

Optimization In AI

Class 02: Mathematical Prerequisites

Instructor: Jianhao Ma

Department of Industrial Engineering, Tsinghua University

Today's Topics

- 01** Linear Algebra
- 02** Taylor Expansion
- 03** Matrix Calculus
- 04** Convexity
- 05** Optimality
- 06** Convergence
- 07** Probability

PART 01

Linear Algebra

Eigenvalues And Eigenvectors

Definition (Eigenvalue And Eigenvector)

For $A \in \mathbb{R}^{n \times n}$, a scalar λ is an eigenvalue if some $v \neq 0$ satisfies

$$Av = \lambda v.$$

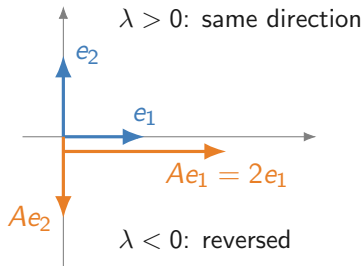
v : eigenvector. λ : scaling factor.

$$A = \begin{pmatrix} 2 & 0 \\ 0 & -1 \end{pmatrix}$$

$$Ae_1 = 2e_1, \quad Ae_2 = -e_2.$$

Invariant Line

$$\lambda = 0 \implies Av = 0.$$



Diagonalization And The Spectral Theorem

General Case: Eigenbasis

$V = [v_1, \dots, v_n]$ with n independent eigenvectors:

$$AV = V\Lambda \implies A = V\Lambda V^{-1}.$$

Theorem (Real Symmetric Spectral Theorem)

Every real symmetric matrix admits

$$\boxed{A = Q\Lambda Q^T}, \quad Q^T Q = I, \quad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_n),$$

with **real eigenvalues** and an **orthonormal eigenbasis**.

In eigen-coordinates: $z = Q^T x$ gives $Q^T Ax = \Lambda z$.

What can fail? $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ has only one independent eigenvector.

Singular Value Decomposition

Theorem (SVD)

Every $A \in \mathbb{R}^{m \times n}$ can be written as

$$\underbrace{A}_{m \times n} = \underbrace{U}_{m \times m} \underbrace{\Sigma}_{m \times n} \underbrace{V^T}_{n \times n}, \quad U^T U = I, \quad V^T V = I.$$

The diagonal entries satisfy $\sigma_1 \geq \dots \geq \sigma_p \geq 0$, $p = \min(m, n)$.

$$x \xrightarrow{\text{coordinates}} V^T x \xrightarrow{\text{scale}} \Sigma V^T x \xrightarrow{\text{rotate / reflect}} Ax$$

$$Av_i = \sigma_i u_i \text{ for } i \leq p. \quad \text{rank}(A) = \#\{i : \sigma_i > 0\}.$$

Proof Of The SVD

1. Right Singular Vectors

$A^T A \succeq 0$; orthonormal eigenbasis $v_1, \dots, v_n$ (spectral theorem).

$$A^T A v_i = \lambda_i v_i, \quad \lambda_1 \geq \dots \geq \lambda_n \geq 0.$$

$\sigma_i = \sqrt{\lambda_i}$, $r = \text{rank}(A) \leq m$; $\ker(A^T A) = \ker(A)$.

2. Left Singular Vectors

$$u_i = A v_i / \sigma_i \quad (i \leq r), \quad u_i^T u_j = \frac{v_i^T A^T A v_j}{\sigma_i \sigma_j} = \delta_{ij}.$$

3. Complete The Basis

Extend $u_1, \dots, u_r$ to an orthonormal basis of $\mathbb{R}^m$. $i > r$: $\|A v_i\|_2^2 = \lambda_i = 0 \implies A v_i = 0$.

$U = [u_1, \dots, u_m]$, $V = [v_1, \dots, v_n]$; $\Sigma \in \mathbb{R}^{m \times n}$ diagonal, $\Sigma_{ii} = \sigma_i$ ($i \leq \min(m, n)$).

$$AV = U\Sigma \implies \boxed{A = U\Sigma V^T}.$$

Eigenvalues Versus Singular Values

Multiply The SVD

$$\begin{aligned}A^{\top}A &= V(\Sigma^{\top}\Sigma)V^{\top}, \\AA^{\top} &= U(\Sigma\Sigma^{\top})U^{\top}.\end{aligned}$$

Their nonzero eigenvalues are σ_i^2 .

Symmetric A : singular values = $|\lambda_i(A)|$,
reordered.

Can singular values determine definiteness?

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 2 & 0 \\ 0 & -1 \end{pmatrix}.$$

	A	B
Eigenvalues	(2, 1)	(2, -1)
Singular values	(2, 1)	(2, 1)
	PD	Indefinite

No: singular values discard signs.

Vector Norms

For $x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$:

Definition (Vector p -Norm)

For $1 \leq p < \infty$, $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}$.

At the endpoint, $\|x\|_\infty = \lim_{p \rightarrow \infty} \|x\|_p = \max_{1 \leq i \leq n} |x_i|$.

Three common cases

$$p = 1 \quad \|x\|_1 = \sum_{i=1}^n |x_i|$$

$$p = 2 \quad \|x\|_2 = \left(\sum_{i=1}^n x_i^2 \right)^{1/2}$$

$$p = \infty \quad \|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

Matrix Norms

Definition (Induced Matrix p -Norm, $1 \leq p \leq \infty$)

$$\|A\|_{p \rightarrow p} = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}. \quad \|A\|_p := \|A\|_{p \rightarrow p}.$$

Spectral ($p = 2$) $\|A\|_2 = \sigma_1(A)$

Maximum column sum ($p = 1$) $\|A\|_1 = \max_j \sum_i |A_{ij}|$

Maximum row sum ($p = \infty$) $\|A\|_\infty = \max_i \sum_j |A_{ij}|$

Frobenius $\|A\|_F = (\sum_{ij} A_{ij}^2)^{1/2} = (\sum_i \sigma_i^2)^{1/2}$

Nuclear $\|A\|_* = \sum_i \sigma_i$

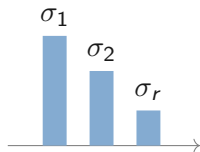
Comparing Matrix Norms

Let $r = \text{rank}(A) > 0$ and $\sigma(A) = (\sigma_1, \dots, \sigma_r)$.

$\|A\|_2 = \|\sigma(A)\|_\infty$ largest singular value,

$\|A\|_F = \|\sigma(A)\|_2$ Euclidean length,

$\|A\|_* = \|\sigma(A)\|_1$ sum of singular values.



Lemma (Matrix-Norm Comparisons)

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{r} \|A\|_2, \quad \|A\|_F \leq \|A\|_* \leq \sqrt{r} \|A\|_F.$$

Cauchy–Schwarz. $A = 0$: all norms vanish.

Induced norms: $\|A\|_2 \leq \sqrt{\|A\|_1 \|A\|_\infty}$.

Check I_d : $\|I_d\|_2 = 1$, $\|I_d\|_F = \sqrt{d}$, $\|I_d\|_* = d$.

Proof Details: Comparing Matrix Norms

Let $r = \text{rank}(A) > 0$ and $s = (\sigma_1, \dots, \sigma_r)$.

Singular-Value Inequalities

Since $0 \leq \sigma_i \leq \sigma_1$,

$$\sigma_1^2 \leq \sum_i \sigma_i^2 \leq r\sigma_1^2.$$

Also, expanding the square and applying Cauchy–Schwarz gives

$$\sum_i \sigma_i^2 \leq \left(\sum_i \sigma_i \right)^2 \leq r \sum_i \sigma_i^2.$$

Take square roots and use $\|A\|_2 = \|s\|_\infty$, $\|A\|_F = \|s\|_2$, $\|A\|_* = \|s\|_1$.

