General math toolkit

A collection of definitions, theorems, and formulae

March 2, 2026

Contents

1 Algebra

1.1 Definition of a group

A *group* is a set G together with an operation $\cdot$ satisfying:

(a) Closure: $\forall a, b \in G,\ a \cdot b \in G$.

(b) Associativity: $(a \cdot b) \cdot c = a \cdot (b \cdot c)$.

(c) Identity: $\exists e \in G$ such that $\forall a \in G,\ e \cdot a = a \cdot e = a$.

(d) Inverses: $\forall a \in G,$ there exists $a^{-1} \in G$ such that $a \cdot a^{-1} = a^{-1} \cdot a = e$.

A group $(G,\ \cdot\ )$ which is is commutative is known as an Abelian group.

1.2 Manifold Hypothesis

"Data often concentrates around a lower-dimensional manifold." Real-world datasets $\mathcal{X} \subset \mathbb{R}^n$ often lie (approximately) on a much lower-dimensional smooth manifold $\mathcal{M} \subset \mathbb{R}^n$:

$$\mathcal{X} \ \subset\ \mathcal{M} \ \subset\ \mathbb{R}^n,$$

where $\mathcal{M}$ has intrinsic dimension $d \ll n$. Equivalently, there exists a *smooth, injective* mapping

$$\phi\colon \mathbb{R}^d \ \longrightarrow\ \mathbb{R}^n$$

such that for each data point $x \in \mathcal{X}$, one can write

$$x \ =\ \phi(z) \ +\ \epsilon,$$

with $z \in \mathbb{R}^d$ and ϵ a small perturbation (noise). The manifold hypothesis justifies using the low intrinsic dimensionality d to build models and algorithms for:

- Dimensionality reduction (e.g. PCA, autoencoders)
- Data visualization in 2D or 3D
- Improved generalization by exploiting the manifold structure

1.3 Homomorphisms of groups

Let $(G, \cdot)$ and $(H, *)$ be groups. A function $f\colon G \to H$ is a *homomorphism* if for all $x, y \in G$,

$$f(x \cdot y) \ =\ f(x) * f(y).$$

Thus f "preserves" the group operation between G and H.

1.4 Topological Manifold.

An n-dimensional set M is a *manifold* if every point $p \in M$ has a neighborhood $U \subset M$ that contains p such that there exists a homeomorphism

$$\phi : U \ \longrightarrow\ V \subset \mathbb{R}^n.$$

Namely, M locally "looks like" Euclidean space.

2 Linear algebra

2.1 Inverse of a 2×2 Matrix

For a matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{with} \quad \det(A) = ad - bc \neq 0,$$

its inverse is

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

2.2 Determinant

For $A \in \mathbb{R}^{n \times n}$ the determinant of A is

$$\det(A) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)}$$

where $S_n = \{\text{all permutations } \sigma \text{ of } [n]\}$. Geometrically, $|\det A|$ is the n-dimensional volume-scaling factor of the linear map $x \mapsto Ax$, and $\operatorname{sgn}(\det A)$ encodes orientation (over $\mathbb{R}$).

The determinant has the following useful identities:

- $\det(AB) = \det(A)\det(B)$
- $\det(A^\top) = \det(A)$
- $\det(P^{-1}AP) = \det(A)$
- A invertible $\iff \det(A) \neq 0$
- $\det(A^{-1}) = (\det A)^{-1}$
- Row swap: $\det \mapsto -\det$
- Scale a row by c: $\det \mapsto c \det$
- Add multiple of one row to another: $\det$ unchanged
- Eigenvalues (λ_i) : $\det(A) = \prod_{i=1}^{n} \lambda_i$ hence a zero eigenvalue implies singularity
- Block triangular $A = \begin{bmatrix} B & * \\ 0 & C \end{bmatrix}$: $\det(A) = \det(B)\det(C)$ for any matrix $*$
- Jacobi's formula: $\dfrac{d}{dt}\det(A(t)) = \det(A(t))\ \operatorname{tr}(A(t)^{-1}\dot{A}(t))$
- Matrix exponential: $\det(e^A) = e^{\operatorname{tr} A}$

2.3 Matrix inverse: core identities and intuition

Let $A, B \in \mathbb{R}^{n \times n}$ be invertible and $\alpha \in \mathbb{R} \setminus \{0\}$.

1. *Definition (two-sided inverse).* A^{-1} is the unique matrix satisfying

$$AA^{-1} = I \quad \text{and} \quad A^{-1}A = I.$$

2. *Product rule (reverse order).*

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

Intuition: applying AB means do B then A; undoing reverses the steps.

3. *Transpose / adjoint.*

$$(A^\top)^{-1} = (A^{-1})^\top, \qquad \text{and (over } \mathbb{C}) \qquad (A^*)^{-1} = (A^{-1})^*.$$

4. *Scalar multiple.*

$$(\alpha A)^{-1} = \alpha^{-1} A^{-1}.$$

5. *Inverse of an inverse.*

$$(A^{-1})^{-1} = A.$$

6. *When does A^{-1} exist?*

   A^{-1} exists $\iff$ $\det(A) \neq 0$ $\iff$ A is full rank $\iff$ $Ax = b$ has a unique solution for all b.

7. *Geometric intuition.* A is an invertible linear change of coordinates: it maps the unit cube to a parallelepiped with nonzero volume, and A^{-1} maps back.

2.4 Special identities for $(I - A)^{-1}$

Let $A \in \mathbb{R}^{n \times n}$.

1. *Neumann series (geometric series).* If $\|A\| < 1$ for any submultiplicative norm, then

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k \qquad \text{and}$$

   Intuition: $(I - A)^{-1}$ aggregates repeated effects $I + A + A^2 + \cdots$.

2. *Resolvent identity.* If $I - A$ and $I - B$ are invertible, then

$$(I - A)^{-1} - (I - B)^{-1} = (I - A)^{-1}(A - B)(I - B)^{-1}.$$

   Useful for perturbations: small $A - B$ gives small change in the inverse (scaled by the two resolvents).

3. *Eigenvalue mapping.* If λ is an eigenvalue of A, then (when defined) $\frac{1}{1-\lambda}$ is an eigenvalue of $(I - A)^{-1}$.

2.5 Column and row pictures of matrix multiplication.

- *Column picture:* If $AB = C$, then each column of C is a linear combination of the columns of A, with coefficients given by the corresponding column of B.
- *Row picture:* If $AB = C$, then each row of C is a linear combination of the rows of B, with weights from the corresponding row of A.

2.6 Positive semidefinite (PSD) matrices over $\mathbb{R}$

For any $A \in \mathbb{R}^{n \times n}$,

$$A \succeq 0 \quad \iff \quad x^\top A x \geq 0 \ \ \forall x \in \mathbb{R}^n.$$

If $A \succeq 0$, then A is automatically symmetric (when we're working in $\mathbb{R}$). To see this, let $q(x) = x^\top A x$, then

$$x^\top A y = \tfrac{1}{4}(q(x + y) - q(x - y)) = \tfrac{1}{4}(q(y + x) - q(y - x)) = y^\top A x,$$

which shows that $x^\top Ay = y^\top Ax$ for all x, y, hence $A = A^\top$.

