or "take the derivative." The matrix appears only after we commit to a *basis*, which installs coordinates and turns the abstract rule into multiplication by an array.

Two ideas follow from that one shift in viewpoint, and they organize the chapter. First, because the matrix depends on the basis we chose, the *same* transformation has many matrices — one per basis — and they are all linked by the relation $B = M^{-1}AM$. That is the precise meaning of *similar matrices*, and it is why a clever basis (eigenvectors, Fourier vectors, wavelets) can replace a complicated matrix by a simple one. We will see this pay off in image compression. Second, we return to the matrices that have no honest inverse — rectangular ones, and square ones that are singular — and ask for the best substitute. Full column rank buys a *left* inverse; full row rank buys a *right* inverse; and in the hardest case, where the rank is short on both sides, the singular value decomposition hands us the *pseudoinverse* A^+, which inverts A on the part of space where inversion is possible and gives the shortest least-squares solution everywhere else. The four fundamental subspaces, met in Chapter 4, run through all of it.

10.1 Linear Transformations Without Coordinates

In older courses linear algebra began here, with transformations, and only later brought in matrices. The geometric approach has a real advantage: it avoids coordinates until computation forces them on us, and it keeps the eye on what the map *does* rather than on a grid of numbers.

A *transformation* T is just a rule that assigns to each input vector v an output vector $T(v)$. We write $T : \mathbb{R}^n \to \mathbb{R}^m$ when inputs live in $\mathbb{R}^n$ and outputs in $\mathbb{R}^m$. The transformations worth studying are the ones that interact properly with the two operations that define

a vector space — adding and scaling.

Definition 10.1: Linear Transformation

A transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ is **linear** if it respects vector addition and scalar multiplication:

$$T(v + w) = T(v) + T(w) \qquad \text{and} \qquad T(cv) = cT(v)$$

for all vectors v, w and all scalars c. The two conditions combine into the single requirement

$$T(cv + dw) = cT(v) + dT(w),$$

which says that T carries every linear combination of inputs to the same linear combination of outputs.

Notice what the definition does *not* mention: no rows, no columns, no numbers. It is a statement about combinations, and it is exactly the property that lets a transformation be pinned down by its action on a few vectors. One small consequence is worth recording: a linear transformation must fix the origin,

$$T(0) = T(0 \cdot v) = 0 \cdot T(v) = 0.$$

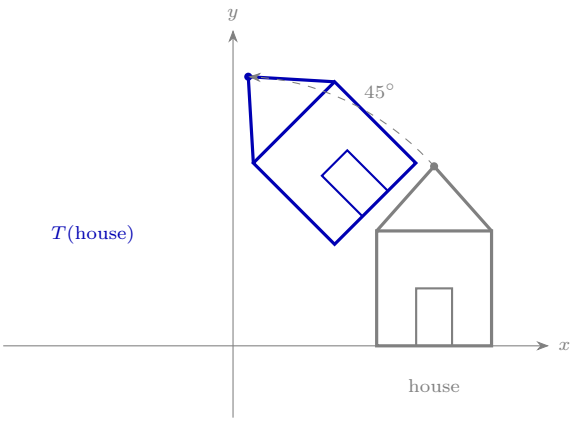
Any rule that moves the origin cannot be linear.

Example (Projection is linear).

Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ send each vector v to its perpendicular projection $T(v)$ onto a fixed line through the origin. Geometrically this drops v straight down onto the line. It is linear: projecting the sum $v + w$ gives the same point as projecting v and w separately and adding, and scaling v scales its shadow. We have not chosen any basis, written any matrix, or computed anything — yet we already know T is linear and that $T(0) = 0$.

Example (Rotation is linear).

Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ rotate every vector counterclockwise by 45° about the origin. Rotating a parallelogram rotates its diagonal, so $T(v + w) = T(v) + T(w)$; rotating a stretched vector stretches the rotated vector, so $T(cv) = cT(v)$. Again the rule is described, and seen to be linear, with no coordinates at all. This is the picture Strang likes: ask what happens to a drawing of a house in the plane, and the answer (a tilted house) is obvious from the geometry, while it would be hard to guess from a matrix of numbers.



It is just as important to see what is *not* linear, because the failures are instructive.

Example (Two non-examples).