For the induced-norm comparison, spectral radius is bounded by any induced norm:

$$\|A\|_2^2 = \rho(A^\top A) \leq \|A^\top A\|_1 \leq \|A^\top\|_1 \|A\|_1 = \|A\|_\infty \|A\|_1.$$

When $r = 0$, $A = 0$ and all inequalities hold with equality.

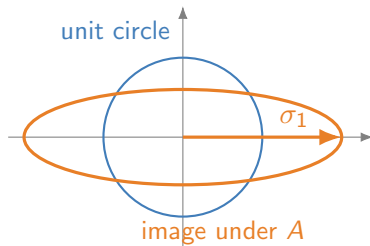
Why The Spectral Norm Is The Largest Stretch

Singular-Vector Coordinates

$$z = V^T x, \quad \|z\|_2 = \|x\|_2.$$

$$\begin{aligned}\|Ax\|_2^2 &= \|\Sigma z\|_2^2 \\ &= \sum_{i=1}^p \sigma_i^2 z_i^2 \\ &\leq \sigma_1^2 \|z\|_2^2 = \sigma_1^2 \|x\|_2^2.\end{aligned}$$

Equality: $x = v_1$. $\|A\|_2 = \sigma_1$.



σ_1 : largest stretching factor

$$\|Ax\|_2 \leq \|A\|_2 \|x\|_2, \quad \|AB\|_2 \leq \|A\|_2 \|B\|_2.$$

Definiteness Of Symmetric Matrices

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

Type	Quadratic form	Eigenvalues
Positive definite (PD)	$x^T A x > 0$ for all $x \neq 0$	all $\lambda_i > 0$
Positive semidefinite (PSD)	$x^T A x \geq 0$ for all x	all $\lambda_i \geq 0$
Negative definite (ND)	$x^T A x < 0$ for all $x \neq 0$	all $\lambda_i < 0$
Negative semidefinite (NSD)	$x^T A x \leq 0$ for all x	all $\lambda_i \leq 0$
Indefinite	takes both signs	positive and negative

Eigen-Coordinates: $x^T A x = \sum_i \lambda_i z_i^2$, where $z = Q^T x$.

$$x = q_i \implies x^T A x = \lambda_i.$$

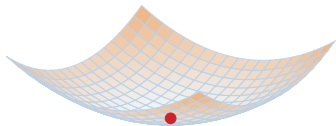
PSD is not entrywise positivity.

Definiteness And Optimization

$$q(x) = \frac{1}{2}x^\top Ax, \quad A = A^\top, \quad \nabla q(0) = 0, \quad \nabla^2 q = A.$$

Positive definite

$$A = \text{diag}(1, 1)$$



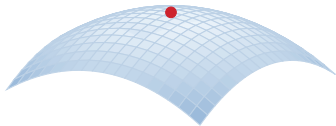
Strict local minimum

Positive curvature in every direction.

For $f \in C^2$ on an open domain: $A = \nabla^2 f(x^*)$ at $\nabla f(x^*) = 0$.

Negative definite

$$A = \text{diag}(-1, -1)$$

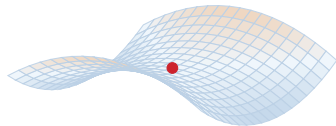


Strict local maximum

Negative curvature in every direction.

Indefinite

$$A = \text{diag}(1, -1)$$



Saddle point

Curvature of both signs.

The Semidefinite Order

Definition (Semidefinite Order)

For symmetric $A, B \in \mathbb{R}^{n \times n}$, $A \succeq B$ means $A - B$ is PSD; equivalently, $x^\top A x \geq x^\top B x$ for every $x \in \mathbb{R}^n$.

Gram matrices

For any rectangular M ,

$$x^\top M^\top M x = \|Mx\|_2^2 \geq 0.$$

$$M^\top M \succeq 0.$$

$M^\top M \succ 0$ if and only if M has full column rank.

Uniform bounds

Equivalent bounds:

$$\mu I \preceq A \preceq LI,$$

$$\mu \|x\|_2^2 \leq x^\top A x \leq L \|x\|_2^2 \quad \forall x,$$

$$\mu \leq \lambda_i(A) \leq L \quad \forall i.$$

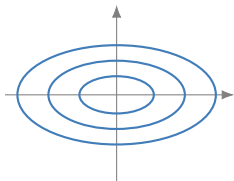


Quadratic Forms And Conditioning

$$q(x) = \frac{1}{2}x^\top Ax, \quad A = A^\top.$$

Positive definite

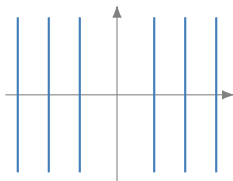
$$A = \text{diag}(1, 4)$$



Elliptical level sets.

Positive semidefinite

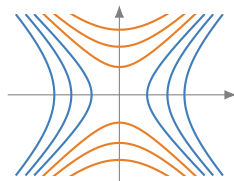
$$A = \text{diag}(1, 0)$$



Parallel level sets; flat in x_2 .

Indefinite

$$A = \text{diag}(1, -1)$$



Hyperbolic level sets.

Blue contours: $q(x) > 0$. Orange contours: $q(x) < 0$.

Condition Number And Ellipse Geometry

For an invertible matrix M , $\kappa(M) = \frac{\sigma_{\max}(M)}{\sigma_{\min}(M)}$.

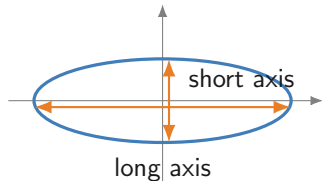
$A \succ 0$, $q(x) = \frac{1}{2}x^\top Ax$:

$$\kappa(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

Level $q(x) = c > 0$: axis lengths $\propto$

$$\frac{1}{\sqrt{\lambda_1}}, \dots, \frac{1}{\sqrt{\lambda_n}}.$$

Larger $\kappa \rightarrow$ More Elongated Level Sets.



$$\kappa(A) = \left(\frac{\text{long axis}}{\text{short axis}} \right)^2.$$

PART 02

Taylor Expansion

Taylor Expansion In One Variable

Theorem (Finite-order Taylor formula)

Let $f \in C^{k+1}$ on an open interval containing the segment from a to $a + h$. For some $\theta \in (0, 1)$,

$$f(a + h) = \sum_{j=0}^k \frac{f^{(j)}(a)}{j!} h^j + \underbrace{\frac{f^{(k+1)}(a + \theta h)}{(k+1)!} h^{k+1}}_{R_k(h)}.$$

Remainder Bound

If $|f^{(k+1)}(t)| \leq M$ on that segment, then

$$|R_k(h)| \leq \frac{M}{(k+1)!} |h|^{k+1}.$$

Finite-order identity; no infinite-series claim.

Comparing Taylor Approximations

For $f(h) = e^h$ about zero:

$$T_1(h) = 1 + h, \quad T_2(h) = 1 + h + \frac{1}{2}h^2.$$

$$|h| \leq r: \quad |f^{(j)}(h)| \leq e^r.$$

$$|e^h - T_1(h)| \leq \frac{e^r}{2}h^2,$$

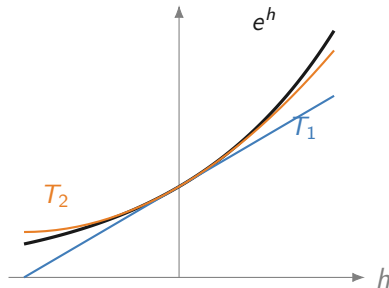
$$|e^h - T_2(h)| \leq \frac{e^r}{6}|h|^3.$$

First order: $O(h^2)$

Second order: $O(h^3)$

At $r = h = 0.1$: bounds 0.00553 and 0.000185.

Check: $|h| \mapsto |h|/2$, fixed $r \implies$ bounds $\div 4$ and $\div 8$.