One canonical example is the normal form, which says that for any $C \in \mathbb{R}^{m\times n}$,

$$x^\top (C^\top C)x = \|Cx\|_2^2 \geq 0 \qquad \text{therefore} \qquad C^\top C \succeq 0.$$

For any $A \in \mathbb{R}^{n\times n}$, the following are equivalent:

(1) $A \succeq 0$
(2) $\exists\, B$ (not necessarily square) : $A = B^\top B$ hence "Cholesky" square roots always exist
(3) All eigenvalues of A are nonnegative: $\lambda_i(A) \geq 0$ for all i
(4) $A = Q\Lambda Q^\top$ with Q orthogonal, $\Lambda = \text{diag}(\lambda_i \geq 0)$
(5) $x^\top Ax = \|A^{1/2}x\|_2^2$ for the unique PSD square root $A^{1/2}$

We also have quadratic-form viewpoints that characterize PSD in terms of non-negativity of polynomials: A is PSD if

$$x^\top Ax = \|A^{1/2}x\|_2^2 = \text{Tr}(A\, xx^\top) \geq 0.$$

Additionally, if A is PSD, then it's eigenvalues satisfy

$$\lambda_{\min}(A) = \min_{\|x\|=1} x^\top Ax, \quad \lambda_{\max}(A) = \max_{\|x\|=1} x^\top Ax.$$

As mentioned above, $A \succeq 0$ implies non-negative eigenvalues $\lambda_i(A) \geq 0$ for all i. Zero eigenvalues are allowed (recall that zero eigenvalues imply singular matrices); in fact

$$x^\top Ax = 0 \iff A^{1/2}x = 0 \iff x \in \ker(A).$$

Counting the dimension of the subspace of x such that the above holds proves that $\text{rank}(A) = \text{rank}(A^{1/2})$.

Lastly, we have the following facts and useful properties:

- Convex cone: if $A, B \succeq 0$ and $\alpha, \beta \geq 0$, then $\alpha A + \beta B \succeq 0$.
- Congruence invariance: for any M, $A \succeq 0 \;\Rightarrow\; M^\top AM \succeq 0$.
- Trace inner product: if $A, B \succeq 0$, then $\text{Tr}(AB) \geq 0$; moreover $x^\top ABx \geq 0$ for all x if A and B commute and are PSD.
- Square roots and factorizations: there is a unique PSD $A^{1/2}$ with $A^{1/2}A^{1/2} = A$, and a (possibly rectangular) Cholesky factorization $A = B^\top B$. if $A \succ 0$, the Cholesky factor can be taken upper triangular.
- Monotonicity under sums/products: $A \succeq 0,\ B \succeq 0 \;\Rightarrow\; A + B \succeq 0$ and $BAB^\top \succeq 0$.

2.7 Conjugate transpose

The conjugate transpose of a complex matrix $A \in \mathbb{C}^{n\times m}$ is

$$A^* := \overline{A}^{\mathsf{T}} \qquad \text{entrywise: } (A^*)_{ij} = \overline{A_{ji}}$$

so we first by first taking the complex conjugate of each element (changing the sign of the imaginary part) and then performing a matrix transpose. For any $A, B \in \mathbb{C}^{n\times n}$, some useful properties include

$$(A + B)^* = A^* + B^*$$

$$(\alpha A)^* = \overline{\alpha}\, A^*$$

$$(AB)^* = B^* A^*,$$

$$(A^{-1})^* = (A^*)^{-1} \quad \text{(assuming that } A \text{ is invertible)}$$

Additionally, properties related to trace and determinant include that For any $A \in \mathbb{C}^{n\times n}$

$$\operatorname{tr}(A^*) = \overline{\operatorname{tr}(A)}, \qquad \det(A^*) = \overline{\det(A)}$$

Lastly,

$$A^*A \text{ is Hermitian and positive semidefinite for all } A \in \mathbb{C}^{n\times n}$$

$$A = A^* \implies A \text{ is known as Hermitian,} \qquad A^*A = I \implies A \text{ is known as unitary.}$$

2.8 Unitary matrix

A unitary matrix is a square matrix of complex numbers whose inverse is equal to its conjugate transpose (also called its Hermitian adjoint)

$$U \in \mathbb{C}^{n\times n} \text{ is unitary } \iff U^*U = UU^* = I \iff U^{-1} = U^*.$$

$$\text{Columns/rows of } U \text{ form an orthonormal basis;} \qquad \langle Ux, Uy\rangle = \langle x, y\rangle, \quad \forall x, y.$$

$$\boxed{\|Ux\|_2 = \|x\|_2 \quad \text{(norm/energy preserving)}}.$$

$$|\det U| = 1, \qquad \det U = e^{i\theta}.$$

- Unitary matrices are length- and angle-preserving linear maps in $\mathbb{C}^n$; in the real case they are orthogonal matrices (rotations/reflections).
- They preserve inner products, probabilities, and energy.
- Any matrix A admits a polar decomposition $A = UH$ with U unitary and H Hermitian positive semidefinite.

2.9 Vandermonde's Matrix

The Vandermonde matrix on n variables is

$$V(x_0, \ldots, x_{n-1}) = \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^{n-1} \\ 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n-1} & x_{n-1}^2 & \cdots & x_{n-1}^{n-1} \end{bmatrix}$$

The determinant has the useful factorization

$$\det V(x_0, \ldots, x_{n-1}) = \prod_{0\le i<j\le n-1} (x_j - x_i) \quad \Rightarrow \quad V \text{ invertible} \iff x_i \text{ are pairwise distinct.}$$

Lastly, two useful facts are

- Permuting the x_i permutes the rows, and det changes by the sign of the permutation.
- The DFT matrix is a Vandermonde with $x_j = \omega^j$ (roots of unity)

2.10 Direction Vector Between Two Points

Identity. To obtain a vector pointing from y towards x in $\mathbb{R}^n$, simply take

$$x - y.$$

Explanation. This "to minus from" convention yields a vector whose tail is at y and whose head is at x.

2.11 Linearity of the Dot Product

Identity. For any vectors $d, x, y \in \mathbb{R}^n$,

$$\langle d,\ x - y\rangle \ =\ \langle d, x\rangle \ -\ \langle d, y\rangle.$$

Explanation. This follows from bilinearity of the inner product: it distributes over vector addition and scalar multiplication.

2.12 Add-subtract trick for vectors

Add–Subtract (Zero-Insertion) Identity. For any objects a, b, c in an inner-product space (including real numbers, vectors, or random variables in L^2)

$$\|a - b\|^2 = \|a - c\|^2 + \|c - b\|^2 + 2\langle a - c,\ c - b\rangle$$

which, in the scalar case, is

$$(a - b)^2 = (a - c)^2 + (c - b)^2 + 2(a - c)(c - b)$$

Note: there are several cases under which the cross term above will be 0, such as in a probabilistic context.

2.13 Eigenvalue and Spectral Decompositions

Eigenvalue Decomposition. A square matrix $A \in \mathbb{R}^{n\times n}$ is *diagonalizable* if it admits a factorization

$$A = V\,\Lambda\,V^{-1},$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ contains the eigenvalues of A on its diagonal, and $V = [v_1\ \ldots\ v_n]$ has the corresponding (linearly independent) eigenvectors as columns. This decomposition exists if and only if A has a full set of n independent eigenvectors. Note that in the general (non-normal) case, V need not be orthogonal.

Trace Identity for Rank-One Updates. For any $X \in \mathbb{R}^{n\times n}$ and vectors $y, z \in \mathbb{R}^n$,

$$\mathrm{Tr}\big(X\,(y\,z^\top)\big) \ =\ z^\top X\,y.$$

Here $y\,z^\top$ is a rank-one matrix, and the cyclic property of the trace gives the identity.

Spectral Decomposition (Normal/Symmetric Case). If A is normal (in particular, real symmetric or complex Hermitian), then there is an *orthonormal* basis of eigenvectors v_i with real eigenvalues λ_i. One can write

$$A \ =\ \sum_{i=1}^{n} \lambda_i\, v_i\, v_i^\top, \quad V = [v_1\ \cdots\ v_n] \text{ orthogonal},$$

so that equivalently

$$A = V\,\Lambda\,V^\top,$$

with $V^\top V = I$ and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$.

