Linear Algebra for Optimization

V. Leclère (ENPC)

February 13th, 2026

What is this course about?

- **Goal:** build a complete toolkit to model, analyze, and solve optimization problems.
- **Foundations (why it works):**
  - numerical linear algebra for optimization;
  - convex geometry and convex analysis;
  - first-order optimality and constraint qualifications $\Rightarrow$ KKT and certificates.
- **Duality as a unifying lens:** Lagrangians, strong duality (Slater), multipliers as sensitivity prices, KKT $\Leftrightarrow$ saddle points (convex).
- **Algorithms (how to compute):**
  - gradient methods, steepest descent, conjugate gradient; rates and conditioning;
  - Newton and quasi-Newton (BFGS/L-BFGS): fast local accuracy with globalization;
  - constrained methods: a brief overview;
  - a quick look at stochastic gradient methods.
- **Throughout:** emphasize certificates (gaps, gradient mappings, dual bounds), assumptions, and when guarantees do or do not apply.

Course structure (12 sessions)

1. Numerical linear algebra
2. Convex sets
3. Convex functions
4. Optimality conditions
5. Convex duality
6. Optimization zoo
7. Gradient descent
8. Newton's method
9. Constrained optimization
10. Interior-point methods
11. Stochastic gradient
12. Exam

Why start with linear algebra?

- Your algorithms will be written in terms of norms, inner products, and linear systems.
- Convergence of Gradient Descent, Conjugate Gradient, Newton, Interior-Point Methods is essentially governed by spectra: $\lambda_{\min}, \lambda_{\max}$ and conditioning.
- In this course we will use mostly symmetric matrices (PSD order, spectral theorem), but we need Singular Value Decomposition (SVD) for operator norms and conditioning.
- Goal for today: fix a common vocabulary and a few key facts we will reuse all semester.

Convention in these slides

Icons / tags:

- ♡ : really important results
- ♢ : more advanced content
- ♣ : very simple in-class exercise
- ♠ : training exercise (harder)
- BV x.y : *Convex Optimization*, S. Boyd, L. Vanderberghe, Ch. x, Sec. y
- JCG x.y : *Fragments d'Optimisation Différentiable* (J.-Ch. Gilbert), Ch. x, Sec. y

➥ both are available online for free.

➥ Boyd's classes on convex optimization are on YouTube.

Color coding[a]:

- In text:
  - violet for definitions,
  - red for key results/theorems/assumptions.
- In math:
  - x primal variables,
  - λ multipliers,
  - θ parameters,
  - L smoothness constants,
  - $x^{(k)}$ iterates,
  - $\tau^{(k)}$ step sizes,
  - d directions.

[a] Best effort

Contents

Contents

Eigenvalues, eigenvectors, eigenspaces, spectrum (recall) ♡

Let $A \in \mathbb{R}^{n \times n}$ (we will always consider real matrices).

- A nonzero vector $x \neq 0$ is an eigenvector of A associated with eigenvalue $\lambda \in \mathbb{R}$ if

$$Ax = \lambda x.$$

- The eigenspace associated with λ is

$$E_\lambda(A) := \ker(A - \lambda I) = \{x \in \mathbb{R}^n : Ax = \lambda x\}.$$

- The spectrum of A is the set of its eigenvalues:

$$\mathrm{sp}(A) := \{\lambda \in \mathbb{R} : \det(A - \lambda I) = 0\}.$$

♣ Exercise: Show that A is invertible iff $0 \notin \mathrm{sp}(A)$.

Eigenvalues, eigenvectors, eigenspaces, spectrum (recall) ♡

Let $A \in \mathbb{R}^{n \times n}$ (we will always consider real matrices).

- A nonzero vector $x \neq 0$ is an eigenvector of A associated with eigenvalue $\lambda \in \mathbb{R}$ if

$$Ax = \lambda x.$$

- The eigenspace associated with λ is

$$E_\lambda(A) := \ker(A - \lambda I) = \{x \in \mathbb{R}^n : Ax = \lambda x\}.$$

- The spectrum of A is the set of its eigenvalues:

$$\mathrm{sp}(A) := \{\lambda \in \mathbb{R} : \det(A - \lambda I) = 0\}.$$

♣ Exercise: Show that A is invertible iff $0 \notin \mathrm{sp}(A)$.

Diagonalization (general case): definition and criterion

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

- A is diagonalizable (over $\mathbb{R}$) if there exist an invertible matrix P and a diagonal matrix D such that
$$A = PDP^{-1}.$$
- In that case, the diagonal entries of D are eigenvalues of A, and the columns of P form a basis of eigenvectors.

Diagonalization criterion (via eigenspaces)

A is diagonalizable over $\mathbb{R}$ iff:

1. all eigenvalues are real (so $\mathrm{sp}(A) \subset \mathbb{R}$), and
2. the eigenvectors span $\mathbb{R}^n$, equivalently
$$n = \sum_{\lambda \in \mathrm{sp}(A)} \dim\left(E_\lambda(A)\right).$$

♣ Exercise: Assume A has n *distinct* real eigenvalues. Show that A is diagonalizable.

Diagonalization (general case): definition and criterion

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

- A is diagonalizable (over $\mathbb{R}$) if there exist an invertible matrix P and a diagonal matrix D such that
$$A = PDP^{-1}.$$
- In that case, the diagonal entries of D are eigenvalues of A, and the columns of P form a basis of eigenvectors.

Diagonalization criterion (via eigenspaces)

A is diagonalizable over $\mathbb{R}$ iff:

1. all eigenvalues are real (so $\mathrm{sp}(A) \subset \mathbb{R}$), and
2. the eigenvectors span $\mathbb{R}^n$, equivalently
$$n = \sum_{\lambda \in \mathrm{sp}(A)} \dim\big(E_\lambda(A)\big).$$

♣ Exercise: Assume A has n *distinct* real eigenvalues. Show that A is diagonalizable.

(Semi)definite matrices and PSD order

Let S_n be the set of real symmetric $n \times n$ matrices.

- Positive semidefinite (PSD): $A \succeq 0$ (equivalently $A \in S_n^+$) iff

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

- Positive definite (PD): $A \succ 0$ (equivalently $A \in S_n^{++}$) iff

$$x^\top A x > 0 \quad \forall x \neq 0.$$

Loewner (PSD) order on S_n

For $A, B \in S_n$, define

$$A \preceq B \iff B - A \succeq 0.$$

Equivalently, $A \preceq B$ iff $x^\top A x \leq x^\top B x$ for all $x \in \mathbb{R}^n$.

♣ Exercise: Show that $\preceq$ is a partial order on S_n (reflexive, antisymmetric, transitive).

(Semi)definite matrices and PSD order

Let S_n be the set of real symmetric $n \times n$ matrices.

