
CS/MATH 3414 — Numerical Methods

Lecture 5 Linear Systems of Equations, Part 2

Virginia Tech Department of Computer Science, Fall 2026

Instructor: Sumeet Khatri

Date: September 17, 2026

1	Solving a system using the inverse	2
1.1	Inverse of a matrix	2
1.2	Obtaining an inverse by row reduction	4
2	When an ordinary inverse cannot be used: the pseudoinverse	6
2.1	Singular values: orthogonal coordinates for a linear map	6
2.2	Deriving the solution directly from the SVD	8
2.3	Definition and properties of the Moore–Penrose pseudoinverse	8
2.4	Useful formulas under rank assumptions	11
3	Computing generalized solutions reliably	15
3.1	Small singular values amplify errors	16
3.2	Numerical rank and truncated SVD	17
4	Error analysis	19
4.1	Residual, forward error, and a backward-error interpretation	19
4.2	Conditioning measures sensitivity of the problem	20
A	Linear algebra background	21
A.1	Inner products, lengths, and matrix norms	21
A.2	Identity, inverses, and equivalent conditions	23
A.3	Orthonormal coordinates and orthogonal projection	25
A.4	Symmetry, positivity, and the spectral theorem	28
A.5	Singular values and the SVD	29
A.6	The pseudoinverse ingredients in these coordinates	31

Learning Objectives

- Understand and explain the definition of the inverse of a matrix.
- Explain why a matrix may not be invertible, and how such linear systems could be solved anyway.
- Understand and explain the definition of a pseudoinverse of a matrix.
- Understand how to numerically take the inverse of a matrix using Python packages.

Recap from Lecture 4, Part 1 of Linear Systems

- Linear systems of equations are ubiquitous in science and engineering.
- Gaussian elimination is a formal, systematic method for implementing the idea: “solve for one of the variables, substitute into the remaining equations, and repeat”. We perform clever row operations (“forward elimination”) to arrive at a triangular system that can be solved in exactly this way (“backward substitution”).

Primary References

- [1] T. A. Driscoll and R. J. Braun, *Fundamentals of Numerical Computation*, Includes Python, Julia, and MATLAB examples (Society for Industrial and Applied Mathematics, Philadelphia, PA, 2022), <https://fncbook.com/>.
- [2] T. Sauer, *Numerical Analysis*, Third (Pearson, 2018), <https://www.pearson.com/en-ca/subject-catalog/p/numerical-analysis/P200000006340/9780134696454>.
- [3] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, Second (Society for Industrial and Applied Mathematics, Philadelphia, PA, 2002), <https://doi.org/10.1137/1.9780898718027>.
- [4] R. L. Burden, J. D. Faires, and A. M. Burden, *Numerical Analysis*, 10th ed. (Cengage Learning, 2016), <https://www.cengage.com/c/numerical-analysis-10e-burden/9781305253667/?filterBy=Student>.
- [5] E. W. Cheney and D. R. Kincaid, *Numerical Mathematics and Computing*, 7th ed. (Brooks/Cole, Cengage Learning, 2013), <https://web.ma.utexas.edu/CNA/NMC7/>.
- [6] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed. (Johns Hopkins University Press, Baltimore, MD, 2013), <https://www.press.jhu.edu/books/title/10678/matrix-computations>.
- [7] R. Penrose, “A Generalized Inverse for Matrices”, *Proceedings of the Cambridge Philosophical Society* **51**, 406 (1955).
- [8] L. N. Trefethen and D. Bau III, *Numerical Linear Algebra* (Society for Industrial and Applied Mathematics, Philadelphia, PA, 1997), <https://epubs.siam.org/doi/book/10.1137/1.9780898719574>.
- [9] G. Strang, *Introduction to Linear Algebra*, 5th ed. (Wellesley-Cambridge Press, Wellesley, MA, 2016), <https://math.mit.edu/~gs/linearalgebra/ila5/indexila5.html>.

Note: much of the text here comes directly from, or is a summarized form of, the text in the above two references.

1 Solving a system using the inverse

Gaussian elimination is an algorithm: it tells us how to transform a system and calculate its solution. The inverse provides a complementary description of the answer. It also prepares us to ask a more general question: what should “solving” mean when a matrix is rectangular or singular?

1.1 Inverse of a matrix

Definition 1 (Inverse of a square matrix). Let $A \in \mathbb{R}^{n \times n}$. An inverse of A is a matrix $A^{-1} \in \mathbb{R}^{n \times n}$ satisfying

$$A^{-1}A = AA^{-1} = \mathbb{I}_n. \quad (1.1)$$

If such a matrix exists, A is called *invertible* or *nonsingular*. Otherwise, A is *singular*.

Multiplication by A transforms an input vector into an output vector. Multiplication by A^{-1} reverses this transformation. This is a matrix operation: A^{-1} is generally not obtained by replacing every entry of A by its reciprocal.

For a general 2×2 matrix, the familiar formula is

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}, \quad ad - bc \neq 0. \quad (1.2)$$

It follows by multiplying the two displayed matrices: their product before dividing by $ad - bc$ is $(ad - bc)\mathbb{1}_2$. When $ad - bc = 0$, the columns are linearly dependent and no inverse exists. This small formula is useful for hand calculations. It does not suggest using determinant and cofactor formulas as general algorithms for large matrices.

Proposition 2 (Uniqueness of an inverse). Let $A \in \mathbb{R}^{n \times n}$. If $B, C \in \mathbb{R}^{n \times n}$ both satisfy the two inverse identities for A , then $B = C$.

Proof. The equality $AC = \mathbb{1}_n$ lets us insert AC to the right of B . The equality $BA = \mathbb{1}_n$ then lets us simplify:

$$B = B\mathbb{1}_n = B(AC) = (BA)C = \mathbb{1}_n C = C. \quad (1.3)$$

The middle step uses associativity of matrix multiplication, not commutativity. Thus the notation A^{-1} refers to one determined matrix whenever an inverse exists. ■

Theorem 1 (Solution of an invertible system). If $A \in \mathbb{R}^{n \times n}$ is invertible, then for every $\mathbf{b} \in \mathbb{R}^n$ the system $A\mathbf{x} = \mathbf{b}$ has exactly one solution, namely

$$\mathbf{x} = A^{-1}\mathbf{b}. \quad (1.4)$$

Proof. **Existence.** The dimensions are compatible: $A^{-1}\mathbf{b}$ is in $\mathbb{R}^n$, the space of unknown vectors. Substitute this candidate into the system and use associativity:

$$A(A^{-1}\mathbf{b}) = (AA^{-1})\mathbf{b} = \mathbb{1}_n\mathbf{b} = \mathbf{b}. \quad (1.5)$$

Therefore at least one solution exists for the given $\mathbf{b}$.

Uniqueness. Suppose $\mathbf{y}$ is any solution. Multiplication on the left by A^{-1} preserves the equality $A\mathbf{y} = \mathbf{b}$ and gives

$$A^{-1}(A\mathbf{y}) = A^{-1}\mathbf{b}, \quad (A^{-1}A)\mathbf{y} = A^{-1}\mathbf{b}, \quad \mathbf{y} = A^{-1}\mathbf{b}. \quad (1.6)$$

Every solution must equal the candidate already verified, so there is exactly one. This proof applies to every right-hand side; there is no assumption on the entries of $\mathbf{b}$. ■

The theorem combines two statements that must be distinguished for more general systems: a solution *exists*, and it is *unique*. For a square matrix, invertibility, rank n , a trivial null space, and the existence of a pivot in every column are equivalent conditions. These equivalences and the relevant subspaces are reviewed in the appendix. They also explain an elimination test for invertibility: reducing an invertible square matrix to reduced row-echelon form produces the identity matrix.

Recap: The rank of a matrix

Let $A \in \mathbb{R}^{m \times n}$, with rows $\mathbf{r}_1, \dots, \mathbf{r}_m \in \mathbb{R}^n$ and columns $\mathbf{c}_1, \dots, \mathbf{c}_n \in \mathbb{R}^m$.

- The **row rank** of A is the dimension of its row space, $\text{span}\{\mathbf{r}_1, \dots, \mathbf{r}_m\}$.
- The **column rank** of A is the dimension of its column space, $\text{span}\{\mathbf{c}_1, \dots, \mathbf{c}_n\}$.

These dimensions are always equal; their common value is the **rank** of A . Thus,

$$\text{rank}(A) = \text{row rank of } A = \text{column rank of } A \leq \min\{m, n\}. \quad (1.7)$$

The matrix has **full row rank** if $\text{rank}(A) = m$, **full column rank** if $\text{rank}(A) = n$, and **full rank** if $\text{rank}(A) = \min\{m, n\}$. It has both full row rank and full column rank precisely when $m = n = \text{rank}(A)$.

Example 1 (An inverse and its action). Let

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \quad A^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 5 \\ 3 \end{pmatrix}. \quad (1.8)$$

The two products can be checked entry by entry:

$$\begin{aligned} AA^{-1} &= \begin{pmatrix} 2(1) + 1(-1) & 2(-1) + 1(2) \\ 1(1) + 1(-1) & 1(-1) + 1(2) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \\ A^{-1}A &= \begin{pmatrix} 1(2) - 1(1) & 1(1) - 1(1) \\ -1(2) + 2(1) & -1(1) + 2(1) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \end{aligned}$$

Thus $AA^{-1} = A^{-1}A = \mathbb{I}_2$. Therefore

$$\mathbf{x} = A^{-1}\mathbf{b} = \begin{pmatrix} 5 - 3 \\ -5 + 6 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}. \quad (1.9)$$

The first column of A^{-1} solves $A\mathbf{x} = (1, 0)^\top$; its second column solves $A\mathbf{x} = (0, 1)^\top$. Linearity combines these two responses to solve for any right-hand side.

1.2 Obtaining an inverse by row reduction

To compute A^{-1} , we can solve n systems simultaneously. Write the unknown inverse as $X = (\mathbf{x}_1 \cdots \mathbf{x}_n)$. The matrix equation $AX = \mathbb{I}_n$ says

$$A\mathbf{x}_j = \mathbf{e}_j, \quad j = 1, \dots, n, \quad (1.10)$$

where $\mathbf{e}_j$ is the j th standard basis vector. Each column of the inverse is therefore the solution for one particular right-hand side.

Theorem 2 (Why Gauss–Jordan inversion works). Let $A \in \mathbb{R}^{n \times n}$. A sequence of elementary row operations transforms $(A \mid \mathbb{1}_n)$ into $(\mathbb{1}_n \mid X)$ if and only if A is invertible. Whenever this reduction succeeds, $X = A^{-1}$.

Proof. Every allowed row operation is reversible: a row exchange reverses itself; multiplication of a row by a nonzero number is reversed by multiplication by its reciprocal; and adding c times one row to another is reversed by adding $-c$ times the same row. Thus the elementary matrix representing each row operation is invertible. If the operations are represented, in order, by $E_1, \dots, E_k$, their combined effect is multiplication on the left by the invertible matrix

$$E = E_k \cdots E_2 E_1. \quad (1.11)$$

Applying the same operations to both blocks gives

$$E \begin{pmatrix} A & \mid & \mathbb{1}_n \end{pmatrix} = \begin{pmatrix} EA & \mid & E\mathbb{1}_n \end{pmatrix} = \begin{pmatrix} EA & \mid & E \end{pmatrix}. \quad (1.12)$$

If the left block has become $\mathbb{1}_n$, then $EA = \mathbb{1}_n$. Because E is invertible, multiplying this equality on the left by E^{-1} gives $A = E^{-1}$. It follows that A is invertible and its inverse is E , which is exactly the final right block X . This argument does not presume that A has an inverse before the row reduction succeeds.

Conversely, suppose A is invertible. Its null space contains only zero, so its n columns are independent. Row reduction therefore has a pivot in every column. For an $n \times n$ matrix, a pivot in every column is also a pivot in every row. In reduced row-echelon form each pivot is one and all other entries in a pivot's column are zero, so the reduced matrix is $\mathbb{1}_n$. Apply those same row operations to $(A \mid \mathbb{1}_n)$; the preceding argument shows that the other block is A^{-1} . ■

The resulting procedure is

$$\begin{pmatrix} A & \mid & \mathbb{1}_n \end{pmatrix} \longrightarrow \begin{pmatrix} \mathbb{1}_n & \mid & A^{-1} \end{pmatrix}. \quad (1.13)$$

This is *Gauss–Jordan elimination*: after creating zeros below each pivot, we also create zeros above it and scale the pivot to one. An elementary operation must always act on the entire augmented row. Changing the coefficient block while leaving the right block untouched would no longer describe the same systems.