2.14 Vector norms

Let $p \geq 1$, the p-norm (also called the ℓ^p-norm of a vector $\mathbf{x} = (x_1, \ldots, x_n)$ is

$$||\mathbf{x}||_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

For $i = 2$ we get the Euclidean norm, which is induced by the canonical inner product

$$||\mathbf{x}||_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$$

For $i \to \infty$,

$$||\mathbf{x}||_\infty = \max_i |x_i|$$

The above norms have the ordering

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \|x\|_1 \leq \sqrt{n}\, \|\mathbf{x}\|_2 \leq n\, \|\mathbf{x}\|_\infty.$$

Lastly, two common identitites are

$$|\mathbf{x}^\top \mathbf{y}| \leq \|\mathbf{x}\|_p \|\mathbf{y}\|_q \text{ (Hölder)}, \qquad |\mathbf{x}^\top \mathbf{y}| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \text{ (Cauchy–Schwarz)}.$$

2.15 SVD decomposition

Any matrix $A \in \mathbb{R}^{m \times n}$ can be decomposed as

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

where

$$U \in \mathbb{R}^{m \times m} \text{ and } V \in \mathbb{R}^{n \times n} \text{ are orthogonal}, \qquad \Sigma = \begin{bmatrix} \operatorname{diag}(\sigma_1, \ldots, \sigma_r) & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{m \times n},$$

with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ and $r = \operatorname{rank}(A)$. The columns of U are the *left singular vectors* and the columns of V are the *right singular vectors*.

Additionally, we have corresponence between singular values of A and eigenvalues of $A^\top A$ and $AA^\top$:

$$\sigma_i = \sqrt{\lambda_i(A^\top A)} = \sqrt{\lambda_i(AA^\top)} \quad \text{(nonzero parts match)},$$

In particular, when A is symmetric positive semidefinite, its singular values coincide with its (nonnegative) eigenvalues.

2.16 Matrix norms

We now define matrix norms.

$$\|A\|_{p\to q} := \sup_{x\neq 0} \frac{\|Ax\|_q}{\|x\|_p} \quad \text{(induced/operator norm)},$$

$$\|A\|_2 = \sigma_{\max}(A) = \sqrt{\lambda_{\max}(A^\top A)} \quad \text{using correspondence between sv's and eigenvalues},$$

$$\|A\|_F := \sqrt{\sum_{i,j} a_{ij}^2} = \sqrt{\operatorname{tr}(A^\top A)} = \|\operatorname{vec}(A)\|_2 \qquad \text{(Frobenius norm)},$$
$$\|A\|_* := \sum_i \sigma_i(A) \quad \text{(nuclear/trace norm)}$$

Useful facts

1. For all matrix norms $\cdot$

$$\|AB\|_\cdot \leq \|A\|_\cdot \|B\|_\cdot \quad \text{(submultiplicativity for any induced norm)}.$$

2. We have the following relation between L2 and Frobenious matrix norms

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{\operatorname{rank}(A)}\, \|A\|_2.$$

3. For all orthogonal U, V,

$$\|UAV\|_2 = \|A\|_2, \quad \|UAV\|_F = \|A\|_F, \quad \|UAV\|_* = \|A\|_*$$

4. ℓ_1: promotes sparsity; ℓ_2: energy/Euclidean length; ℓ_∞: worst-case component.

5. Spectral norm: largest stretch of A; Frobenius: RMS of entries; Nuclear: sum of singular values (rank surrogate).

2.17 Metric spaces

At its core, a metric space is simply a set of points and a function—called a metric—which us how far apart any two points are. In particular, a *metric space* is a pair (X, d) where $d : X \times X \to [0, \infty)$ satisfies the following for all $x, y, z \in X$:

$$\begin{aligned} &\text{(i) } d(x,y) = 0 \iff x = y \quad \text{(identity of indiscernibles)} \\ &\text{(ii) } d(x,y) = d(y,x) \quad \text{(symmetry)} \\ &\text{(iii) } d(x,z) \leq d(x,y) + d(y,z) \quad \text{(triangle inequality)}. \end{aligned}$$

where the intuition for the triangle inequality is that "the distance of going straight can be no longer than making a detour:" $d(x, z) \leq d(x, y) + d(y, z)$.

As a consequence of this definition we have

- *Reverse triangle:* $|d(x, z) - d(y, z)| \leq d(x, y)$.
- *Continuity of d:* $(x_n, y_n) \to (x, y) \implies d(x_n, y_n) \to d(x, y)$.
- A sequence $x_n \to x$ iff $d(x_n, x) \to 0$.

Metric spaces also produce open sets of X using the metric $d(x)$:

- All open sets of X can be written as a union of open balls, where an open ball is $B_r(x) = \{y \in X : \ d(x,y) < r\}$.
- Since open sets are unions of open balls, if the families of balls from d_1 and d_2 generate the *same* collection of open sets, then we say that d_1, d_2 on X are *topologically equivalent*
- A function $f : X \to X$ is an *isometry* if $d_Y(f(x), f(y)) = d_X(x,y)$. Isometries preserve all metric properties.

Canonical examples of metric spaces include:

- On $\mathbb{R}^n$:
  - ℓ^2 (Euclidean) $d_2(x,y) = \left(\sum_i (x_i - y_i)^2\right)^{1/2}$
  - ℓ^1 $d_1(x,y) = \sum_i |x_i - y_i|$
  - ℓ^∞ $d_\infty(x,y) = \max_i |x_i - y_i|$

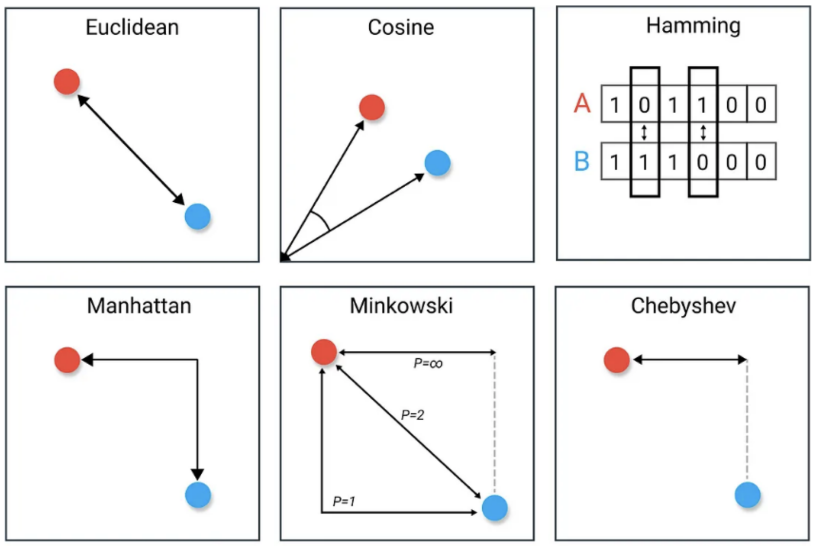

Figure 1: Common distances used in metric spaces

2.18 Null space and rank of $A^\top A$

Let $A \in \mathbb{R}^{n \times d}$. Then

$$\text{null}(A^\top A) = \text{null}(A)$$

from which it follows that

$$\text{rank}(A^\top A) = \text{rank}(A), \qquad \dim \text{null}(A^\top A) = d - \text{rank}(A).$$

PROOF:

$\subseteq$ If $A^\top A v = 0$, then $0 = v^\top A^\top A v = \|Av\|_2^2$, so $Av = 0$ and $v \in \text{null}(A)$.

$\supseteq$ If $v \in \text{null}(A)$, then $Av = 0$, hence $A^\top A v = 0$ and $v \in \text{null}(A^\top A)$.

2.19 Lower triangular

A matrix L is *lower triangular* if $L_{ij} = 0$ for $j > i$; it has *positive diagonal* if $L_{ii} > 0$.

2.20 Cholesky factorization

If $A \in \mathbb{R}^{n\times n}$ is PSD, then there exists a lower triangular L with positive diagonal such that

$$A = LL^{\mathsf{T}}.$$

If $A \succ 0$ then the factor L is *unique*. In either case, this gives us a easy expression for the determinant because $A = LL^{\mathsf{T}}$ is triangular times triangular,

$$\det(A) = \det(L)\det(L^{\mathsf{T}}) = \Big(\prod_{i=1}^{n} L_{ii}\Big)^2,$$

where we used that the determinant of a triangular matrix is the product of its diagonal entries, and $\det(L^{\mathsf{T}}) = \det(L)$.