Higher-Order Directional Derivatives

Let $f : U \rightarrow \mathbb{R}$ be C^j on an open set $U \subseteq \mathbb{R}^d$, with $x \in U$ and $h \in \mathbb{R}^d$.

Definition (Higher-Order Directional Derivative)

For $j \geq 1$, evaluate the j th derivative on j copies of the direction h :

$$D^j f(x) \underbrace{[h, \dots, h]}_{j \text{ copies}} = \left. \frac{d^j}{dt^j} f(x + th) \right|_{t=0}.$$

By convention, the order-zero term is $D^0 f(x) = f(x)$.

The first two orders

$$Df(x)[h] = \nabla f(x)^\top h, \quad D^2 f(x)[h, h] = h^\top \nabla^2 f(x) h.$$

$$D^j f(x): j\text{-linear map.} \quad D^j f(x)[h, \dots, h] \in \mathbb{R}.$$

Multivariable Taylor Expansion

Fix x, h and restrict f to the line segment:

$$\phi(t) = f(x + th), \quad \phi^{(j)}(t) = D^j f(x + th)[h, \dots, h].$$

Theorem (Directional form)

For an integer $k \geq 0$, suppose $f \in C^{k+1}$ near the segment $[x, x + h]$. Then

$$f(x + h) = \sum_{j=0}^k \frac{1}{j!} D^j f(x)[h, \dots, h] + R_k(x, h),$$

where, for some $\theta \in (0, 1)$,

$$R_k(x, h) = \frac{D^{k+1} f(x + \theta h)[h, \dots, h]}{(k+1)!}.$$

Proof: apply scalar Taylor to ϕ about 0, evaluated at 1.

Scalar Taylor Along A Line

Proof Details: Taylor Along A Line

Fix h and define $\phi(t) = f(x + th)$ for $0 \leq t \leq 1$.

Chain Rule

The chain rule gives

$$\phi'(t) = Df(x + th)[h], \quad \phi''(t) = D^2f(x + th)[h, h].$$

Inductively, h is constant, so

$$\phi^{(j)}(t) = D^j f(x + th)[h, \dots, h].$$

Scalar Taylor between 0 and 1 yields, for some $\theta \in (0, 1)$,

$$\phi(1) = \sum_{j=0}^k \frac{\phi^{(j)}(0)}{j!} + \frac{\phi^{(k+1)}(\theta)}{(k+1)!}.$$

Substitute $\phi(1) = f(x + h)$ and the derivative identities. For $j = 0$, the term is $f(x)$. The assumption C^{k+1} near the segment ensures that the scalar Taylor theorem applies.

The First-Order Model

Definition (Differentiability at x)

$$f(x+h) = f(x) + \nabla f(x)^\top h + o(\|h\|_2).$$

$$o(\|h\|_2)/\|h\|_2 \rightarrow 0 \text{ as } h \rightarrow 0.$$

Directional derivative

$$Df(x)[v] = \nabla f(x)^\top v.$$

Minimize over $\|v\|_2 = 1$, with $\nabla f(x) \neq 0$:

$$v = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}.$$

Differentiable is not enough

For $f(t) = |t|^{3/2}$ at zero,

$$f'(0) = 0, \quad R_1(h) = |h|^{3/2}.$$

This is $o(|h|)$, but not $O(h^2)$.

A Quantitative First-Order Remainder

Definition (L -smoothness)

On an open convex set D , a differentiable f is L -smooth if

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq L\|y - x\|_2 \quad \text{for all } x, y \in D.$$

Lemma (First-order error bound)

Under this assumption, for every segment $[x, x + h] \subset D$,

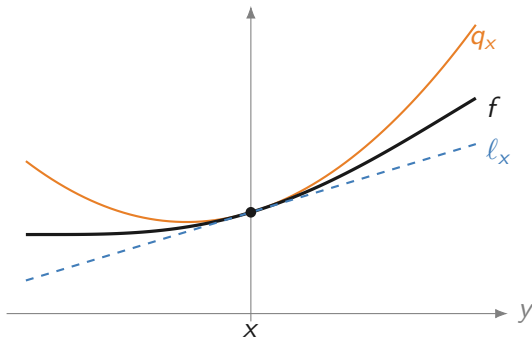
$$\left| f(x + h) - f(x) - \nabla f(x)^\top h \right| \leq \frac{L}{2} \|h\|_2^2.$$

Integrate $\longrightarrow$ Subtract $\longrightarrow$ Bound

Smoothness Gives A Quadratic Upper Model

For L -smooth f on an open convex domain, x, y in that domain:

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2} \|y - x\|_2^2.$$



Same value and slope at x

Tangent model ℓ_x

Quadratic envelope q_x

Tangent: no bound in general.

$q_x(y) \geq f(y)$ throughout the domain.

Proof: Integrate Along The Segment

Subtract The Linear Term

Fundamental Theorem Of Calculus

$$R_1(h) = \int_0^1 [\nabla f(x + th) - \nabla f(x)]^\top h \, dt.$$

Cauchy–Schwarz:

$$|R_1(h)| \leq \int_0^1 \|\nabla f(x + th) - \nabla f(x)\|_2 \|h\|_2 \, dt.$$

L -smoothness along the segment:

$$|R_1(h)| \leq \int_0^1 Lt \|h\|_2^2 \, dt = \frac{L}{2} \|h\|_2^2.$$

L -Smoothness On The Whole Segment.

The Second-Order Model

Theorem (Second-order Taylor expansion)

If $f \in C^2$ near x , then

$$f(x + h) = f(x) + \nabla f(x)^\top h + \frac{1}{2} h^\top \nabla^2 f(x) h + o(\|h\|_2^2).$$

Lipschitz Hessian: Cubic Remainder

If $\|\nabla^2 f(x + th) - \nabla^2 f(x)\|_2 \leq \rho t \|h\|_2$ on the segment,

$$|R_2(h)| \leq \frac{\rho}{6} \|h\|_2^3.$$

$$R_2(h) = \int_0^1 (1-t) h^\top [\nabla^2 f(x + th) - \nabla^2 f(x)] h dt,$$

Constant: $\int_0^1 t(1-t) dt = 1/6$.

Differentiability: $o(\|h\|)$ Lipschitz gradient: $O(\|h\|^2)$ Lipschitz Hessian: $O(\|h\|^3)$

Proof Details: The Integral Remainder

Assume $f \in C^2$ near $[x, x + h]$ and write $H(z) = \nabla^2 f(z)$. Twice applying the fundamental theorem of calculus gives

$$\begin{aligned} f(x + h) - f(x) - \nabla f(x)^\top h &= \int_0^1 [\nabla f(x + th) - \nabla f(x)]^\top h \, dt \\ &= \int_0^1 \int_0^t h^\top H(x + sh) h \, ds \, dt \\ &= \int_0^1 (1 - s) h^\top H(x + sh) h \, ds. \end{aligned}$$

The last equality integrates over $0 \leq s \leq t \leq 1$ in the opposite order. Since $\int_0^1 (1 - s) \, ds = 1/2$, subtracting the quadratic term yields

$$R_2(h) = \int_0^1 (1 - s) h^\top [H(x + sh) - H(x)] h \, ds.$$

This exact identity gives both remainder bounds on the next page.

Proof Details: Two Remainder Bounds

Continuity: Little- o Remainder

Let $\omega(h) = \sup_{0 \leq s \leq 1} \|H(x + sh) - H(x)\|_2$. Continuity of H at x implies $\omega(h) \rightarrow 0$ as $h \rightarrow 0$.

$$|R_2(h)| \leq \int_0^1 (1-s)\omega(h)\|h\|_2^2 ds = \frac{1}{2}\omega(h)\|h\|_2^2 = o(\|h\|_2^2).$$

Lipschitz Hessian: Cubic Bound

If $\|H(x + sh) - H(x)\|_2 \leq \rho s\|h\|_2$, then

$$|R_2(h)| \leq \rho\|h\|_2^3 \int_0^1 s(1-s) ds = \frac{\rho}{6}\|h\|_2^3.$$

Here $|h^\top B h| \leq \|B\|_2\|h\|_2^2$ follows from Cauchy-Schwarz and the definition of the induced norm.