- Positive semidefinite (PSD): $A \succeq 0$ (equivalently $A \in S_n^+$) iff

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

- Positive definite (PD): $A \succ 0$ (equivalently $A \in S_n^{++}$) iff

$$x^\top A x > 0 \quad \forall x \neq 0.$$

Loewner (PSD) order on S_n

For $A, B \in S_n$, define

$$A \preceq B \iff B - A \succeq 0.$$

Equivalently, $A \preceq B$ iff $x^\top A x \leq x^\top B x$ for all $x \in \mathbb{R}^n$.

♣ Exercise: Show that $\preceq$ is a partial order on S_n (reflexive, antisymmetric, transitive).

Contents

Spectral theorem (real symmetric matrices)

Let S_n be the set of real symmetric $n \times n$ matrices.

Theorem (Spectral theorem)

Any symmetric matrix $A \in S_n$ has n real eigenvalues (counted with multiplicity) $\lambda_1, \dots, \lambda_n$ and an orthonormal basis of eigenvectors $(q_i)_{i \in [n]}$ i.e.,such that:

$$Aq_i = \lambda_i q_i, \quad \text{and} \quad q_i^\top q_j = \delta_{ij}, \qquad \forall i, j.$$

In other words, there exists an orthogonal matrix Q (i.e. $Q^\top Q = I$) such that

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

where $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ and $\lambda_1, \dots, \lambda_n$ are the eigenvalues of A.

Consequences for (semi)definite matrices

For $A \in S_n$ with spectral decomposition $A = Q\Lambda Q^\top$, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$:

- $A \in S_n^+ \iff A \succeq 0 \iff \lambda_i \geq 0$ for all i.
- $A \in S_n^{++} \iff A \succ 0 \iff \lambda_i > 0$ for all i.
- For any $A \in S_n^+$, with $A = Q\Lambda Q^\top$, we have

$$A^{1/2} = Q\Lambda^{1/2}Q^\top, \quad \text{where} \quad \Lambda^{1/2} = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}),$$

  where $A^{1/2}$ is the unique PSD square root of A (i.e. $A^{1/2}A^{1/2} = A$).

♠ Exercise: Prove the above statements using the spectral decomposition.

Rayleigh quotient and eigenvalue extrema

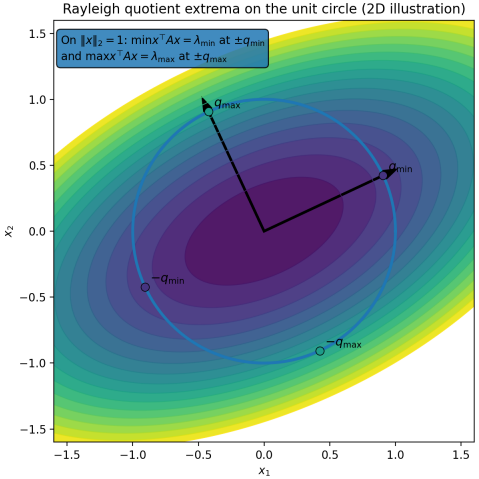
For $A \in S_n$ and $x \neq 0$, define $R_A(x) = \frac{x^\top A x}{x^\top x}$.

- $\lambda_{\min}(A) = \min_{\|x\|_2=1} x^\top A x$,
  $\lambda_{\max}(A) = \max_{\|x\|_2=1} x^\top A x$.

➥ min and max are attained at eigenvectors associated with $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$.

- If $A \succ 0$, then $\|x\|_A^2 = x^\top A x$ satisfies

$$\lambda_{\min}(A)\,\|x\|_2^2 \leq \|x\|_A^2 \leq \lambda_{\max}(A)\,\|x\|_2^2.$$



Contents

Orthogonal projectors

A matrix P is an orthogonal projector iff $P^2 = P$ and $P = P^\top$.

- Then $\mathbb{R}^n = \mathrm{Im}(P) \oplus \ker(P)$, and Px is the closest (for $\|\cdot\|_2$) point to x in $\mathrm{Im}(P)$.
- Eigenvalues of P are 0 or 1, so $P \succeq 0$ and $\|P\|_2 = 1$ (unless $P = 0$).
- If $\|q\|_2 = 1$, then $P = qq^\top$ is the orthogonal projector onto $\mathrm{span}(q)$.
- More generally, if $Q \in \mathbb{R}^{n \times k}$ has orthonormal columns, then $P = QQ^\top$ projects onto $\mathrm{Im}(Q)$.

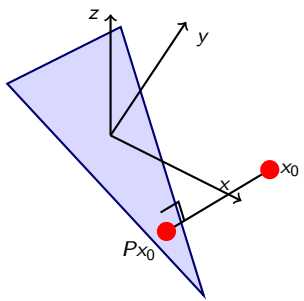
Px_0 is the Euclidean closest point to x_0 in $\mathrm{Im}(P)$.

Orthogonal projectors

A matrix P is an orthogonal projector iff $P^2 = P$ and $P = P^\top$.

- Then $\mathbb{R}^n = \mathrm{Im}(P) \oplus \ker(P)$, and Px is the closest (for $\|\cdot\|_2$) point to x in $\mathrm{Im}(P)$.
- Eigenvalues of P are 0 or 1, so $P \succeq 0$ and $\|P\|_2 = 1$ (unless $P = 0$).

- If $\|q\|_2 = 1$, then $P = qq^\top$ is the orthogonal projector onto $\mathrm{span}(q)$.
- More generally, if $Q \in \mathbb{R}^{n \times k}$ has orthonormal columns, then $P = QQ^\top$ projects onto $\mathrm{Im}(Q)$.



Px_0 is the Euclidean closest point to x_0 in $\mathrm{Im}(P)$.

Spectral decomposition as a sum of rank-one matrices

Let $A \in S_n$ with spectral decomposition $A = Q\Lambda Q^\top$, where $Q = [q_1 \cdots q_n]$ is orthogonal and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$.

Rank-one expansion

$$A = \sum_{i=1}^{n} \lambda_i \, q_i q_i^\top .$$

Each term $q_i q_i^\top$ is the orthogonal projector onto $\mathrm{span}(q_i)$ (rank one).

Explanation: Since $Q\Lambda Q^\top = \sum_{i=1}^{n} \lambda_i \, Q e_i e_i^\top Q^\top$ and $Qe_i = q_i$,

$$Q\Lambda Q^\top = \sum_{i=1}^{n} \lambda_i \, q_i q_i^\top .$$

Spectral decomposition as a sum of rank-one matrices

Let $A \in S_n$ with spectral decomposition $A = Q\Lambda Q^\top$, where $Q = [q_1 \cdots q_n]$ is orthogonal and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$.

Rank-one expansion

$$A = \sum_{i=1}^{n} \lambda_i \, q_i q_i^\top.$$

Each term $q_i q_i^\top$ is the orthogonal projector onto $\mathrm{span}(q_i)$ (rank one).