2.21 Matrix exponential in terms of power series

Let $A \in \mathbb{R}^{n\times n}$ be a matrix, then its exponential is

$$\exp(A) \;:=\; e^A \;=\; \sum_{k=0}^{\infty} \frac{A^k}{k!}$$

Some useful facts are

- (Always exists). The series converges absolutely for every matrix. More precisely, for any submultiplicative norm
$$\sum_{k\geq 0} \|A\|^k/k! = e^{\|A\|} < \infty$$
- (Link with eigenvalues). If A is diagonalizable, $A = V\Lambda V^{-1}$, then
$$e^A = V\, e^{\Lambda}\, V^{-1}, \qquad e^{\Lambda} = \operatorname{diag}(e^{\lambda_1}, \ldots, e^{\lambda_n}),$$
i.e. it exponentiates eigenvalues with a corresponding change of basis.
- (Trace/determinant link). For all A, use the trace-determinant link to write
$$\det(e^A) = e^{\operatorname{tr}(A)} \neq 0$$
therefore for all A
$$e^A \text{ is always invertible, with } (e^A)^{-1} = e^{-A}.$$
- (Commuting case). If $AB = BA$, then
$$e^{A+B} = e^A e^B.$$
- (Linear ODE flow). The linear ODE $X'(t) = AX(t)$ with initial condition $X(0) = I$ has the unique solution
$$X(t) = e^{tA}, \qquad \text{so } e^{tA} \text{ is the time-}t \text{ evolution operator.}$$

2.22 Idempotent operators

An operator (or matrix) P is called idempotent if

$$P^2 = P.$$

Important facts and intuition:

- Idempotent operators represent projections: P projects any vector onto a subspace $\mathcal{R}(P)$.
- For any vector x, $Px = x$ if and only if $x \in \mathcal{R}(P)$, and $Px = 0$ if $x \in \text{Null}(P)$.
- The space decomposes as a direct sum $\mathbb{R}^n = \mathcal{R}(P) \oplus \text{Null}(P)$.
- The eigenvalues of an idempotent matrix are only 0 and 1.

2.23 Bilinear form

For vectors $u_t \in \mathbb{R}^m$ and $v_t \in \mathbb{R}^m$ for $t = 1, \ldots, T$, the bilinear form is written as

$$u^\mathsf{T} X v = \sum_{i,j} u_i X_{ij} v_j$$

2.24 Frobenius inner product as a sum of bilinear forms

Using the Frobenius inner product, defined as

$$\langle A, X \rangle_F := \text{tr}(A^\mathsf{T} X) = \sum_{i,j} A_{ij} X_{ij},$$

Let $u_t \in \mathbb{R}^m$ and $v_t \in \mathbb{R}^n$ for $t = 1, \ldots, T$ we have the identity

$$\Big\langle \sum_{t=1}^{T} u_t v_t^\mathsf{T},\, X \Big\rangle_F = \sum_{t=1}^{T} \langle u_t v_t^\mathsf{T}, X \rangle_F = \sum_{t=1}^{T} u_t^\mathsf{T} X v_t.$$

Hence we have an easy formula for the Frobenius norm of a sum of rank 1 matrices.

3 Functions

3.1 Derivative of the Natural Logarithm Function

For a differentiable function $f(u)$, the derivative of $\ln(f(u))$ with respect to u is

$$\frac{d}{du}[\ln(f(u))] = \frac{f'(u)}{f(u)}.$$

This identity follows from the chain rule and the basic derivative of the natural logarithm.

3.2 Entropy of a Discrete Random Variable

Let X be a discrete random variable taking values $\{x_i\}_{i=1}^n$ with probability mass function $p(x_i)$. The (base-b) entropy of X is

$$H(X) \ = \mathbb{E}[-\log_b p(x)] = \ -\sum_{i=1}^{n} p(x_i) \log_b p(x_i).$$

Entropy is maximized when the distribution is uniform, i.e. $p(x_i) = 1/n$ for all i.

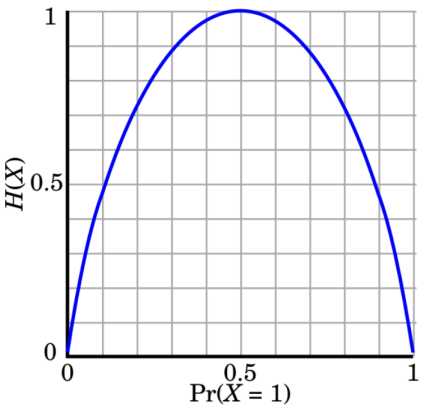

Figure 2: Entropy for a two sided coin as a function of probability of heads; entropy is maximized at most random.

If X is a continuous random variable with density function $f(x)$, then entropy extends as expected

$$H(X) \ = \mathbb{E}[-\log_b f(x)] = \ -\int_{\mathcal{X}} f(x) \log_b f(x) dx.$$

and again, the

3.3 Divergence of a Vector Field

Let $\mathbf{F}\colon \mathbb{R}^n \to \mathbb{R}^n$ be a differentiable vector field,

$$\mathbf{F}(x) = (F_1(x), F_2(x), \ldots, F_n(x)).$$

(notice that $\mathbf{F}(x)$ could be the gradient of a function, as above). The *divergence* of vector field $\mathbf{F}$ is the scalar field defined by

$$\nabla \cdot \mathbf{F}(x) \ = \sum_{i=1}^{n} \frac{\partial F_i(x)}{\partial x_i}.$$

Remarks:

- In $\mathbb{R}^3$, if $\mathbf{F}(x, y, z) = (F_1, F_2, F_3)$, then

$$\nabla \cdot \mathbf{F} = \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z}.$$

- Intuitively, the divergence measures the net rate of "flow out" of a point. A positive divergence means the point acts like a source, while a negative divergence means it acts like a sink.

3.4 Green's theorem

Green's theorem relates the line integral of a function to a surface integral of the divergence of that function. In particular, let $D \subset \mathbb{R}^2$ be a nice (e.g. bounded) region with positively oriented boundary $C = \partial D$ (counterclockwise), and let $\mathbf{F} = (P, Q)$ be a C^1 vector field on a neighborhood of D. Then Green's theorem states that

$$\oint_C P\,dx + Q\,dy \;=\; \iint_D \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) dA.$$

In 2D, this takes the form using the divergence,

$$\oint_C \mathbf{F} \cdot \mathbf{n}\,ds \;=\; \iint_D \nabla \cdot \mathbf{F}\,dA \qquad \text{where} \qquad \nabla \cdot \mathbf{F} = \frac{\partial P}{\partial x} + \frac{\partial Q}{\partial y}.$$

The intuition of this result is that the net outward flow across C is equal to the total divergence (local source minus sink density) inside D.

3.5 Sigmoid (logistic) function

The (general form) sigmoid function, for constants L, x_0, is given by

$$\sigma : \mathbb{R} \to (0, L)$$

$$\sigma(x) = \frac{L}{1 + e^{-k(x - x_0)}}$$

The standard logistic function is obtained by setting $L = 1,\ x_0 = 0,\ k = 1$ is given by

$$\sigma(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{1 + e^x} \qquad \in (0, 1).$$

We briefly remark that we often use this function to model probabilities, which we cover in the Analytics chapter. Useful facts about the sigmoid function include

- Complement / symmetry: $1 - \sigma(x) = \sigma(-x)$.
- Derivative: $\sigma'(x) = \sigma(x)(1 - \sigma(x))$.
- From the sigmoid function we get the logit function, which is simply the log odds

$$\text{logit}(p) \;=\; \log\left(\frac{p}{1-p}\right)$$

$$p \;=\; \text{logit}^{-1}(\ell) \;=\; \frac{1}{1 + e^{-\ell}} \;=\; \frac{e^{\ell}}{1 + e^{\ell}}$$

- The special case that is used in prediction is

$$\mathbb{P}(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \mathbf{x}^\top \boldsymbol{\beta})}}$$

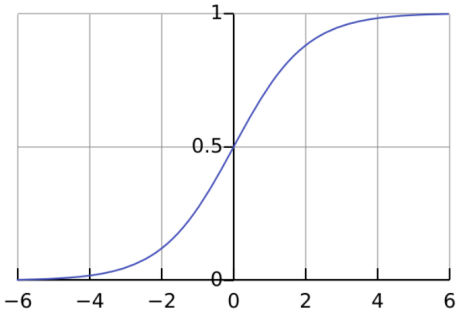

Figure 3: Standard logistic function where $L = 1, k = 1, x_0 = 0$

3.6 Softmax function

The softmax function generalizes the logistic sigmoid to K dimensions. It is used in multinomial logistic regression and commonly as the final activation of a neural network to convert scores (logits) into a probability distribution over K classes.