PART 03

Matrix Calculus

Derivatives, Jacobians, And Gradients

For $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ differentiable at x ,

$$g(x+h) = g(x) + J_g(x)h + o(\|h\|_2), \quad [J_g]_{ij} = \frac{\partial g_i}{\partial x_j}.$$

A vector output

$$J_g(x) \in \mathbb{R}^{m \times n}, \quad dg = J_g dx.$$

For $g(x) = Ax$, the Jacobian is A .

Shape Check: $dx \in \mathbb{R}^n \mapsto dg \in \mathbb{R}^m$.

A scalar output

Column Convention

$$\nabla f \in \mathbb{R}^n, \quad J_f = \nabla f^\top.$$

For $f(x) = x^\top x$, $\nabla f = 2x$ but $J_f = 2x^\top$.

Gradients With Respect To Matrices

Definition (Matrix Gradient)

For differentiable scalar f and $W \in \mathbb{R}^{m \times n}$, define $\nabla_W f \in \mathbb{R}^{m \times n}$ by

$$df = \langle \nabla_W f, dW \rangle_F, \quad \langle A, B \rangle_F = \sum_{i,j} A_{ij} B_{ij}.$$

A practical recipe

$$df \xrightarrow{\text{isolate } dW} \langle G, dW \rangle_F \implies \nabla_W f = G.$$

Compatible Dimensions:

$$\langle G, AB \rangle_F = \langle GB^\top, A \rangle_F = \langle A^\top G, B \rangle_F.$$

Basic Matrix Derivatives

Function	Differential	Gradient
$\langle C, W \rangle_F$	$\langle C, dW \rangle_F$	C
$\frac{1}{2} \ W\ _F^2$	$\langle W, dW \rangle_F$	W
$x^\top Ax$	$(dx)^\top Ax + x^\top A dx$	$(A + A^\top)x$

The Transpose Matters

$$d(x^\top Ax) = \langle Ax, dx \rangle + \langle x, A dx \rangle = \langle (A + A^\top)x, dx \rangle.$$

Only when $A = A^\top$ can we write $\nabla(x^\top Ax) = 2Ax$.

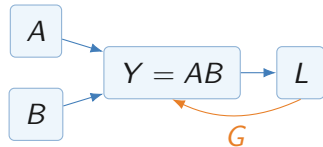
Try: for $f(x) = \frac{1}{2}x^\top Ax - b^\top x$ and symmetric A , find ∇f and $\nabla^2 f$.

$$\nabla f(x) = Ax - b, \quad \nabla^2 f(x) = A.$$

Matrix Multiplication: The Reverse Rules

$$A \in \mathbb{R}^{m \times k}, B \in \mathbb{R}^{k \times n}, Y = AB.$$

$$\text{Upstream gradient: } G = \nabla_Y L \in \mathbb{R}^{m \times n}.$$



$$dY = (dA)B + A dB.$$

Transpose Across The Inner Product

$$\begin{aligned} dL &= \langle G, (dA)B \rangle_F + \langle G, A dB \rangle_F \\ &= \langle GB^\top, dA \rangle_F + \langle A^\top G, dB \rangle_F. \end{aligned}$$

$$\nabla_A L = GB^\top \in \mathbb{R}^{m \times k}, \quad \nabla_B L = A^\top G \in \mathbb{R}^{k \times n}.$$

Classical dense products: $O(mkn)$ each (forward / reverse).

Proof Details: Reading Off Matrix Gradients

For $Y = AB$ and $G = \nabla_Y L$, linearity of the inner product gives

$$dL = \langle G, dY \rangle_F = \langle G, (dA)B \rangle_F + \langle G, A dB \rangle_F.$$

Transpose Identity

For any $H \in \mathbb{R}^{m \times k}$, expand the entrywise inner product:

$$\langle G, HB \rangle_F = \sum_{i,j,\ell} G_{ij} H_{i\ell} B_{\ell j} = \sum_{i,\ell} (GB^\top)_{i\ell} H_{i\ell} = \langle GB^\top, H \rangle_F.$$

Similarly, $\langle G, AK \rangle_F = \langle A^\top G, K \rangle_F$. Apply these identities with $H = dA$ and $K = dB$.

Compare with the definition

$$dL = \langle \nabla_A L, dA \rangle_F + \langle \nabla_B L, dB \rangle_F.$$

Because dA and dB are arbitrary, their coefficients are $\nabla_A L = GB^\top$ and $\nabla_B L = A^\top G$.

Worked Example: Matrix Least Squares

Let $X \in \mathbb{R}^{b \times d}$, $W \in \mathbb{R}^{d \times m}$ and $Y \in \mathbb{R}^{b \times m}$.

$$f(W) = \frac{1}{2} \|XW - Y\|_F^2, \quad E = XW - Y. \quad \text{Examples are rows.}$$

Differentiate The Residual

$$dE = X dW, \quad df = \langle E, X dW \rangle_F = \langle X^\top E, dW \rangle_F.$$

$$\nabla_W f = X^\top E \in \mathbb{R}^{d \times m}. \quad \nabla^2 f(W)[\Delta W] = X^\top X \Delta W.$$

Gradient Cost: $O(mdb)$.

Hessian Action: $X^\top(X\Delta W)$; no full Hessian needed.

Proof Details: Least-Squares Curvature

For $f(W) = \frac{1}{2} \|XW - Y\|_F^2$, the gradient is $\nabla f(W) = X^\top(XW - Y)$.

Differentiate The Gradient

For an arbitrary $\Delta W \in \mathbb{R}^{d \times m}$,

$$\nabla f(W + t\Delta W) = \nabla f(W) + tX^\top X\Delta W.$$

Therefore

$$\nabla^2 f(W)[\Delta W] = \left. \frac{d}{dt} \nabla f(W + t\Delta W) \right|_{t=0} = X^\top X\Delta W.$$

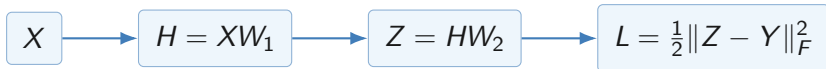
Its quadratic form is nonnegative:

$$\langle \Delta W, X^\top X\Delta W \rangle_F = \langle X\Delta W, X\Delta W \rangle_F = \|X\Delta W\|_F^2 \geq 0.$$

It is positive for every nonzero ΔW exactly when X has full column rank. Each output column has the same curvature matrix $X^\top X$.

A Two-Layer Linear Network

$X \in \mathbb{R}^{b \times d}$, $W_1 \in \mathbb{R}^{d \times h}$, $W_2 \in \mathbb{R}^{h \times c}$, $Y \in \mathbb{R}^{b \times c}$. Examples are rows.



Forward: $X \longrightarrow H \longrightarrow Z \longrightarrow L$ **Backward:** $L \longrightarrow Z \longrightarrow H \longrightarrow W_1$.

Question: To find $\nabla_{W_1} L$, what must first reach H from the loss?

Backpropagate Through The Output Layer

Residual $E = Z - Y$.

$$G_Z := \nabla_Z L = E.$$

differentiate $Z = HW_2$

$$dZ = (dH)W_2 + H dW_2, \quad dL = \langle E, dZ \rangle_F.$$

$$dL = \langle EW_2^\top, dH \rangle_F + \langle H^\top E, dW_2 \rangle_F.$$

$$\boxed{\nabla_{W_2} L = H^\top E}$$

and

$$\boxed{G_H := \nabla_H L = EW_2^\top}.$$

$G_H \longrightarrow$ First Layer

Backpropagate Through The First Layer

$$G_H = EW_2^\top, \quad H = XW_1.$$

Differentiate $H = XW_1$

$$dH = X dW_1, \quad dL = \langle G_H, X dW_1 \rangle_F = \langle X^\top G_H, dW_1 \rangle_F.$$

$$\boxed{\nabla_{W_1} L = X^\top G_H = X^\top E W_2^\top.}$$

Shape check: $(d \times b)(b \times h) = d \times h$, the shape of W_1 .

