Review: Linear Algebra

SDS 391P.6, Spring 2026

Pratik Patil

1 Vector norms and their properties

In this section, we will briefly recap some properties of vector norms. Throughout, we work in $\mathbb{R}^n$ with vectors $x = (x_1, \ldots, x_n)^\top$ and $y = (y_1, \ldots, y_n)^\top$.

1.1 Standard inner product and Euclidean norm

The standard inner product on $\mathbb{R}^n$ is $\langle x, y\rangle = x^\top y = \sum_{i=1}^n x_i y_i$. We say x and y are *orthogonal* if $\langle x, y\rangle = 0$. The Euclidean (or ℓ_2) norm is defined by

$$\|x\|_2 \;=\; \sqrt{\langle x, x\rangle} \;=\; \Big(\sum_{i=1}^n x_i^2\Big)^{1/2}.$$

A basic inequality relating the inner product and the Euclidean norm is Cauchy–Schwarz's inequality:

$$|\langle x, y\rangle| \;\leqslant\; \|x\|_2\,\|y\|_2, \qquad \text{with equality iff } x \text{ and } y \text{ are collinear (i.e., } x = \alpha y).$$

Geometrically, one can write $\langle x, y\rangle = \|x\|_2\|y\|_2 \cos\theta$ for an angle θ between x and y: the dot product is the product of lengths times an alignment factor.

1.2 General norms and the ℓ_p family

There are several other norms besides the Euclidean norm that are useful in practice. In general, a function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}_{\geqslant 0}$ is called a *norm* if it satisfies: (i) $\|x\| = 0$ iff $x = 0$ (definiteness), (ii) $\|\alpha x\| = |\alpha|\|x\|$ (homogeneity), and (iii) $\|x + y\| \leqslant \|x\| + \|y\|$ (triangle inequality).

The most important family of norms on $\mathbb{R}^n$ is the ℓ_p family. For $p \in [1, \infty]$,

$$\|x\|_p \;=\; \begin{cases} \Big(\sum_{i=1}^n |x_i|^p\Big)^{1/p}, & 1 \leqslant p < \infty, \\ \max_{1\leqslant i\leqslant n} |x_i|, & p = \infty. \end{cases}$$

The triangle inequality for ℓ_p norms is known as *Minkowski's inequality*:

$$\|x + y\|_p \;\leqslant\; \|x\|_p + \|y\|_p.$$

This guarantees that each $\|\cdot\|_p$ is indeed a norm for $p \geqslant 1$.

Three special cases appear constantly in statistics/ML:

- ℓ_2 (least squares, Gaussian geometry, smoothness, rotation invariance);
- ℓ_1 (sparsity and robustness; corners in the unit ball drive sparse solutions in many convex programs);
- ℓ_∞ (uniform/worst-case control across coordinates; max constraints and sup-norm error bounds).

A useful way to visualize a norm is through its unit ball

$$B_p^n \;:=\; \{x \in \mathbb{R}^n : \|x\|_p \leqslant 1\}.$$

For example, $B_\infty^n = [-1,1]^n$ is the hypercube, while $B_1^n = \operatorname{conv}(\{\pm e_1, \ldots, \pm e_n\})$ is a cross-polytope.

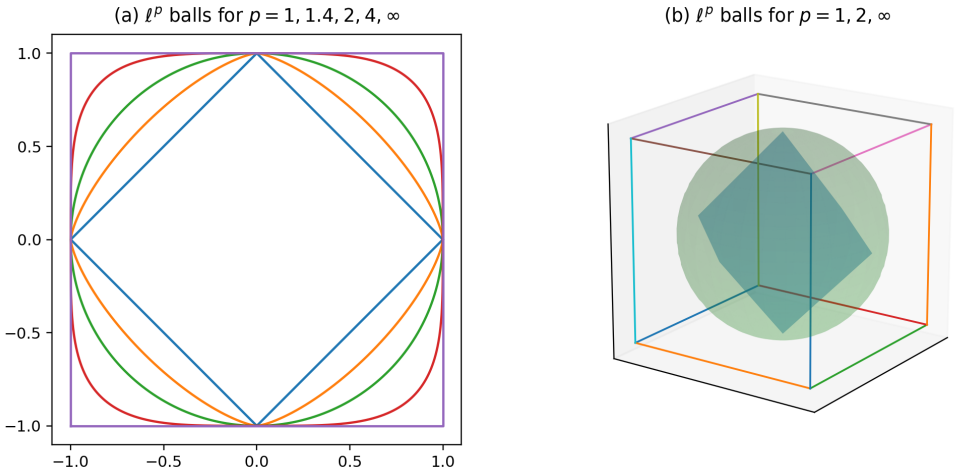

Figure 1: Unit ℓ^p balls in low dimensions become "rounder" as p increases: $p = 1$ (diamond), $p = 2$ (circle/sphere), $p = \infty$ (box).

1.3 Comparing ℓ_p norms in finite dimensions

In finite-dimensional spaces, all norms are equivalent up to constants depending on dimension. For ℓ_p norms, one has a sharp and extremely useful comparison: for $1 \leqslant p \leqslant q \leqslant \infty$ and all $x \in \mathbb{R}^n$,

$$\|x\|_q \;\leqslant\; \|x\|_p \;\leqslant\; n^{\frac{1}{p}-\frac{1}{q}}\,\|x\|_q.$$

Two immediate consequences are worth keeping in mind:

- Monotonicity: $\|x\|_p$ is non-increasing in p (larger p means smaller norm value on a fixed vector).
- Nested unit balls: $B_p^n \subseteq B_q^n$ whenever $p \leqslant q$.