For $z \in \mathbb{R}^K$ and $i = 1, \ldots, K$, the softmax function

$$\text{softmax} : \mathbb{R}^K \to (0, 1)^K$$

where the i^{th} element of the output is

$$\text{softmax}(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}.$$

This maps z to the probability simplex (non-negative entries summing to 1). Basic properties of the include:

$$\text{softmax}(z + c\mathbf{1}) = \text{softmax}(z) \quad \text{(shift invariance)}$$

$$\text{softmax}_\tau(z)_i = \frac{e^{z_i/\tau}}{\sum_j e^{z_j/\tau}}, \ \tau > 0,$$

where smaller τ makes the distribution peakier and larger τ makes it flatter (uniform in the limit). Formally, as $\tau \downarrow 0$ we approach one-hot at $\arg\max_i z_i$; as $\tau \uparrow \infty$ we approaches uniform over all categories.

The softmax function reduces to a special case of the sigmoid when $K = 2$. More formally, let $z = (z_1, z_2)$, then

$$\text{softmax}(z)_1 = \frac{e^{z_1}}{e^{z_1} + e^{z_2}} = \frac{1}{1 + e^{-(z_1 - z_2)}} = \sigma(z_1 - z_2),$$

and $\text{softmax}(z)_2 = 1 - \text{softmax}(z)_1$. Thus, for two classes, softmax equals a sigmoid applied to the *logit difference* $z_1 - z_2$.

Lastly, we often use a more numerically stable form which takes advantage of the softmax

function's shirt invariance:

$$\text{softmax}(z)_i = \frac{e^{\,z_i - \max_j z_j}}{\sum_k e^{\,z_k - \max_j z_j}}.$$

which avoids exploding exponentials.

3.7 Stirling's approximation

For large $n \in \mathbb{N}$,

$$n! \;\approx\; \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

3.8 Laplacian of a Function

Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function. The *Laplacian* of f is the scalar-valued differential operator defined by

$$\Delta f \;=\; \nabla \cdot \nabla f \;=\; \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}.$$

Equivalently, $\Delta = \nabla^2$ is the divergence of the gradient of f.

3.9 Inverses of functions

$$f \text{ strictly increasing} \quad \Longrightarrow \quad f^{-1} \text{ strictly increasing.}$$

$$f \text{ strictly decreasing} \quad \Longrightarrow \quad f^{-1} \text{ strictly decreasing.}$$

$$f \text{ non-decreasing (increasing but not strictly)} \quad \Longrightarrow \quad f^{-1} \text{ is also non-decreasing.}$$

$$f \text{non-increasing (decreasing but not strictly)} \quad \Longrightarrow \quad f^{-1} \text{ is also non-increasing.}$$

- The graph of f^{-1} is the reflection of the graph of f across the line $y = x$.
- Thus, monotonicity is preserved under inversion.
- If f is strictly increasing/decreasing, then f^{-1} exists as a true function and is strictly increasing/decreasing.
- If f is merely non-decreasing/non-increasing, the "inverse" is not a single-valued function but a set-valued map:

  $$f^{-1}(y) = \{\, x : f(x) = y \,\},$$

  which is an interval. But monotonicity is still preserved in the sense that larger y correspond to sets of larger x values.

3.10 Taylor Series

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}\,(x-a)^n, \quad \text{valid near } x = a.$$

Common Maclaurin series ($a = 0$):

$$e^x = 1 + x + \tfrac{x^2}{2!} + \tfrac{x^3}{3!} + \cdots = \sum_{n=0}^{\infty} \frac{x^n}{n!},$$

$$\sin x = x - \tfrac{x^3}{3!} + \tfrac{x^5}{5!} - \tfrac{x^7}{7!} + \cdots = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!},$$

$$\cos x = 1 - \tfrac{x^2}{2!} + \tfrac{x^4}{4!} - \tfrac{x^6}{6!} + \cdots = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!},$$

$$\ln(1+x) = x - \tfrac{x^2}{2} + \tfrac{x^3}{3} - \tfrac{x^4}{4} + \cdots = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n}, \quad |x| < 1,$$

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots = \sum_{n=0}^{\infty} x^n, \quad |x| < 1.$$

3.11 Generating functions

The generating function of a sequence $\{a_n\}_{n \in \mathbb{N}_0}$ is a formal power series with coefficients that are the elements in the infinite sequence.

$$A(x) = \sum_{n \geq 0} a_n x^n$$

$$\sum_{n \geq 0} x^n = \frac{1}{1-x}, \qquad (1-x)^{-\alpha} = \sum_{n \geq 0} \binom{\alpha + n - 1}{n} x^n = \sum_{n \geq 0} \binom{\alpha}{n} (-x)^n.$$

$$A(x)B(x) = \sum_{n \geq 0} \Big(\sum_{k=0}^{n} a_k b_{n-k} \Big) x^n$$

$$xA(x) = \sum_{n \geq 1} a_{n-1} x^n, \qquad A'(x) = \sum_{n \geq 1} n a_n x^{n-1}$$

$$\frac{A(x) - A(0)}{x} = \sum_{n \geq 0} a_{n+1} x^n, \qquad \int_0^x A(t)\, dt = \sum_{n \geq 0} \frac{a_n}{n+1} x^{n+1}.$$

$$B(0) = 0 \;\Rightarrow\; A(B(x)) \text{ well-defined (OGF)}; \quad [x^n]A(B) = \sum_{k \geq 0} a_k \, [x^n] B(x)^k.$$

$$a_n = \sum_{j=1}^{d} c_j a_{n-j} \;\; (n \geq d) \iff A(x) = \frac{P(x)}{1 - \sum_{j=1}^{d} c_j x^j} \;\; (\deg P < d).$$

Generating functions for sequences are used find explicit formula for sequences (often generated by recurrences)

3.12 The bowl function

The bowl function in $\mathbb{R}^3$ is

$$f(x, y) = (x - a)^2 + (y - b)^2 + c$$

3.13 Operators of functions

An operator (on functions) is a map that takes a function as input and returns another function:

$$T : \mathcal{F} \to \mathcal{G}, \qquad f \mapsto Tf$$

Equivalently, for each x in the underlying domain, $(Tf)(x)$ is a rule built from f.

Common intuition:

- Many familiar operators in analysis and PDE are linear (derivatives, integrals, convolutions): $T(af + bg) = a\,Tf + b\,Tg$
- Composition: operators can be chained; $(S \circ T)f = S(Tf)$. This is the basic way to build complex transformations from simple ones
- Identity and inverse: the identity operator satisfies $(If)(x) = f(x)$; when T is bijective, an inverse T^{-1} satisfies $T^{-1}(Tf) = f$.

Two canonical examples are the derivative and the integral

$$(Df)(x) = f'(x)$$

$$(Kf)(x) = \int_a^b k(x, y)\, f(y)\, dy$$

3.14 Continuity

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ and let $x_0 \in \mathbb{R}^n$. The function f is continuous at x_0 iff

$$\lim_{x \to x_0} f(x) = f(x_0),$$

equivalently (the ε–δ identity),

$$\forall \varepsilon > 0\ \exists \delta > 0 \text{ s.t. } \|x - x_0\| < \delta \implies \|f(x) - f(x_0)\| < \varepsilon.$$

Furthermore, we say that f is continuous on $A \subseteq \mathbb{R}^n$ iff it is continuous at every $x_0 \in A$.