Explanation: Since $Q\Lambda Q^\top = \sum_{i=1}^n \lambda_i \, Q e_i e_i^\top Q^\top$ and $Q e_i = q_i$,

$$Q\Lambda Q^\top = \sum_{i=1}^{n} \lambda_i \, q_i q_i^\top.$$

PSD order and basic algebra

For $A, B \in S_n$, recall that $A \preceq B$ iff $B - A \succeq 0$.

- Order is compatible with quadratic forms: $A \preceq B \Rightarrow x^\top Ax \leq x^\top Bx$ for all x.
- If $0 \prec A \preceq B$, then $B^{-1} \preceq A^{-1}$ (order reverses under inversion).

- $A \preceq B$ implies $\lambda_{\min}(A) \leq \lambda_{\min}(B)$ and $\lambda_{\max}(A) \leq \lambda_{\max}(B)$.
- $A \preceq MI$ iff $\lambda_{\max}(A) \leq M$, and $A \succeq mI$ iff $\lambda_{\min}(A) \geq m$.

- We will constantly use $mI \preceq \nabla^2 f(x) \preceq MI$ to control GD and Newton.

PSD order and basic algebra

For $A, B \in S_n$, recall that $A \preceq B$ iff $B - A \succeq 0$.

- Order is compatible with quadratic forms: $A \preceq B \Rightarrow x^\top Ax \leq x^\top Bx$ for all x.
- If $0 \prec A \preceq B$, then $B^{-1} \preceq A^{-1}$ (order reverses under inversion).

- $A \preceq B$ implies $\lambda_{\min}(A) \leq \lambda_{\min}(B)$ and $\lambda_{\max}(A) \leq \lambda_{\max}(B)$.
- ➥ $A \preceq MI$ iff $\lambda_{\max}(A) \leq M$, and $A \succeq mI$ iff $\lambda_{\min}(A) \geq m$.

- ➥ We will constantly use $mI \preceq \nabla^2 f(x) \preceq MI$ to control GD and Newton.

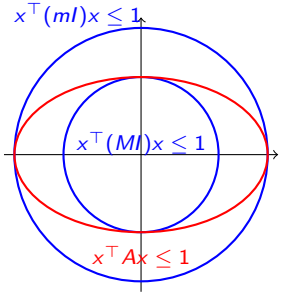
PSD order and basic algebra

For $A, B \in S_n$, recall that $A \preceq B$ iff $B - A \succeq 0$.

- Order is compatible with quadratic forms: $A \preceq B \Rightarrow x^\top Ax \leq x^\top Bx$ for all x.
- If $0 \prec A \preceq B$, then $B^{-1} \preceq A^{-1}$ (order reverses under inversion).

- $A \preceq B$ implies $\lambda_{\min}(A) \leq \lambda_{\min}(B)$ and $\lambda_{\max}(A) \leq \lambda_{\max}(B)$.
- ➥ $A \preceq MI$ iff $\lambda_{\max}(A) \leq M$, and $A \succeq mI$ iff $\lambda_{\min}(A) \geq m$.

- ➥ We will constantly use $mI \preceq \nabla^2 f(x) \preceq MI$ to control GD and Newton.



Contents

Contents

Norms: a quick recall

A norm on $\mathbb{R}^n$ is a map $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}_+$ such that:

- Positive definiteness: $\|x\| \geq 0$ and $\|x\| = 0 \iff x = 0$.
- Absolute homogeneity: $\|\alpha x\| = |\alpha| \, \|x\|$ for all $\alpha \in \mathbb{R}$.
- Triangle inequality: $\|x + y\| \leq \|x\| + \|y\|$.
- Classical examples:

$$\|x\|_2 = \sqrt{\sum_i x_i^2}, \ \|x\|_1 = \sum_i |x_i|, \ \|x\|_\infty = \max_i |x_i|.$$

- Useful inequalities (for $x \in \mathbb{R}^n$):

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

Norms: a quick recall

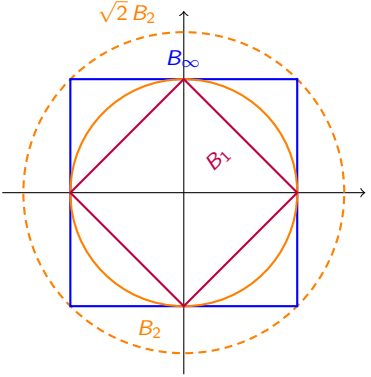
A norm on $\mathbb{R}^n$ is a map $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}_+$ such that:

- Positive definiteness: $\|x\| \geq 0$ and $\|x\| = 0 \iff x = 0$.
- Absolute homogeneity: $\|\alpha x\| = |\alpha| \, \|x\|$ for all $\alpha \in \mathbb{R}$.
- Triangle inequality: $\|x + y\| \leq \|x\| + \|y\|$.
- Classical examples:

$$\|x\|_2 = \sqrt{\sum_i x_i^2}, \ \|x\|_1 = \sum_i |x_i|, \ \|x\|_\infty = \max_i |x_i|.$$

- Useful inequalities (for $x \in \mathbb{R}^n$):

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



Inner products and geometry

- An inner product on $\mathbb{R}^n$ is a map $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ such that:
  - Bilinear: $\langle ax_1 + bx_2,\, y \rangle = a\langle x_1, y \rangle + b\langle x_2, y \rangle$ and $\langle x,\, ay_1 + by_2 \rangle = a\langle x, y_1 \rangle + b\langle x, y_2 \rangle$.
  - Symmetric: $\langle x, y \rangle = \langle y, x \rangle$.
  - Positive definite: $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0 \iff x = 0$.
- It defines a norm: $\|x\| := \sqrt{\langle x, x \rangle}$.
- Cauchy–Schwarz: $|\langle x, y \rangle| \leq \|x\|\,\|y\|$.
  - Equality case: if $x \neq 0$ and $y \neq 0$, equality holds iff $y = \alpha x$ for some $\alpha \in \mathbb{R}$ (colinearity).
  - (Trivial cases: if $x = 0$ or $y = 0$, equality holds with both sides $= 0$.)
- Orthogonality depends on the inner product: $x \perp y \iff \langle x, y \rangle = 0$.