Specializing to $(p,q) \in \{(2,1), (\infty,2), (\infty,1)\}$ gives the familiar bounds

$$\|x\|_2 \leqslant \|x\|_1 \leqslant \sqrt{n}\,\|x\|_2, \qquad \|x\|_\infty \leqslant \|x\|_2 \leqslant \sqrt{n}\,\|x\|_\infty, \qquad \|x\|_\infty \leqslant \|x\|_1 \leqslant n\,\|x\|_\infty.$$

These inequalities are not artifacts: they are tight up to constants. Two extremal regimes are helpful intuition. If x is 1-sparse (only one nonzero entry), then $\|x\|_p = \|x\|_q$ for all p, q. If x is "flat" (e.g., $x = (1, \ldots, 1)$), then $\|x\|_p = n^{1/p}\|x\|_\infty$ and the dimension factor becomes active.

A related and often-used fact is that ℓ_∞ is the limiting case of ℓ_p:

$$\|x\|_\infty \;\leqslant\; \|x\|_p \;\leqslant\; n^{1/p}\,\|x\|_\infty, \qquad \text{so} \qquad \lim_{p\to\infty}\|x\|_p = \|x\|_\infty.$$

A practical rule of thumb: if $p \geqslant \ln n$, then $n^{1/p} \leqslant e$, hence $\|x\|_p$ is within a factor e of $\|x\|_\infty$.

1.4 Hölder inequality and dual norms

A key generalization of Cauchy–Schwarz is Hölder's inequality. Let $p \in [1,\infty]$ and define its *conjugate exponent* $p' \in [1,\infty]$ by $\frac{1}{p} + \frac{1}{p'} = 1$ (with $1/\infty = 0$). Then for all $x, y \in \mathbb{R}^n$,

$$|\langle x, y\rangle| \;\leqslant\; \|x\|_p\,\|y\|_{p'}.$$

The case $p = p' = 2$ recovers Cauchy–Schwarz. Hölder is also *tight*: for any fixed x one can choose $y \neq 0$ so that equality holds (informally, y should concentrate where x is large, with the matching power-law scaling).

This tightness is cleanly expressed through *dual norms*. Given any norm $\|\cdot\|$ on $\mathbb{R}^n$, its dual norm is

$$\|y\|_* \;:=\; \sup_{\|x\|\leqslant 1} \langle y, x\rangle.$$

Equivalently, $\|y\|_*$ is the support function of the unit ball of $\|\cdot\|$. For ℓ_p norms, duality takes a particularly simple form: the dual of $\|\cdot\|_p$ is $\|\cdot\|_{p'}$, and

$$\|x\|_p \;=\; \sup\{\langle x, y\rangle :\; \|y\|_{p'} \leqslant 1\}.$$

This geometric relationship is illustrated in Figure 2 for the dual pairs of ℓ_1 and ℓ_∞, where the faces of the primal ball correspond to the vertices of the dual ball.

2 Basic matrix results and some special matrix classes

In this section, we recall basic matrix results and some special matrix classes. Throughout the section, we use dimension conventions that are common in statistics: typically $X \in \mathbb{R}^{n\times p}$, where n is the sample size (rows = observations) and p is the number of features (columns = covariates).

2.1 Notation and basic operations

For a positive integer n, write $[n] := \{1, 2, \ldots, n\}$. For $A \in \mathbb{R}^{n\times p}$, the (i,j) entry is A_{ij} for $i \in [n]$, $j \in [p]$. We sometimes write $A_{i,:}$ for the i-th row and $A_{:,j}$ for the j-th column.

For $A \in \mathbb{R}^{n\times p}$, the transpose is $A^\top \in \mathbb{R}^{p\times n}$ with $(A^\top)_{ij} = A_{ji}$. For conformable matrices A, B, $(AB)^\top = B^\top A^\top$. A square matrix $A \in \mathbb{R}^{p\times p}$ is *symmetric* if $A = A^\top$; the set of real symmetric $p \times p$ matrices is denoted $\mathbb{S}^p$.

For a square matrix $A \in \mathbb{R}^{p\times p}$, the trace is $\operatorname{tr}(A) := \sum_{j=1}^p A_{jj}$. Trace is invariant under cyclic permutations: whenever products are well-defined,

$$\operatorname{tr}(AB) = \operatorname{tr}(BA), \qquad \text{and more generally} \qquad \operatorname{tr}(ABC) = \operatorname{tr}(CAB) = \operatorname{tr}(BCA).$$

This simple identity is a workhorse for rewriting quadratic forms and simplifying expressions in optimization and statistics.