Example 2 (Computing an inverse and using it). Let

$$A = \begin{pmatrix} 2 & 1 \\ 5 & 3 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}. \quad (1.14)$$

Beginning with $(A \mid \mathbb{1}_2)$, subtract $5/2$ times the first row from the second, and then multiply the second row by 2:

$$\left(\begin{array}{cc|cc} 2 & 1 & 1 & 0 \\ 5 & 3 & 0 & 1 \end{array} \right) \longrightarrow \left(\begin{array}{cc|cc} 2 & 1 & 1 & 0 \\ 0 & 1/2 & -5/2 & 1 \end{array} \right) \longrightarrow \left(\begin{array}{cc|cc} 2 & 1 & 1 & 0 \\ 0 & 1 & -5 & 2 \end{array} \right). \quad (1.15)$$

Subtract the second row from the first and divide the first row by 2. We obtain

$$\left(\begin{array}{cc|cc} 1 & 0 & 3 & -1 \\ 0 & 1 & -5 & 2 \end{array} \right), \quad A^{-1} = \begin{pmatrix} 3 & -1 \\ -5 & 2 \end{pmatrix}. \quad (1.16)$$

Consequently,

$$\mathbf{x} = A^{-1}\mathbf{b} = \begin{pmatrix} 3 & -1 \\ -5 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}. \quad (1.17)$$

Substitution gives $2(1) + (-1) = 1$ and $5(1) + 3(-1) = 2$. Notice that the direct elimination method would have needed just one augmented column. Computing the inverse required two augmented columns because it solved for every column of $\mathbb{I}_2$.

The next example constructs the inverse used in the earlier solution-operator example. It also illustrates choosing a convenient pivot by interchanging rows.

Example 3 (Gauss–Jordan inversion, step by step). For $A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ from the first inverse example,

$$\begin{pmatrix} 2 & 1 & | & 1 & 0 \\ 1 & 1 & | & 0 & 1 \end{pmatrix} \xrightarrow{R_1 \leftrightarrow R_2} \begin{pmatrix} 1 & 1 & | & 0 & 1 \\ 2 & 1 & | & 1 & 0 \end{pmatrix} \xrightarrow{R_2 \leftarrow R_2 - 2R_1} \begin{pmatrix} 1 & 1 & | & 0 & 1 \\ 0 & -1 & | & 1 & -2 \end{pmatrix} \\ \xrightarrow{R_2 \leftarrow -R_2} \begin{pmatrix} 1 & 1 & | & 0 & 1 \\ 0 & 1 & | & -1 & 2 \end{pmatrix} \xrightarrow{R_1 \leftarrow R_1 - R_2} \begin{pmatrix} 1 & 0 & | & 1 & -1 \\ 0 & 1 & | & -1 & 2 \end{pmatrix}.$$

Every row operation acts on both blocks. If A is singular, its left block cannot be reduced to $\mathbb{I}_n$; an ordinary inverse does not exist.

2 When an ordinary inverse cannot be used: the pseudoinverse

We now allow $A \in \mathbb{R}^{m \times n}$, with m equations and n unknowns. The right-hand side is $\mathbf{b} \in \mathbb{R}^m$, and a candidate solution is $\mathbf{x} \in \mathbb{R}^n$. A two-sided inverse cannot exist when $m \neq n$. A square matrix can also fail to have an inverse because its columns are dependent. We need to separate three questions:

1. Does any $\mathbf{x}$ satisfy $A\mathbf{x} = \mathbf{b}$ exactly?
2. If exact solutions exist, is there one or are there infinitely many?
3. If no exact solution exists, what criterion should replace exact equality?

2.1 Singular values: orthogonal coordinates for a linear map

Let $r = \text{rank}(A) > 0$. A *singular value decomposition* (SVD) is

$$A = U_r \Sigma_r V_r^T = \sum_{j=1}^r \sigma_j \mathbf{u}_j \mathbf{v}_j^T, \quad (2.1)$$

where

$$U_r \in \mathbb{R}^{m \times r}, \quad V_r \in \mathbb{R}^{n \times r}, \\ \Sigma_r = \text{diag}(\sigma_1, \dots, \sigma_r), \quad \sigma_1 \geq \dots \geq \sigma_r > 0.$$

The columns $\mathbf{u}_1, \dots, \mathbf{u}_r$ of U_r are orthonormal, as are the columns $\mathbf{v}_1, \dots, \mathbf{v}_r$ of V_r . In particular,

$$U_r^\top U_r = \mathbb{I}_r, \quad V_r^\top V_r = \mathbb{I}_r. \quad (2.2)$$

The numbers σ_j are the positive *singular values*. The vectors $\mathbf{u}_j$ form an orthonormal basis of $\text{Col}(A)$; the vectors $\mathbf{v}_j$ form an orthonormal basis of the row space, regarded as a subspace of $\mathbb{R}^n$. These statements follow, for example, from

$$A\mathbf{v}_j = \sigma_j \mathbf{u}_j, \quad A^\top \mathbf{u}_j = \sigma_j \mathbf{v}_j, \quad j = 1, \dots, r. \quad (2.3)$$

Thus in suitable orthonormal coordinates, applying A simply multiplies the j th retained coordinate by σ_j .

Theorem 3 (Existence and subspaces of the SVD). Every real $m \times n$ matrix A of rank $r > 0$ admits (2.1) with the stated positive singular values and orthonormal columns. Moreover,

$$\text{Col}(A) = \text{span}\{\mathbf{u}_1, \dots, \mathbf{u}_r\}, \quad \text{Col}(A^\top) = \text{span}\{\mathbf{v}_1, \dots, \mathbf{v}_r\}, \quad (2.4)$$

and $\text{Null}(A)$ is the orthogonal complement of the latter space. For $A = 0$, the corresponding expressions are empty sums.

Proof. The matrix $A^\top A$ is real symmetric. By the spectral theorem, it has an orthonormal eigenbasis $\mathbf{v}_1, \dots, \mathbf{v}_n$. All its eigenvalues are nonnegative: if $A^\top A\mathbf{v} = \mu\mathbf{v}$ and $\|\mathbf{v}\|_2 = 1$, then

$$\mu = \mathbf{v}^\top A^\top A\mathbf{v} = \|A\mathbf{v}\|_2^2 \geq 0. \quad (2.5)$$

Also $A^\top A$ and A have the same null space. One inclusion follows from $A\mathbf{z} = 0 \Rightarrow A^\top A\mathbf{z} = 0$; the reverse follows by multiplying $A^\top A\mathbf{z} = 0$ on the left by $\mathbf{z}^\top$, which gives $\|A\mathbf{z}\|_2^2 = 0$. Rank-nullity now shows that exactly r eigenvalues are positive. Order them as $\sigma_1^2 \geq \dots \geq \sigma_r^2 > 0$, with zero eigenvalues after these, and define

$$\mathbf{u}_j = \frac{A\mathbf{v}_j}{\sigma_j}, \quad j = 1, \dots, r. \quad (2.6)$$

For $1 \leq i, j \leq r$,

$$\mathbf{u}_i^\top \mathbf{u}_j = \frac{\mathbf{v}_i^\top A^\top A\mathbf{v}_j}{\sigma_i \sigma_j} = \frac{\sigma_j^2 \mathbf{v}_i^\top \mathbf{v}_j}{\sigma_i \sigma_j}. \quad (2.7)$$

This is one if $i = j$ and zero otherwise, so the $\mathbf{u}_j$ are orthonormal. For $j > r$, the corresponding eigenvalue is zero, hence $A\mathbf{v}_j = 0$. Expand an arbitrary vector $\mathbf{x}$ in the orthonormal basis $\mathbf{v}_1, \dots, \mathbf{v}_n$ and apply A :

$$A\mathbf{x} = \sum_{j=1}^n (\mathbf{v}_j^\top \mathbf{x}) A\mathbf{v}_j = \sum_{j=1}^r \sigma_j (\mathbf{v}_j^\top \mathbf{x}) \mathbf{u}_j. \quad (2.8)$$

Since this identity holds for every $\mathbf{x}$, it is exactly the matrix equality (2.1).

Every output in this expression belongs to the span of the $\mathbf{u}_j$, while each $\mathbf{u}_j = A\mathbf{v}_j/\sigma_j$ belongs to $\text{Col}(A)$. This proves the column-space equality. Transposing the SVD gives the row-space equality in the same way; in particular, $A^\top \mathbf{u}_j = \sigma_j \mathbf{v}_j$. Finally, the orthonormal expansion shows that $A\mathbf{x} = 0$ precisely when the first r coordinates $\mathbf{v}_j^\top \mathbf{x}$ are zero. The remaining directions form the orthogonal complement of the row space, proving the null-space statement. ■

The spectral theorem, including its proof, appears in Appendix A as Theorem 14. This construction explains the SVD in exact arithmetic. It is not a recommendation to form $A^\top A$ to compute an SVD numerically: doing so can destroy information about small singular values, as a later example demonstrates.

2.2 Deriving the solution directly from the SVD

The following calculation proves existence of least-squares solutions a second way. It also identifies the freedom that the ordinary inverse cannot resolve.

Decompose the data vector into its component in the column space and its orthogonal remainder:

$$\mathbf{b} = \sum_{j=1}^r \beta_j \mathbf{u}_j + \mathbf{b}_\perp, \quad \beta_j = \mathbf{u}_j^\top \mathbf{b}, \quad U_r^\top \mathbf{b}_\perp = 0. \quad (2.9)$$

Similarly, every candidate input can be written uniquely as

$$\mathbf{x} = \sum_{j=1}^r \alpha_j \mathbf{v}_j + \mathbf{z}, \quad \alpha_j = \mathbf{v}_j^\top \mathbf{x}, \quad \mathbf{z} \in \text{Null}(A). \quad (2.10)$$

To justify the last condition, the row space is the orthogonal complement of the null space. Equivalently, (2.1) shows that $V_r^\top \mathbf{z} = 0$ exactly when $A\mathbf{z} = 0$.

Applying A and using orthogonality gives

$$A\mathbf{x} = \sum_{j=1}^r \sigma_j \alpha_j \mathbf{u}_j, \quad (2.11)$$

$$\|\mathbf{b} - A\mathbf{x}\|_2^2 = \sum_{j=1}^r (\beta_j - \sigma_j \alpha_j)^2 + \|\mathbf{b}_\perp\|_2^2. \quad (2.12)$$

This is the central calculation. Each scalar α_j affects only one nonnegative term, and $\mathbf{z}$ affects none of them. Therefore the residual is minimized exactly when

$$\alpha_j = \frac{\beta_j}{\sigma_j}, \quad j = 1, \dots, r. \quad (2.13)$$

The component $\mathbf{b}_\perp$ is the unavoidable residual: it points outside the column space, so no choice of $\mathbf{x}$ can reproduce it. The null-space component $\mathbf{z}$ remains arbitrary because the transformation cannot detect it.

Among all the minimizing vectors, which has the smallest Euclidean norm? Orthogonality in (2.10) yields

$$\|\mathbf{x}\|_2^2 = \sum_{j=1}^r \left(\frac{\beta_j}{\sigma_j} \right)^2 + \|\mathbf{z}\|_2^2. \quad (2.14)$$

This expression is uniquely minimized by $\mathbf{z} = 0$. Thus the unique minimum-norm least-squares solution is

$$\mathbf{x}_* = \sum_{j=1}^r \frac{\mathbf{u}_j^\top \mathbf{b}}{\sigma_j} \mathbf{v}_j. \quad (2.15)$$

This derivation supplies both the desired solution and an algorithmic interpretation: project onto the left singular vectors, divide by the positive singular values, and recombine using the right singular vectors.

2.3 Definition and properties of the Moore–Penrose pseudoinverse

Definition 3 (Moore–Penrose pseudoinverse). For the SVD (2.1), the pseudoinverse is the $n \times m$ matrix

$$A^+ = V_r \Sigma_r^{-1} U_r^\top = \sum_{j=1}^r \frac{1}{\sigma_j} \mathbf{v}_j \mathbf{u}_j^\top. \quad (2.16)$$

For the zero matrix, define $A^+ = 0$ with the transposed dimensions.

In particular, one never divides by a zero singular value. The plus sign in A^+ is notation for this particular generalized inverse; it does not denote an ordinary arithmetic operation. The definition is independent of the choice of singular vectors in repeated-singular-value subspaces. The uniqueness characterization below also proves this independence.

Multiplying the SVD factors shows that

$$AA^+ = U_r U_r^\top, \quad A^+A = V_r V_r^\top. \quad (2.17)$$

The first matrix is the orthogonal projector onto $\text{Col}(A) \subseteq \mathbb{R}^m$. The second is the orthogonal projector onto the row space in $\mathbb{R}^n$. Therefore these products are generally *projectors*, not identity matrices. They become identities precisely when the corresponding subspace is the whole space.

Theorem 4 (Penrose characterization). For every $A \in \mathbb{R}^{m \times n}$, its pseudoinverse is the unique matrix $X \in \mathbb{R}^{n \times m}$ satisfying

$$AXA = A, \quad XAX = X, \quad (2.18)$$

$$(AX)^\top = AX, \quad (XA)^\top = XA. \quad (2.19)$$

Proof. Existence. Assume first that $r > 0$ and take $X = V_r \Sigma_r^{-1} U_r^\top$. Multiplying and using $U_r^\top U_r = V_r^\top V_r = \mathbb{1}_r$ gives

$$AX = U_r U_r^\top, \quad XA = V_r V_r^\top. \quad (2.20)$$

Both products are symmetric. Their actions on the corresponding SVD factors give

$$AXA = (U_r U_r^\top)(U_r \Sigma_r V_r^\top) = U_r \Sigma_r V_r^\top = A,$$

$$XAX = (V_r V_r^\top)(V_r \Sigma_r^{-1} U_r^\top) = V_r \Sigma_r^{-1} U_r^\top = X.$$

Thus the constructed matrix satisfies all four conditions. If $A = 0$, choosing $X = 0$ verifies all four directly.