The *shift* $T(v) = v + v_0$ that slides the whole plane by a fixed nonzero vector v_0 is not linear: it moves the origin, $T(0) = v_0 \neq 0$. Concretely $T(2v) = 2v + v_0$ while $2T(v) = 2v + 2v_0$, and these differ. (A map of this affine form is useful in graphics, but it is not a linear transformation.) The *length* map $T(v) = \|v\|$ also fails: it sends a vector to a single number, and $T(cv) = |c|\,\|v\|$, which is not $cT(v)$ when $c < 0$. From here on T always means a linear transformation.

Example (Multiplication by a matrix).

Every matrix $A \in \mathbb{R}^{m\times n}$ *defines* a linear transformation $T(v) = Av$, because matrix multiplication distributes over addition and pulls out scalars:

$$A(v + w) = Av + Aw, \qquad A(cv) = c\,Av.$$

For instance $A = \left[\begin{smallmatrix} 1 & 0 \\ 0 & -1 \end{smallmatrix}\right]$ leaves the x-component of a vector alone and flips the sign of the y-component — it reflects the plane across the x-axis. A 2×3 matrix gives a linear map $\mathbb{R}^3 \to \mathbb{R}^2$, squashing space down a dimension. The plan for the rest of the section is the converse: *every* linear transformation, once we pick a basis, is multiplication by some matrix.

10.1.1 A transformation is determined by its action on a basis

How much do we need to know about T to know it completely — to predict $T(v)$ for *every* input v? Far less than you might fear. Linearity is a powerful constraint.

Suppose we know $T(v_1)$ for one vector v_1. Then we know $T(cv_1) = c\,T(v_1)$ for every scalar multiple along that line. If we also know $T(v_2)$ for a second, independent vector, then for any combination we get

$$T(c_1v_1 + c_2v_2) = c_1T(v_1) + c_2T(v_2),$$

so T is determined on the entire plane they span. Push this to a full basis and the conclusion is complete.

Theorem 10.2: A Transformation Is Determined by a Basis

Let $v_1, \dots, v_n$ be a basis for the input space $\mathbb{R}^n$, and let $T : \mathbb{R}^n \to \mathbb{R}^m$ be linear. Then T is completely determined by the n output vectors $T(v_1), \dots, T(v_n)$. Indeed, every input has a unique expansion $v = c_1 v_1 + \cdots + c_n v_n$, and

$$T(v) = c_1\, T(v_1) + c_2\, T(v_2) + \cdots + c_n\, T(v_n).$$

Proof. Because $v_1, \dots, v_n$ is a basis, every $v \in \mathbb{R}^n$ is a combination $v = c_1 v_1 + \cdots + c_n v_n$, and the coefficients c_i are unique (Chapter 4: spanning gives existence, independence gives uniqueness). Applying linearity term by term,

$$T(v) = T\Big(\sum_i c_i v_i\Big) = \sum_i c_i\, T(v_i).$$

The right-hand side uses only the known vectors $T(v_i)$ and the coordinates c_i of v. So $T(v)$ is determined for every v. □

This theorem is the whole bridge from the coordinate-free world to matrices. The unique coefficients $c_1, \dots, c_n$ are the *coordinates* of v in the basis $v_1, \dots, v_n$. Coordinates come *from* a basis; change the basis and the same vector gets new coordinates. The transformation itself does not change — only its bookkeeping does.

10.2 The Matrix of a Linear Transformation

We now make the bridge explicit. To turn a linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$ into a matrix, we must choose *two* bases: an input basis $v_1, \dots, v_n$ for $\mathbb{R}^n$ to give coordinates to inputs, and an output basis $w_1, \dots, w_m$ for $\mathbb{R}^m$ to give coordinates to outputs. The recipe for the matrix is forced on us by the theorem above.

How to build the matrix of T

Fix an input basis $v_1, \dots, v_n$ and an output basis $w_1, \dots, w_m$. To find column j of the matrix A, apply T to the j-th input basis vector and expand the result in the output basis:

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

The coefficients $a_{1j}, \dots, a_{mj}$ are exactly column j of A. With this A, if v has coordinates $c = (c_1, \dots, c_n)$ in the input basis, then $T(v)$ has coordinates Ac in the output basis.