Inner products and geometry

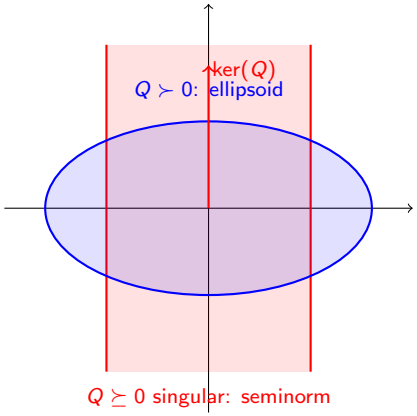
- An inner product on $\mathbb{R}^n$ is a map $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ such that:
  - Bilinear: $\langle ax_1 + bx_2,\, y \rangle = a\langle x_1, y \rangle + b\langle x_2, y \rangle$ and $\langle x,\, ay_1 + by_2 \rangle = a\langle x, y_1 \rangle + b\langle x, y_2 \rangle$.
  - Symmetric: $\langle x, y \rangle = \langle y, x \rangle$.
  - Positive definite: $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0 \iff x = 0$.
- It defines a norm: $\|x\| := \sqrt{\langle x, x \rangle}$.
- Cauchy–Schwarz: $|\langle x, y \rangle| \leq \|x\|\,\|y\|$.
  - Equality case: if $x \neq 0$ and $y \neq 0$, equality holds iff $y = \alpha x$ for some $\alpha \in \mathbb{R}$ (colinearity).
  - (Trivial cases: if $x = 0$ or $y = 0$, equality holds with both sides $= 0$.)
- Orthogonality depends on the inner product: $x \perp y \iff \langle x, y \rangle = 0$.

The $\|\cdot\|_Q$ norm (and when it is not a norm)

Let $Q \in S_n^+$.

- Define $\|x\|_Q := \sqrt{x^\top Q x}$.
- If $Q \succ 0$: $\|\cdot\|_Q$ is a norm and its unit ball is an ellipsoid.
- If $Q \succeq 0$ but singular: $\|\cdot\|_Q$ is a seminorm, with nullspace $\ker(Q) = \{x : Qx = 0\}$.
- Change of variables (SPD case): $\|x\|_Q = \|Q^{1/2}x\|_2$.

♠ Exercise: Show $\|x\|_Q \le \sqrt{\lambda_{\max}(Q)}\,\|x\|_2$ and $\|x\|_2 \le \frac{1}{\sqrt{\lambda_{\min}(Q)}}\,\|x\|_Q$ when $Q \succ 0$.



Unit "ball": ellipse (SPD) vs unbounded strip (singular PSD).

Contents

Gradient (recap)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable at x.

Gradient = the unique vector defining the first-order approximation

There exists a unique vector $g \in \mathbb{R}^n$ such that

$$f(x+h) = f(x) + g^\top h + o(\|h\|_2) \qquad \text{as } h \to 0.$$

We denote it $g = \nabla f(x)$.

Coordinate formula

In the canonical basis,

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix}.$$

Gradient (recap)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable at x.

Gradient = the unique vector defining the first-order approximation

There exists a unique vector $g \in \mathbb{R}^n$ such that

$$f(x+h) = f(x) + g^\top h + o(\|h\|_2) \qquad \text{as } h \to 0.$$

We denote it $g = \nabla f(x)$.

Coordinate formula

In the canonical basis,

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix}.$$

Hessian and second-order Taylor expansion (recap)

If f is twice differentiable at x:

Hessian = the unique matrix defining the second-order approximation

There exists a unique symmetric matrix $H \in S_n$ such that

$$f(x+h) = f(x) + \nabla f(x)^\top h + \frac{1}{2} h^\top H h + o(\|h\|_2^2) \qquad \text{as } h \to 0.$$

We denote it $H = \nabla^2 f(x)$.

Coordinate formula

$$\nabla^2 f(x) = \left[\frac{\partial^2 f}{\partial x_i \, \partial x_j}(x) \right]_{i,j=1}^{n}.$$

Hessian and second-order Taylor expansion (recap)

If f is twice differentiable at x:

Hessian = the unique matrix defining the second-order approximation

There exists a unique symmetric matrix $H \in S_n$ such that

$$f(x+h) = f(x) + \nabla f(x)^\top h + \frac{1}{2} h^\top H h + o(\|h\|_2^2) \qquad \text{as } h \to 0.$$

We denote it $H = \nabla^2 f(x)$.

Coordinate formula

$$\nabla^2 f(x) = \left[\frac{\partial^2 f}{\partial x_i \, \partial x_j}(x) \right]_{i,j=1}^{n}.$$

Hessian bounds and strong convexity

Assume $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable.

PSD order $\iff$ eigenvalue bounds

For any symmetric matrix $H \in S_n$ and any $m \in \mathbb{R}$,

$$H \succeq mI \quad \iff \quad \lambda_{\min}(H) \geq m \quad \iff \quad x^\top Hx \geq m\|x\|_2^2 \ \ \forall x \in \mathbb{R}^n.$$

Strong convexity via the Hessian

If there exists $m > 0$ such that

$$\nabla^2 f(x) \succeq mI \qquad \forall x \in \mathbb{R}^n,$$

then f is m-strongly convex (w.r.t. $\|\cdot\|_2$), i.e.

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{m}{2}\|y - x\|_2^2 \qquad \forall x, y \in \mathbb{R}^n.$$

Equivalently, $\lambda_{\min}(\nabla^2 f(x)) \geq m$ for all x.

Hessian bounds and strong convexity

Assume $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable.

PSD order $\iff$ eigenvalue bounds

For any symmetric matrix $H \in S_n$ and any $m \in \mathbb{R}$,

$$H \succeq mI \quad \iff \quad \lambda_{\min}(H) \geq m \quad \iff \quad x^\top Hx \geq m\|x\|_2^2 \ \ \forall x \in \mathbb{R}^n.$$

Strong convexity via the Hessian

If there exists $m > 0$ such that

$$\nabla^2 f(x) \succeq mI \qquad \forall x \in \mathbb{R}^n,$$

then f is m-strongly convex (w.r.t. $\|\cdot\|_2$), i.e.

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{m}{2}\|y - x\|_2^2 \qquad \forall x, y \in \mathbb{R}^n.$$

Equivalently, $\lambda_{\min}(\nabla^2 f(x)) \geq m$ for all x.

Contents

Dual norm: definition and two key examples

Given a norm $\|\cdot\|$, its dual norm is

$$\|y\|_\star := \sup_{\|x\| \leq 1} y^\top x.$$

- Generalized Cauchy-Schwarz: $y^\top x \leq \|y\|_\star \|x\|$.
- Examples: $(\|\cdot\|_2)^\star = \|\cdot\|_2$, $(\|\cdot\|_1)^\star = \|\cdot\|_\infty$, $(\|\cdot\|_\infty)^\star = \|\cdot\|_1$.

Induced operator norm and the spectral norm

For matrix A, the operator norm induced by $\|\cdot\|_{op}$ is

$$\|A\|_{op} := \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

- For $\|\cdot\|_2$: $\|A\|_{2,op}$ is the spectral norm.
- Equivalent characterizations:
  - $\|A\|_{2,op} = \sqrt{\lambda_{\max}(A^\top A)}$.
  - $\|A\|_{2,op} = \sigma_{\max}(A)$ (largest singular value – see later).
- For $A \in S_n$: $\|A\|_{2,op} = \max_i |\lambda_i(A)|$
- For $A \in S_n^+$, $\|A\|_{2,op} = \lambda_{\max}(A)$.