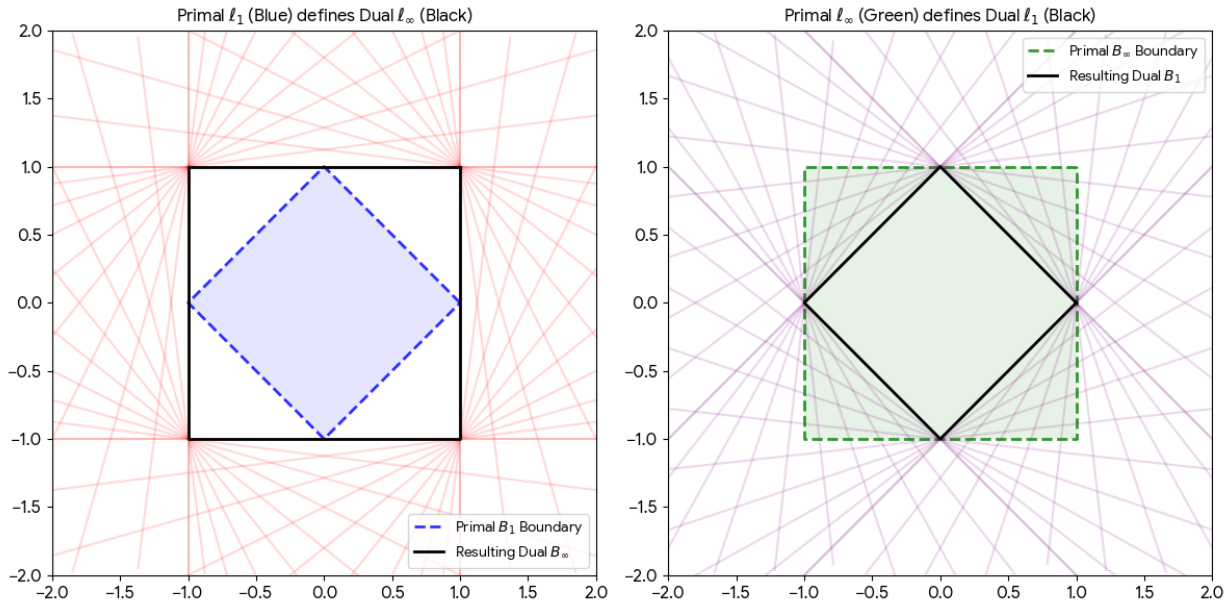

Figure 2: Visualizing dual norms. Note that the unit ball of the dual norm is the intersection of half-spaces defined by the primal unit ball. Left panel: The primal ℓ_1 ball (blue diamond) generates the dual ℓ_∞ ball (black square) via its supporting hyperplanes (red lines). Right panel: The primal ℓ_∞ ball (green square) generates the dual ℓ_1 ball (black diamond) via its supporting hyperplanes (purple lines).

We will also use the Frobenius inner product for matrices $A, B \in \mathbb{R}^{n\times p}$, $\langle A, B\rangle_F := \operatorname{tr}(A^\top B) = \sum_{i,j} A_{ij}B_{ij}$, and the corresponding Frobenius norm $\|A\|_F^2 = \operatorname{tr}(A^\top A)$. (We will come back to these in Section 5 on matrix norms below.)

A side note: For an excellent visual guide on the different ways to view matrix multiplication (including the dot product view, column view, and row view), we refer the reader to Hiranabe (2021).

2.2 Linear maps, fundamental subspaces, rank and nullity

A matrix $A \in \mathbb{R}^{n\times p}$ defines a linear map $A : \mathbb{R}^p \to \mathbb{R}^n$, $x \mapsto Ax$. The associated subspaces organize much of linear algebra (and many identifiability questions in statistics):

- The *column space* (range) is $\operatorname{range}(A) := \{Ax : x \in \mathbb{R}^p\} \subseteq \mathbb{R}^n$.
- The *row space* is $\operatorname{range}(A^\top) \subseteq \mathbb{R}^p$.
- The *null space* (kernel) is $\operatorname{null}(A) := \{x \in \mathbb{R}^p : Ax = 0\} \subseteq \mathbb{R}^p$.
- The *left null space* is $\operatorname{null}(A^\top) \subseteq \mathbb{R}^n$.

The *rank* of A is $\operatorname{rank}(A) := \dim(\operatorname{range}(A))$, and one also has $\operatorname{rank}(A) = \dim(\operatorname{range}(A^\top))$ (column rank equals row rank). Always $\operatorname{rank}(A) \leqslant \min\{n, p\}$. We say A has *full column rank* if $\operatorname{rank}(A) = p$ (injective map), and *full row rank* if $\operatorname{rank}(A) = n$ (surjective map).

Two basic facts are worth recalling:

$$\dim(\operatorname{null}(A)) + \operatorname{rank}(A) = p \qquad \text{(rank–nullity)},$$

and the orthogonality relations

$$\text{range}(A)^{\perp} = \text{null}(A^{\top}), \qquad \text{range}(A^{\top})^{\perp} = \text{null}(A).$$

These identities express a useful duality: directions in $\text{null}(A)$ are parameter perturbations that do not change Ax, while $\text{null}(A^{\top})$ contains the linear functionals that annihilate the range of A.

A recurring construction in statistics is the Gram matrix $A^{\top}A$ (e.g. $X^{\top}X$ in least squares). It is always symmetric positive semidefinite, and it has the same rank as A:

$$\text{rank}(A^{\top}A) = \text{rank}(AA^{\top}) = \text{rank}(A).$$

In particular, $X^{\top}X$ is invertible if and only if X has full column rank.

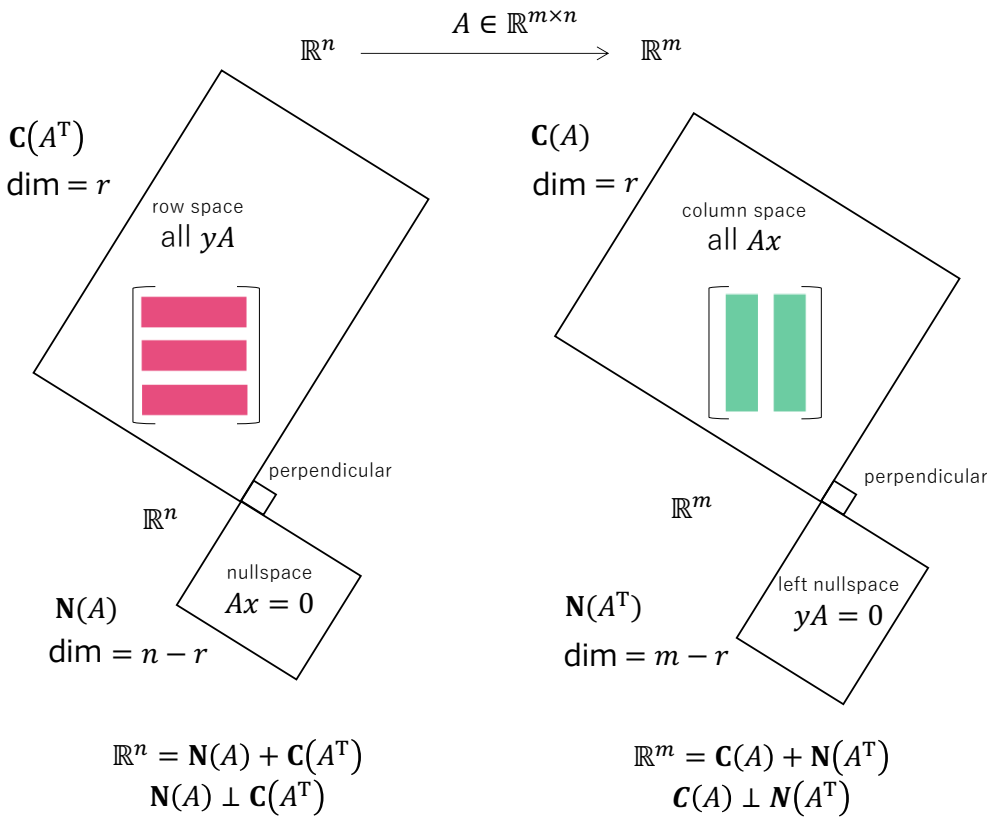

Figure 3: The four fundamental subspaces. Figure courtesy of Hiranabe (2021).