The rule looks circular at first, but it is the only thing it could be. Column j records "where the j-th input basis vector goes," written in the output basis. Linearity then propagates that information to every input, because $Ac = \sum_j c_j(\text{column } j)$ reproduces $\sum_j c_j T(v_j) = T(v)$. The product Ac being a combination of the columns — the column picture from Chapter 1 — is exactly what makes this work.

10.2.1 Projection: a clever basis versus the standard one

The single best illustration is projection, because the right basis makes the matrix almost trivial and the wrong basis makes it messy — for the very same transformation.

Example (Projection onto a line, in two bases).

Let $T : \mathbb{R}^2 \to \mathbb{R}^2$ project onto a line L through the origin, and use the same basis for input and output ($n = m = 2$).

Smart basis. Choose $v_1 = w_1$ to be a unit vector *along* L and $v_2 = w_2$ a unit vector *perpendicular* to L. Projection keeps the part along L and kills the perpendicular part:

$$T(v_1) = v_1, \qquad T(v_2) = 0.$$

Reading off coordinates in the basis $\{w_1, w_2\}$, the matrix is the diagonal

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad A \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} c_1 \\ 0 \end{bmatrix}.$$

A perfect matrix: it simply zeroes the second coordinate. This is no accident — v_1, v_2 are eigenvectors of the projection, with eigenvalues 1 and 0, and **in a basis of eigenvectors the matrix of any transformation is the diagonal matrix Λ of eigenvalues** (Chapter 7).

Standard basis. Now project onto the line at 45° but insist on the standard basis $e_1 = (1, 0)$, $e_2 = (0, 1)$. With $a = (1, 1)/\sqrt{2}$ a unit vector along the line, the projection matrix is $P = aa^{\mathsf{T}}$ (Chapter 5),

$$P = \frac{a\, a^{\mathsf{T}}}{a^{\mathsf{T}} a} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

Same transformation, different basis, and a fuller matrix. The lesson Strang draws is that the difficulty lives in the basis, not in the transformation. The diagonal $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$ and the dense $\left[\begin{smallmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{smallmatrix}\right]$ describe one and the same projection.

10.2.2 The derivative is a matrix

Linear transformations are not confined to arrows in $\mathbb{R}^n$. The space of polynomials is a vector space (Chapter 4), and differentiation is linear on it — a striking example because calculus turns into a matrix multiply.

Example (The derivative as a matrix).

Let the input space be quadratic polynomials $c_1 + c_2 x + c_3 x^2$, with basis $v_1 = 1$, $v_2 = x$, $v_3 = x^2$. The derivative

$$T(c_1 + c_2 x + c_3 x^2) = c_2 + 2c_3 x$$

lands in the space of linear polynomials, with output basis $w_1 = 1$, $w_2 = x$. To build the matrix, differentiate each input basis vector and expand:

$$T(1) = 0, \qquad T(x) = 1 = w_1, \qquad T(x^2) = 2x = 2w_2.$$

The three results, in output coordinates, are the columns of

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \qquad A \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} c_2 \\ 2c_3 \end{bmatrix}.$$

Differentiating $c_1 + c_2x + c_3x^2$ has become multiplying its coordinate vector by A, and the output coordinates $(c_2, 2c_3)$ are exactly the coefficients of $c_2 + 2c_3x$. The matrix *is* the derivative, once a basis is fixed.

Remark (Where matrix multiplication comes from).

Two facts close the loop and explain operations we have used all along. If T is invertible, then the inverse transformation T^{-1} has matrix A^{-1} — undoing the map undoes the matrix. And if we do one transformation after another, $T_2 \circ T_1$ (first $T_1 : v \mapsto A_1v$, then $T_2 : w \mapsto A_2w$), the combined transformation has matrix A_2A_1. **Matrix multiplication is defined the way it is precisely so that it matches composition of transformations.** The row-times-column rule is not arbitrary; it is forced by "do A_1, then A_2."

10.3 Change of Basis and Similar Matrices

We have seen one transformation wear two different matrices, depending on the basis. The relationship between those matrices is not random — it is a precise algebraic tie called *similarity*, and it is the engine behind eigenvalues, diagonalization, and the compression scheme in the next section. First we settle how the *coordinates of a vector* change, then how the *matrix of a transformation* changes.