Uniqueness. Suppose now that some $X \in \mathbb{R}^{n \times m}$ satisfies the four conditions. Complete the orthonormal columns of U_r and V_r to square orthogonal matrices U and V . Orthogonal changes of coordinates preserve both multiplication identities and symmetry: for example, $(U^\top A X U)^\top = U^\top (AX)^\top U$. In these coordinates put

$$D_0 = U^\top A V = \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix}, \quad Y = V^\top X U = \begin{pmatrix} B & C \\ D & E \end{pmatrix}. \quad (2.21)$$

The row and column sizes of D_0 are respectively $(r, m - r)$ and $(r, n - r)$. The sizes for Y are reversed; in particular, B is $r \times r$, C is $r \times (m - r)$, D is $(n - r) \times r$, and E is $(n - r) \times (m - r)$.

The identity $AXA = A$ becomes $D_0 Y D_0 = D_0$. Its upper-left block is

$$\Sigma_r B \Sigma_r = \Sigma_r. \quad (2.22)$$

Multiply on the left and right by the invertible diagonal matrix Σ_r^{-1} to obtain $B = \Sigma_r^{-1}$. Next compute

$$D_0 Y = \begin{pmatrix} \Sigma_r B & \Sigma_r C \\ 0 & 0 \end{pmatrix}, \quad Y D_0 = \begin{pmatrix} B \Sigma_r & 0 \\ D \Sigma_r & 0 \end{pmatrix}. \quad (2.23)$$

Symmetry of AX requires $D_0 Y$ to be symmetric. Its lower-left block is zero, so its upper-right block must be zero as well: $\Sigma_r C = 0$, hence $C = 0$. Similarly symmetry of XA forces $D \Sigma_r = 0$, hence $D = 0$. Therefore Y has the form

$$Y = \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & E \end{pmatrix}. \quad (2.24)$$

Finally, the remaining identity $XAX = X$ becomes $Y D_0 Y = Y$, but

$$Y D_0 Y = \begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix}. \quad (2.25)$$

Comparison of the lower-right blocks gives $E = 0$. Consequently

$$X = V Y U^\top = V_r \Sigma_r^{-1} U_r^\top = A^+. \quad (2.26)$$

When $r = m$ or $r = n$, the blocks of size zero are omitted, and the same argument holds. When $r = 0$, $A = 0$ and $XAX = X$ reads $0 = X$, so uniqueness also holds in that case. Every matrix has therefore exactly one matrix satisfying the four identities. ■

The first equality says that applying A , then its generalized inverse, and then A again preserves the original output. The symmetry equalities select orthogonal projections rather than oblique projections. All four requirements together distinguish the Moore–Penrose pseudoinverse from other generalized inverses [7].

Proposition 4 (The four orthogonal projections associated with A). For $A \in \mathbb{R}^{m \times n}$, the matrices AA^+ and $\mathbb{1}_m - AA^+$ are the orthogonal projectors onto $\text{Col}(A)$ and its orthogonal complement. The matrices A^+A and $\mathbb{1}_n - A^+A$ are the orthogonal projectors onto $\text{Col}(A^\top)$ and $\text{Null}(A)$, respectively.

Proof. For any matrix Q with orthonormal columns, $QQ^\top y$ belongs to its column space, because it is a linear combination of those columns. The remainder satisfies

$$Q^\top (y - QQ^\top y) = Q^\top y - (Q^\top Q) Q^\top y = 0, \quad (2.27)$$

so it is orthogonal to that space. This proves that $QQ^\top$ gives the orthogonal projection. It also shows that $\mathbb{1} - QQ^\top$ gives the complementary orthogonal projection. Algebraically, $QQ^\top$ is symmetric and idempotent, since $(QQ^\top)^2 = Q(Q^\top Q)Q^\top = QQ^\top$.

Use $Q = U_r$ and $Q = V_r$ together with (2.17) and Theorem 3. The latter theorem identifies their column spaces and says that the row space's orthogonal complement is $\text{Null}(A)$. These are the four stated subspaces. For $A = 0$, AA^+ and A^+A are zero, and the complementary projectors are identity matrices, as required. ■

Theorem 5 (What A^+b solves). Let $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Then

$$x_* = A^+b \quad (2.28)$$

is the unique vector of smallest Euclidean norm among all minimizers of $\|b - Ax\|_2$. More precisely:

1. All least-squares minimizers have the form

$$x = A^+b + z, \quad z \in \text{Null}(A), \quad (2.29)$$

or, equivalently,

$$x = A^+b + (\mathbb{1}_n - A^+A)w, \quad w \in \mathbb{R}^n. \quad (2.30)$$

2. Their common fitted output and residual are

$$Ax = AA^+b, \quad r_* = (\mathbb{1}_m - AA^+)b. \quad (2.31)$$

3. The system is consistent if and only if $AA^+b = b$. In that case A^+b is the unique minimum-norm *exact* solution.

Proof. For $r > 0$, equation (2.12) proves that all minimizers have exactly the positive-singular-value coordinates in (2.15), with an arbitrary null-space component. The fixed component is A^+b by (2.16). Equation (2.17) shows that $\mathbb{1}_n - A^+A$ projects onto the null space, so its range is exactly the set of allowable z .

The two components are orthogonal. Consequently, for every null-space vector z ,

$$\|A^+b + z\|_2^2 = \|A^+b\|_2^2 + \|z\|_2^2. \quad (2.32)$$

The norm is smallest uniquely when $z = 0$. Applying A to any minimizer eliminates z , proving the expressions for the fitted output and residual. The residual vanishes exactly when $AA^+b = b$, which proves the consistency statement. If $A = 0$, every vector has residual b ; the unique vector with smallest norm is zero, and the same formulas hold. ■

When A is square and invertible, all singular values are positive and both projectors in (2.17) are identity matrices. Thus $A^+ = A^{-1}$ and Theorem 5 reduces to Theorem 1. A pseudoinverse is a genuine extension of the inverse, with an additional choice of solution when uniqueness is lost.

2.4 Useful formulas under rank assumptions

Theorem 6 (Full-rank formulas). For a real matrix $A \in \mathbb{R}^{m \times n}$:

1. If A has full column rank n (so $m \geq n$), then

$$A^+ = (A^T A)^{-1} A^T, \quad A^+ A = \mathbb{1}_n. \quad (2.33)$$

2. If A has full row rank m (so $m \leq n$), then

$$A^+ = A^T (A A^T)^{-1}, \quad A A^+ = \mathbb{1}_m. \quad (2.34)$$

Proof. Full column rank. Here $r = n$. Let us prove that $A^\top A$ is invertible. The transpose of $A^\top A$ is itself, so it is symmetric. If $\mathbf{z} \neq 0$, full column rank gives $A\mathbf{z} \neq 0$, and therefore

$$\mathbf{z}^\top A^\top A \mathbf{z} = (A\mathbf{z})^\top (A\mathbf{z}) = \|A\mathbf{z}\|_2^2 > 0. \quad (2.35)$$

This is positive definiteness. If $A^\top A$ had a nonzero null-space vector, substituting that vector into the left side would instead give zero, a contradiction. Thus $A^\top A$ is invertible. This means that the proposed formula is defined. In the SVD, V_r is now an $n \times n$ orthogonal matrix. Multiplication gives

$$A^\top A = V_r \Sigma_r U_r^\top U_r \Sigma_r V_r^\top = V_r \Sigma_r^2 V_r^\top. \quad (2.36)$$

Because $V_r^\top V_r = V_r V_r^\top = \mathbb{1}_n$, direct multiplication verifies that its inverse is $V_r \Sigma_r^{-2} V_r^\top$. Consequently

$$\begin{aligned} (A^\top A)^{-1} A^\top &= (V_r \Sigma_r^{-2} V_r^\top) (V_r \Sigma_r U_r^\top) \\ &= V_r \Sigma_r^{-2} \Sigma_r U_r^\top = V_r \Sigma_r^{-1} U_r^\top = A^+. \end{aligned}$$

Multiplying the formula on the right by A yields $A^+ A = (A^\top A)^{-1} (A^\top A) = \mathbb{1}_n$, so it is a left inverse.

Full row rank. Here $r = m$. The rows of A are independent, so $A^\top$ has trivial null space. For nonzero $\mathbf{y} \in \mathbb{R}^m$,

$$\mathbf{y}^\top A A^\top \mathbf{y} = \|A^\top \mathbf{y}\|_2^2 > 0. \quad (2.37)$$

Thus $AA^\top$ is symmetric positive definite and invertible. In the SVD, U_r is now square orthogonal, and

$$AA^\top = U_r \Sigma_r^2 U_r^\top, \quad (AA^\top)^{-1} = U_r \Sigma_r^{-2} U_r^\top. \quad (2.38)$$

It follows that

$$\begin{aligned} A^\top (AA^\top)^{-1} &= (V_r \Sigma_r U_r^\top) (U_r \Sigma_r^{-2} U_r^\top) \\ &= V_r \Sigma_r \Sigma_r^{-2} U_r^\top = V_r \Sigma_r^{-1} U_r^\top = A^+. \end{aligned}$$

Multiplication on the left gives $AA^+ = (AA^\top)(AA^\top)^{-1} = \mathbb{1}_m$, so this formula is a right inverse. ■

The hypotheses matter. If the columns are dependent, $A^\top A$ is singular. If the rows are dependent, $AA^\top$ is singular. The SVD formula remains valid in both cases. Also, for a tall full-column-rank matrix, $A^+ A = \mathbb{1}_n$ does not imply $AA^+ = \mathbb{1}_m$; the latter is only a projector. For a wide full-row-rank matrix, the roles are reversed.

There is a second useful derivation of the full-row-rank solution. Any minimum-norm exact solution must be orthogonal to $\text{Null}(A)$: otherwise removing its nonzero null-space component would preserve the equations and decrease its norm. It therefore lies in $\text{Col}(A^\top)$ and can be written $\mathbf{x} = A^\top \mathbf{y}$. The equation becomes

$$AA^\top \mathbf{y} = \mathbf{b}, \quad (2.39)$$

so $\mathbf{y} = (AA^\top)^{-1} \mathbf{b}$ and $\mathbf{x} = A^\top (AA^\top)^{-1} \mathbf{b}$. Conversely, this vector solves the equations and lies in the row space; adding any nonzero null-space vector increases its squared norm. This verifies the minimum-norm property directly, without confusing it with the choice of setting free variables to zero.

For the one-row matrix $A = (1 \ 1)$, we have $AA^\top = (2)$ and

$$A^+ = \frac{1}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \quad (2.40)$$

For $b = 2$, this recovers $(1, 1)^\top$ from the first of the three motivational systems. The projector A^+A is $\frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$, while AA^+ is the 1×1 identity $\mathbb{1}_1$.

Example 4 (Fitting a line to three measurements: the solution operator). Revisit the fit $y \approx \alpha + \beta t$ to the points $(0, 1), (1, 2), (2, 2)$. Then

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}, \quad A^\top A = \begin{pmatrix} 3 & 3 \\ 3 & 5 \end{pmatrix}, \quad A^\top \mathbf{b} = \begin{pmatrix} 5 \\ 6 \end{pmatrix}. \quad (2.41)$$

Solving the 2×2 normal equations gives $\alpha = \frac{7}{6}, \beta = \frac{1}{2}$. The fitted values and residual are

$$A\mathbf{x} = \begin{pmatrix} 7/6 \\ 5/3 \\ 13/6 \end{pmatrix}, \quad \mathbf{r} = \begin{pmatrix} -1/6 \\ 1/3 \\ -1/6 \end{pmatrix}, \quad \|\mathbf{r}\|_2^2 = \frac{1}{6}. \quad (2.42)$$

Check that the entries of $\mathbf{r}$ sum to zero and that $0(-1/6) + 1(1/3) + 2(-1/6) = 0$. These are exactly the two equations $A^\top \mathbf{r} = 0$; the fit satisfies least-squares optimality even though its residual is not zero.

Now compute the entire pseudoinverse and the data-space projector, so that the same fit can be reused for other measurements at $t = 0, 1, 2$. Since the columns are independent,

$$\begin{aligned} (A^\top A)^{-1} &= \frac{1}{6} \begin{pmatrix} 5 & -3 \\ -3 & 3 \end{pmatrix}, \\ A^+ &= \frac{1}{6} \begin{pmatrix} 5 & 2 & -1 \\ -3 & 0 & 3 \end{pmatrix}, \\ AA^+ &= \frac{1}{6} \begin{pmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{pmatrix}. \end{aligned}$$

Thus every data vector $(b_1, b_2, b_3)^\top$ is assigned the line coefficients

$$\alpha = \frac{5b_1 + 2b_2 - b_3}{6}, \quad \beta = \frac{-b_1 + b_3}{2}. \quad (2.43)$$

For $(1, 2, 2)^\top$ these formulas reproduce $\alpha = 7/6$ and $\beta = 1/2$. The vector $(1, -2, 1)^\top$ is perpendicular to both columns of A , and direct subtraction gives

$$\mathbb{1}_3 - AA^+ = \frac{1}{6} \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} \begin{pmatrix} 1 & -2 & 1 \end{pmatrix}. \quad (2.44)$$

Applying this complementary projector to $(1, 2, 2)^\top$ gives $(-1/6, 1/3, -1/6)^\top$, the same residual obtained by subtraction. This is a complete operator-level version of the earlier normal-equation fit: A^+ returns its coefficients, AA^+ returns the fitted data, and $\mathbb{1}_3 - AA^+$ returns the unattainable data component.