Frobenius norm (matrix "Euclidean" norm)

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

$$\|A\|_F := \sqrt{\sum_{i,j} A_{ij}^2} = \sqrt{\mathrm{Tr}(A^\top A)}.$$

If $A = U\Sigma V^\top$ is an SVD with singular values (σ_i), then

$$\|A\|_F^2 = \sum_{i=1}^{p} \sigma_i^2, \qquad \|A\|_2 \leq \|A\|_F \leq \sqrt{\mathrm{rank}(A)}\, \|A\|_2.$$

Contents

Preview: Newton as repeated quadratic minimization

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable. Use second order Taylor expansion around the current iterate $x^{(k)}$:

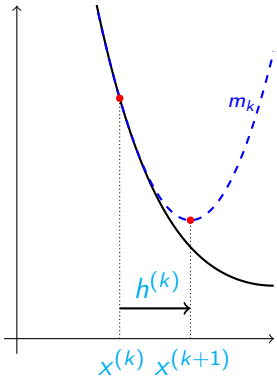
$$f(x^{(k)} + h) = m_k(h) + o(\|h\|_2^2),$$
$$m_k(h) := f(x^{(k)}) + \nabla f(x^{(k)})^\top h + \tfrac{1}{2}\, h^\top \nabla^2 f(x^{(k)})\, h.$$

Newton idea

Newton[a] is: build a quadratic model, minimize it, repeat.

$$h^{(k)} \leftarrow \arg\min_h m_k(h)$$
$$x^{(k+1)} = x^{(k)} + h^{(k)}$$

[a] Often the fastest *local* method under strong regularity assumptions.



Newton step as quadratic model minimization

Preview: Newton as repeated quadratic minimization

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable. Use second order Taylor expansion around the current iterate $x^{(k)}$:

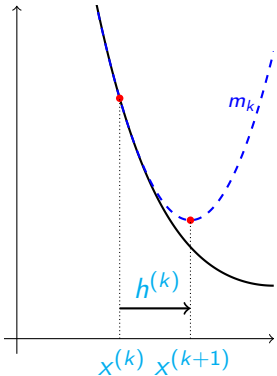
$$f(x^{(k)} + h) = m_k(h) + o(\|h\|_2^2),$$
$$m_k(h) := f(x^{(k)}) + \nabla f(x^{(k)})^\top h + \tfrac{1}{2}\, h^\top \nabla^2 f(x^{(k)})\, h.$$

Newton idea

Newton[a] is: build a quadratic model, minimize it, repeat.

$$h^{(k)} \leftarrow \arg\min_h m_k(h)$$
$$x^{(k+1)} = x^{(k)} + h^{(k)}$$

[a] Often the fastest *local* method under strong regularity assumptions.



Newton step as quadratic model minimization

Minimizing the quadratic model gives a linear system

Recall the local quadratic model at $x^{(k)}$:

$$m_k(h) := f(x^{(k)}) + \nabla f(x^{(k)})^\top h + \tfrac{1}{2}\, h^\top \nabla^2 f(x^{(k)})\, h.$$

Optimality condition

A minimizer $h^{(k)}$ satisfies $\nabla m_k(h^{(k)}) = 0$. Since

$$\nabla m_k(h) = \nabla f(x^{(k)}) + \nabla^2 f(x^{(k)})\, h,$$

we obtain the (symmetric) linear system

$$\nabla^2 f(x^{(k)})\, h^{(k)} = -\nabla f(x^{(k)}).$$

If $\nabla^2 f(x^{(k)}) \succ 0$, the minimizer is unique.

Minimizing the quadratic model gives a linear system

Recall the local quadratic model at $x^{(k)}$:

$$m_k(h) := f(x^{(k)}) + \nabla f(x^{(k)})^\top h + \tfrac{1}{2}\, h^\top \nabla^2 f(x^{(k)})\, h.$$

Optimality condition

A minimizer $h^{(k)}$ satisfies $\nabla m_k(h^{(k)}) = 0$. Since

$$\nabla m_k(h) = \nabla f(x^{(k)}) + \nabla^2 f(x^{(k)})\, h,$$

we obtain the (symmetric) linear system

$$\nabla^2 f(x^{(k)})\, h^{(k)} = -\nabla f(x^{(k)}).$$

If $\nabla^2 f(x^{(k)}) \succ 0$, the minimizer is unique.

Contents

How to solve $Ax = b$?

Goal: solve one linear system $Ax = b$.

Triangular systems are easy

If L is lower triangular, $Lx = b$ is solved by forward substitution (one pass).
If U is upper triangular, $Ux = b$ is solved by backward substitution (one pass).
$\rightsquigarrow O(n^2)$ operations each.

Computing A^{-1} means solving n systems

Let $e_1, \dots, e_n$ be the canonical basis. Then

$$A^{-1} = \begin{bmatrix} x^{(1)} & \cdots & x^{(n)} \end{bmatrix} \quad \text{with} \quad Ax^{(i)} = e_i \ \ (i = 1, \dots, n).$$

So forming A^{-1} amounts to solving n right-hand sides.
$\rightsquigarrow O(n^3)$ operations total.

Conclusion: solve $Ax = b$ with dedicated tools, do not invert.

How to solve $Ax = b$?

Goal: solve one linear system $Ax = b$.

Triangular systems are easy

If L is lower triangular, $Lx = b$ is solved by forward substitution (one pass).
If U is upper triangular, $Ux = b$ is solved by backward substitution (one pass).
$\rightsquigarrow O(n^2)$ operations each.

Computing A^{-1} means solving n systems

Let $e_1, \ldots, e_n$ be the canonical basis. Then

$$A^{-1} = \begin{bmatrix} x^{(1)} & \cdots & x^{(n)} \end{bmatrix} \quad \text{with} \quad Ax^{(i)} = e_i \ \ (i = 1, \ldots, n).$$

So forming A^{-1} amounts to solving n right-hand sides.
$\rightsquigarrow O(n^3)$ operations total.

Conclusion: solve $Ax = b$ with dedicated tools, do not invert.

Why factorize?

In optimization, we repeatedly solve linear systems $Ax = b$ (Newton steps, least-squares, KKT blocks, ...).

Principle

Factorize once, solve many times.

- We reuse the same factorization, we do not compute A^{-1}.
- The expensive part is the factorization; each new right-hand side is cheap (triangular solves).

♣ Exercise: Give two different vectors $b^{(1)}, b^{(2)}$. Explain why reusing the same factorization of A is efficient.