10.3.1 Coordinates of a vector

Suppose we install a new basis whose vectors are the columns of a matrix $W = [w_1\ w_2\ \cdots\ w_n]$. A vector x, written in the *standard* basis, has new coordinates $c = (c_1, \ldots, c_n)$ in the w-basis meaning $x = c_1w_1 + \cdots + c_nw_n$. That sum is exactly Wc:

Change of basis for vectors

If the columns of W are the new basis vectors, then standard coordinates x and new coordinates c are related by

$$x = Wc, \qquad \text{equivalently} \qquad c = W^{-1}x.$$

To *express* x in the new basis we solve for its coordinates $c = W^{-1}x$; to *reconstruct* x from coordinates we multiply $x = Wc$.

The matrix W is invertible precisely because its columns are a basis (independent and spanning), so W^{-1} exists and the dictionary between x and c goes both ways.

10.3.2 The matrix of a transformation under change of basis

Now the main event. One transformation T, two bases, two matrices — how are they related? Suppose T has matrix A in the old basis and matrix B in the new basis, with M the change-of-basis matrix linking the two coordinate systems. Tracking a vector through both descriptions forces the relationship.

Theorem 10.3: Change of Basis Produces Similar Matrices

Let T be a linear transformation with matrix A in one basis and matrix B in another. If M is the change-of-basis matrix from the new coordinates to the old (so an input written u in the new basis is Mu in the old), then

$$B = M^{-1} A\, M.$$

Two matrices related this way are called **similar**.

Proof. Take an input with new coordinates u. Its old coordinates are Mu. Apply T in old coordinates: the output, in old coordinates, is $A(Mu)$. Convert that output back to new coordinates by multiplying by M^{-1}: the output in new coordinates is $M^{-1}AMu$. But by definition B is the matrix that sends new input coordinates straight to new output coordinates, so $Bu = M^{-1}AMu$ for every u. Hence $B = M^{-1}AM$. □

Similarity is the algebraic fingerprint of "same transformation, different basis." Similar matrices share everything that belongs to the transformation rather than to the basis: the same eigenvalues, the same determinant, the same rank, the same trace, the same characteristic polynomial. They differ only in the coordinates used to write them down.

Example (Diagonalization is a change of basis).

The eigenvalue factorization $A = S\Lambda S^{-1}$ from Chapter 7 is exactly this theorem read backwards. Rearranged, $\Lambda = S^{-1}AS$, which is $B = M^{-1}AM$ with $M = S$ the eigenvector matrix and $B = \Lambda$. The message: *in the basis of eigenvectors, the transformation's matrix is the diagonal* Λ. Diagonalizing a matrix means choosing the basis in which the transformation is simplest — pure stretching along the eigenvector axes, with no mixing. When the basis is orthonormal eigenvectors (the symmetric case, Chapter 8), $M = Q$ is orthogonal and $M^{-1} = Q^{\mathsf{T}}$, so the change of basis costs only a transpose.

10.4 Application: Image Compression

Here is why a good basis is worth real money. The whole content of compression is: *store the same information in a basis where most coordinates are nearly zero, then throw the small ones away.*

Picture one black-and-white video frame at 512×512 pixels. The camera records a brightness level for each of the 512^2 pixels, so a single frame is a vector in a 512^2-dimensional space. The *standard basis* for that space has one vector per pixel, and transmitting all 512^2 coordinates frame after frame would swamp any channel. The trouble is that the standard

basis is blind to structure: there is no efficient way for it to say "the whole frame is black," because that statement still costs one number per pixel.

Change the basis and the costs change. Suppose one basis vector is the all-ones vector $(1, 1, \ldots, 1)$ — every pixel at the same brightness. Then a blank gray board is a *single* coordinate, not 512^2 of them. Add a vector that is half $+1$ and half -1 (bright on the left, dark on the right), a vector that alternates $+1, -1, +1, -1$ (fine stripes), and so on for the various patterns an image is built from. Smooth, large-scale features now live in just a few coordinates.

Example (Two good bases: Fourier and Haar).

The best-known choice is the **Fourier basis**, whose vectors are sampled cosines and sines (the real parts of the powers ω^{jk} from the Fourier matrix). The JPEG standard cuts each frame into 8×8 blocks of 64 pixels and rewrites every block in this cosine basis. A competing choice is the **Haar wavelet basis** for $\mathbb{R}^8$,

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}, \ldots$$

The first vector measures overall brightness; later vectors measure finer and finer differences, many of their entries being 0. These vectors are orthogonal and can be scaled to orthonormal, which is what makes the next step cheap. JPEG2000 uses an improved wavelet basis of this kind.