Example 5 (An underdetermined system and the shortest exact solution). Consider

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \quad (2.45)$$

The equations are $x_1 + x_3 = 1$ and $x_2 + x_3 = 1$. Setting $x_3 = t$ gives all exact solutions:

$$\mathbf{x}(t) = \begin{pmatrix} 1-t \\ 1-t \\ t \end{pmatrix}, \quad t \in \mathbb{R}. \quad (2.46)$$

Setting the free variable to zero gives $(1, 1, 0)^\top$, but there is no reason for that choice to have smallest norm. In fact,

$$\|\mathbf{x}(t)\|_2^2 = 2(1-t)^2 + t^2 = 3\left(t - \frac{2}{3}\right)^2 + \frac{2}{3}. \quad (2.47)$$

The minimum occurs at $t = 2/3$, giving $\mathbf{x}_* = (1/3, 1/3, 2/3)^\top$.

The rows are independent, so the full-row-rank formula gives the same answer:

$$AA^\top = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad (AA^\top)^{-1} = \frac{1}{3} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}, \quad (2.48)$$

$$A^+ = \frac{1}{3} \begin{pmatrix} 2 & -1 \\ -1 & 2 \\ 1 & 1 \end{pmatrix}, \quad A^+\mathbf{b} = \begin{pmatrix} 1/3 \\ 1/3 \\ 2/3 \end{pmatrix}. \quad (2.49)$$

The null space is spanned by $(-1, -1, 1)^\top$. Its inner product with $\mathbf{x}_*$ is $-1/3 - 1/3 + 2/3 = 0$, verifying that the minimum-norm solution has no null-space component. The squared norm is $2/3$, compared with 2 for the basic solution obtained by setting the free variable to zero.

Example 6 (A singular matrix with consistent and inconsistent data). Let

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}. \quad (2.50)$$

Its second row is twice its first, so its rank is one. A SVD is

$$\mathbf{u}_1 = \mathbf{v}_1 = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \sigma_1 = 5, \quad A = 5\mathbf{u}_1\mathbf{v}_1^\top. \quad (2.51)$$

Consequently,

$$A^+ = \frac{1}{5}\mathbf{v}_1\mathbf{u}_1^\top = \frac{1}{25} \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}. \quad (2.52)$$

Neither $A^\top A$ nor $AA^\top$ is invertible, so the special full-rank formulas cannot be used.

For $\mathbf{b} = (3, 6)^\top$, the system is consistent and the pseudoinverse gives

$$\mathbf{x}_* = A^+ \mathbf{b} = \begin{pmatrix} 3/5 \\ 6/5 \end{pmatrix}, \quad A\mathbf{x}_* = \begin{pmatrix} 3 \\ 6 \end{pmatrix}. \quad (2.53)$$

Every exact solution is $\mathbf{x}_* + t(-2, 1)^\top$, with $t \in \mathbb{R}$. Because $\mathbf{x}_*^\top(-2, 1)^\top = 0$, its squared norm is increased by exactly $5t^2$ when a null-space component is added.

Now change the data to $\mathbf{b} = (3, 5)^\top$. Elimination produces a zero row with right-hand side -1 , so an exact solution is impossible. Nevertheless,

$$\mathbf{x}_* = A^+ \mathbf{b} = \frac{1}{25} \begin{pmatrix} 13 \\ 26 \end{pmatrix}, \quad A\mathbf{x}_* = \begin{pmatrix} 13/5 \\ 26/5 \end{pmatrix}, \quad \mathbf{r}_* = \begin{pmatrix} 2/5 \\ -1/5 \end{pmatrix}. \quad (2.54)$$

The residual is orthogonal to $(1, 2)^\top$, the direction spanning the column space. Its squared norm is $1/5$, the smallest achievable value. All least-squares minimizers are again $\mathbf{x}_* + t(-2, 1)^\top$, and $t = 0$ selects the unique one with smallest norm. The same matrix A^+ handles both data vectors; consistency determines whether the minimum residual is zero.

Example 7 (A singular matrix: exact solutions and best fits). Return to $A = \begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix}$. It has rank one and the SVD

$$\mathbf{u}_1 = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \mathbf{v}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \sigma_1 = \sqrt{10}. \quad (2.55)$$

Consequently

$$A^+ = \frac{1}{10} \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}, \quad AA^+ = \frac{1}{5} \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}, \quad A^+A = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}. \quad (2.56)$$

For the consistent right-hand side $\mathbf{b} = (3, 6)^\top$, $\mathbf{x}_* = (\frac{3}{2}, \frac{3}{2})^\top$, and every exact solution is $\mathbf{x}_* + t(1, -1)^\top$. For the inconsistent right-hand side $\mathbf{b} = (1, 0)^\top$,

$$\mathbf{x}_* = \begin{pmatrix} 1/10 \\ 1/10 \end{pmatrix}, \quad A\mathbf{x}_* = \begin{pmatrix} 1/5 \\ 2/5 \end{pmatrix}, \quad \mathbf{r}_* = \begin{pmatrix} 4/5 \\ -2/5 \end{pmatrix}. \quad (2.57)$$

The residual is perpendicular to $(1, 2)^\top$, and $\|\mathbf{r}_*\|_2^2 = 4/5$, agreeing with the earlier scalar minimization. Adding $t(1, -1)^\top$ keeps this residual but increases the squared solution norm by $2t^2$.

3 Computing generalized solutions reliably

The preceding formulas describe exact mathematical objects. Numerical implementations must also decide which matrix directions can be reliably resolved. This section connects the geometry of the pseudoinverse to conditioning, floating-point calculations, and practical software choices.

Once a SVD is available, compute A^+b in three stages:

$$c = U_r^\top b, \quad d_j = c_j / \sigma_j \ (j = 1, \dots, r), \quad x = V_r d. \quad (3.1)$$

The cost is $O(mr + nr)$ operations for one right-hand side, excluding the factorization. Explicitly forming A^+ costs an additional dense matrix product and stores nm entries. The factorized application is therefore analogous to solving with LU factors instead of first constructing an ordinary inverse.

For a general dense $m \times n$ matrix, standard algorithms for an economy-size SVD have cost $O(mn \min(m, n))$: $O(mn^2)$ for a tall matrix and $O(nm^2)$ for a wide one. These statements concern the usual factorization without unnecessarily constructing full square orthogonal factors. Requested factors, implementation details, and numerical accuracy affect constants. Although QR and SVD can have the same asymptotic order for a given shape, the SVD generally requires more arithmetic; it also provides singular values and more direct control over rank decisions [6].

3.1 Small singular values amplify errors

Theorem 7 (Sensitivity to a change in the right-hand side). Fix $A \in \mathbb{R}^{m \times n}$ with rank $r > 0$ and smallest positive singular value σ_r . If $x_* = A^+b$ and $\tilde{x}_* = A^+(b + \delta b)$, then

$$\delta x := \tilde{x}_* - x_* = \sum_{j=1}^r \frac{u_j^\top \delta b}{\sigma_j} v_j, \quad \|\delta x\|_2 \leq \frac{\|\delta b\|_2}{\sigma_r}. \quad (3.2)$$

The bound is sharp: $\|A^+\|_2 = 1/\sigma_r$. For $A = 0$, every minimum-norm solution is zero and the change is zero.

Proof. Linearity gives $\delta x = A^+ \delta b$, and substituting the SVD formula gives the displayed sum. Since the right singular vectors are orthonormal,

$$\|\delta x\|_2^2 = \sum_{j=1}^r \frac{|u_j^\top \delta b|^2}{\sigma_j^2} \leq \frac{1}{\sigma_r^2} \sum_{j=1}^r |u_j^\top \delta b|^2 \leq \frac{1}{\sigma_r^2} \|\delta b\|_2^2. \quad (3.3)$$

The first inequality uses $\sigma_j \geq \sigma_r$; the second is Pythagoras after decomposing δb into its column-space component and perpendicular remainder. Taking square roots gives the bound. It is attained for the unit data perturbation $\delta b = u_r$, whose response is v_r/σ_r and has norm $1/\sigma_r$. This also proves the asserted operator norm. For $A = 0$, the definition $A^+ = 0$ gives the conclusion directly. ■

This absolute error bound applies to singular matrices as well, provided A is held fixed. An error along u_j is multiplied by $1/\sigma_j$; an error perpendicular to the column space does not change A^+b at all, although it changes the residual. Large responses to errors in weak but retained directions can reflect sensitivity of the mathematical problem rather than a defect of the algorithm. Perturbations in A itself are more delicate because they can change its rank and singular subspaces.

Rank itself is discontinuous under perturbations. For example, $\text{diag}(1, \varepsilon)$ has rank two for every $\varepsilon \neq 0$, but rank one when $\varepsilon = 0$. Its second inverse entry is $1/\varepsilon$ before the rank drops, whereas the corresponding pseudoinverse entry at zero is zero. Consequently, deciding whether a tiny computed singular value represents meaningful information requires a tolerance.

3.2 Numerical rank and truncated SVD

Choose a nonnegative threshold τ . For example, a scale-aware rule is $\tau = \tau_{\text{abs}} + \tau_{\text{rel}}\sigma_1$, where the absolute threshold has the same units as the singular values and the relative threshold is dimensionless. Define

$$\begin{aligned} r_\tau &= \#\{j \in \{1, \dots, r\} : \sigma_j > \tau\}, \\ A_\tau &= \sum_{\substack{j=1 \\ \sigma_j > \tau}}^r \sigma_j \mathbf{u}_j \mathbf{v}_j^\top, \\ \mathbf{x}_\tau &= A_\tau^+ \mathbf{b} = \sum_{\substack{j=1 \\ \sigma_j > \tau}}^r \frac{\mathbf{u}_j^\top \mathbf{b}}{\sigma_j} \mathbf{v}_j. \end{aligned}$$

The count r_τ is a numerical rank at the selected threshold. The truncated SVD solves a modified problem in which weak singular directions are discarded. It is the exact minimum-norm least-squares solution for A_τ , but generally not for the original A .

Proposition 5 (What truncation changes). Use the exact SVD of A and a threshold $\tau \geq 0$. Let $J = \{j \in \{1, \dots, r\} : \sigma_j > \tau\}$ and set $\beta_j = \mathbf{u}_j^\top \mathbf{b}$. The truncated vector $\mathbf{x}_\tau$ is the unique minimum-norm least-squares solution for A_τ . Its residual for the original matrix A satisfies

$$\|\mathbf{b} - A\mathbf{x}_\tau\|_2^2 = \|\mathbf{b}_\perp\|_2^2 + \sum_{\substack{j=1 \\ j \notin J}}^r \beta_j^2, \quad (3.4)$$

where the omitted indices range over the positive singular values of A . Thus it is also a least-squares solution for the original A if and only if every discarded coefficient β_j is zero. If at least one direction is retained, the norm of the truncated inverse is $1/\min_{j \in J} \sigma_j$; if no direction is retained, that operator is zero.

Proof. The retained terms form a SVD of A_τ , so Theorem 5 gives its minimum-norm solution as

$$\mathbf{x}_\tau = \sum_{\substack{j=1 \\ j \in J}}^r \frac{\beta_j}{\sigma_j} \mathbf{v}_j. \quad (3.5)$$

Applying the original A gives $A\mathbf{x}_\tau = \sum_{j \in J} \beta_j \mathbf{u}_j$, since $A\mathbf{v}_j = \sigma_j \mathbf{u}_j$. Consequently

$$\mathbf{b} - A\mathbf{x}_\tau = \mathbf{b}_\perp + \sum_{\substack{j=1 \\ j \notin J}}^r \beta_j \mathbf{u}_j. \quad (3.6)$$

These summands are orthogonal, proving (3.4). The exact minimum residual for A has squared norm $\|\mathbf{b}_\perp\|_2^2$. Equality with the truncated residual therefore holds exactly when every discarded scalar coefficient is zero. The operator-norm statement follows by the same orthonormal calculation as Theorem 7, now using only retained singular values. An empty sum gives the zero operator and the zero solution, covering the case J is empty. ■

A commonly used scale for roundoff-level rank decisions is

$$\tau = \max(m, n) \epsilon_{\text{mach}} \sigma_1, \quad (3.7)$$

where ϵ_{mach} denotes the spacing immediately above one in the working precision. This is a convention for a numerical decision, not a theorem that smaller singular values of the exact matrix vanish. A threshold based on measurement error, units, or modeling uncertainty can be more appropriate. If no singular values are retained, the output is the zero vector.

Example 8 (A tiny singular value is not the same as zero). Let $A = \text{diag}(1, 10^{-8})$ and $\mathbf{b} = (1, 10^{-8})^\top$. The exact solution is $(1, 1)^\top$. Changing the second data entry by 10^{-8} changes the second solution entry by one. The inverse amplifies that perturbation by 10^8 . With a chosen threshold $\tau = 10^{-6}$, the truncated SVD instead gives $\mathbf{x}_\tau = (1, 0)^\top$ and residual $(0, 10^{-8})^\top$. The small residual does not imply a small solution error. The cutoff makes a deliberate approximation that may be appropriate for noisy data; it does not discover that this mathematically invertible matrix is exactly singular.