2.3 Special classes of matrices

Identity and diagonal matrices. The $n \times n$ identity matrix is denoted I_n (or simply I when the dimension is clear). Given $x \in \mathbb{R}^p$, $\text{diag}(x) \in \mathbb{R}^{p\times p}$ denotes the diagonal matrix with diagonal entries $x_1, \ldots, x_p$, i.e. $\text{diag}(x) = \text{diag}(x_1, \ldots, x_p)$.

Invertible matrices. A square matrix $A \in \mathbb{R}^{p\times p}$ is *invertible* if there exists A^{-1} such that $AA^{-1} = A^{-1}A = I_p$. Equivalently: A is invertible iff $\text{rank}(A) = p$, iff $\text{null}(A) = \{0\}$, iff $Ax = b$ has a unique solution for every $b \in \mathbb{R}^p$. Two basic identities (when inverses exist) are $(AB)^{-1} = B^{-1}A^{-1}$ and $(A^{\top})^{-1} = (A^{-1})^{\top}$.

Orthogonal matrices. A matrix $Q \in \mathbb{R}^{p\times p}$ is *orthogonal* if $Q^\top Q = QQ^\top = I_p$, equivalently $Q^{-1} = Q^\top$. Orthogonal matrices preserve Euclidean geometry: for all $u, v \in \mathbb{R}^p$, $\langle Qu, Qv\rangle = \langle u, v\rangle$ and $\|Qu\|_2 = \|u\|_2$. Thus they represent rotations/reflections (change of orthonormal basis).

Positive semidefinite and positive definite matrices. A symmetric matrix $A \in \mathbb{S}^p$ is *positive semidefinite* (PSD) if $x^\top Ax \geqslant 0$ for all $x \in \mathbb{R}^p$; we write $A \succeq 0$. It is *positive definite* (PD) if $x^\top Ax > 0$ for all $x \neq 0$; we write $A \succ 0$. More generally, for symmetric A, B, $A \succeq B$ means $A - B \succeq 0$. PSD matrices include covariance matrices and kernel/Gram matrices; PD matrices arise as strictly convex Hessians and as covariance matrices of non-degenerate Gaussian models.

Projection matrices. A square matrix $P \in \mathbb{R}^{p\times p}$ is a *projection* if $P^2 = P$ (idempotent). It is an *orthogonal projection* if additionally $P^\top = P$. Orthogonal projections satisfy a clean geometric decomposition: for every $x \in \mathbb{R}^p$, $x = Px + (I-P)x$ with $Px \in \text{range}(P)$ and $(I-P)x \in \text{range}(P)^\perp$. Eigenvalues of any projection are in $\{0, 1\}$; for an orthogonal projection one has $\text{rank}(P) = \text{tr}(P)$.

Two projections that appear constantly in least squares are:

- If $X \in \mathbb{R}^{n\times p}$ has full column rank, the orthogonal projector onto $\text{range}(X) \subseteq \mathbb{R}^n$ is

$$P_X \;=\; X(X^\top X)^{-1}X^\top,$$

  and $I_n - P_X$ projects onto the residual subspace $\text{range}(X)^\perp = \text{null}(X^\top)$.

- If $A \in \mathbb{R}^{m\times p}$ has full row rank ($m \leqslant p$), the orthogonal projector onto $\text{null}(A) \subseteq \mathbb{R}^p$ is

$$P_{\text{null}(A)} \;=\; I_p - A^\top(AA^\top)^{-1}A.$$

3 Eigenvalues and eigenvectors

Eigenvalues and eigenvectors connect linear algebra to geometry and optimization. We will see this connection in this section. Notationally, we adopt dimension conventions common in statistics. Typically a covariance, Gram, or Hessian matrix lives in $\mathbb{R}^{p\times p}$, where p is the number of features/parameters.

3.1 Eigenvalues of a square matrix

Let $A \in \mathbb{R}^{p\times p}$. A scalar $\lambda \in \mathbb{C}$ is an *eigenvalue* of A if there exists a nonzero $u \in \mathbb{C}^p$ such that

$$Au = \lambda u.$$

The vector u is then an eigenvector associated with λ. (The zero vector is never an eigenvector, although $\lambda = 0$ can be an eigenvalue.)

Eigenvalues are characterized as the roots of the characteristic polynomial

$$p_A(\lambda) \;:=\; \det(\lambda I - A).$$

Since p_A is a degree-p polynomial, every $A \in \mathbb{R}^{p\times p}$ has exactly p eigenvalues in $\mathbb{C}$, counting algebraic multiplicities. Real matrices can have complex eigenvalues; non-real eigenvalues come in conjugate pairs.