The mechanics are nothing but change of basis. To rewrite a signal x (standard, pixel coordinates) in a new basis with vectors as the columns of W, we want coordinates c with

$$x = c_1 w_1 + \cdots + c_n w_n = Wc, \qquad \text{so} \qquad c = W^{-1} x.$$

This is the change-of-basis formula again. Two properties make a basis good for compression:

- **Fast transforms.** Multiplying by W and by W^{-1} must be cheap. The Fourier basis is ideal because the FFT computes Wc and $W^{-1}x$ in $O(n \log n)$ time, and a wavelet transform is even faster. An *orthonormal* basis is best of all: then $W^{-1} = W^{\mathsf{T}}$, so the inverse transform is free.
- **Sparse coordinates.** Most of the c_i should come out small, safely set to zero without a visible change. Storing only the few large coefficients is the compression.

The pipeline is: take x, compute its 64 coefficients c losslessly in the new basis, then *lossily* discard the coefficients below a threshold, leaving $\hat{c}$ with many zeros, and reconstruct $\hat{x} = \sum \hat{c}_i w_i \approx x$.

Remark (Why not the eigenvector basis?).

> In principle the perfect basis is the eigenvectors of the relevant transformation — it diagonalizes everything. In practice computing that basis is far more expensive than a JPEG could ever afford. Fourier and wavelet bases are the engineering compromise: not optimal, but *fixed* (the same for every image) and equipped with a fast transform. The best basis for a slide of algebra notes differs from the best basis for an action movie, but both pay off for the same reason — they concentrate the picture into a handful of coordinates.

10.5 Left and Right Inverses

We return to inverses, but now for matrices that are not square — or that are square but singular. A genuine, two-sided inverse A^{-1} satisfies $AA^{-1} = I = A^{-1}A$ and exists only when A is square with full rank, $r = m = n$. Then $N(A) = \{0\}$ and $N(A^\mathsf{T}) = \{0\}$: nothing is collapsed in either direction. When A is rectangular, we cannot have a two-sided inverse, because either A or A^T has a nontrivial null space. But we can sometimes invert from *one* side, and which side depends on which kind of full rank A has.

10.5.1 Full column rank: a left inverse

Suppose A is $m \times n$ with *full column rank* $r = n$ (tall, independent columns). Then $N(A) = \{0\}$: the equation $Ax = b$ has *at most* one solution — exactly one when $b \in C(A)$, none otherwise. From Chapter 5 we know the crucial fact that makes least squares work: when the columns of A are independent, the $n \times n$ matrix $A^\mathsf{T}A$ is invertible.

The left inverse

If A has full column rank $r = n$, then $A^\mathsf{T}A$ is invertible and

$$A_{\text{left}}^{-1} = (A^\mathsf{T}A)^{-1}A^\mathsf{T}$$

is a **left inverse**: it satisfies $A_{\text{left}}^{-1}\,A = (A^\mathsf{T}A)^{-1}A^\mathsf{T}A = I_n$. Multiplying A on the *left* by this matrix recovers the identity.

This is precisely the matrix that solved least squares: $\hat{x} = (A^\mathsf{T}A)^{-1}A^\mathsf{T}b$ is the best solution of an unsolvable $Ax = b$. The left inverse undoes A from the left, but the *other* product,

$$A\,A_{\text{left}}^{-1} = A(A^\mathsf{T}A)^{-1}A^\mathsf{T} = P,$$

is the projection matrix onto the column space $C(A)$ (Chapter 5), not the identity — it equals I_m only when $m = n$. A tall matrix can be undone after it acts, but it cannot be undone before: there is no way to recover the directions of $\mathbb{R}^m$ outside the column space, because A never reached them.

10.5.2 Full row rank: a right inverse

The mirror image. Suppose A is $m \times n$ with *full row rank* $r = m$ (wide, independent rows). Then $N(A^\mathsf{T}) = \{0\}$ and $C(A) = \mathbb{R}^m$, so $Ax = b$ is solvable for *every* b — in fact, with $n - m$

free variables, there are infinitely many solutions when $n > m$. Now AA^T is the invertible one (it is $m \times m$ and full rank).