Key facts / intuition:

- Intuition: small changes in input produce small changes in output (no jumps).
- Compactness principle: if $K \subseteq \mathbb{R}^n$ is compact and f is continuous on K, then $f(K)$ is compact; in particular f attains a maximum and minimum on K and is uniformly continuous on K.

4 Tools

4.1 Triangle inequality for vectors

The triangle inequality states that for all $x, y \in \mathbb{R}^n$

$$\|x + y\| \;\leq\; \|x\| + \|y\| \quad \forall\, x, y.$$

4.2 Triangle Inequality for Integrals

The triangle inequality also has an integral version:

$$\left| \int_a^b f(x)\,\mathrm{d}x \right| \;\leq\; \int_a^b \left| f(x) \right| \mathrm{d}x$$

for *any* $f(x)$.

When Are They Equal? The equality holds if and only if $f(x)$ does not change sign on $[a, b]$:

- If $f(x) \geq 0$ for all $x \in [a, b]$, then

$$\left| \int_a^b f(x)\,\mathrm{d}x \right| = \int_a^b f(x)\,\mathrm{d}x = \int_a^b \left| f(x) \right| \mathrm{d}x.$$

- If $f(x) \leq 0$ for all $x \in [a, b]$, then

$$\left| \int_a^b f(x)\,\mathrm{d}x \right| = -\int_a^b f(x)\,\mathrm{d}x = \int_a^b \left| f(x) \right| \mathrm{d}x.$$

4.3 Cauchy-Schwarz for vectors

For all vectors $x, y \in \mathbb{R}^n$ $\langle x, y \rangle \;\leq\; \|x\|\,\|y\|$

4.4 Cauchy–Schwarz inequality for integrals

For any measurable functions $f(x)$ and $g(x)$ and any set $X \subseteq \mathbb{R}^d$

$$\left(\int_X f(x)\, g(x)\,\mathrm{d}x \right)^2 \;\leq\; \left(\int_X \|f(x)\|_2^2\,\mathrm{d}x \right) \left(\int_X \|g(x)\|_2^2\,\mathrm{d}x \right).$$

4.5 Zero-mean error and L^2 loss minimization

For a square-integrable random variable Y and observed features x, define the regression function

$$m(x) = \mathbb{E}[Y \mid x].$$

then

$$\mathbb{E}[Y - m(x) \mid x] = 0$$

which implies

$$\mathbb{E}[(Y - m(x))\, h(x)] = 0 \quad \text{for all } \; h(x) \text{ measurable w.r.t. } x.$$

This fact has some useful applications for the L^2 loss. For any predictor $g(x)$,

$$\mathbb{E}\Big[(Y - g(x))^2\Big] = \mathbb{E}\Big[(Y - m(x))^2\Big] + \mathbb{E}\Big[(m(x) - g(x))^2\Big],$$

so the mean-squared-error is minimized uniquely at $g(x) = m(x)$.

The interpretation here is that $m(x)$ is the L^2 projection of Y onto the space of functions of x; in regression, it is the Bayes optimal predictor under squared loss (e.g., linear regression estimates $m(x)$ within a linear class).

4.6 Absolute difference with opposite signs

If a and b have opposite signs (i.e. $a \cdot b < 0$), then

$$|b - a| = |a - b| \;=\; |a| + |b|.$$

4.7 Jensen's inequality for random variables

If X a random variable, then Jensen's inequality says that for any convex function φ

$$\varphi(\mathbb{E}[X]) \;\leq\; \mathbb{E}[\varphi(X)].$$

"The function of the expectation is less than the expectation of the function"

4.8 Jensen's inequality for integrals

Let (X, μ) be a measure space, $f\colon X \to \mathbb{R}$ integrable, and ϕ a convex function on an interval containing the range of f. Then

$$\phi\Big(\int_X f(x)\,\mathrm{d}\mu(x)\Big) \;\leq\; \int_X \phi(f(x))\,\mathrm{d}\mu(x).$$

Jensen's inequality asserts that when ϕ is convex, applying ϕ to the average (integral) of f does not exceed the average of $\phi \circ f$.

4.9 Derivative of matrix functions

Let $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$. Then

$$\frac{\partial}{\partial x}(Ax) = A,$$

and

$$\nabla_x \|Ax\| = \frac{A^T A x}{\|Ax\|},$$

and

$$\nabla_x \|Ax\|^2 = \nabla_x \big((Ax)^T (Ax)\big) = 2\,A^T A x.$$

4.10 Chain Rule for Nested Dependencies

Identity. Suppose $f(u, v, w)$, is a function where u, v, w depend on β, so $f\big(u(\beta), v(\beta), w(\beta)\big)$ Then

$$\frac{df}{d\beta} \;=\; \frac{\partial f}{\partial u}\,\frac{du}{d\beta} \;+\; \frac{\partial f}{\partial v}\,\frac{dv}{d\beta} \;+\; \frac{\partial f}{\partial w}\,\frac{dw}{d\beta}.$$

Explanation. The chain rule breaks the total rate of change of f with respect to β into

contributions through each intermediate variable and uses the basic chain rule on each of those intermediate variables.

4.11 Derivative of a Squared Inner-Product Error

Let $f(x) = \left(x^\top y - c\right)^2$ for fixed $y \in \mathbb{R}^n$ and scalar c. Then

$$\frac{\partial}{\partial x}\,(x^\top y - c)^2 \;=\; 2\,(x^\top y - c)\,y.$$

View $u = x^\top y - c$ as an intermediate scalar; by the chain rule, $\partial f/\partial x = 2u\,\partial u/\partial x = 2(x^\top y - c)\,y$.

4.12 Fundamental Theorem of Calculus

Theorem. Let f be continuous on $[a, b]$ and define

$$F(x) = \int_a^x f(t)\,\mathrm{d}t.$$

Then:

(1) F is differentiable on (a, b) with $F'(x) = f(x)$.

(2) If G is any antiderivative of f on $[a, b]$, then $\displaystyle\int_a^b f(t)\,\mathrm{d}t = G(b) - G(a)$.

4.13 Fundamental Theorem of Algebra

Theorem. Every nonzero, single-variable, degree-n polynomial with complex coefficients has exactly n complex roots when counted with multiplicity. Equivalently, the field of complex numbers $\mathbb{C}$ is algebraically closed.

4.14 The quadratic formula

For a quadratic equation

$$ax^2 + bx + c = 0, \quad a \neq 0,$$

the solutions x are given by

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

4.15 Change of variables in integrals

Case 1 (1D change of variables). Let $g : [a, b] \to \mathbb{R}$ be continuously differentiable and (for simplicity) monotone, and let f be integrable on $g([a, b])$. Then we can express an integral of the form

$$\int_a^b f(g(t))\,g'(t)\,dt \;=\; \int_{g(a)}^{g(b)} f(u)\,du.$$

Use this when you substitute a single variable $u = g(t)$ in a one-dimensional integral. The above implies that $g'(t)\,dt = du$. For example,

$$\int_0^1 \cos(t^2)\,2t\,dt.$$

Case 2 (2D linear change of variables). Let $A \in \mathbb{R}^{2\times 2}$ with $\det(A) \neq 0$ and define the

linear map

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \qquad T(u,v) = A\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix},$$

i.e.

$$x = au + bv, \qquad y = cu + dv.$$

If $\Omega \subset \mathbb{R}^2$ is a region in the (u,v)-plane and F is integrable on $T(\Omega)$, then

$$\iint_{T(\Omega)} F(x,y)\,dx\,dy \;=\; \iint_{\Omega} F\big(au+bv,\; cu+dv\big)\,|\det(A)|\,du\,dv.$$

Use this when (x,y) depends linearly on (u,v) (rotations, scalings, shears), so the area scaling is constant.