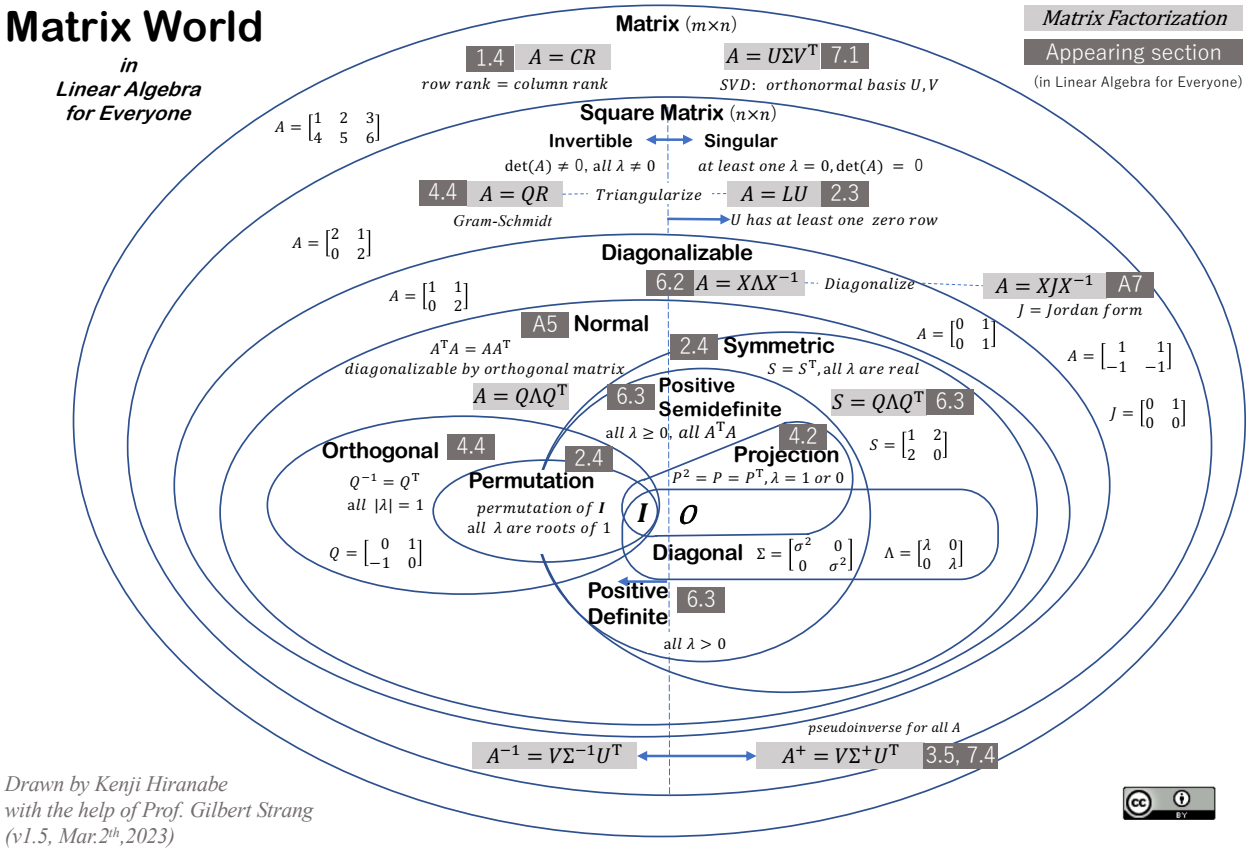

Figure 4: The "Matrix World": A map of special matrix classes. Figure courtesy of Hiranabe (2021).

Two global quantities, trace and determinant, can be read off from the eigenvalues $\lambda_1, \ldots, \lambda_p$ (counting multiplicities):

$$\operatorname{tr}(A) = \sum_{i=1}^{p} \lambda_i, \qquad \det(A) = \prod_{i=1}^{p} \lambda_i.$$

A few basic transformations are worth remembering: $A^\top$ has the same eigenvalues as A; adding a multiple of the identity shifts the spectrum ($A + cI$ has eigenvalues $\lambda_i + c$); powers map eigenvalues to powers (A^k has eigenvalues λ_i^k); and if A is invertible then A^{-1} has eigenvalues λ_i^{-1}.

Finally, eigenvalues are invariant under a change of coordinates: if S is invertible, then A and $S^{-1}AS$ have the same eigenvalues. This reflects that eigenvalues are intrinsic to the linear map, not to a particular basis.

For an eigenvalue $\lambda \in \mathbb{C}$, the associated *eigenspace* is $\mathcal{E}_\lambda(A) := \operatorname{null}(A - \lambda I) \subseteq \mathbb{C}^p$. If λ is repeated, $\mathcal{E}_\lambda(A)$ can be multi-dimensional, and any nonzero vector in that subspace is an eigenvector. Thus eigenvectors are generally non-unique: even in a one-dimensional eigenspace the sign/scale is arbitrary, and in a higher-dimensional eigenspace any basis works.

For a general (non-symmetric) real matrix, eigenvalues may be complex and there may not exist a basis of eigenvectors (i.e., the matrix may fail to be diagonalizable). In statistics/ML, the most important case is the symmetric one: covariances, Gram matrices, and Hessians are typically symmetric, and the spectral theory is especially clean. We review this case next.

3.2 Real symmetric matrices and the spectral theorem

From here on, assume $A \in \mathbb{S}^p$ is real symmetric. Then all eigenvalues of A are real, and A admits an orthonormal eigenbasis. Concretely, there exists an orthogonal matrix $U = [u_1|\cdots|u_p] \in \mathbb{R}^{p\times p}$ and a diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p) \in \mathbb{R}^{p\times p}$ such that

$$A = U\Lambda U^\top, \qquad \text{equivalently} \qquad A = \sum_{i=1}^{p} \lambda_i \, u_i u_i^\top.$$

The multiset of eigenvalues $\{\lambda_1, \ldots, \lambda_p\}$ is uniquely determined by A (as the roots of p_A), up to permutation. Eigenvectors are not unique: one can flip signs, and if an eigenvalue has multiplicity $m > 1$, any orthonormal basis of its eigenspace yields a valid set of eigenvectors (i.e., one can rotate within that invariant subspace without changing A).

A useful structural corollary is that, for symmetric A, the rank is the number of nonzero eigenvalues: $\mathrm{rank}(A) = \#\{i : \lambda_i \neq 0\}$.