The right inverse

If A has full row rank $r = m$, then AA^T is invertible and

$$A_{\text{right}}^{-1} = A^\mathsf{T}(AA^\mathsf{T})^{-1}$$

is a **right inverse**: it satisfies $A\,A_{\text{right}}^{-1} = AA^\mathsf{T}(AA^\mathsf{T})^{-1} = I_m$. Multiplying A on the *right* by this matrix recovers the identity.

Symmetrically, the opposite product

$$A_{\text{right}}^{-1}\,A = A^\mathsf{T}(AA^\mathsf{T})^{-1}A$$

is the projection onto the row space $C(A^\mathsf{T})$, not the identity. The right inverse picks out, among the infinitely many solutions of $Ax = b$, the one lying in the row space.

Remark (One table to remember).

The two one-sided inverses are transposes of each other's construction, and each lives where its $A^\mathsf{T}A$ or AA^T is invertible.

case	condition	inverse
two-sided	$r = m = n$	A^{-1}
left	$r = n < m$ (full column rank)	$(A^\mathsf{T}A)^{-1}A^\mathsf{T}$
right	$r = m < n$ (full row rank)	$A^\mathsf{T}(AA^\mathsf{T})^{-1}$

The remaining and hardest case — $r < m$ *and* $r < n$, short rank on both sides — has neither a left nor a right inverse. That is the case the pseudoinverse is built for.

10.6 The Pseudoinverse A^+

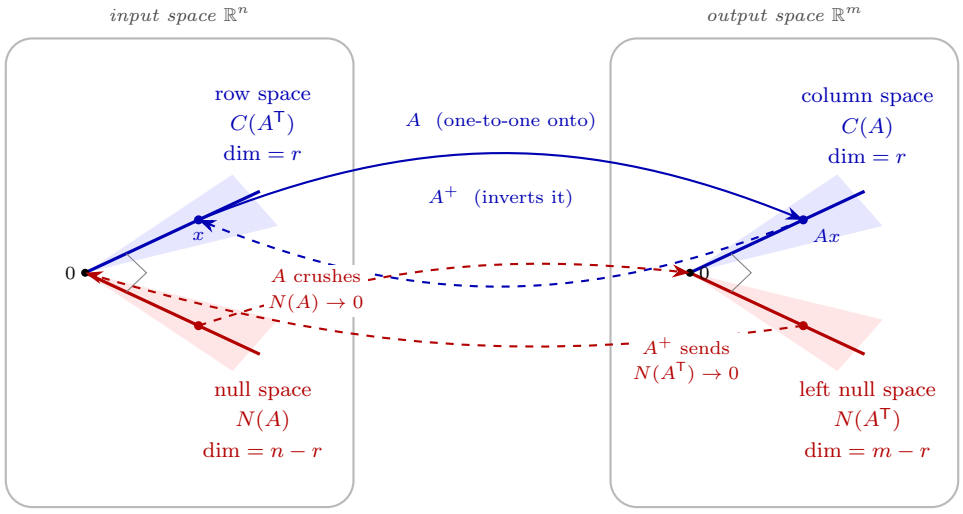
When A has $r < m$ and $r < n$, both $N(A)$ and $N(A^\mathsf{T})$ are nontrivial. A nonzero x with $Ax = 0$ kills information that no inverse can resurrect: there is no matrix sending 0 back to x. So we cannot ask for a true inverse. But there is a clean part of the action that *is* invertible, and the pseudoinverse inverts exactly that.

The clean part is the heart of the four-subspaces picture. Recall (Chapter 4) that A crushes the null space $N(A)$ to zero, and on the row space $C(A^\mathsf{T})$ — the orthogonal complement of $N(A)$ inside $\mathbb{R}^n$ — it acts one-to-one onto the column space $C(A)$. Both of those spaces have dimension r. So restricted to the row space, A is a genuine bijection onto the column space, and *that* restriction can be inverted.

Theorem 10.4: Row Space and Column Space Are in One-to-One Correspondence

Let A be $m \times n$ of rank r. The map $x \mapsto Ax$ sends the row space $C(A^\mathsf{T})$ one-to-one onto the column space $C(A)$. That is, if $x \neq y$ are both in the row space, then $Ax \neq Ay$.