Contents

LU factorization (square matrices)

For a general square matrix $A \in \mathbb{R}^{n \times n}$, we aim for

$$PA = LU,$$

where L is lower triangular (often with ones on the diagonal) and U is upper triangular, and P is a permutation matrix (pivoting for numerical stability).

♣ Exercise: Assume $A = LU$. Show that solving $Ax = b$ reduces to two triangular solves. Show that it stays the same if $PA = LU$.

Cholesky factorization (symmetric positive definite) ♢

If $A \in S_n^{++}$ (*i.e.*,$A \succ 0$), then there exists a unique lower triangular L with positive diagonal such that

$$A = LL^\top.$$

- Solve $Ax = b$ via:

$$Ly = b, \qquad L^\top x = y.$$

- This is the standard path for Hessian systems in strongly convex optimization.

♣ Exercise: Show that if $A = LL^\top$ and L is invertible, then $A \succ 0$.

Why $LL^\top$ (Cholesky) over LU?

Assume $A \in S_n$ and $A \succ 0$.

- Faster (dense): Cholesky costs about $\frac{1}{3}n^3$ flops vs about $\frac{2}{3}n^3$ for LU.
- Less memory: store essentially one triangular factor (symmetry is exploited).
- More robust here: for SPD matrices, Cholesky does not require pivoting.
- Matches optimization geometry: $x^\top Ax = \|L^\top x\|_2^2$.

♠ Exercise: Assume $0 \prec A \preceq B$. Explain why this suggests that solving with A is typically harder than with B (conditioning viewpoint).

Worked example: Cholesky on a 2×2 SPD matrix

Let

$$A = \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}, \qquad b = \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

We look for $A = LL^\top$ with $L = \begin{pmatrix} \ell_{11} & 0 \\ \ell_{21} & \ell_{22} \end{pmatrix}$.

Matching entries:

$$\ell_{11} = 2, \qquad \ell_{21} = 1, \qquad \ell_{22} = \sqrt{2}, \quad \Rightarrow \quad L = \begin{pmatrix} 2 & 0 \\ 1 & \sqrt{2} \end{pmatrix}.$$

Solve $Ly = b$: $y_1 = 1$, $y_2 = 0$. Then solve $L^\top x = y$: $x_2 = 0$, $x_1 = \frac{1}{2}$.

$$x = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}.$$

♣ Exercise: Change b to $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Reuse the same L and solve again.

Worked example: Cholesky on a 2×2 SPD matrix

Let

$$A = \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}, \qquad b = \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

We look for $A = LL^\top$ with $L = \begin{pmatrix} \ell_{11} & 0 \\ \ell_{21} & \ell_{22} \end{pmatrix}$.

Matching entries:

$$\ell_{11} = 2, \qquad \ell_{21} = 1, \qquad \ell_{22} = \sqrt{2}, \quad \Rightarrow \quad L = \begin{pmatrix} 2 & 0 \\ 1 & \sqrt{2} \end{pmatrix}.$$

Solve $Ly = b$: $y_1 = 1$, $y_2 = 0$. Then solve $L^\top x = y$: $x_2 = 0$, $x_1 = \frac{1}{2}$.

$$x = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}.$$

♣ Exercise: Change b to $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Reuse the same L and solve again.

Worked example: Cholesky on a 2×2 SPD matrix

Let

$$A = \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}, \qquad b = \begin{pmatrix} 2 \\ 1 \end{pmatrix}.$$

We look for $A = LL^\top$ with $L = \begin{pmatrix} \ell_{11} & 0 \\ \ell_{21} & \ell_{22} \end{pmatrix}$.

Matching entries:

$$\ell_{11} = 2, \qquad \ell_{21} = 1, \qquad \ell_{22} = \sqrt{2}, \quad \Rightarrow \quad L = \begin{pmatrix} 2 & 0 \\ 1 & \sqrt{2} \end{pmatrix}.$$

Solve $Ly = b$: $y_1 = 1$, $y_2 = 0$. Then solve $L^\top x = y$: $x_2 = 0$, $x_1 = \frac{1}{2}$.

$$x = \begin{pmatrix} \frac{1}{2} \\ 0 \end{pmatrix}.$$

♣ Exercise: Change b to $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$. Reuse the same L and solve again.

Takeaways (what to remember)

- Solving $Ax = b$ is done via factorization + triangular solves, not via A^{-1}.
- Use Cholesky when A is symmetric positive definite: $A = LL^\top$.
- Use LU as the general-purpose tool (often with a permutation $PA = LU$).
- Reuse factorizations whenever the matrix A stays the same and only b changes.

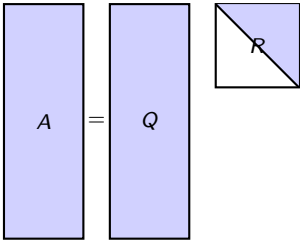
♠ Exercise: (Optimization link) Consider $f(x) = \frac{1}{2}x^\top Ax - b^\top x$ with $A \succ 0$. Show that the minimizer solves $Ax = b$, and explain how Cholesky gives the minimizer efficiently.

Contents

QR factorization (thin QR)

For a matrix $A \in \mathbb{R}^{m \times n}$, with $m \geq n$, there exist matrices $Q \in \mathbb{R}^{m \times n}$ and $R \in \mathbb{R}^{n \times n}$ such that

$$A = QR,$$
$$Q^\top Q = I_n,$$
R upper triangular.



tall matrix $m \geq n$

Blue = potentially nonzero entries .

In particular:

- Q has orthonormal columns. If $\mathrm{rank}(A) = n$, they form a basis of $\mathrm{Im}(A)$.
- R is a change of coordinates in that basis (triangular $\Rightarrow$ easy solves).
- If A has full column rank, then R is invertible.
- Note that $P := QQ^\top$ is the orthogonal projector onto $\mathrm{Im}(Q)$.

Least squares and why we avoid normal equations

Given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, the (linear) least-squares problem is

$$\min_{x \in \mathbb{R}^n} \ \|Ax - b\|_2^2.$$

Optimality condition (normal equations)

If A has full column rank, the objective is differentiable and strictly convex, and the unique minimizer $x^\star$ satisfies

$$A^\top (Ax^\star - b) = 0 \qquad \Longleftrightarrow \qquad A^\top A\, x^\star = A^\top b.$$

But: normal equations can be numerically bad

Forming $A^\top A$ typically squares the conditioning:

$$\kappa_2(A^\top A) = \kappa_2(A)^2 \quad \text{(when } A \text{ has full column rank).}$$

Conclusion: solve LS via QR (or SVD), not via $A^\top A$.