3.3 Rayleigh quotient and variational characterizations

A central bridge between eigenvalues and optimization is the Rayleigh quotient. For $A \in \mathbb{S}^p$ and $x \neq 0$,

$$\mathcal{R}_A(x) \;:=\; \frac{x^\top A x}{x^\top x}.$$

If we order the eigenvalues as $\lambda_1 \geqslant \cdots \geqslant \lambda_p$, then the extreme eigenvalues solve simple constrained optimization problems:

$$\lambda_1 = \max_{\|x\|_2=1} x^\top A x, \qquad \lambda_p = \min_{\|x\|_2=1} x^\top A x,$$

and the maximizer/minimizer can be chosen as a top/bottom unit eigenvector. In particular, if $\Sigma \succeq 0$ is a covariance matrix, then $v^\top \Sigma v$ is the variance of the projection onto direction v; thus PCA can be viewed as repeatedly maximizing $v^\top \Sigma v$ over unit vectors (with orthogonality constraints to extract multiple components).

More generally, the k-th eigenvalue can be characterized through orthogonality constraints or through subspace min–max formulations. A useful statement (Courant–Fischer) is:

$$\lambda_k(A) = \max_{\substack{E\subseteq\mathbb{R}^p \\ \dim(E)=k}} \min_{\substack{x\in E \\ \|x\|_2=1}} x^\top A x \;=\; \min_{\substack{E\subseteq\mathbb{R}^p \\ \dim(E)=p-k+1}} \max_{\substack{x\in E \\ \|x\|_2=1}} x^\top A x.$$

Intuitively, the first equality says: among all k-dimensional subspaces, choose the one on which the quadratic form is as large as possible in its *worst* direction; the optimal subspace is $\mathrm{span}(u_1, \ldots, u_k)$.

3.4 Positive semidefinite matrices through eigenvalues

Eigenvalues give especially clean characterizations of positive semidefinite (PSD) and positive definite (PD) matrices. For $A \in \mathbb{S}^p$,

$$A \succeq 0 \quad\Longleftrightarrow\quad \lambda_i(A) \geqslant 0 \text{ for all } i, \qquad A \succ 0 \quad\Longleftrightarrow\quad \lambda_i(A) > 0 \text{ for all } i.$$

Equivalently, $A \succeq 0$ iff $A = B^\top B$ for some (possibly rectangular) matrix B; and $A \succ 0$ iff A is invertible and $A^{-1} \succ 0$.

When $A \succeq 0$, the spectral theorem also provides a canonical square root: if $A = U\,\mathrm{diag}(\lambda_1, \dots, \lambda_p)U^\top$ with $\lambda_i \geqslant 0$, then

$$A^{1/2} \;=\; U\,\mathrm{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_p})U^\top$$

is PSD and satisfies $A^{1/2}A^{1/2} = A$. This square root is unique among PSD matrices.

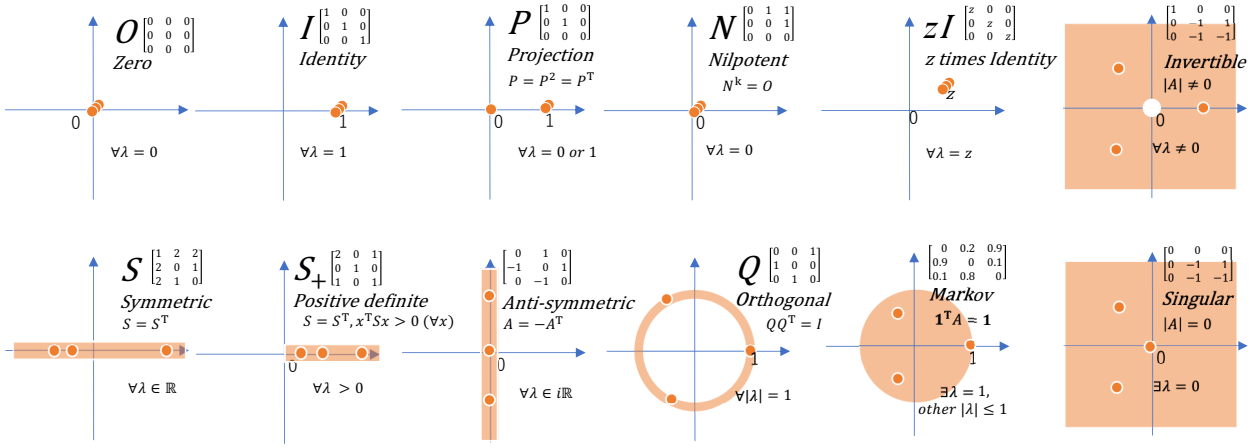

Figure 5: Map of eigenvalues. Figure courtesy of Hiranabe (2021).

4 Singular values and singular vectors

Eigenvalue decompositions apply cleanly to symmetric matrices. The singular value decomposition (SVD) extends the same geometric and variational ideas to *arbitrary* rectangular data matrices. Throughout, let $A \in \mathbb{R}^{n\times p}$, let $q := \min\{n, p\}$, and let $r := \mathrm{rank}(A)$.

4.1 Singular value decomposition and basic consequences

The SVD states that any matrix can be factorized using orthogonal changes of basis on the left and right:

$$A = U\Sigma V^\top,$$

where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{p\times p}$ are orthogonal, and $\Sigma \in \mathbb{R}^{n\times p}$ is rectangular diagonal with nonnegative diagonal entries $\sigma_1(A) \geqslant \sigma_2(A) \geqslant \cdots \geqslant \sigma_q(A) \geqslant 0$. The numbers $\sigma_i(A)$ are the *singular values* of A, and the columns of U and V are called the *left* and *right* singular vectors, respectively.

It is often convenient to keep only the nonzero singular values. Writing $\Sigma_r = \mathrm{diag}(\sigma_1, \dots, \sigma_r)$ and letting $U_r = [u_1|\cdots|u_r] \in \mathbb{R}^{n\times r}$, $V_r = [v_1|\cdots|v_r] \in \mathbb{R}^{p\times r}$, we have the *compact SVD* $A = U_r\Sigma_r V_r^\top$. In particular,

$$A = \sum_{i=1}^{r} \sigma_i\, u_i v_i^\top, \qquad Av_i = \sigma_i u_i, \qquad A^\top u_i = \sigma_i v_i \quad (i = 1, \dots, r).$$

The rank r is exactly the number of nonzero singular values.