Cache $H \rightarrow$ Reuse In $\nabla_{W_2} L = H^\top E$.

PART 04

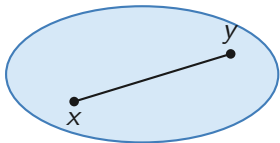
Convexity

Convex Sets

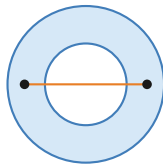
Definition (Convex Set)

A set $C \subseteq \mathbb{R}^d$ is convex if, for all $x, y \in C$ and $t \in [0, 1]$,

$$(1 - t)x + ty \in C.$$



Convex: every segment stays inside.



Nonconvex: a segment crosses the hole.

Examples: norm balls, affine sets, halfspaces, and the PSD cone.

Convex Combinations And Set Operations

Proposition

C is convex if and only if it contains every finite convex combination of its points:

$$\sum_{i=1}^r \alpha_i x_i \in C \quad \text{whenever } x_i \in C, \alpha_i \geq 0, \sum_i \alpha_i = 1.$$

Proof: induction on r ; converse: $r = 2$.

Operations preserving convexity

Arbitrary intersections.

Affine images and affine preimages.

Example: $\{x : Ax \leq b\}$ is an intersection of halfspaces.

A useful failure

Unions need not be convex:

$$[-2, -1] \cup [1, 2].$$

The midpoint of -1 and 1 is missing.



Convex Functions

Definition (Convex Function)

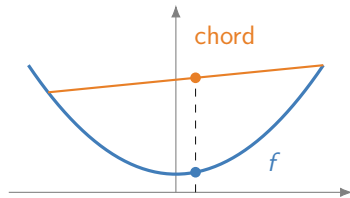
Let C be convex. A function $f : C \rightarrow \mathbb{R}$ is convex if

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y)$$

for all $x, y \in C$ and $t \in [0, 1]$.

The graph lies below each chord.

Examples: affine functions, norms, e^x , and $\frac{1}{2}x^\top Ax$ when $A \succeq 0$.



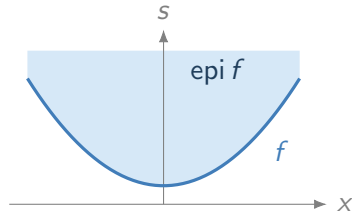
Epigraphs And Sublevel Sets

Definition (Epigraph)

$$\text{epi } f = \{(x, s) : x \in \text{dom } f, s \geq f(x)\}.$$

Proposition

f is convex if and only if $\text{epi } f$ is a convex set.



Forward: for $s \geq f(x)$ and $r \geq f(y)$,

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y) \leq (1-t)s + tr.$$

Reverse: apply epigraph convexity to $(x, f(x))$ and $(y, f(y))$.

Every sublevel set $\{x : f(x) \leq \alpha\}$ of a convex function is convex.

Converse fails: x^3 has interval sublevel sets but is not convex on $\mathbb{R}$.

Proof Details: Epigraphs And Sublevels

Let $D = \text{dom } f$ and $z = (1 - t)x + ty$, $t \in [0, 1]$.

Epigraph Equivalence

If f is convex and $(x, s), (y, r) \in \text{epi } f$, then

$$f(z) \leq (1 - t)f(x) + tf(y) \leq (1 - t)s + tr.$$

Thus $(z, (1 - t)s + tr) \in \text{epi } f$. Conversely, use the epigraph points $(x, f(x))$ and $(y, f(y))$. Their convex combination lies in the epigraph, giving the convexity inequality. The same argument ensures $z \in D$.

If $f(x), f(y) \leq \alpha$, then $f(z) \leq (1 - t)\alpha + t\alpha = \alpha$: each sublevel set is convex. For

$f(x) = x^3$, every sublevel set is $(-\infty, \sqrt[3]{\alpha}]$. But $f''(x) = 6x < 0$ for $x < 0$, so convex sublevels do not imply convexity.

First-Order Characterization Of Convexity

Theorem

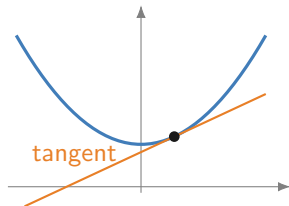
Let f be differentiable on an open convex set $D \subseteq \mathbb{R}^d$. Then f is convex if and only if

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) \quad \text{for all } x, y \in D.$$

Global Supporting Planes

⇒ Secant Limit

⇐ Combine Two Lower Bounds



Proof: Convexity Gives A Supporting Plane

$x, y \in D, \quad 0 < t \leq 1, \quad [x, y] \subset D.$

Short Secant

Convexity gives

$$f(x + t(y - x)) \leq (1 - t)f(x) + tf(y).$$

Rearrange:

$$\frac{f(x + t(y - x)) - f(x)}{t} \leq f(y) - f(x).$$

Let $t \downarrow 0$ and use differentiability:

$$\nabla f(x)^\top (y - x) \leq f(y) - f(x).$$

Local Slope $\longrightarrow$ Global Lower Bound

Proof: Supporting Planes Give Convexity

Fix $x, y \in D$, $t \in [0, 1]$, and $z = (1 - t)x + ty$.

Two Lower Bounds At z

$$\begin{aligned}f(x) &\geq f(z) + \nabla f(z)^\top (x - z), \\f(y) &\geq f(z) + \nabla f(z)^\top (y - z).\end{aligned}$$

Weighted Sum: $(1 - t) \times \text{first} + t \times \text{second}$

$$(1 - t)f(x) + tf(y) \geq f(z) + \nabla f(z)^\top \underbrace{((1 - t)x + ty - z)}_0.$$

$$f((1 - t)x + ty) \leq (1 - t)f(x) + tf(y).$$

Second-Order Characterization Of Convexity

Theorem

For $f \in C^2(D)$ on an open convex set D ,

$$f \text{ is convex} \iff \nabla^2 f(x) \succeq 0 \quad \text{for every } x \in D.$$

Restrict To Lines

$$\phi(t) = f(x + tv), \quad \phi''(t) = v^\top \nabla^2 f(x + tv) v.$$

$$\text{Convexity} \implies \frac{\phi(t+h) + \phi(t-h) - 2\phi(t)}{h^2} \geq 0 \xrightarrow{h \rightarrow 0} \phi''(t) \geq 0.$$

$$\phi'' \geq 0 \implies \phi' \text{ nondecreasing} \implies \phi(b) - \phi(a) \geq \phi'(a)(b-a), \quad a < b.$$

$$t \in \mathbb{R} \longrightarrow x + tv \longrightarrow \phi(t) = f(x + tv)$$

Every Line Segment In The Domain.

Proof Details: The Hessian Test

Assume $f \in C^2(D)$ and D is open and convex.

Convexity $\Rightarrow$ PSD Hessian

For $x \in D$, any v , and sufficiently small s , midpoint convexity gives

$$f(x + sv) + f(x - sv) - 2f(x) \geq 0.$$

Second-order Taylor makes the left side $s^2 v^\top \nabla^2 f(x) v + o(s^2)$. Divide by s^2 and let $s \rightarrow 0$. Thus $v^\top \nabla^2 f(x) v \geq 0$ for every v .

PSD Hessian $\Rightarrow$ Supporting Planes

Fix $x, y \in D$, $h = y - x$, and $\phi(t) = f(x + th)$. Since $\phi''(t) = h^\top \nabla^2 f(x + th) h \geq 0$, $\phi'(t) \geq \phi'(0)$.

$$f(y) - f(x) = \int_0^1 \phi'(t) dt \geq \phi'(0) = \nabla f(x)^\top (y - x).$$

The first-order characterization now gives convexity.