Example 9 (Truncation trades data fit for reduced noise amplification). Let

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 10^{-8} \end{pmatrix}, \quad \mathbf{x}_{\text{true}} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \mathbf{b}_{\text{true}} = \begin{pmatrix} 1 \\ 10^{-8} \end{pmatrix}. \quad (3.8)$$

Suppose the measured right-hand side has an additive error $(0, 10^{-6})^\top$. Then

$$\mathbf{b}_{\text{obs}} = \begin{pmatrix} 1 \\ 1.01 \times 10^{-6} \end{pmatrix}, \quad A^{-1} \mathbf{b}_{\text{obs}} = \begin{pmatrix} 1 \\ 101 \end{pmatrix}. \quad (3.9)$$

This solution fits the observed data exactly, but its error relative to $\mathbf{x}_{\text{true}}$ has norm 100. With threshold $\tau = 10^{-7}$, the second singular value is discarded and

$$\mathbf{x}_\tau = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \|\mathbf{b}_{\text{obs}} - A\mathbf{x}_\tau\|_2 = 1.01 \times 10^{-6}, \quad \|\mathbf{x}_\tau - \mathbf{x}_{\text{true}}\|_2 = 1. \quad (3.10)$$

Truncation improves the reconstruction in this example by declining to fit the data along a direction dominated by noise. It also introduces bias: the true second coordinate is not recovered. Its usefulness depends on the noise scale and on what information can reasonably be recovered from the data. A tolerance should reflect those considerations, not just a desire for a smaller solution norm.

There are two distinct sources of a rank threshold: floating-point uncertainty in the matrix calculation and measurement or modeling uncertainty in the data. A threshold based only on machine precision addresses the former. It need not be appropriate when the data are substantially noisier than floating-point arithmetic.

4 Error analysis

4.1 Residual, forward error, and a backward-error interpretation

Let $\hat{\mathbf{x}}$ be a computed approximation to the exact solution $\mathbf{x}$ of a nonsingular system. Its *residual* is

$$\mathbf{r} = \mathbf{b} - A\hat{\mathbf{x}}. \quad (4.1)$$

The *forward error* is $\hat{\mathbf{x}} - \mathbf{x}$, which is usually unknown because the exact solution is unknown. These quantities are related but are not interchangeable.

For a dense square matrix, explicitly computing $A\hat{\mathbf{x}}$ uses $2n^2 - n$ flops and subtracting it from $\mathbf{b}$ uses n more, so a residual costs $2n^2$ flops under our convention. It is inexpensive compared with a cubic factorization and should use the original data, not the overwritten factor storage. The identity

$$A\hat{\mathbf{x}} = \mathbf{b} - \mathbf{r} \quad (4.2)$$

says that the computed vector is the exact solution for a perturbed right-hand side. Thus $\|\mathbf{r}\|/\|\mathbf{b}\|$, for $\mathbf{b} \neq 0$, is a relative backward error if A is held fixed and perturbations to $\mathbf{b}$ alone are permitted.

A useful scaled residual, using the Euclidean vector norm and its induced matrix norm, is

$$\eta = \frac{\|\mathbf{b} - A\hat{\mathbf{x}}\|_2}{\|A\|_2\|\hat{\mathbf{x}}\|_2 + \|\mathbf{b}\|_2}. \quad (4.3)$$

Theorem 8 (Exact interpretation of the scaled residual). Let $A \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, and $\hat{\mathbf{x}} \in \mathbb{R}^n$. Define $d = \|A\|_2\|\hat{\mathbf{x}}\|_2 + \|\mathbf{b}\|_2$. If $d > 0$, the value $\eta = \|\mathbf{b} - A\hat{\mathbf{x}}\|_2/d$ is the smallest $\epsilon \geq 0$ for which perturbations can satisfy

$$(A + \Delta A)\hat{\mathbf{x}} = \mathbf{b} + \Delta \mathbf{b}, \quad \|\Delta A\|_2 \leq \epsilon \|A\|_2, \quad \|\Delta \mathbf{b}\|_2 \leq \epsilon \|\mathbf{b}\|_2. \quad (4.4)$$

This statement uses the spectral matrix norm and the Euclidean vector norm. If $d = 0$, the residual is zero and we define $\eta = 0$.

Proof. Write $\mathbf{r} = \mathbf{b} - A\hat{\mathbf{x}}$. Any admissible perturbations obey $\mathbf{r} = \Delta A\hat{\mathbf{x}} - \Delta \mathbf{b}$. The triangle inequality and the induced-norm inequality give

$$\|\mathbf{r}\|_2 \leq \|\Delta A\|_2\|\hat{\mathbf{x}}\|_2 + \|\Delta \mathbf{b}\|_2 \leq \epsilon d. \quad (4.5)$$

For $d > 0$, every admissible ϵ is therefore at least η .

To show that the bound can be attained, first suppose $\hat{\mathbf{x}} \neq 0$ and choose

$$\Delta A = \frac{\|A\|_2}{d} \frac{\mathbf{r}\hat{\mathbf{x}}^\top}{\|\hat{\mathbf{x}}\|_2}, \quad \Delta \mathbf{b} = -\frac{\|\mathbf{b}\|_2}{d} \mathbf{r}. \quad (4.6)$$

The spectral norm of an outer product $\mathbf{u}\mathbf{v}^\top$ equals $\|\mathbf{u}\|_2\|\mathbf{v}\|_2$: Cauchy–Schwarz bounds its action on a unit vector, and equality is attained in the direction of $\mathbf{v}$ when it is nonzero. Thus these choices have norms $\eta\|A\|_2$ and $\eta\|\mathbf{b}\|_2$. Furthermore,

$$\Delta A\hat{\mathbf{x}} - \Delta \mathbf{b} = \frac{\|A\|_2\|\hat{\mathbf{x}}\|_2 + \|\mathbf{b}\|_2}{d} \mathbf{r} = \mathbf{r}, \quad (4.7)$$

so the perturbed equation holds. If $\hat{\mathbf{x}} = 0$ and $d > 0$, then $\mathbf{b} \neq 0$ and $\eta = 1$; choose $\Delta A = 0$ and $\Delta \mathbf{b} = -\mathbf{b}$. Finally, $d = 0$ implies $\mathbf{b} = 0$ and $\|A\|_2\|\hat{\mathbf{x}}\|_2 = 0$, hence $A\hat{\mathbf{x}} = 0$. No perturbation is needed, so the defined value $\eta = 0$ is appropriate. ■

A similarly scaled residual can be a useful diagnostic in other induced norms. The explicit attaining construction above is written for the Euclidean/spectral pair. Residual computation itself incurs rounding; using higher precision for that calculation can be valuable when very small backward errors matter [3].

4.2 Conditioning measures sensitivity of the problem

Definition 6 (Condition number). For a nonsingular square matrix and an induced matrix norm, define

$$\kappa(A) = \|A\| \|A^{-1}\|. \quad (4.8)$$

In the spectral norm, $\kappa_2(A) = \sigma_{\max}(A)/\sigma_{\min}(A)$, where the numerator and denominator are the largest and smallest singular values. A large condition number means that some small relative perturbations to the data can cause much larger relative perturbations to the solution.

The condition number is at least one because $1 = \|\mathbb{1}\| = \|AA^{-1}\| \leq \|A\| \|A^{-1}\|$. Multiplying the entire matrix by a nonzero scalar does not change it: the scalar factors in $\|cA\| \|(cA)^{-1}\|$ cancel. This is one reason conditioning is more informative than an unscaled determinant or an absolute pivot threshold. Its appearance in a formula does not require us to form an inverse numerically; practical condition estimators can work with existing factors.

Theorem 9 (Residual-to-error bound). Let $A \in \mathbb{R}^{n \times n}$ be nonsingular, let $\mathbf{b} \neq 0$, and let $\mathbf{x} = A^{-1}\mathbf{b}$. For any approximation $\hat{\mathbf{x}} \in \mathbb{R}^n$ with exact residual $\mathbf{r} = \mathbf{b} - A\hat{\mathbf{x}}$, every vector norm and its induced matrix norm satisfy

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \kappa(A) \frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}. \quad (4.9)$$

Proof. The hypotheses imply $\mathbf{x} \neq 0$, so both denominators are nonzero. Subtract the equations defining the residual and the exact solution: $A(\mathbf{x} - \hat{\mathbf{x}}) = \mathbf{r}$. Applying A^{-1} and using the induced-norm inequality gives

$$\|\mathbf{x} - \hat{\mathbf{x}}\| \leq \|A^{-1}\| \|\mathbf{r}\|. \quad (4.10)$$

The exact equation also gives $\|\mathbf{b}\| \leq \|A\| \|\mathbf{x}\|$, or $1/\|\mathbf{x}\| \leq \|A\|/\|\mathbf{b}\|$. Multiplying the first bound by this upper bound on $1/\|\mathbf{x}\|$ gives the claimed relative inequality. No property of the algorithm producing $\hat{\mathbf{x}}$ was used; this is a statement about the problem and the approximation itself. ■

Example 10 (A tiny residual with a large solution error). Let

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 10^{-8} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ 10^{-8} \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \hat{\mathbf{x}} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \quad (4.11)$$

The residual is $(0, 10^{-8})^\top$, so its relative norm is about 10^{-8} . Nevertheless, $\|\hat{\mathbf{x}} - \mathbf{x}\|_2 / \|\mathbf{x}\|_2 = 1/\sqrt{2}$. Here $\kappa_2(A) = 10^8$. The small coefficient in the second equation makes a large error in x_2 almost invisible to the residual. A residual must be interpreted together with the sensitivity of the problem.

For the same diagonal example, the infinity norm gives an especially transparent equality: the relative solution error is 1, the relative residual is 10^{-8} , and $\kappa_\infty(A) = 10^8$. The right side of (4.9) is exactly 1. This supplements the Euclidean calculation in the example and shows why all norms in a particular bound must be chosen consistently.

Theorem 10 (Simultaneous perturbations of the matrix and data). Let A be nonsingular, $\mathbf{b} \neq 0$, and $A\mathbf{x} = \mathbf{b}$. Use a vector norm and its induced matrix norm, and define

$$\epsilon_A = \frac{\|\Delta A\|}{\|A\|}, \quad \epsilon_b = \frac{\|\Delta \mathbf{b}\|}{\|\mathbf{b}\|}. \quad (4.12)$$

If $\kappa(A)\epsilon_A < 1$, then $A + \Delta A$ is nonsingular and the unique perturbed solution $\mathbf{x} + \Delta \mathbf{x}$ satisfies

$$\frac{\|\Delta \mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\kappa(A)}{1 - \kappa(A)\epsilon_A} (\epsilon_A + \epsilon_b). \quad (4.13)$$

Proof. First, $\|A^{-1}\Delta A\| \leq \kappa(A)\epsilon_A < 1$. If $(A + \Delta A)\mathbf{z} = 0$, then $\mathbf{z} = -A^{-1}\Delta A\mathbf{z}$, so $\|\mathbf{z}\| \leq \|A^{-1}\Delta A\|\|\mathbf{z}\|$. The strict bound forces $\mathbf{z} = 0$. A square matrix with trivial nullspace is invertible; hence the perturbed solution exists uniquely.

Expand its equation and subtract $A\mathbf{x} = \mathbf{b}$:

$$A\Delta \mathbf{x} = \Delta \mathbf{b} - \Delta A\mathbf{x} - \Delta A\Delta \mathbf{x}. \quad (4.14)$$

Apply A^{-1} and the norm inequalities, and then collect the term containing $\|\Delta \mathbf{x}\|$ on the left:

$$(1 - \|A^{-1}\|\|\Delta A\|)\|\Delta \mathbf{x}\| \leq \|A^{-1}\|(\|\Delta \mathbf{b}\| + \|\Delta A\|\|\mathbf{x}\|). \quad (4.15)$$

The coefficient on the left is positive. Divide by it and by $\|\mathbf{x}\| > 0$. Finally use $\|\mathbf{b}\|/\|\mathbf{x}\| \leq \|A\|$ and substitute the definitions of ϵ_A , ϵ_b , and $\kappa(A)$. This gives (4.13). ■

For matrix perturbations small enough that the denominator is near one, this is approximately a condition-number multiple of the relative data perturbation. When that denominator approaches zero, no such small-error guarantee survives. This explains the separate roles of problem sensitivity and the size of the numerical perturbation.

A Linear algebra background

A.1 Inner products, lengths, and matrix norms

The Euclidean inner product of two vectors of length n is

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{j=1}^n x_j y_j. \quad (A.1)$$

It is symmetric, linear in each argument, and satisfies $\mathbf{x}^\top \mathbf{x} \geq 0$ with equality only for $\mathbf{x} = 0$. Two vectors are *orthogonal* if their inner product is zero. Their Euclidean length is $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^\top \mathbf{x}}$; a vector with length one is a unit vector. A family is *orthonormal* when its vectors are unit vectors and distinct members are orthogonal.