A useful way to remember the geometry is: A maps an orthonormal basis (v_i) in $\mathbb{R}^p$ to orthogonal directions (u_i) in $\mathbb{R}^n$, stretching the i-th direction by σ_i. Equivalently, A maps the unit sphere in $\mathbb{R}^p$ to an ellipsoid in $\mathbb{R}^n$ whose principal axes are u_i with semi-axis lengths σ_i.

4.2 Connection to Gram matrices and variational characterizations

The SVD is tightly connected to the spectral theory of symmetric positive semidefinite matrices. From $A = U\Sigma V^\top$,

$$A^\top A = V(\Sigma^\top \Sigma)V^\top = \sum_{i=1}^{r} \sigma_i^2\, v_i v_i^\top, \qquad AA^\top = U(\Sigma\Sigma^\top)U^\top = \sum_{i=1}^{r} \sigma_i^2\, u_i u_i^\top.$$

Thus the right singular vectors v_i are eigenvectors of $A^\top A$, the left singular vectors u_i are eigenvectors of $AA^\top$, and the nonzero eigenvalues of both Gram matrices are σ_i^2. In particular, for $k \leqslant q$, $\sigma_k(A) = \sqrt{\lambda_k(A^\top A)} = \sqrt{\lambda_k(AA^\top)}$ (with eigenvalues ordered from largest to smallest).

The largest singular value has a simple optimization interpretation: it is the maximum Euclidean stretch factor of the linear map $x \mapsto Ax$,

$$\|A\|_2 \;:=\; \max_{\|x\|_2=1} \|Ax\|_2 \;=\; \sigma_1(A).$$

More generally, the Courant–Fischer min–max theorem applied to $A^\top A$ yields variational formulas for $\sigma_k(A)$. One convenient consequence is a sequential characterization: after extracting the top $k-1$ right singular directions, $\sigma_k(A)$ is the maximum of $\|Ax\|_2$ over unit vectors x orthogonal to $\operatorname{span}(v_1, \ldots, v_{k-1})$.

In algorithms it is common to work with leading singular subspaces. Let $U_k = [u_1|\cdots|u_k]$ and $V_k = [v_1|\cdots|v_k]$ for $k \leqslant r$. The corresponding orthogonal projectors are $P_{U_k} = U_k U_k^\top$ onto $\operatorname{span}(U_k) \subseteq \mathbb{R}^n$ and $P_{V_k} = V_k V_k^\top$ onto $\operatorname{span}(V_k) \subseteq \mathbb{R}^p$, with complementary projectors $I - P_{U_k}$ and $I - P_{V_k}$.

4.3 Pseudoinverse and least squares

The SVD also provides the canonical generalized inverse. Using the compact SVD $A = U_r \Sigma_r V_r^\top$, define the Moore–Penrose pseudoinverse

$$A^\dagger \;:=\; V_r \Sigma_r^{-1} U_r^\top, \qquad \Sigma_r^{-1} = \operatorname{diag}(\sigma_1^{-1}, \ldots, \sigma_r^{-1}).$$

When A is square and invertible, $A^\dagger = A^{-1}$, and the inverse has the SVD-form $A^{-1} = V\Sigma^{-1}U^\top = \sum_{i=1}^{p} \sigma_i^{-1} v_i u_i^\top$.

The pseudoinverse is characterized by the Moore–Penrose identities $AA^\dagger A = A$, $A^\dagger AA^\dagger = A^\dagger$, and symmetry of $AA^\dagger$ and $A^\dagger A$. Geometrically, $AA^\dagger$ is the orthogonal projector onto $\operatorname{range}(A) \subseteq \mathbb{R}^n$, while $A^\dagger A$ projects onto $\operatorname{range}(A^\top) \subseteq \mathbb{R}^p$.

In least squares, $x^\star = A^\dagger b$ is the canonical solution of $\min_x \|Ax - b\|_2$. If the minimizer is not unique (e.g. $p > n$ or A is rank-deficient), then $x^\star$ is the unique minimizer with minimum Euclidean norm $\|x\|_2$.

4.4 Low-rank approximation

SVD is the canonical tool for low-rank structure. Let $A = U\Sigma V^\top$ and define the rank-k truncation $A_k := U_k \Sigma_k V_k^\top$, where $\Sigma_k = \operatorname{diag}(\sigma_1, \ldots, \sigma_k)$. The Eckart–Young–Mirsky theorem states that A_k is a best rank-k approximation to A in both Frobenius and spectral norms, and the approximation error is controlled exactly by the tail singular values:

$$\|A - A_k\|_2 = \sigma_{k+1}, \qquad \|A - A_k\|_F^2 = \sum_{i=k+1}^{r} \sigma_i^2.$$

Finally, if $M \in \mathbb{R}^{p\times p}$ is symmetric positive semidefinite, then its SVD coincides with its eigendecomposition: $M = U\Lambda U^\top$ with $\Lambda \succeq 0$, the left and right singular vectors coincide ($U = V$), and $\sigma_i(M) = \lambda_i(M)$. In this case $\operatorname{tr}(M) = \sum_{i=1}^{r} \sigma_i(M)$, and the top singular/eigenvector maximizes the Rayleigh quotient $x^\top Mx/(x^\top x)$.

A side note: For a comprehensive graphic summary of the five main matrix decompositions ($A = LU$, $A = QR$, $A = CR$, $S = U\Lambda U^\top$, and $A = U\Sigma V^\top$) (we only recapped the last two), see Hiranabe (2021).

5 Matrix norms and their properties

Matrix norms quantify the size of a matrix viewed as (i) a *linear operator* (how much it can stretch vectors), or (ii) a *collection of entries/columns* (useful for regularization and structured sparsity). As before, let $A \in \mathbb{R}^{n\times p}$, $q := \min\{n, p\}$, and $r := \operatorname{rank}(A)$.