Least squares and why we avoid normal equations

Given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, the (linear) least-squares problem is

$$\min_{x \in \mathbb{R}^n} \ \|Ax - b\|_2^2.$$

Optimality condition (normal equations)

If A has full column rank, the objective is differentiable and strictly convex, and the unique minimizer $x^\star$ satisfies

$$A^\top (Ax^\star - b) = 0 \qquad \Longleftrightarrow \qquad A^\top A\, x^\star = A^\top b.$$

But: normal equations can be numerically bad

Forming $A^\top A$ typically squares the conditioning:

$$\kappa_2(A^\top A) = \kappa_2(A)^2 \quad \text{(when } A \text{ has full column rank).}$$

Conclusion: solve LS via QR (or SVD), not via $A^\top A$.

Least squares and why we avoid normal equations

Given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$, the (linear) least-squares problem is

$$\min_{x \in \mathbb{R}^n} \ \|Ax - b\|_2^2.$$

Optimality condition (normal equations)

If A has full column rank, the objective is differentiable and strictly convex, and the unique minimizer $x^\star$ satisfies

$$A^\top (Ax^\star - b) = 0 \qquad \Longleftrightarrow \qquad A^\top A\, x^\star = A^\top b.$$

But: normal equations can be numerically bad

Forming $A^\top A$ typically squares the conditioning:

$$\kappa_2(A^\top A) = \kappa_2(A)^2 \quad \text{(when } A \text{ has full column rank).}$$

Conclusion: solve LS via QR (or SVD), not via $A^\top A$.

Solving least squares via QR (thin QR)

Assume $m \geq n$. Let $A = QR$ be a thin QR factorization with

$$Q \in \mathbb{R}^{m \times n}, \quad Q^\top Q = I_n, \qquad R \in \mathbb{R}^{n \times n} \text{ upper triangular.}$$

Key step: orthogonal decomposition

Let $P := QQ^\top$ (orthogonal projector onto $\mathrm{Im}(Q)$). Then for any x,

$$\|Ax - b\|_2^2 = \|QRx - b\|_2^2 = \|Rx - Q^\top b\|_2^2 + \|(I - P)b\|_2^2.$$

Consequence (full column rank)

If A has full column rank, then R is invertible and the unique minimizer satisfies

$$Rx^\star = Q^\top b,$$

solved by backward substitution.

Solving least squares via QR (thin QR)

Assume $m \geq n$. Let $A = QR$ be a thin QR factorization with

$$Q \in \mathbb{R}^{m \times n}, \quad Q^\top Q = I_n, \qquad R \in \mathbb{R}^{n \times n} \text{ upper triangular.}$$

Key step: orthogonal decomposition

Let $P := QQ^\top$ (orthogonal projector onto $\mathrm{Im}(Q)$). Then for any x,

$$\|Ax - b\|_2^2 = \|QRx - b\|_2^2 = \|Rx - Q^\top b\|_2^2 + \|(I - P)b\|_2^2.$$

Consequence (full column rank)

If A has full column rank, then R is invertible and the unique minimizer satisfies

$$Rx^\star = Q^\top b,$$

solved by backward substitution.

Contents

Contents

Singular Value Decomposition (SVD): definition

SVD extends spectral decomposition to rectangular matrices.

Existence (main requirement)

For any $A \in \mathbb{R}^{m\times n}$, there exist orthogonal matrices $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$, and a rectangular diagonal matrix $\Sigma \in \mathbb{R}^{m\times n}$ such that

$$A = U\Sigma V^\top, \qquad \Sigma = \begin{cases} \mathrm{diag}(\sigma_1, \dots, \sigma_n) & \text{if } m \geq n, \\ \mathrm{diag}(\sigma_1, \dots, \sigma_m) & \text{if } m < n, \end{cases}$$

with $p = \min(m, n)$ and $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$.

- U, V: orthogonal changes of basis (rotations/reflections).
- Σ: axis scalings; zeros encode directions collapsed by A.
- If $r = \mathrm{rank}(A)$: $\sigma_1 \geq \cdots \geq \sigma_r > 0$ and $\sigma_{r+1} = \cdots = \sigma_p = 0$.
- Geometry: $A(B_2)$ is an ellipsoid with semi-axes σ_i (directions = columns of U).

Singular Value Decomposition (SVD): definition

SVD extends spectral decomposition to rectangular matrices.

Existence (main requirement)

For any $A \in \mathbb{R}^{m\times n}$, there exist orthogonal matrices $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$, and a rectangular diagonal matrix $\Sigma \in \mathbb{R}^{m\times n}$ such that

$$A = U\Sigma V^\top, \qquad \Sigma = \begin{cases} \mathrm{diag}(\sigma_1, \dots, \sigma_n) & \text{if } m \geq n, \\ \mathrm{diag}(\sigma_1, \dots, \sigma_m) & \text{if } m < n, \end{cases}$$

with $p = \min(m, n)$ and $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$.

- U, V: orthogonal changes of basis (rotations/reflections).
- Σ: axis scalings; zeros encode directions collapsed by A.
- If $r = \mathrm{rank}(A)$: $\sigma_1 \geq \cdots \geq \sigma_r > 0$ and $\sigma_{r+1} = \cdots = \sigma_p = 0$.
- Geometry: $A(B_2)$ is an ellipsoid with semi-axes σ_i (directions = columns of U).

Singular Value Decomposition (SVD): definition

SVD extends spectral decomposition to rectangular matrices.

Existence (main requirement)

For any $A \in \mathbb{R}^{m\times n}$, there exist orthogonal matrices $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$, and a rectangular diagonal matrix $\Sigma \in \mathbb{R}^{m\times n}$ such that

$$A = U\Sigma V^\top, \qquad \Sigma = \begin{cases} \mathrm{diag}(\sigma_1, \dots, \sigma_n) & \text{if } m \geq n, \\ \mathrm{diag}(\sigma_1, \dots, \sigma_m) & \text{if } m < n, \end{cases}$$

with $p = \min(m, n)$ and $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$.

- U, V: orthogonal changes of basis (rotations/reflections).
- Σ: axis scalings; zeros encode directions collapsed by A.
- If $r = \mathrm{rank}(A)$: $\sigma_1 \geq \cdots \geq \sigma_r > 0$ and $\sigma_{r+1} = \cdots = \sigma_p = 0$.
- Geometry: $A(B_2)$ is an ellipsoid with semi-axes σ_i (directions = columns of U).

SVD: key properties

Let $A = U\Sigma V^\top$ be an SVD with singular values $\sigma_1 \geq \cdots \geq \sigma_p$.

- σ_i^2 are the positive eigenvalues of $A^\top A$ (and of $AA^\top$).
- $\|A\|_2 = \sigma_1$ (spectral/operator norm).
- $$\min_{\|x\|_2=1} \|Ax\|_2 = \begin{cases} \sigma_p & \text{if } \ker(A) = \{0\} \quad \text{(full column rank)}, \\ 0 & \text{otherwise.} \end{cases}$$
- If A is square invertible: $\|A^{-1}\|_2 = 1/\sigma_n$ and $\kappa_2(A) = \sigma_1/\sigma_n$.
- The truncated SVD gives the best rank-k approximation of A in spectral and Frobenius norms.