Proof. Suppose instead that $x \neq y$ lie in the row space but $Ax = Ay$. Then $A(x - y) = 0$, so $x - y \in N(A)$. But the row space is a subspace, so the difference $x - y$ of two row-space vectors is again in the row space. Thus $x - y$ lies in both $N(A)$ and $C(A^\mathsf{T})$. By the Fundamental Theorem these are orthogonal complements (Chapter 4), meeting only at the origin, so $x - y = 0$, i.e. $x = y$ — contradicting $x \neq y$. Hence distinct row-space vectors have distinct images: the map is one-to-one, and since its image is all of $C(A)$ (every Ax lies in $C(A)$, and the row-space part of any x already produces it), it is onto. □



This bijection is what the pseudoinverse inverts. It returns each Ax in the column space to the unique x in the row space that produced it, and it sends everything in the left null space $N(A^\mathsf{T})$ to zero (those directions never came from any input we can recover).

Definition 10.5: Pseudoinverse

The **pseudoinverse** A^+ (also called the Moore–Penrose inverse) of an $m \times n$ matrix A is the $n \times m$ matrix that inverts the row-space-to-column-space bijection: $A^+(Ax) = x$ for every x in the row space $C(A^\mathsf{T})$, and $A^+ y = 0$ for every y in the left null space $N(A^\mathsf{T})$. In symbols, $N(A^+) = N(A^\mathsf{T})$.

10.6.1 Building A^+ from the SVD

The definition is geometric; the construction is the singular value decomposition (Chapter 9). The SVD is the perfect tool because it presents A already split along the four subspaces.

Write $A = U\Sigma V^{\mathsf{T}}$, where U $(m \times m)$ and V $(n \times n)$ are orthogonal and Σ $(m \times n)$ holds the singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$ on its first r diagonal entries, zeros elsewhere. The first r columns of V are an orthonormal basis for the row space, the first r columns of U an orthonormal basis for the column space, and A stretches the k-th row-space axis by σ_k onto the k-th column-space axis.

Because U and V are orthogonal, $U^{-1} = U^{\mathsf{T}}$ and $V^{-1} = V^{\mathsf{T}}$, so inverting A reduces to inverting Σ. And Σ is "diagonal," so we invert it the only sensible way: reciprocate the nonzero singular values and leave the zeros alone.

The pseudoinverse via SVD

If $A = U\Sigma V^{\mathsf{T}}$, then

$$A^{+} = V\,\Sigma^{+}\,U^{\mathsf{T}},$$

where Σ^{+} is the $n \times m$ matrix with $1/\sigma_1, \ldots, 1/\sigma_r$ on its first r diagonal entries and zeros elsewhere. The product $\Sigma\Sigma^{+}$ is the $m \times m$ diagonal matrix with r ones (then zeros); $\Sigma^{+}\Sigma$ is the $n \times n$ version. When A is square and invertible ($r = m = n$), every $\sigma_i \neq 0$, $\Sigma^{+} = \Sigma^{-1}$, and $A^{+} = A^{-1}$ collapses to the ordinary inverse.

The reason we reciprocate σ_k but leave the zeros as zeros is the geometry: on the row space, A multiplies the k-th axis by σ_k, so undoing it divides by σ_k; in the null-space directions, A multiplied by 0, and there is nothing to undo, so A^{+} multiplies by 0 there too. That is the only choice consistent with $N(A^{+}) = N(A^{\mathsf{T}})$.

Example (A rank-one pseudoinverse).

Take the rank-one $A = \left[\begin{smallmatrix} 1 & 2 \\ 2 & 4 \end{smallmatrix}\right]$, both columns multiples of $(1, 2)$, so $r = 1$, $m = n = 2$. A two-sided inverse is impossible (the matrix is singular), yet A^{+} exists. Writing $A = \sigma_1 u_1 v_1^{\mathsf{T}}$ with unit vectors $u_1 = v_1 = (1, 2)/\sqrt{5}$ and $\sigma_1 = 5$ (since $\|(1, 2)\|^2 = 5$ and $A = (1, 2)(1, 2)^{\mathsf{T}}$), the pseudoinverse reciprocates the one nonzero singular value:

$$A^{+} = \tfrac{1}{\sigma_1}\, v_1 u_1^{\mathsf{T}} = \tfrac{1}{5} \cdot \frac{1}{5} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \begin{bmatrix} 1 & 2 \end{bmatrix} = \frac{1}{25} \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}.$$