Example (Case 2: linear change of variables). Compute

$$\iint_R (x+y)\,dx\,dy, \qquad R = \{(x,y) : 0 \le x-y \le 1,\;\; 0 \le x+y \le 2\}.$$

The region R is awkward in (x,y) but becomes a rectangle after the linear substitution

$$u = x - y, \qquad v = x + y.$$

In matrix form,

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}, \qquad \text{so} \qquad \begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}.$$

Thus

$$x = \frac{u+v}{2}, \qquad y = \frac{v-u}{2}.$$

The Jacobian determinant for $(u,v) \mapsto (x,y)$ is

$$\left|\det\left(\frac{\partial(x,y)}{\partial(u,v)}\right)\right| = \left|\det\left(\frac{1}{2}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}\right)\right| = \frac{1}{2}.$$

Also, the constraints defining R become

$$0 \le u \le 1, \qquad 0 \le v \le 2,$$

so the (u,v)-region is the rectangle $\Omega = [0,1] \times [0,2]$. Finally, since $x + y = v$, the integral becomes

$$\iint_R (x+y)\,dx\,dy = \iint_{\Omega} v \cdot \frac{1}{2}\,du\,dv = \frac{1}{2}\int_0^1 \int_0^2 v\,dv\,du = \frac{1}{2}\cdot 1 \cdot \left[\frac{v^2}{2}\right]_0^2 = 1.$$

Case 3 (2D nonlinear change of variables: polar coordinates). Define the polar-coordinate map $T : (0,\infty) \times \mathbb{R} \to \mathbb{R}^2$ by

$$T(r,\theta) = \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r\cos\theta \\ r\sin\theta \end{pmatrix}, \qquad \text{i.e.} \qquad x = r\cos\theta, \quad y = r\sin\theta.$$

Its Jacobian matrix is

$$DT(r,\theta) = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}, \qquad |\det DT(r,\theta)| = r.$$

Hence the area element transforms as

$$dx\,dy = r\,dr\,d\theta,$$

and for any region $\Omega \subset (0,\infty)\times\mathbb{R}$ in the (r,θ)-plane and any integrable F on $T(\Omega)$,

$$\iint_{T(\Omega)} F(x,y)\,dx\,dy \;=\; \iint_{\Omega} F\big(r\cos\theta,\; r\sin\theta\big)\,r\,dr\,d\theta.$$

Use this when the region/integrand has circular symmetry (disks, annuli, sectors), or when x^2+y^2 appears naturally; choose Ω so that $T(\Omega)$ is exactly your region in the (x,y)-plane.

4.16 Change of variables in integrals

Case 1 (1D substitution).

$$\int_{g(a)}^{g(b)} f(u)\,du \;=\; \int_a^b f(g(t))\,g'(t)\,dt.$$

Here $u = g(t)$ is the substitution map, and $g'(t)$ is the *stretching factor* converting du to dt (i.e. $du = g'(t)\,dt$). *Use this when you substitute a single variable in a one-dimensional integral.*

Case 2 (2D linear change of variables).

$$\iint_{T(\Omega)} F(x,y)\,dx\,dy \;=\; \iint_{\Omega} F(T(u,v))\,|\det(A)|\,du\,dv.$$

Here $T(u,v) = A\binom{u}{v}$ is a linear coordinate map with

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \qquad T(u,v) = (au+bv,\; cu+dv),$$

and $|\det(A)|$ is the *constant area-scaling factor* (so $dx\,dy = |\det(A)|\,du\,dv$). *Use this when (x,y) depends linearly on (u,v) (rotations, scalings, shears), so the scaling factor is constant.*

Case 3 (2D nonlinear change of variables: polar coordinates).

$$\iint_{T(\Omega)} F(x,y)\,dx\,dy \;=\; \iint_{\Omega} F(T(r,\theta))\,|\det DT(r,\theta)|\,dr\,d\theta \;=\; \iint_{\Omega} F(r\cos\theta,\; r\sin\theta)\,r\,dr\,d\theta.$$

Here the polar map is $T(r,\theta) = (r\cos\theta,\; r\sin\theta)$, and

$$DT(r,\theta) = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}, \qquad |\det DT(r,\theta)| = r,$$

so the area element becomes $dx\,dy = r\,dr\,d\theta$. *Use this when the region/integrand is naturally described in terms of radius and angle (disks, annuli, sectors, or expressions involving x^2+y^2). Choose Ω in the (r,θ)-plane so that $T(\Omega)$ is exactly your region in the (x,y)-plane.*

4.17 Helmholtz decomposition of a vector field

Let $\Omega \subseteq \mathbb{R}^3$ be a sufficiently nice domain. For a sufficiently smooth vector field $u : \Omega \to \mathbb{R}^3$, one can decompose

$$u \;=\; \nabla\phi \;+\; \nabla \times A \;+\; h,$$

where

- $\phi : \Omega \to \mathbb{R}$ is a scalar potential
- $A : \Omega \to \mathbb{R}^3$ is a vector potential
- h is a harmonic field satisfying

$$\nabla \cdot h = 0, \qquad \nabla \times h = 0.$$

In many settings (e.g. $\Omega = \mathbb{R}^3$ with u decaying fast enough at infinity, or a simply connected bounded domain with boundary conditions that kill harmonic fields), one has $h \equiv 0$.

1. Intuition: $\nabla\phi$ is the irrotational (curl-free) part, while $\nabla \times A$ is the divergence-free part (sources/sinks), and h are vortices

4.18 Explicit potentials on $\mathbb{R}^3$ (decay at infinity)

If $\Omega = \mathbb{R}^3$ and u decays sufficiently fast at infinity, one can write

$$\phi(x) = \frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{\nabla \cdot u(y)}{\|x - y\|}\, dy, \qquad A(x) = \frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{\nabla \times u(y)}{\|x - y\|}\, dy,$$

so that

$$u(x) = \nabla\phi(x) + \nabla \times A(x).$$

5 Misc.

5.1 Sign of a permutation

The sign of a permutation is a value, either $+1$ (even) or -1 (odd), that indicates the parity of a permutation. It is defined as $(-1)^m$, where m is the number of transpositions (two-element swaps) needed to express the permutation in terms of the original ordering

5.2 Population growth

The simple model for population growth is

$$\frac{dN}{dt} = a(t)N(t)$$

where $N(t)$ is the population at time t, $N(0) = N_0$ (a known constant), and $a(t)$ is the relative rate of growth at time t. Clearly this is solved by

$$N(t) = e^{a(t)}N_0$$

5.3 Compact sets

For any $K \subset \mathbb{R}$, the Heine–Borel identity states that

$$K \text{ compact} \iff K \text{ is closed and bounded.}$$

(Extreme Value Theorem) If K is compact and f is continuous on K, then f attains its max and min on K.

5.4 Trees

A tree on n nodes has exactly $n - 1$ edges.

5.5 Galois theory on closed-form solutions

- A polynomial is *solvable* if its roots can be written using a finite number of $+$, $-$, $\times$, $\div$, and nth roots $\sqrt[n]{\ }$ as a function of its coefficients.
- Theorem: a polynomial of degree n is solvable by radicals *iff* its Galois group S_n is a *solvable group.*
- Consequences.
  - Degrees 1–4 (linear, quadratic, cubic, quartic) are solvable by radicals (closed form solutions)
  - The degree-n polynomial is *not* solvable in closed form is not solvable in closed form for $n \geq 5$

5.6 Countably infinite $\subset$ countable

A set is *countable* if either it is finite or it can be made in one to one correspondence with the set of natural numbers. Equivalently, a set is countable if there exists an injective function from it into the natural numbers; this means that each element in the set may be associated to a unique natural number, or that the elements of the set can be counted one at a time, although the counting may never finish due to an infinite number of elements.