5.1 Induced operator norms

A *matrix norm* is simply a norm on the vector space $\mathbb{R}^{n\times p}$: a mapping $\|\cdot\| : \mathbb{R}^{n\times p} \to \mathbb{R}_{\geqslant 0}$ such that $\|A\| = 0$ iff $A = 0$, $\|\alpha A\| = |\alpha|\,\|A\|$, and $\|A + B\| \leqslant \|A\| + \|B\|$. In analysis, it is often important that a matrix norm be *submultiplicative* (also called *consistent*):

$$\|MN\| \;\leqslant\; \|M\|\,\|N\| \qquad \text{(whenever the product } MN \text{ is defined).}$$

This property is the algebraic form of stability: it lets us control how perturbations propagate through products.

A standard way to build submultiplicative matrix norms is to start from a vector norm. Given $\|\cdot\|_p$ on vectors, the induced operator norm is

$$\|A\|_{p\to p} \;:=\; \sup_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p} \;=\; \sup_{\|x\|_p=1} \|Ax\|_p.$$

Induced norms are automatically submultiplicative. Two entrywise-friendly special cases (useful in concentration bounds and worst-case analyses) are

$$\|A\|_1 = \max_{1\leqslant j\leqslant p} \sum_{i=1}^{n} |A_{ij}| \quad \text{(maximum absolute column sum),}$$

$$\|A\|_\infty = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{p} |A_{ij}| \quad \text{(maximum absolute row sum).}$$

5.2 Frobenius norm

The Frobenius norm treats a matrix as a vector of its entries:

$$\|A\|_F \;:=\; \Big(\sum_{i=1}^{n}\sum_{j=1}^{p} A_{ij}^2\Big)^{1/2}.$$

It is induced by the Frobenius inner product $\langle A, B\rangle_F := \sum_{i,j} A_{ij}B_{ij} = \operatorname{tr}(A^\top B)$, so $\|A\|_F^2 = \operatorname{tr}(A^\top A)$. If $\sigma_1(A) \geqslant \cdots \geqslant \sigma_r(A) > 0$ are the nonzero singular values of A (where $r = \operatorname{rank}(A)$),

then

$$\|A\|_F^2 \;=\; \operatorname{tr}(A^\top A) \;=\; \sum_{i=1}^{r} \sigma_i(A)^2.$$

This singular-value expression is often the most convenient way to reason about $\|A\|_F$ in high-dimensional problems.

A small but useful probabilistic identity is that if $Z \in \mathbb{R}^p$ is isotropic ($\mathbb{E}[ZZ^\top] = I_p$), then $\mathbb{E}\|AZ\|_2^2 = \|A\|_F^2$. This shows up repeatedly in random matrix theory and trace estimation.

5.3 Spectral operator norm

The most important induced norm in spectral methods is the ℓ_2-operator norm, also called the *spectral norm*:

$$\|A\|_2 \;:=\; \max_{\|x\|_2=1} \|Ax\|_2 \;=\; \sigma_1(A) \;=\; \sqrt{\lambda_{\max}(A^\top A)} \;=\; \max_{\|x\|_2=\|y\|_2=1} |y^\top A x|.$$

It measures the maximum stretch factor of the linear map $x \mapsto Ax$. In the symmetric case $A \in \mathbb{S}^p$, the singular values are $|\lambda_i(A)|$, so $\|A\|_2 = \max_i |\lambda_i(A)|$.

Both $\|\cdot\|_F$ and $\|\cdot\|_2$ are invariant under orthogonal changes of basis: if $Q \in \mathbb{R}^{n\times n}$ and $R \in \mathbb{R}^{p\times p}$ are orthogonal, then

$$\|QAR\|_F = \|A\|_F, \qquad \|QAR\|_2 = \|A\|_2.$$

Equivalently, these norms depend only on the singular values, not on the singular vectors.

A basic comparison that is often used when translating between entrywise and spectral control is

$$\|A\|_2 \;\leqslant\; \|A\|_F \;\leqslant\; \sqrt{r}\,\|A\|_2 \qquad \text{where } r = \operatorname{rank}(A).$$

The lower bound is tight for rank-one matrices; the upper bound is tight when the nonzero singular values are all equal.

Two quick sanity checks that are handy in calculations: (i) if $A = uv^\top$ is rank one, then $\|A\|_2 = \|A\|_F = \|u\|_2\|v\|_2$; (ii) if D is diagonal with entries d_i, then $\|D\|_2 = \max_i |d_i|$.

5.4 Schatten norms

Since $\|\cdot\|_2$ and $\|\cdot\|_F$ are both functions of the singular values, it is useful to name the whole family. Let $\sigma_1(A) \geqslant \cdots \geqslant \sigma_q(A) \geqslant 0$ be the singular values of A. For $s \in [1,\infty)$, define the Schatten s-norm by

$$\|A\|_{S_s} \;:=\; \Big(\sum_{i=1}^{q} \sigma_i(A)^s\Big)^{1/s}, \qquad \|A\|_{S_\infty} := \max_i \sigma_i(A) = \|A\|_2.$$

Key special cases are:

$$\|A\|_{S_2} = \|A\|_F \quad \text{(Frobenius)}, \qquad \|A\|_{S_\infty} = \|A\|_2 \quad \text{(spectral)}, \qquad \|A\|_{S_1} = \sum_i \sigma_i(A) \quad \text{(nuclear norm)}.$$

The nuclear norm is widely used as a convex surrogate for rank in low-rank estimation problems (e.g. matrix completion and trace-norm regularization).

Source material

Parts of this lecture are based on references: Vershynin (2018); Strang (2020); Hiranabe (2021), in addition to the author's accumulated experience working on related topics.

References

Hiranabe, K. (2021). The art of linear algebra. https://github.com/kenjihiranabe/The-Art-of-Linear-Algebra. Graphic notes on "Linear Algebra for Everyone"' by Gilbert Strang.

Strang, G. (2020). *Linear Algebra for Everyone*. SIAM.

Vershynin, R. (2018). *High-dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press.