Proposition 7 (Cauchy–Schwarz and Pythagoras). For vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$|\mathbf{x}^\top \mathbf{y}| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2. \quad (\text{A.2})$$

If the vectors are orthogonal, then

$$\|\mathbf{x} + \mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2. \quad (\text{A.3})$$

Proof. If $\mathbf{y} = 0$, the first inequality holds with both sides zero. Otherwise set $c = (\mathbf{x}^\top \mathbf{y})/(\mathbf{y}^\top \mathbf{y})$. Expanding a nonnegative squared length gives

$$0 \leq \|\mathbf{x} - c\mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2 - \frac{(\mathbf{x}^\top \mathbf{y})^2}{\|\mathbf{y}\|_2^2}. \quad (\text{A.4})$$

Multiply by the positive number $\|\mathbf{y}\|_2^2$ and take square roots. For the second claim, expand $\|\mathbf{x} + \mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2 + 2\mathbf{x}^\top \mathbf{y} + \|\mathbf{y}\|_2^2$; orthogonality removes the cross term. ■

A *norm* is a function measuring vector size. It is nonnegative and vanishes only at zero, satisfies $\|c\mathbf{x}\| = |c|\|\mathbf{x}\|$, and obeys the triangle inequality $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$. The norms used in this note are

$$\|\mathbf{x}\|_1 = \sum_{j=1}^n |x_j|, \quad \|\mathbf{x}\|_2 = \left(\sum_{j=1}^n |x_j|^2 \right)^{1/2}, \quad \|\mathbf{x}\|_\infty = \max_{1 \leq j \leq n} |x_j|. \quad (\text{A.5})$$

Nonnegativity and homogeneity follow directly from the definitions. The triangle inequality for the 1- and infinity norms follows by applying $|x_j + y_j| \leq |x_j| + |y_j|$ in each coordinate. For the 2-norm, Cauchy–Schwarz bounds the cross term in the preceding expansion by $2\|\mathbf{x}\|_2 \|\mathbf{y}\|_2$, yielding $\|\mathbf{x} + \mathbf{y}\|_2^2 \leq (\|\mathbf{x}\|_2 + \|\mathbf{y}\|_2)^2$.

Definition 8 (Induced matrix norms). For $p \in \{1, 2, \infty\}$ and $A \in \mathbb{R}^{m \times n}$, define

$$\|A\|_p = \max_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_p} = \max_{\|\mathbf{x}\|_p=1} \|A\mathbf{x}\|_p. \quad (\text{A.6})$$

This is the largest amplification of a vector by A . The input norm is on $\mathbb{R}^n$, and the output norm of the same type is on $\mathbb{R}^m$. The Frobenius norm is a different matrix norm:

$$\|A\|_F = \left(\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2 \right)^{1/2}. \quad (\text{A.7})$$

It treats the entries as one long vector.

The two maxima in the induced-norm definition agree because rescaling a nonzero vector to unit length does not change the ratio. They are attained: a continuous function on the finite-dimensional compact unit sphere attains its maximum. We use this familiar extreme-value fact again in the spectral theorem proof below.

Proposition 9 (Entrywise formulas and norm inequalities). For a real $m \times n$ matrix,

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^m |a_{ij}|, \quad \|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}|. \quad (\text{A.8})$$

For conformable matrices and vectors and $p \in \{1, 2, \infty\}$,

$$\|A\mathbf{x}\|_p \leq \|A\|_p \|\mathbf{x}\|_p, \quad \|AB\|_p \leq \|A\|_p \|B\|_p. \quad (\text{A.9})$$

The induced norms are matrix norms, and $\|AB\|_F \leq \|A\|_F \|B\|_F$.

Proof. For the 1-norm, sum the componentwise triangle inequalities:

$$\|A\mathbf{x}\|_1 \leq \sum_{i=1}^m \sum_{j=1}^n |a_{ij}| |x_j| = \sum_{j=1}^n \left(\sum_{i=1}^m |a_{ij}| \right) |x_j| \leq \left(\max_j \sum_{i=1}^m |a_{ij}| \right) \|\mathbf{x}\|_1. \quad (\text{A.10})$$

This proves an upper bound on $\|A\|_1$. Choose $\mathbf{x} = \mathbf{e}_{j_*}$, where column j_* has maximal absolute column sum. It has unit 1-norm, and $A\mathbf{x}$ is that column, so the upper bound is attained.

For the infinity norm, each row satisfies $|(A\mathbf{x})_i| \leq (\sum_{j=1}^n |a_{ij}|) \|\mathbf{x}\|_\infty$. Choose a row i_* of maximal absolute row sum and take $x_j = 1$ when $a_{i_*j} \geq 0$ and $x_j = -1$ when $a_{i_*j} < 0$. Then $\|\mathbf{x}\|_\infty = 1$ and $(A\mathbf{x})_{i_*} = \sum_{j=1}^n |a_{i_*j}|$, proving equality.

The bound on $A\mathbf{x}$ is the definition when $\mathbf{x} \neq 0$ and is trivial when $\mathbf{x} = 0$. Applying it twice gives $\|AB\mathbf{x}\|_p \leq \|A\|_p \|B\|_p \|\mathbf{x}\|_p$; maximize over unit vectors to obtain submultiplicativity. Positivity and homogeneity of the induced norm follow from those of the vector norm. If $A \neq 0$, some column $A\mathbf{e}_j$ is nonzero, so $\|A\|_p > 0$. Finally, $\|(A+B)\mathbf{x}\|_p \leq (\|A\|_p + \|B\|_p) \|\mathbf{x}\|_p$ proves its triangle inequality.

For the Frobenius product inequality, let $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times \ell}$. Apply Cauchy–Schwarz to each row–column dot product:

$$\sum_{i=1}^m \sum_{j=1}^\ell |(AB)_{ij}|^2 \leq \sum_{i=1}^m \sum_{j=1}^\ell \left(\sum_{k=1}^n |a_{ik}|^2 \right) \left(\sum_{k=1}^n |b_{kj}|^2 \right) = \|A\|_F^2 \|B\|_F^2. \quad (\text{A.11})$$

Taking square roots proves the claim. ■

Every induced norm of $\mathbb{1}_n$ is 1, while $\|\mathbb{1}_n\|_F = \sqrt{n}$. Thus for $n > 1$ the Frobenius norm is not induced by any single vector norm on $\mathbb{R}^n$. It remains a useful matrix norm. The SVD will also give $\|A\|_2 \leq \|A\|_F \leq \sqrt{r} \|A\|_2$ for rank r . Different norms measure different aspects of an error, so the norm must be specified when reporting accuracy or a condition number.

A.2 Identity, inverses, and equivalent conditions

A square matrix $A \in \mathbb{R}^{n \times n}$ is *invertible*, or *nonsingular*, if a matrix A^{-1} satisfies $A^{-1}A = AA^{-1} = \mathbb{1}_n$. The inverse is a matrix describing the reverse linear map, not an entrywise reciprocal. Its uniqueness follows directly: if B and C are both inverses, then $B = B(AC) = (BA)C = C$.

Theorem 11 (Equivalent descriptions of invertibility). For $A \in \mathbb{R}^{n \times n}$, the following are equivalent:

1. A is invertible.
2. $\text{rank}(A) = n$.
3. The columns of A are independent.
4. The columns of A form a basis of $\mathbb{R}^n$.

5. $\text{Null}(A) = \{0\}$.
6. $\text{Col}(A) = \mathbb{R}^n$.
7. For every $\mathbf{b} \in \mathbb{R}^n$, $A\mathbf{x} = \mathbf{b}$ has exactly one solution.

In this case that solution is $\mathbf{x} = A^{-1}\mathbf{b}$.

Proof. If A has an inverse and $A\mathbf{x} = 0$, multiplying by A^{-1} gives $\mathbf{x} = 0$. Thus (1) implies (5). Conversely, rank-nullity proves (5) $\iff$ (2), because the number of columns is n . The columns are independent exactly when their only zero linear combination is the trivial one, which is (5); thus (3) $\iff$ (5). There are exactly n columns, so an independent family of this size in $\mathbb{R}^n$ is a basis by the basis facts above. This proves the equivalence with (4). The column space has dimension n precisely when it is all of $\mathbb{R}^n$: a proper subspace could have its basis extended by a vector outside it, so could not already have dimension n . This proves (2) $\iff$ (6).

If (4) holds, every $\mathbf{b}$ has a unique representation in the column basis, which is exactly the unique solution of $A\mathbf{x} = \mathbf{b}$. Thus (4) implies (7). If (7) holds, its instance $\mathbf{b} = 0$ says the nullspace is $\{0\}$, so (7) also implies (5).

To finish the only direction not yet established, suppose (7) holds. For each coordinate vector $\mathbf{e}_j$, let $\mathbf{c}_j$ be its unique solution, and assemble $C = (\mathbf{c}_1 \cdots \mathbf{c}_n)$. Then $AC = \mathbb{1}_n$. For any $\mathbf{x}$, the two vectors $C A \mathbf{x}$ and $\mathbf{x}$ have the same image under A , since $A(C A \mathbf{x}) = (AC) A \mathbf{x} = A \mathbf{x}$. Uniqueness, applied to right-hand side $A \mathbf{x}$, gives $C A \mathbf{x} = \mathbf{x}$ for every vector. In particular this holds for every $\mathbf{e}_j$, so $CA = \mathbb{1}_n$ and C is the inverse. Finally, multiplying $A\mathbf{x} = \mathbf{b}$ by A^{-1} gives the solution formula, and direct substitution verifies it. ■

The square hypothesis is essential. A rectangular matrix can have full row rank without full column rank, or conversely. These cases support a right inverse or a left inverse respectively, but not a two-sided inverse between equal-dimensional spaces.

If A, B are invertible square matrices of the same size, then

$$(AB)^{-1} = B^{-1}A^{-1}, \quad (A^T)^{-1} = (A^{-1})^T. \quad (\text{A.12})$$

For the first identity, multiply AB by $B^{-1}A^{-1}$ in either order; the adjacent inverse pairs cancel to give $\mathbb{1}$. For the second, transpose $AA^{-1} = A^{-1}A = \mathbb{1}$ and use the reversal rule for a transposed product. Uniqueness of the inverse completes both arguments.

Proposition 10 (Invertibility of a triangular matrix). A triangular square matrix is invertible exactly when every diagonal entry is nonzero. In particular, a unit triangular matrix is invertible.

Proof. First let U be upper triangular with nonzero diagonal. For any right-hand side, its last equation determines x_n by division by u_{nn} . Once $x_{i+1}, \dots, x_n$ have been determined, row i determines $x_i = (b_i - \sum_{j=i+1}^n u_{ij}x_j)/u_{ii}$. This gives one and only one solution for every right-hand side, so U is invertible by Theorem 11.

If a diagonal entry is zero, let k be the first such index. Set $x_k = 1$ and $x_j = 0$ for $j > k$. Rows $k, \dots, n$ of $U\mathbf{x} = 0$ are then already satisfied: the k th diagonal coefficient vanishes, and later rows only involve the later zero coordinates. Solve rows $k-1, \dots, 1$ upward; their diagonal entries are nonzero by the choice of k . The result is a nonzero nullspace vector because $x_k = 1$. Thus U is not invertible. A lower triangular matrix has an upper triangular transpose with the same diagonal, and invertibility is unchanged by transposition, proving the lower triangular case. ■

The determinant: an algebraic criterion, with a proof. For a square matrix, the determinant can be defined by the finite sum

$$\det(A) = \sum_{\pi \in S_n} \text{sgn}(\pi) \prod_{i=1}^n a_{i, \pi(i)}, \quad (\text{A.13})$$

where S_n is the finite set of all permutations of $\{1, \dots, n\}$ and $\text{sgn}(\pi)$ is $+1$ for an even permutation and -1 for an odd permutation. This definition makes the determinant linear in any one row when the other rows are fixed. Interchanging two rows changes its sign: reindex the permutations by the same transposition, which changes their signs and preserves the corresponding entry products. Equal rows therefore force the determinant to be its own negative, hence zero. A zero row also makes it zero directly.

Proposition 11 (Determinant and invertibility). For $A \in \mathbb{R}^{n \times n}$, A is invertible if and only if $\det(A) \neq 0$. For a triangular matrix, the determinant is the product of its diagonal entries.

Proof. If U is upper triangular, a nonzero product in the determinant sum would require $\pi(i) \geq i$ for every i . But $\sum_{i=1}^n \pi(i) = \sum_{i=1}^n i$, so all these inequalities must be equalities. Only the identity permutation contributes, yielding $\det(U) = \prod_i u_{ii}$. The same argument with inequalities reversed applies to a lower triangular matrix.

Now row-reduce A to echelon form. A row swap changes the determinant's sign, and scaling a row by $c \neq 0$ multiplies it by c . Adding a multiple of one row to another leaves it unchanged: linearity in the changed row gives the original determinant plus a term with two identical rows, and that second term is zero. Thus these operations preserve whether the determinant vanishes. If the echelon form has rank n , all n columns are pivot columns, so it is upper triangular with nonzero diagonal and has nonzero determinant. If it has rank less than n , it has a zero row and determinant zero. Since the echelon rank is $\text{rank}(A)$, the criterion follows from Theorem 11. ■

Geometrically, $|\det(A)|$ is the volume scaling factor of the square linear map. Its magnitude alone is not a reliable numerical measure of nearness to singularity. For example, $A = \alpha \mathbb{1}_n$ has $\det(A) = \alpha^n$ and condition number one for every nonzero α , because its induced norm is $|\alpha|$ and that of its inverse is $|\alpha|^{-1}$. Conversely, $A = \text{diag}(\varepsilon, \varepsilon^{-1})$ has determinant one but 2-norm condition number ε^{-2} when $0 < \varepsilon < 1$: the largest stretching factor of both A and A^{-1} is ε^{-1} . Numerical rank and sensitivity should be assessed with suitable factorizations and singular values, rather than by comparing a determinant with an arbitrary absolute tolerance.