Strong Convexity

Let $f : D \rightarrow \mathbb{R}$, where $D \subseteq \mathbb{R}^d$ is convex, and let $\mu > 0$.

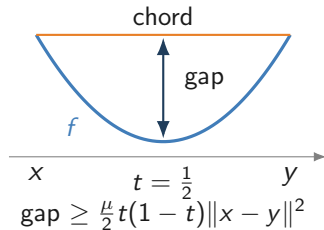
Definition (μ -strong convexity)

For every $x, y \in D$ and $t \in [0, 1]$,

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y) - \frac{\mu}{2}t(1-t)\|x - y\|_2^2.$$

Uniform Quadratic Gap

Same $\mu > 0$ throughout D .



Equivalent Definitions Of Strong Convexity

Let $\mu > 0$ and D be open and convex. These four forms are equivalent:

1 Jensen form

For every x, y and $t \in [0, 1]$,

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y) - \frac{\mu}{2}t(1-t)\|x - y\|_2^2.$$

3 First-order lower model

If f is differentiable, for every x, y ,

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|_2^2.$$

2 Subtract a quadratic

$$g(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$$

is convex on D .

4 Hessian lower bound

If $f \in C^2(D)$, for every x ,

$$\nabla^2 f(x) \succeq \mu I.$$

Proof: Apply Convexity Tests To $g = f - \frac{\mu}{2}\|\cdot\|_2^2$.

Why Subtracting A Quadratic Works

Quadratic Identity

$$\|(1-t)x + ty\|_2^2 = (1-t)\|x\|_2^2 + t\|y\|_2^2 - t(1-t)\|x - y\|_2^2.$$

For $g = f - \frac{\mu}{2}\|\cdot\|_2^2$, $z = (1-t)x + ty$,

$$\begin{aligned}(1-t)g(x) + tg(y) - g(z) \\ = (1-t)f(x) + tf(y) - f(z) - \frac{\mu}{2}t(1-t)\|x - y\|_2^2.\end{aligned}$$

For the differentiable and twice-differentiable cases,

$$\nabla g(x) = \nabla f(x) - \mu x, \quad \nabla^2 g(x) = \nabla^2 f(x) - \mu I.$$

Apply To g : supporting-plane test / PSD Hessian test.

Proof Details: Expanding The Quadratic

Put $z = (1 - t)x + ty$ and $g(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$. Expanding the square shows

$$\begin{aligned}\|z\|_2^2 &= (1 - t)^2\|x\|_2^2 + 2t(1 - t)\langle x, y \rangle + t^2\|y\|_2^2, \\ (1 - t)\|x\|_2^2 + t\|y\|_2^2 - \|z\|_2^2 &= t(1 - t)\|x - y\|_2^2.\end{aligned}$$

Convexity Of g

$$f(z) - \frac{\mu}{2}\|z\|_2^2 \leq (1 - t)\left[f(x) - \frac{\mu}{2}\|x\|_2^2\right] + t\left[f(y) - \frac{\mu}{2}\|y\|_2^2\right].$$

Move the quadratic terms to the right and use the identity:

$$f(z) \leq (1 - t)f(x) + tf(y) - \frac{\mu}{2}t(1 - t)\|x - y\|_2^2.$$

Every step is reversible, proving the equivalence.

Proof Details: The Lower Model And Hessian

For differentiable f , set $h = y - x$ and $g(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$. The supporting-plane inequality for g is

$$f(y) - \frac{\mu}{2}\|y\|_2^2 \geq f(x) - \frac{\mu}{2}\|x\|_2^2 + (\nabla f(x) - \mu x)^\top h.$$

Collect Quadratic Terms

Since $\|x + h\|_2^2 - \|x\|_2^2 - 2x^\top h = \|h\|_2^2$,

$$f(y) \geq f(x) + \nabla f(x)^\top h + \frac{\mu}{2}\|h\|_2^2.$$

The algebra is reversible, so this lower model is equivalent to convexity of g .

If $f \in C^2(D)$, then

$$g \text{ convex} \iff \nabla^2 g(x) \succeq 0 \iff \nabla^2 f(x) \succeq \mu I \quad (x \in D).$$

Both characterizations use an open convex domain and the stated differentiability.

Upper And Lower Quadratic Models

For $f \in C^2(D)$ on open convex D , take $x, y \in D$ and write $h = y - x$,
 $\ell_x(h) = f(x) + \nabla f(x)^\top h$.

L -smoothness: an upper model

$\|\nabla^2 f(x)\|_2 \leq L$ for every x , equivalently
 $-LI \preceq \nabla^2 f(x) \preceq LI$.

$$f(x+h) \leq \ell_x(h) + \frac{L}{2} \|h\|_2^2.$$

Combined Bounds

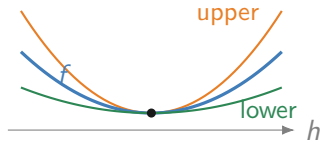
$$\ell_x(h) + \frac{\mu}{2} \|h\|_2^2 \leq f(x+h) \leq \ell_x(h) + \frac{L}{2} \|h\|_2^2.$$

Same value and slope at $h = 0$.

μ -strong convexity: a lower model

$\nabla^2 f(x) \succeq \mu I$ for every x :
uniform positive curvature.

$$f(x+h) \geq \ell_x(h) + \frac{\mu}{2} \|h\|_2^2.$$



Example: Ridge Regression Curvature

$$f(x) = \frac{1}{2} \|Ax - b\|_2^2 + \frac{\lambda}{2} \|x\|_2^2, \quad \lambda \geq 0.$$

$$\nabla^2 f = A^\top A + \lambda I.$$

$$L = \lambda_{\max}(A^\top A) + \lambda,$$

$$\mu = \lambda_{\min}(A^\top A) + \lambda.$$

$$\lambda > 0 \implies \mu > 0.$$

Ridge shifts every eigenvalue by λ .

For $A = \text{diag}(1, 0)$:

	L	μ
$\lambda = 0$	1	0
$\lambda > 0$	$1 + \lambda$	λ

PART 05

Optimality

Global And Local Minimizers

Consider $\min_{x \in C} f(x)$, and let $x^* \in C$.

Definition (Global And Local Minimizers)

Global minimizer: $f(x^*) \leq f(x)$ for every $x \in C$.

Local minimizer: there exists $r > 0$ such that

$$f(x^*) \leq f(x) \quad \text{for every } x \in C \cap B_r(x^*).$$

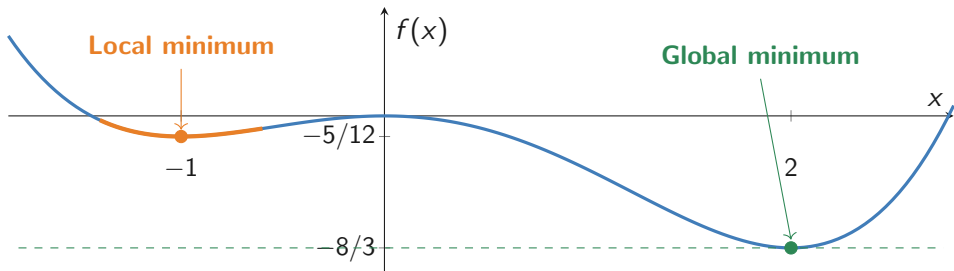
It is **strict** if the inequality is strict whenever $x \neq x^*$ in that neighborhood.

$$f^* = \inf_{x \in C} f(x), \quad X^* = \{x \in C : f(x) = f^*\}.$$

Existence is separate: $f(x) = e^x$ on $\mathbb{R}$ has $f^* = 0$ but $X^* = \emptyset$.