Check the defining property on the row space: the row space is the line through v_1, and $A^{+}A$ projects onto it. Indeed $A^{+}A = \frac{1}{25}\left[\begin{smallmatrix} 1 & 2 \\ 2 & 4 \end{smallmatrix}\right]\left[\begin{smallmatrix} 1 & 2 \\ 2 & 4 \end{smallmatrix}\right] = \frac{1}{25}\left[\begin{smallmatrix} 5 & 10 \\ 10 & 20 \end{smallmatrix}\right] = \frac{1}{5}\left[\begin{smallmatrix} 1 & 2 \\ 2 & 4 \end{smallmatrix}\right]$, the projection onto the line through $(1, 2)$, exactly as the theory predicts.

10.6.2 The pseudoinverse as the shortest least-squares solution

The pseudoinverse is not a curiosity. It is the single formula that solves $Ax = b$ "as well as possible" in *every* case at once — the unifying answer that ordinary, left, and right inverses each handled only in their special situation.

What $A^{+}b$ computes

For any A and any b, the vector

$$\hat{x} = A^{+}b$$

is the **minimum-norm least-squares solution** of $Ax = b$:

- among all x that minimize $\|Ax - b\|$ (the least-squares condition), $\hat{x}$ is the one of smallest length $\|x\|$;
- it always lies in the row space $C(A^\mathsf{T})$, the part of $\mathbb{R}^n$ with no wasted null-space component.

Each earlier inverse is a special case of this one. When A is square and invertible, $A^+ = A^{-1}$ and $\hat{x}$ is the exact solution. When A has full column rank, $A^+ = (A^\mathsf{T}A)^{-1}A^\mathsf{T} = A_{\text{left}}^{-1}$, and $\hat{x}$ is the least-squares solution of Chapter 5. When A has full row rank, $A^+ = A^\mathsf{T}(AA^\mathsf{T})^{-1} = A_{\text{right}}^{-1}$, and $\hat{x}$ is the shortest of the infinitely many exact solutions. In the hardest case — short rank on both sides, no left or right inverse — A^+b still returns the shortest best-fit. This is why statisticians reach for A^+ in regression: they cannot always promise their design matrix has independent columns, but A^+ returns a sensible answer regardless.

Remark (Everything ties together here).

Look at what the pseudoinverse gathers up. The *four fundamental subspaces* (Chapter 4) provide the stage: A^+ inverts the row-space-to-column-space bijection and zeroes the left null space. The *SVD* (Chapter 9) provides the construction: orthogonal U, V and a reciprocated Σ. *Projection and least squares* (Chapter 5) provide the meaning: AA^+ projects onto $C(A)$, A^+A projects onto $C(A^\mathsf{T})$, and A^+b is the shortest best-fit solution. The pseudoinverse is the place where the four subspaces, the SVD, orthogonal matrices, and least squares are revealed to be one circle of ideas — the right note on which to close the book.

10.7 What to Carry Forward

A matrix was never the primary object; it is the coordinate shadow of a *linear transformation*, a coordinate-free rule satisfying $T(v + w) = Tv + Tw$ and $T(cv) = cTv$. Choosing an input basis and an output basis turns T into a matrix A whose j-th column is $T(v_j)$ written in the output basis. Choose the basis well — eigenvectors, Fourier vectors, wavelets — and the matrix becomes simple (diagonal, or fast to apply); that single fact powers diagonalization and image compression alike.

Because the matrix depends on the basis, one transformation has many matrices, all linked by $B = M^{-1}AM$: they are *similar*, sharing eigenvalues, determinant, rank, and trace while differing only in coordinates. And when a matrix has no honest inverse, the rank decides the substitute: full column rank gives the left inverse $(A^\mathsf{T}A)^{-1}A^\mathsf{T}$, full row rank gives the right inverse $A^\mathsf{T}(AA^\mathsf{T})^{-1}$, and the general case gives the pseudoinverse $A^+ = V\Sigma^+U^\mathsf{T}$, which inverts A on the row space, kills the left null space, and returns the shortest least-squares solution $\hat{x} = A^+b$ for any A and any b. Four subspaces, one number r, and the SVD — the threads of the whole book pulled into a single knot.