Contents

Condition number and why optimization cares

- For A invertible, $\kappa_2(A) := \|A\|_2 \, \|A^{-1}\|_2 = \sigma_1/\sigma_n \geq 1$.
- For $A \in S_n^{++}$: $\kappa_2(A) = \lambda_{\max}(A)/\lambda_{\min}(A)$.
- Linear systems: relative error can be amplified by $\kappa_2(A)$.
- GD on quadratics: rates depend on κ (you already use this later).

♠ Exercise: Preview: CG improves κ to $\sqrt{\kappa}$ in the convergence factor.

Stability vs conditioning (linear solves)

Two different notions:

- Conditioning = sensitivity of the *problem*.

$$Ax = b \quad \Rightarrow \quad \text{how much does } x \text{ change if } b \text{ changes a bit?}$$

- Stability = behavior of the *algorithm* in finite precision.

$$\text{does the computed } \hat{x} \text{ solve a nearby linear system?}$$

Rule of thumb

Even with a very good algorithm, the best relative accuracy you can expect is often on the order of

$$\frac{\|\hat{x} - x\|}{\|x\|} \approx O(\kappa_2(A)\, u)\,,$$

where u is machine precision.

♣ Exercise: Take $A = \mathrm{diag}(1, \varepsilon)$, $b = (1, 1)$. What happens to the solution when ε is very small? Interpret with $\kappa_2(A)$.

Stability vs conditioning (linear solves)

Two different notions:

- Conditioning = sensitivity of the *problem*.

$$Ax = b \quad \Rightarrow \quad \text{how much does } x \text{ change if } b \text{ changes a bit?}$$

- Stability = behavior of the *algorithm* in finite precision.

$$\text{does the computed } \hat{x} \text{ solve a nearby linear system?}$$

Rule of thumb

Even with a very good algorithm, the best relative accuracy you can expect is often on the order of

$$\frac{\|\hat{x} - x\|}{\|x\|} \approx O(\kappa_2(A)\, u)\,,$$

where u is machine precision.

♣ Exercise: Take $A = \mathrm{diag}(1, \varepsilon)$, $b = (1, 1)$. What happens to the solution when ε is very small? Interpret with $\kappa_2(A)$.

Backward stability (the ideal guarantee)

Definition (informal)

An algorithm for solving $Ax = b$ is backward stable if it returns $\hat{x}$ such that

$$(A + \Delta A)\,\hat{x} = b \quad \text{with} \quad \frac{\|\Delta A\|}{\|A\|} \text{ small (typically } O(u)).$$

- Meaning: the algorithm behaves as if it solved *exactly* a slightly perturbed problem.
- If the problem is well-conditioned, a small backward error implies a small forward error.

Practical takeaway

Good solvers are designed to be (close to) backward stable:

- SPD $\Rightarrow$ Cholesky + triangular solves.
- General $A \Rightarrow$ LU with pivoting.
- Least-squares $\Rightarrow$ QR (avoid normal equations).

Backward stability (the ideal guarantee)

Definition (informal)

An algorithm for solving $Ax = b$ is backward stable if it returns $\hat{x}$ such that

$$(A + \Delta A)\,\hat{x} = b \quad \text{with} \quad \frac{\|\Delta A\|}{\|A\|} \text{ small (typically } O(u)).$$

- Meaning: the algorithm behaves as if it solved *exactly* a slightly perturbed problem.
- If the problem is well-conditioned, a small backward error implies a small forward error.

Practical takeaway

Good solvers are designed to be (close to) backward stable:

- SPD $\Rightarrow$ Cholesky + triangular solves.
- General $A \Rightarrow$ LU with pivoting.
- Least-squares $\Rightarrow$ QR (avoid normal equations).

Backward stability (the ideal guarantee)

Definition (informal)

An algorithm for solving $Ax = b$ is backward stable if it returns $\hat{x}$ such that

$$(A + \Delta A)\,\hat{x} = b \quad \text{with} \quad \frac{\|\Delta A\|}{\|A\|} \text{ small (typically } O(u)).$$

- Meaning: the algorithm behaves as if it solved *exactly* a slightly perturbed problem.
- If the problem is well-conditioned, a small backward error implies a small forward error.

Practical takeaway

Good solvers are designed to be (close to) backward stable:

- SPD $\Rightarrow$ Cholesky + triangular solves.
- General $A \Rightarrow$ LU with pivoting.
- Least-squares $\Rightarrow$ QR (avoid normal equations).

What you have to know

- Definitions: eigenvalues/eigenvectors; spectrum; PSD/PD; Loewner order $A \preceq B$.
- How to read $mI \preceq A \preceq MI$ as a "geometry sandwich" (ellipsoid between balls) and as eigenvalue bounds.
- Spectral theorem for symmetric matrices: $A = Q\Lambda Q^\top$ and consequences ($A \succeq 0 \Leftrightarrow \lambda_i \geq 0$).
- Dual norms and induced operator norm; in particular $\|A\|_2 = \sigma_{\max}(A)$ and $\|A\|_2 = \max_i |\lambda_i(A)|$ for $A \in S_n$.

What you really should know

- Rayleigh quotient facts: $\lambda_{\min} = \min_{\|x\|_2=1} x^\top A x$, $\lambda_{\max} = \max_{\|x\|_2=1} x^\top A x$.
- The projector viewpoint: $P^2 = P$, $P = P^\top$, Px is the Euclidean closest point in $\mathrm{Im}(P)$; $P = QQ^\top$.
- Conditioning as the main complexity parameter: $\kappa_2(A) = \lambda_{\max}/\lambda_{\min}$ (SPD), impacts GD/CG/Newton and numerical accuracy.
- Stability vs conditioning: good solvers aim for (near) backward stability; forward error typically scales like $\kappa \cdot u$.

What you have to be able to do

- Use spectral decomposition to prove PSD/PD statements and derive inequalities like $\lambda_{\min}\|x\|_2^2 \leq x^\top Ax \leq \lambda_{\max}\|x\|_2^2$.
- Recognize when $\|\cdot\|_Q$ is a norm vs a seminorm; use $\|x\|_Q = \|Q^{1/2}x\|_2$ when $Q \succ 0$.
- Solve $Ax = b$ efficiently via factorizations (LU / Cholesky): factorize once, then triangular solves for new right-hand sides.

What you should be able to do

- Diagnose ill-posedness: interpret tiny singular values / large κ as instability risk; know when SVD is the right diagnostic tool.
- Give quick "method choices" for linear algebra primitives: SPD $\rightarrow$ Cholesky; general square $\rightarrow$ pivoted LU; LS $\rightarrow$ QR (or SVD).