A Local Minimum Need Not Be Global

$$f(x) = \frac{x^4}{4} - \frac{x^3}{3} - x^2, \quad x \in \mathbb{R}.$$



At $x = -1$: compare nearby points.

A strict local minimum, but $f(2) < f(-1)$.

At $x = 2$: compare every point.

A global minimum is also a local minimum.

Stationary Points And Saddles

For a differentiable function on an open domain:

Definition (Stationary And Saddle Points)

A **stationary point** satisfies $\nabla f(x) = 0$.

Here, a **saddle point** is a stationary point with both larger and smaller function values in every neighborhood.

A quadratic saddle

$f(x, y) = x^2 - y^2$ at $(0, 0)$:

$$\nabla f = 0, \quad \nabla^2 f = \text{diag}(2, -2).$$

Positive and negative curvature coexist.

A flat saddle

$f(x, y) = x^4 - y^4$ at $(0, 0)$:

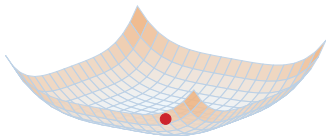
$$\nabla f = 0, \quad \nabla^2 f = 0.$$

The Hessian alone does not detect it.

A Zero Hessian Is Inconclusive

At the origin, all three functions have $\nabla f(0,0) = 0$ and $\nabla^2 f(0,0) = 0$.

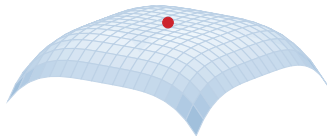
$$f(x,y) = x^4 + y^4$$



Strict local minimum

$f(x,y) > 0$ at every nonzero point.

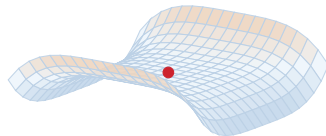
$$f(x,y) = -x^4 - y^4$$



Strict local maximum

$f(x,y) < 0$ at every nonzero point.

$$f(x,y) = x^4 - y^4$$



Flat saddle point

$f(t,0) = t^4 > 0$;
 $f(0,t) = -t^4 < 0$ for $t \neq 0$.

A zero Hessian leaves higher-order terms to decide.

Necessary And Sufficient Local Conditions

At an unconstrained interior point x^* , with $f \in C^2$ nearby:

Theorem (Necessary)

x^* is a local minimizer



$$\nabla f(x^*) = 0$$

$$\nabla^2 f(x^*) \succeq 0$$

Theorem (Sufficient)

$$\nabla f(x^*) = 0, \quad \nabla^2 f(x^*) \succ 0$$



x^* is a strict local minimizer

At Zero: x^4 : strict minimum; $-x^4$: strict maximum.

Both have $f'(0) = f''(0) = 0$: PSD insufficient; PD unnecessary.

Proof Sketch: Test Small Displacements

Necessity: Test Every Direction

x^* local minimum $\implies \phi(t) = f(x^* + tv)$ minimized at 0.

$$\phi'(0) = \nabla f(x^*)^\top v = 0, \quad \phi''(0) = v^\top \nabla^2 f(x^*) v \geq 0 \quad \forall v.$$

Sufficiency: Quadratic Dominates Remainder

Let $m = \lambda_{\min}(\nabla^2 f(x^*)) > 0$. Taylor's formula gives

$$f(x^* + h) - f(x^*) \geq \frac{m}{2} \|h\|_2^2 + o(\|h\|_2^2) \geq \frac{m}{4} \|h\|_2^2 > 0$$

for all sufficiently small nonzero h .

Proof Details: Controlling The Local Remainder

Suppose $\nabla f(x^*) = 0$ and $H = \nabla^2 f(x^*) \succ 0$. Let $m = \lambda_{\min}(H) > 0$.

Uniform Neighborhood

Taylor's theorem gives

$$f(x^* + h) - f(x^*) = \frac{1}{2}h^\top Hh + r(h), \quad \frac{r(h)}{\|h\|_2^2} \longrightarrow 0.$$

Hence there is $\delta > 0$ such that $|r(h)| \leq \frac{m}{4}\|h\|_2^2$ whenever $0 < \|h\|_2 < \delta$.

Because $h^\top Hh \geq m\|h\|_2^2$,

$$f(x^* + h) - f(x^*) \geq \frac{m}{2}\|h\|_2^2 - \frac{m}{4}\|h\|_2^2 = \frac{m}{4}\|h\|_2^2 > 0.$$

The same δ works for every direction. This proves a strict local minimum, not merely positivity along each fixed line.

What Convexity Adds To Optimality

Theorem

For a convex function on a convex feasible set:

- Every local minimizer is a **global minimizer**.
- Strict convexity implies **at most one** global minimizer.

On an open convex domain, if f is differentiable,

$$\nabla f(x^*) = 0 \iff x^* \text{ is a global minimizer.}$$

Contradiction: if $f(y) < f(x^*)$, then

$$f((1-t)x^* + ty) \leq (1-t)f(x^*) + tf(y) < f(x^*)$$

$0 < t \leq 1$; arbitrarily small t contradict local minimality.

Strict Convexity: two minima $\implies$ strictly better midpoint.

Existence Requires A Separate Argument.

Proof Details: Convex Optimality

Local $\Rightarrow$ Global

If x^* is local but $f(y) < f(x^*)$, choose small $t > 0$. Then $z_t = (1 - t)x^* + ty$ lies in its local neighborhood, while

$$f(z_t) \leq (1 - t)f(x^*) + tf(y) < f(x^*),$$

a contradiction. Convexity of the domain keeps z_t feasible.

Uniqueness And Stationarity

For strictly convex f , two distinct minimizers x, y would satisfy

$$f((x + y)/2) < \frac{1}{2}f(x) + \frac{1}{2}f(y) = f^*,$$

which is impossible. For differentiable convex f on an open domain, $\nabla f(x^*) = 0$ makes the supporting-plane inequality $f(y) \geq f(x^*)$. Conversely, a global minimizer is local; the unconstrained first-order necessary condition gives $\nabla f(x^*) = 0$.

PART 06

Convergence

What Error Are We Measuring?

Metric	What it measures
$f(x_k) - f^*$	objective suboptimality
$\text{dist}(x_k, X^*)$	distance to the solution set, if nonempty
$\ \nabla f(x_k)\ _2$	unconstrained first-order stationarity
$\ \nabla f(x_k)\ _2^2$	a different accuracy convention

Stationarity $\neq$ Global Optimality

Output: last / best / averaged / random iterate.

Randomized Guarantees: expectation / high probability.

Sublinear And Linear Convergence

Let $e_k \geq 0$ denote a **specified** error measure.

Sublinear bounds

$$e_k \leq \frac{C}{(k+1)^p}, \quad p > 0.$$

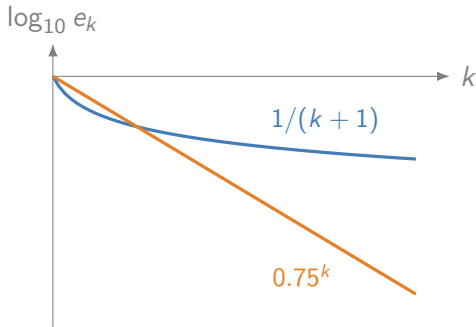
Examples: $O(k^{-1/2})$, $O(k^{-1})$, $O(k^{-2})$.

Geometric / R-linear bound

$$e_k \leq Cq^k, \quad 0 < q < 1.$$

Q-linear recurrence: $e_{k+1} \leq qe_k$.

Geometric Decay: A Straight Line On A Log-Error Plot.



From Rates To Total Work

Target: $e_k \leq \varepsilon$.