A.3 Orthonormal coordinates and orthogonal projection

Lemma 12 (Constructing an orthonormal basis). Every subspace of $\mathbb{R}^m$ has an orthonormal basis. An orthonormal family can be extended to an orthonormal basis of $\mathbb{R}^m$.

Proof. For the zero subspace, the empty family is already an orthonormal basis. For a nonzero subspace, start with a basis $\mathbf{v}_1, \dots, \mathbf{v}_r$ of the subspace. Set $\mathbf{q}_1 = \mathbf{v}_1 / \|\mathbf{v}_1\|_2$. If orthonormal vectors $\mathbf{q}_1, \dots, \mathbf{q}_{j-1}$ spanning the first $j-1$ basis vectors have been constructed, set

$$\mathbf{w}_j = \mathbf{v}_j - \sum_{i=1}^{j-1} (\mathbf{q}_i^\top \mathbf{v}_j) \mathbf{q}_i. \quad (\text{A.14})$$

For each $\ell < j$, taking an inner product with $\mathbf{q}_\ell$ gives $\mathbf{q}_\ell^\top \mathbf{w}_j = 0$, because orthonormality cancels exactly the term with $i = \ell$. The vector $\mathbf{w}_j$ is nonzero; otherwise $\mathbf{v}_j$ would be a combination of the preceding basis

vectors. Therefore $q_j = w_j / \|w_j\|_2$ is well defined. The spans agree at every step because each new q_j is a nonzero multiple of v_j plus a combination of earlier vectors. This is the Gram–Schmidt construction.

To extend an already orthonormal family, first extend it to any basis of $\mathbb{R}^m$ by Lemma 13, then apply this construction in that order. The original orthonormal vectors remain unchanged, because all earlier inner products are already zero and their lengths are one. ■

Lemma 13 (Finite-dimensional basis facts). An independent family in a subspace can be extended to a basis of that subspace. Any two bases of a subspace have the same number of vectors. This number is its *dimension*; the zero subspace has dimension zero.

Proof. First suppose $v_1, \dots, v_k$ are independent and $w_1, \dots, w_\ell$ span a subspace containing them. We show $k \leq \ell$ by successive replacement. Express v_1 using the w 's. At least one coefficient is nonzero, since $v_1 \neq 0$; solving for that w lets us replace it by v_1 without changing the span. Suppose i replacements have produced a spanning family consisting of $v_1, \dots, v_i$ and $\ell - i$ remaining w 's. If $i < k$, the expression for v_{i+1} in this family must use a remaining w with nonzero coefficient. Otherwise it would lie in the span of the first i independent vectors, a contradiction. Solve for that w and replace it. This cannot continue after all ℓ positions are filled, so $k \leq \ell$.

Starting with an independent family in S , add a vector of S outside its span whenever one exists. Each addition preserves independence: a dependence with a nonzero coefficient of the new vector would express it using the previous ones. There can be at most n independent vectors in $\mathbb{R}^n$ by the preceding argument with its standard basis. Thus this process terminates with an independent spanning family of S , a basis. Finally, comparing two bases, each as an independent family against the other as a spanning family, gives both inequalities between their sizes. The sizes are equal. ■

For $Q \in \mathbb{R}^{m \times r}$ with orthonormal columns, $Q^T Q = \mathbb{1}_r$. If $c, d \in \mathbb{R}^r$, then

$$(Qc)^T(Qd) = c^T Q^T Q d = c^T d. \quad (\text{A.15})$$

Thus these coordinates preserve inner products and lengths: $\|Qc\|_2 = \|c\|_2$. When $r = m$, Q is called *orthogonal*; its inverse is Q^T and $QQ^T = \mathbb{1}_m$. An orthogonal matrix therefore preserves Euclidean angles as well as lengths. For $r < m$, QQ^T is not $\mathbb{1}_m$, as the next theorem explains.

Theorem 12 (Projection onto the span of orthonormal columns). Let $Q \in \mathbb{R}^{m \times r}$ satisfy $Q^T Q = \mathbb{1}_r$, and let $S = \text{Col}(Q)$. The unique closest vector in S to $b \in \mathbb{R}^m$ in Euclidean norm is

$$p = QQ^T b. \quad (\text{A.16})$$

The residual $r = b - p$ is orthogonal to every vector in S . The matrix $P = QQ^T$ is symmetric and idempotent: $P^T = P$ and $P^2 = P$. Conversely, every real symmetric idempotent matrix is the orthogonal projection onto its column space.

Proof. The vector $Q^T b$ has r coordinates, so $QQ^T b$ belongs to S . Moreover,

$$Q^T r = Q^T b - Q^T QQ^T b = Q^T b - \mathbb{1}_r Q^T b = 0. \quad (\text{A.17})$$

This means r is orthogonal to each column of Q , and hence to every linear combination of those columns. For any $y \in S$, $p - y$ lies in S , so Pythagoras gives

$$\|b - y\|_2^2 = \|r + (p - y)\|_2^2 = \|r\|_2^2 + \|p - y\|_2^2. \quad (\text{A.18})$$

The final term is nonnegative and vanishes exactly when $\mathbf{y} = \mathbf{p}$. This proves existence and uniqueness of the closest point, not just stationarity of a squared-distance expression.

The transpose rule gives $(QQ^\top)^\top = QQ^\top$, and $P^2 = Q(Q^\top Q)Q^\top = QQ^\top = P$. Conversely, suppose $P^\top = P$ and $P^2 = P$. The candidate $P\mathbf{b}$ lies in its column space. Any vector of that space has the form $\mathbf{y} = P\mathbf{w}$, and

$$\mathbf{y}^\top(\mathbf{b} - P\mathbf{b}) = \mathbf{w}^\top P^\top(\mathbb{1} - P)\mathbf{b} = \mathbf{w}^\top(P - P^2)\mathbf{b} = 0. \quad (\text{A.19})$$

Thus its residual is orthogonal to the column space. The same Pythagorean argument proves that $P\mathbf{b}$ is the unique closest point. For $r = 0$, use the empty-column matrix Q and $P = 0$; its column space is $\{0\}$ and the same conclusion is immediate. ■

For example, $\mathbf{q} = 2^{-1/2}(1, 1)^\top$ gives

$$P = \mathbf{q}\mathbf{q}^\top = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad P \begin{pmatrix} 3 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}. \quad (\text{A.20})$$

The residual $(1, -1)^\top$ is perpendicular to the line spanned by $(1, 1)^\top$. Applying the projection twice makes no further change, because the first output is already on that line.

Theorem 13 (Orthogonal decompositions and the fundamental spaces). For a subspace $S \subseteq \mathbb{R}^d$, let $S^\perp = \{\mathbf{z} : \mathbf{z}^\top \mathbf{y} = 0 \text{ for every } \mathbf{y} \in S\}$. Every vector has a unique decomposition into a vector of S and one of $S^\perp$. Their dimensions sum to d . For $A \in \mathbb{R}^{m \times n}$,

$$\text{Null}(A) = \text{Row}(A)^\perp, \quad \text{Null}(A^\top) = \text{Col}(A)^\perp. \quad (\text{A.21})$$

Consequently,

$$\begin{aligned} \mathbb{R}^n &= \text{Row}(A) \oplus \text{Null}(A), \\ \mathbb{R}^m &= \text{Col}(A) \oplus \text{Null}(A^\top). \end{aligned}$$

The symbol $\oplus$ here denotes a sum with unique, mutually orthogonal components.

Proof. Take an orthonormal basis matrix Q for S . Projection gives $\mathbf{b} = QQ^\top \mathbf{b} + (\mathbb{1} - QQ^\top)\mathbf{b}$, with the first term in S and the second in $S^\perp$. If also $\mathbf{b} = \mathbf{s} + \mathbf{t}$ with $\mathbf{s} \in S$ and $\mathbf{t} \in S^\perp$, subtract the two expressions. The difference of the two S components lies in both S and $S^\perp$. A vector in their intersection is orthogonal to itself, so its squared norm is zero and it must vanish. Both components therefore agree, proving uniqueness.

Extend the $r = \dim S$ orthonormal columns of Q to an orthonormal basis of $\mathbb{R}^d$. The remaining $d - r$ vectors are in $S^\perp$. They span it: a vector orthogonal to the first r basis vectors has zero coefficients along those vectors in its orthonormal expansion. Thus $\dim S^\perp = d - r$, and the same argument also proves $(S^\perp)^\perp = S$.

Finally, $A\mathbf{x} = 0$ says precisely that $\mathbf{x}$ is orthogonal to each row of A . Being orthogonal to those rows is equivalent to being orthogonal to their span, proving the first identity in (A.21). Applying this reasoning to $A^\top$ proves the second. Substitute these identities into the decomposition just proved to obtain the displayed fundamental-space decompositions. ■

The matrix $\mathbb{1} - P$ projects onto $S^\perp$: the decomposition shows its output is the perpendicular component, and its symmetry and idempotence can also be checked directly. The input decomposition for A separates

detectable row-space components from invisible nullspace components. The output decomposition separates attainable data from unattainable data. These are the geometric ingredients of least squares and of the minimum-norm choice among multiple solutions.

A.4 Symmetry, positivity, and the spectral theorem

A real square matrix H is *symmetric* if $H^\top = H$. It is *symmetric positive definite* if it is symmetric and $\mathbf{x}^\top H \mathbf{x} > 0$ for every nonzero $\mathbf{x}$. Replacing $>$ by $\geq$ gives positive semidefiniteness. Symmetry is part of these definitions in this note.

An *eigenpair* of a square matrix consists of a scalar λ and a nonzero vector $\mathbf{v}$ such that $H\mathbf{v} = \lambda\mathbf{v}$. The vector is an eigenvector and the scalar is its eigenvalue. Even real square matrices can have complex eigenvalues, and not every matrix has a basis of eigenvectors. Rectangular matrices do not have eigenvalues in this usual sense. Symmetric real matrices have a particularly useful structure.

Theorem 14 (Real spectral theorem). Every real symmetric $d \times d$ matrix H has an orthonormal basis of real eigenvectors. Equivalently, there are a real orthogonal $W \in \mathbb{R}^{d \times d}$ and real scalars $\lambda_1, \dots, \lambda_d$ such that

$$H = W\Lambda W^\top, \quad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_d). \quad (\text{A.22})$$

Proof. We first find one unit eigenvector using the continuous function $f(\mathbf{x}) = \mathbf{x}^\top H \mathbf{x}$ on the Euclidean unit sphere. The sphere is compact, so the extreme-value theorem gives a maximizing unit vector $\mathbf{u}$. Set $\lambda = \mathbf{u}^\top H \mathbf{u}$, which is real. If $\mathbf{v}$ is any unit vector perpendicular to $\mathbf{u}$, the curve $\mathbf{x}(t) = \mathbf{u} \cos t + \mathbf{v} \sin t$ stays on the unit sphere, because orthogonality gives $\|\mathbf{x}(t)\|_2^2 = \cos^2 t + \sin^2 t = 1$. Therefore the differentiable function

$$g(t) = \mathbf{x}(t)^\top H \mathbf{x}(t) = \lambda \cos^2 t + 2(\mathbf{v}^\top H \mathbf{u}) \sin t \cos t + (\mathbf{v}^\top H \mathbf{v}) \sin^2 t \quad (\text{A.23})$$

has a maximum at $t = 0$; symmetry of H produced the factor 2 in its middle term. Thus $0 = g'(0) = 2\mathbf{v}^\top H \mathbf{u}$. Rescaling shows the same orthogonality for every vector in $\text{span}\{\mathbf{u}\}^\perp$. The vector $H\mathbf{u} - \lambda\mathbf{u}$ is also orthogonal to $\mathbf{u}$, since $\mathbf{u}^\top (H\mathbf{u} - \lambda\mathbf{u}) = \lambda - \lambda = 0$. It is orthogonal to the entire perpendicular subspace as well, so it is orthogonal to every vector in $\mathbb{R}^d$ and must be zero. Consequently $H\mathbf{u} = \lambda\mathbf{u}$.

Next, the perpendicular subspace is invariant under H . If $\mathbf{z}^\top \mathbf{u} = 0$, then

$$\mathbf{u}^\top H \mathbf{z} = (H\mathbf{u})^\top \mathbf{z} = \lambda \mathbf{u}^\top \mathbf{z} = 0. \quad (\text{A.24})$$

The equality using symmetry is essential. Thus H maps $\text{span}\{\mathbf{u}\}^\perp$ back into itself.

Now argue by induction on d . The result for $d = 1$ is immediate. For $d > 1$, choose an orthonormal basis matrix $Q \in \mathbb{R}^{d \times (d-1)}$ for the perpendicular subspace. The restriction of H to it is represented by $B = Q^\top H Q$, which is a real symmetric $(d-1) \times (d-1)$ matrix. By induction, B has an orthonormal eigenbasis $\mathbf{c}_1, \dots, \mathbf{c}_{d-1}$. Invariance gives $HQ = QB$: each column of HQ lies in the column space of Q , so its orthonormal coordinates are the corresponding column of $Q^\top H Q$. Therefore, if $B\mathbf{c}_j = \mu_j \mathbf{c}_j$, then $H(Q\mathbf{c}_j) = QB\mathbf{c}_j = \mu_j Q\mathbf{c}_j$. The vectors $Q\mathbf{c}_j$ are orthonormal because $Q^\top Q = \mathbb{I}$, and they are all perpendicular to $\mathbf{u}$. Together with $\mathbf{u}$ they form the required orthonormal eigenbasis.