A countable set that is not finite is said to be *countably infinite.*

5.7 Binomial Coefficients

- $\binom{n}{k}$ counts the number of possible k-subsets of an n-element set (ways to choose k distinct items without order).
- It is the coefficient of $x^{\,k}$ in $(1+x)^n$ (weight of k "successes").
- In probability, $\mathbb{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ for $X\sim \mathrm{Bin}(n,p)$.

$$\binom{n}{k}=\frac{n!}{k!(n-k)!}$$

$$(x+y)^n=\sum_{k=0}^{n}\binom{n}{k}x^{\,n-k}y^{\,k},$$

$$\binom{n}{k}=\binom{n}{n-k}\quad\text{(symmetry)},$$

$$\binom{n}{k}=\binom{n-1}{k}+\binom{n-1}{k-1}\quad\text{(Pascal)},$$

$$\binom{n}{k}=\frac{n}{k}\binom{n-1}{k-1}.$$

$$\sum_{k=0}^{n}\binom{n}{k}=2^n,\qquad \sum_{k=0}^{n}(-1)^k\binom{n}{k}=0\ (n\geq 1),$$

$$\sum_{k=0}^{r}\binom{m}{k}\binom{n}{r-k}=\binom{m+n}{r}\quad\text{(Vandermonde)}.$$

5.8 Multinomial coefficients

- multinomial coefficient $\mathrm{Mult}(n,k_1,\ldots,k_m)$ counts the number of ways to distribute n labeled (distinct) objects into m bins where bin i must have k_i objects.

$$\binom{n}{k_1,\ldots,k_m}=\frac{n!}{k_1!\cdots k_m!}\qquad\text{for } k_i\in\mathbb{Z}_{\geq 0},\ \sum_{i=1}^{m}k_i=n.$$

$$(x_1+\cdots+x_m)^n=\sum_{\substack{\text{all } k_i \text{ s.t}\\ k_1+\cdots+k_m=n\\ k_i\geq 0}}\binom{n}{k_1,\ldots,k_m}x_1^{k_1}\cdots x_m^{k_m}$$

$$\binom{n}{k_1,\ldots,k_m}=\binom{n}{k_{\sigma(1)},\ldots,k_{\sigma(m)}}\quad\text{(symmetry)},$$

$$\binom{n}{k_1,\ldots,k_m}=\binom{n}{k_1}\binom{n-k_1}{k_2}\cdots\binom{n-k_1-\cdots-k_{m-1}}{k_m}$$

5.9 Solving an optimization problem via a sampling distribution

Consider a generic optimization problem

$$\min_{x\in\mathcal{X}} f(x),$$

(e.g. an LP with $\mathcal{X} = \{x : Ax \le b\}$ and $f(x) = c^\top x$). A standard way to "solve by sampling" is to form a Gibbs/Boltzmann distribution

$$\pi_\beta(x) \;:=\; \frac{1}{Z(\beta)}\,\exp(-\beta f(x))\,\mathbf{1}\{x\in\mathcal{X}\} \qquad\qquad \text{for } \; x\in\mathcal{X}$$

where

$$Z(\beta) \;:=\; \int_{\mathcal{X}} \exp(-\beta f(x))\,dx.$$

Then the optimal objective value and the optimal solution are recovered as a zero-temperature limit:

$$\min_{x\in\mathcal{X}} f(x) \;=\; \lim_{\beta\to\infty}\left(-\frac{1}{\beta}\log Z(\beta)\right),$$

$$\arg\min_{x\in\mathcal{X}} f(x) \;\subseteq\; \lim_{\beta\to\infty} \operatorname{supp}(\pi_\beta).$$

Important facts / intuition:

- Large β concentrates π_β on (near-)minimizers: for any $\varepsilon > 0$ and $f^\star := \min_{x\in\mathcal{X}} f(x)$

  $$\pi_\beta\big(\{x : f(x) \le f^\star + \varepsilon\}\big) \xrightarrow[\beta\to\infty]{} 1$$

  under mild regularity (e.g. $\mathcal{X}$ has positive volume near the minimizers).
- In an LP, minimizers occur at extreme points; as $\beta\to\infty$, π_β concentrates on optimal extreme points (and spreads over multiple optima if the optimum face is not a single point).
- Practically, sampling typically uses MCMC (e.g. hit-and-run / Langevin-type methods) and an annealing schedule $\beta_1 < \beta_2 < \cdots \to \infty$; the core idea is that sampling from π_β becomes easier at small β and becomes more "solution-like" at large β.

6 Reinforcement learning in games

6.1 Discounted Markov decision process (MDP)

Let a discounted Markov decision process (MDP) be a tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ where

- $\mathcal{S}$ is a (finite) state space
- $\mathcal{A}$ is a (finite) action space
- $P(\cdot \mid s, a)$ is a transition kernel
- $r(s, a)$ is a (bounded) immediate reward
- $\gamma \in (0, 1)$ is a discount factor

A (stationary) policy π maps each state s to a distribution $\pi(\cdot \mid s)$ over actions. The value function $V^\pi : \mathcal{S} \to \mathbb{R}$ of such a policy π is

$$V^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \, r(S_t, A_t) \,|\, S_0 = s \right]$$

and the optimal value function is then $V^*(s) = \sup_\pi V^\pi(s)$.

6.2 Bellman operators of a MDP

For a given Markov decision process $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ define the Bellman expectation operator T^π acting on any function $V : \mathcal{S} \to \mathbb{R}$ is

$$(T^\pi V)(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} \Big[r(s, a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)}[V(s')] \Big].$$

and define the Bellman optimality operator T by

$$(TV)(s) = \max_{a \in \mathcal{A}} \Big[r(s, a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)}[V(s')] \Big].$$

6.3 Bellman operators satisfy the Bellman equations, thus have fixed points

For a given Markov decision process $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with the Bellman expectation operator T^π and the Bellman optimality operator T, these two operators obey the Bellman equations, also called the fixed-point identities which state

$$V^\pi = T^\pi V^\pi \qquad \text{value of any policy } \pi$$

$$V^* = TV^* \qquad \text{value of the optimal policy } \pi^*.$$

They also have the contraction property (in the sup norm $\|V\|_\infty = \max_{s \in \mathcal{S}} |V(s)|$):

$$\|T^\pi V - T^\pi W\|_\infty \le \gamma \|V - W\|_\infty, \qquad \|TV - TW\|_\infty \le \gamma \|V - W\|_\infty.$$

Hence, by the Banach fixed-point theorem, each T^π has a unique fixed point V^π, and T has a unique fixed point V^*.

6.4 Find fixed point of Bellman operators by iterating

Given a Markov decision process $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with the Bellman expectation operator T^π and the Bellman optimality operator T. Given any fixed policy π, iterating

$$V_{k+1} = T^\pi V_k,$$

then $V_k \to V^\pi$. This does not change π; it only computes the value of π. Similarly, if you iterate

$$V_{k+1} = TV_k,$$

then $V_k \to V^*$. This produces the optimal value function, not the policy. If we solve this by interarting so that the optimal value function V^* were known, then an optimal policy $\pi^*(s)$ can be obtained immediately by acting greedily with respect to V^*:

$$\pi^*(s) \in \arg\max_{a \in \mathcal{A}} \Big[r(s,a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')] \Big].$$

The difficulty is that V^* is unknown and cannot be computed directly without solving the Bellman optimality fixed-point problem (which is hard).

6.5 Policy iteratation and value iteration to find fixed points

Policy iteration explains how to reach V^* and π^* jointly which is more stable. Concretely, policy iteration alternates:

1. Policy evaluation: compute the unique fixed point

   $$V^{\pi_k} = T^{\pi_k} V^{\pi_k},$$

   which measures the long-run performance of the current policy.

2. Policy improvement: update the policy by acting greedily with respect to that value,

   $$\pi_{k+1}(s) \in \arg\max_{a \in \mathcal{A}} \Big[r(s,a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^{\pi_k}(s')] \Big].$$

This guarantees that each new policy is at least as good as the previous one, and the process converges to an optimal policy in finitely many steps for finite MDPs.

In contrast, value iteration skips explicit policy evaluation and directly iterates

$$V_{k+1} = TV_k,$$

which converges to V^*. The optimal policy is then recovered by a final greedy step. In practice, the most commonly used algorithms are value iteration and its large-scale variants (e.g. approximate value iteration, Q-learning, and deep Q-learning).

6.6 Q-learning (value-based RL)

6.7 PPO (proximal policy optimization; actor–critic)

6.8 CFR (counterfactual regret minimization; extensive-form games)