Bound	Sufficient count
-------	------------------

$$e_k \leq C/(k+1) \quad k+1 \geq C/\varepsilon$$

$$e_k \leq C/(k+1)^2 \quad k+1 \geq \sqrt{C/\varepsilon}$$

$$e_k \leq Cq^k \quad k \geq \frac{\log(C/\varepsilon)}{\log(1/q)}$$

In the last row, $0 < q < 1$ and $0 < \varepsilon < C$.

Second method: $r > 0$ times the per-iteration cost.

Exercise: compare total work ε^{-1} and $r\varepsilon^{-1/2}$.

Resolution: $r\varepsilon^{-1/2} < \varepsilon^{-1} \iff r < \varepsilon^{-1/2}$.

Big- O upper bounds alone do not rank actual runtimes.

Total work

Number of iterations

× work per iteration.

Gradients / samples / arithmetic /
memory / time.

PART 07



Probability

Random Variables And Expectation

Random Variable: random outcome $\mapsto$ numerical value.

Definition (Expectation)

If X takes values x_i with probabilities p_i , where $p_i \geq 0$ and $\sum_i p_i = 1$, then

$$\mathbb{E}[X] = \sum_i p_i x_i.$$

For countably many values, assume $\sum_i p_i |x_i| < \infty$.

Linearity

For integrable X, Y and constants a, b ,

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

No Independence Required. Vectors: coordinatewise expectation.

A Sampled Gradient Is Unbiased

For differentiable component losses, define

$$F(\theta) = \frac{1}{n} \sum_{i=1}^n f_i(\theta), \quad I \sim \text{Unif}\{1, \dots, n\}.$$

Fixed: θ . **Random:** I .

Average Over The Index

$$\mathbb{E}[\nabla f_I(\theta)] = \frac{1}{n} \sum_{i=1}^n \nabla f_i(\theta) = \nabla F(\theta).$$

Unbiased = Correct In Expectation.

Variance And Second Moments

Let X be scalar and G vector-valued, both with finite second moments.

Definition (Scalar And Total Vector Variance)

Scalar variance

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Total variance of vector G

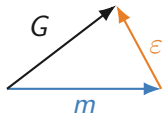
$$\mathbb{E}\|G - \mathbb{E}[G]\|_2^2.$$

Lemma (Second-Moment Decomposition)

$$\mathbb{E}\|G\|_2^2 = \|\mathbb{E}[G]\|_2^2 + \mathbb{E}\|G - \mathbb{E}[G]\|_2^2.$$

$$G = m + \varepsilon, \quad m = \mathbb{E}[G].$$

$$\mathbb{E}[\varepsilon] = 0 \implies \mathbb{E}\langle m, \varepsilon \rangle = 0.$$



Proof Details: Mean And Second Moment

Linearity: Joint Probabilities

For finite-valued X, Y , let $p_{ij} = \Pr(X = x_i, Y = y_j)$. Then

$$\mathbb{E}[aX + bY] = \sum_{i,j} p_{ij}(ax_i + by_j) = a \sum_i x_i \sum_j p_{ij} + b \sum_j y_j \sum_i p_{ij}.$$

The inner sums are the marginal probabilities. No factorization of p_{ij} is used.

Zero-Mean Cross Term

For a vector G with finite second moment, put $m = \mathbb{E}[G]$ and $\varepsilon = G - m$.

$$\begin{aligned}\mathbb{E}\|G\|_2^2 &= \mathbb{E}[\|m\|_2^2 + 2\langle m, \varepsilon \rangle + \|\varepsilon\|_2^2] \\ &= \|m\|_2^2 + 2\langle m, \mathbb{E}[\varepsilon] \rangle + \mathbb{E}\|\varepsilon\|_2^2 \\ &= \|\mathbb{E}[G]\|_2^2 + \mathbb{E}\|G - \mathbb{E}[G]\|_2^2.\end{aligned}$$

Conditional Expectation

$\mathbb{E}[Y \mid X]$: average with X known; a function of X .

Lemma (Conditional Expectation Rules)

Assume Y and $a(X)Y$ integrable.

Tower Property

$$\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y].$$

Known Factor

$$\mathbb{E}[a(X)Y \mid X] = a(X)\mathbb{E}[Y \mid X].$$

Fix Known Information $\longrightarrow$ Average Remaining Randomness.

Proof Details: Conditioning In A Finite Model

Let $p_{ij} = \Pr(X = x_i, Y = y_j)$ and $p_i = \sum_j p_{ij} > 0$. Within the group $X = x_i$, the weights are p_{ij}/p_i , so

$$\mathbb{E}[Y \mid X = x_i] = \sum_j y_j \frac{p_{ij}}{p_i}.$$

Tower Property

$$\mathbb{E}[\mathbb{E}[Y \mid X]] = \sum_i p_i \sum_j y_j \frac{p_{ij}}{p_i} = \sum_{i,j} p_{ij} y_j = \mathbb{E}[Y].$$

Known Factor

$$\mathbb{E}[a(X)Y \mid X = x_i] = \sum_j a(x_i) y_j \frac{p_{ij}}{p_i} = a(x_i) \mathbb{E}[Y \mid X = x_i].$$

Zero-probability groups contribute nothing.

Conditioning On The History In SGD

$$x_{k+1} = x_k - \eta_k g_k.$$

History $\mathcal{F}_k$: x_k and past samples, before the k th fresh sample.

Across runs: x_k is random

Different samples $\longrightarrow$ different iterates.

Given $\mathcal{F}_k$: x_k is fixed

Average over a fresh, independent sample.

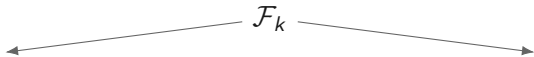
Assumption (Conditional Unbiasedness)

$$\mathbb{E}[g_k \mid \mathcal{F}_k] = \nabla F(x_k).$$

Probability Tools For An SGD Proof

Assume finite second moments and conditional unbiasedness.

Condition on $\mathcal{F}_k$: x_k and $\nabla F(x_k)$ are fixed.



Conditional unbiasedness

$$\mathbb{E}[g_k \mid \mathcal{F}_k] = \nabla F(x_k).$$

Known factor rule $\Downarrow$

$$\begin{aligned}\mathbb{E}[\langle \nabla F(x_k), g_k \rangle \mid \mathcal{F}_k] \\ = \|\nabla F(x_k)\|_2^2.\end{aligned}$$

Bounded conditional variance

$$\mathbb{E}[\|g_k - \nabla F(x_k)\|_2^2 \mid \mathcal{F}_k] \leq \sigma^2.$$

Signal plus noise $\Downarrow$

$$\begin{aligned}\mathbb{E}[\|g_k\|_2^2 \mid \mathcal{F}_k] \\ \leq \|\nabla F(x_k)\|_2^2 + \sigma^2.\end{aligned}$$

Tower Property: Average Over Histories.

Proof Details: Conditional Gradient Identities

Write $m_k = \nabla F(x_k)$ and $\varepsilon_k = g_k - m_k$. Given $\mathcal{F}_k$, m_k is known and conditional unbiasedness gives $\mathbb{E}[\varepsilon_k \mid \mathcal{F}_k] = 0$.

Linear Term: Known Vector

$$\mathbb{E}[\langle m_k, g_k \rangle \mid \mathcal{F}_k] = \langle m_k, \mathbb{E}[g_k \mid \mathcal{F}_k] \rangle = \|m_k\|_2^2.$$

Quadratic Term: Conditional Mean

$$\begin{aligned}\mathbb{E}[\|g_k\|_2^2 \mid \mathcal{F}_k] &= \|m_k\|_2^2 + 2\langle m_k, \mathbb{E}[\varepsilon_k \mid \mathcal{F}_k] \rangle + \mathbb{E}[\|\varepsilon_k\|_2^2 \mid \mathcal{F}_k] \\ &\leq \|m_k\|_2^2 + \sigma^2.\end{aligned}$$

The last step uses the bounded conditional variance assumption. These conditional identities hold almost surely; the tower rule later averages them over histories.