Put those eigenvectors in the columns of W . Their eigenvector equations say $HW = W\Lambda$. Since W is orthogonal, multiplying by $W^\top$ gives $H = W\Lambda W^\top$. ■

The only analysis facts used in this proof are that a continuous function on a compact sphere attains a maximum, and that a differentiable scalar function has zero derivative at an interior maximum. The rest is finite-dimensional linear algebra.

Proposition 14 (Positivity and Gram matrices). A real symmetric matrix is positive definite exactly when all its eigenvalues are positive, and positive semidefinite exactly when all are nonnegative. For any $B \in \mathbb{R}^{m \times n}$, $B^T B$ is symmetric positive semidefinite. It is positive definite exactly when B has independent columns.

Proof. Write the spectral decomposition $H = W\Lambda W^T$ and set $\mathbf{c} = W^T \mathbf{x}$. Orthogonality makes this a bijective change of coordinates and gives

$$\mathbf{x}^T H \mathbf{x} = \mathbf{c}^T \Lambda \mathbf{c} = \sum_{j=1}^d \lambda_j c_j^2. \quad (\text{A.25})$$

If all eigenvalues are positive, this sum is positive for every nonzero vector. If all are nonnegative, it is nonnegative. Conversely, evaluating on each unit eigenvector gives precisely λ_j , proving the necessary signs in both cases.

The transpose rule shows $(B^T B)^T = B^T B$, and $\mathbf{x}^T B^T B \mathbf{x} = \|B\mathbf{x}\|_2^2 \geq 0$. This is strictly positive for every nonzero $\mathbf{x}$ exactly when $B\mathbf{x} = 0$ has no nonzero solution, equivalently when its columns are independent. ■

A.5 Singular values and the SVD

Singular values apply to every real matrix, including rectangular and rank-deficient matrices. They are nonnegative and describe stretching between orthonormal input and output coordinates. The next construction explains why an SVD always exists.

Theorem 15 (Existence and structure of a singular value decomposition). Let $A \in \mathbb{R}^{m \times n}$, let $q = \min(m, n)$, and let $r = \text{rank}(A)$. There are orthogonal matrices U, V such that

$$A = U\Sigma V^T, \quad U \in \mathbb{R}^{m \times m}, \quad \Sigma \in \mathbb{R}^{m \times n}, \quad V \in \mathbb{R}^{n \times n}. \quad (\text{A.26})$$

The matrix Σ has zeros off its main diagonal and entries $\sigma_1 \geq \dots \geq \sigma_q \geq 0$ on that diagonal. Exactly r of these entries are positive. If $\mathbf{u}_j$ and $\mathbf{v}_j$ denote columns of U and V , then

$$A\mathbf{v}_j = \sigma_j \mathbf{u}_j, \quad A^T \mathbf{u}_j = \sigma_j \mathbf{v}_j \quad (1 \leq j \leq r). \quad (\text{A.27})$$

The first r columns of U form a basis of $\text{Col}(A)$; the first r columns of V form a basis of $\text{Row}(A)$. The remaining columns of V form a basis of $\text{Null}(A)$, and the remaining columns of U form a basis of $\text{Null}(A^T)$.

Proof. By the spectral theorem, the symmetric positive semidefinite $n \times n$ matrix $A^T A$ has an orthonormal eigenbasis $\mathbf{v}_1, \dots, \mathbf{v}_n$ with eigenvalues ordered as $\lambda_1 \geq \dots \geq \lambda_n \geq 0$. For any vector $\mathbf{x}$,

$$A^T A \mathbf{x} = 0 \iff A \mathbf{x} = 0. \quad (\text{A.28})$$

The reverse implication is immediate; the forward implication follows by multiplying on the left by $\mathbf{x}^T$ to get $\|A\mathbf{x}\|_2^2 = 0$. Thus the two matrices have the same nullspace. By rank-nullity this nullspace has dimension $n - r$, so exactly r eigenvalues of $A^T A$ are positive. Relabel them first, and set $\sigma_j = \sqrt{\lambda_j} > 0$ for $1 \leq j \leq r$.

For those positive eigenvalues, define $\mathbf{u}_j = A\mathbf{v}_j/\sigma_j \in \mathbb{R}^m$. Their inner products are

$$\mathbf{u}_i^\top \mathbf{u}_j = \frac{\mathbf{v}_i^\top A^\top A \mathbf{v}_j}{\sigma_i \sigma_j} = \frac{\lambda_j \mathbf{v}_i^\top \mathbf{v}_j}{\sigma_i \sigma_j} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \quad (\text{A.29})$$

They are therefore orthonormal, so in particular $r \leq m$. For $j > r$, $\lambda_j = 0$ and $\|A\mathbf{v}_j\|_2^2 = \mathbf{v}_j^\top A^\top A \mathbf{v}_j = 0$; hence $A\mathbf{v}_j = 0$.

Extend $\mathbf{u}_1, \dots, \mathbf{u}_r$ to an orthonormal basis $\mathbf{u}_1, \dots, \mathbf{u}_m$ of $\mathbb{R}^m$. Let U contain these columns and let V contain the eigenvectors above. Form the $m \times n$ matrix Σ with its first r diagonal entries $\sigma_1, \dots, \sigma_r$ and all other entries zero. Column by column, the equations already proved say $AV = U\Sigma$. Multiplying by $V^\top$ gives the full SVD. Its remaining $q - r$ diagonal entries are defined to be zero.

The first r columns of U belong to $\text{Col}(A)$. Conversely, expand any input as $\mathbf{x} = \sum_{j=1}^n c_j \mathbf{v}_j$. Then $A\mathbf{x} = \sum_{j=1}^r c_j \sigma_j \mathbf{u}_j$, so they span that column space. Their orthonormality makes them a basis. The same expansion shows $A\mathbf{x} = 0$ exactly when $c_1 = \dots = c_r = 0$. Thus $\mathbf{v}_{r+1}, \dots, \mathbf{v}_n$ form a nullspace basis. Since the entire V basis is orthonormal, its first r columns span the perpendicular complement of the nullspace, which is $\text{Row}(A)$ by Theorem 13. The remaining $m - r$ columns of U similarly span the perpendicular complement of $\text{Col}(A)$, namely $\text{Null}(A^\top)$. Finally, for $j \leq r$, $A^\top \mathbf{u}_j = A^\top A \mathbf{v}_j / \sigma_j = \lambda_j \mathbf{v}_j / \sigma_j = \sigma_j \mathbf{v}_j$.

If $r = 0$, then $A = 0$ by its zero-dimensional column space. Choose any orthogonal U, V and $\Sigma = 0$; all the statements reduce to the full input and output spaces being nullspaces. ■

This proof constructs the SVD through $A^\top A$ to explain its existence. It is not a recommendation to form $A^\top A$ as a numerical SVD algorithm: its squared conditioning can obscure small singular values.

Full, thin, and compact forms: check every shape. The full form in (A.26) includes complete orthonormal bases, even for directions annihilated by A . Keeping the first $q = \min(m, n)$ columns of U and V gives the *thin* form

$$A = U_q \Sigma_q V_q^\top, \quad U_q \in \mathbb{R}^{m \times q}, \quad \Sigma_q \in \mathbb{R}^{q \times q}, \quad V_q \in \mathbb{R}^{n \times q}. \quad (\text{A.30})$$

Here $\Sigma_q = \text{diag}(\sigma_1, \dots, \sigma_q)$. The product has shape $(m \times q)(q \times q)(q \times n) = m \times n$. If $m < n$, this drops $n - m$ right nullspace vectors; if $m > n$, it drops $m - n$ left nullspace vectors. All their coefficients in A were zero, so the matrix represented is unchanged.

Keeping only the r positive singular values gives the *compact* form

$$A = U_r \Sigma_r V_r^\top = \sum_{j=1}^r \sigma_j \mathbf{u}_j \mathbf{v}_j^\top, \quad (\text{A.31})$$

$$U_r \in \mathbb{R}^{m \times r}, \quad \Sigma_r \in \mathbb{R}^{r \times r}, \quad V_r \in \mathbb{R}^{n \times r}.$$

Now Σ_r is invertible for $r > 0$. When $r < q$, the thin form retains zero singular values and the compact form discards them. Terminology varies between texts and software, so checking the shapes is more informative than relying on the name alone.

For a concrete rectangular example,

$$A = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} (3) \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \quad (\text{A.32})$$

is already a SVD with rank 1. A full SVD uses $U = \mathbb{1}_2$, $V = \mathbb{1}_3$, and $\Sigma = A$. The thin form uses

$$U_q = \mathbb{1}_2, \quad \Sigma_q = \begin{pmatrix} 3 & 0 \\ 0 & 0 \end{pmatrix}, \quad V_q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad q = 2. \quad (\text{A.33})$$

The compact form keeps only the first columns and $\Sigma_r = (3)$. The map preserves one input direction, scaled by 3, and annihilates the other two. Its attainable outputs lie on the first coordinate axis of $\mathbb{R}^2$.

Proposition 15 (Norms and conditioning from singular values). For a rank- r matrix with singular values as above,

$$\|A\|_2 = \sigma_1, \quad \|A\|_F^2 = \sum_{j=1}^q \sigma_j^2, \quad \|A\|_2 \leq \|A\|_F \leq \sqrt{r} \|A\|_2. \quad (\text{A.34})$$

If A is square, it is invertible exactly when $\sigma_n > 0$. For such a matrix,

$$\|A^{-1}\|_2 = \frac{1}{\sigma_n}, \quad \kappa_2(A) = \|A\|_2 \|A^{-1}\|_2 = \frac{\sigma_1}{\sigma_n}. \quad (\text{A.35})$$

Proof. For $r > 0$, express a unit input in the orthonormal V basis as $\mathbf{x} = \sum_{j=1}^n c_j \mathbf{v}_j$, where $\sum_{j=1}^n c_j^2 = 1$. Then

$$\|A\mathbf{x}\|_2^2 = \sum_{j=1}^r \sigma_j^2 c_j^2 \leq \sigma_1^2 \sum_{j=1}^n c_j^2 = \sigma_1^2. \quad (\text{A.36})$$

Equality holds for $\mathbf{x} = \mathbf{v}_1$, proving the induced norm formula. Left multiplication by an orthogonal matrix preserves the Euclidean length of each column and hence preserves the Frobenius norm, whose square is the sum of squared column lengths. Right multiplication preserves it as well: transpose the product and apply the same column argument to its rows. Therefore $\|A\|_F = \|\Sigma\|_F$, which proves the sum-of-squares formula. That sum lies between σ_1^2 and $r\sigma_1^2$, giving the inequalities. If $r = 0$, all singular values and all three norms in these claims are zero, so the conclusions remain valid.

For square A , full rank is equivalent to every singular value being positive, by Theorem 15. In that case $A^{-1} = V\Sigma^{-1}U^\top$, as multiplication in either order verifies. Reversing the preceding stretching argument shows that its largest factor is $1/\sigma_n$, attained on $\mathbf{u}_n$. Multiplying the two norm formulas proves the condition-number identity. ■

The eigenvalues of $A^\top A$ are the squared positive singular values together with zeros, for a total of n eigenvalues. If $n > q$, the last $n - q$ are additional zeros beyond the q entries recorded on the rectangular diagonal of Σ . Singular values are not generally the eigenvalues of A : even the 1×1 matrix $A = (-2)$ has eigenvalue -2 and singular value 2 .

For a symmetric matrix $H = W\Lambda W^\top$, its singular values are the absolute values of its eigenvalues. Indeed, $H^\top H = H^2 = W\Lambda^2 W^\top$, and the singular values are the nonnegative square roots of its eigenvalues. Reordering, if needed, places these absolute values in decreasing order.

A.6 The pseudoinverse ingredients in these coordinates

The SVD shows exactly which directions can be reversed. For $r > 0$, reverse each positive stretching factor and interchange the input and output singular vectors:

$$A^+ = V_r \Sigma_r^{-1} U_r^\top = \sum_{j=1}^r \frac{1}{\sigma_j} \mathbf{v}_j \mathbf{u}_j^\top \in \mathbb{R}^{n \times m}. \quad (\text{A.37})$$

This is the Moore–Penrose pseudoinverse studied in the main text. Its shape is $(n \times r)(r \times r)(r \times m) = n \times m$, so it takes data in $\mathbb{R}^m$ back to an input in $\mathbb{R}^n$. If $A = 0$, the sum is empty and $A^+ = 0$.

Multiplication using $U_r^T U_r = V_r^T V_r = \mathbb{1}_r$ gives $AA^+ = U_r U_r^T$ and $A^+A = V_r V_r^T$. By the projection theorem, these are the orthogonal projections onto the column space and row space respectively. They are symmetric, and applying them to the corresponding factors leaves those factors fixed, so $AA^+A = A$ and $A^+AA^+ = A^+$. These computations verify the four defining pseudoinverse identities directly; the main text proves the uniqueness and minimum-norm least-squares characterization in detail.

For the rectangular example above, $A^+ \mathbf{b} = (b_1/3, 0, 0)^T$. The unattainable component b_2 remains in the residual, while the two invisible input coordinates are set to zero. Any nonzero choices for those coordinates would leave the fit unchanged and increase the squared solution norm.