AIE6003: Optimization for Machine Learning

Xiaopeng Li
CUHK(SZ)
SAI

February 4, 2026

Outline

Introduction

Convexity

Optimality Conditions

Projection

Smoothness

Introduction to myself

My name: Xiaopeng Li

Position: Assistant Professor, School of Artificial Intelligence, CUHK(SZ)

Website: `https://williampixell.github.io`

Education background:

- Ph.D. in Operations Research, Columbia University.
- B.S. in Applied Mathematics, The Chinese University of Hong Kong, Shenzhen.

Research focus: nonconvex optimization theory and algorithms with applications in machine learning; large language model reasoning.

How to address me: You can call me 'Xiaopeng' or 'Dr. Li'.

Contact and Office Hours

Instructor: Xiaopeng Li
Email: `xiaopengli1@cuhk.edu.cn`
Lecture: 10:30am–11:50am, Monday/Wednesday, TA107
Office hour: 2pm–3pm, Wednesday, TA409c

Teaching Assistant: Yiheng Wang
Email: `224040341@link.cuhk.edu.cn`
Office hour: 2:00pm-3:00pm, Tuesday, Letian Building, 1st floor
Role: The TA will grade assignments and exams, host office hours and provide support on questions related to course content.

Communication:
Please email with a clear subject line and include `AIE6003` in the subject. We aim to respond within 48 hours on weekdays.

Who is this course for?

This course is designed for **first-year MPhil/PhD students**.

Industry track:
Preparing for high-impact careers in AI/ML industry roles (e.g., ML Researcher, Data Scientist, Quantitative Researcher).

Research track:
Develop rigorous understanding of optimization theory and algorithms to read and write research papers.

Prerequisites:
Calculus and linear algebra; mathematical analysis; basic probability and statistics.

Why a strong mathematical foundation?

Optimization is the language behind ML training.
A strong foundation helps you choose algorithms, tune/debug training, and understand guarantees and failure modes.

For master students:
Training stability (step size, batch size), efficiency at scale (SGD/momentum/Adam), constraints/robustness (projection, regularization, adversarial training).

For PhD students:
Additionally, read and verify proofs, design new methods, publish in top venue (MP/SIOPT/MOR/JMLR/COLT/NeurIPS, etc.).

Course Agenda

Goal:

- Learn optimization tools that are scalable, stable, and theoretically grounded.
- Understand the connection between them and machine learning problems.

How this course is organized:

- **Preliminary:** convexity, optimality, projection, smoothness
- **Deterministic methods:** gradient descent, acceleration, proximal method, mirror descent;
- **Stochastic methods:** stochastic gradient descent, momentum, adaptive gradients (AdaGrad/Adam);
- **Extra topics:** duality, ADMM, low-rank, variance reduction, minimax, sparse, black-box.

Assessment Scheme

Component	Weight
Homework	20%
Midterm exam	30%
Final exam	50%
Total	100%

Notes:

- Homework emphasizes derivations, proofs, and implementation.
- Exams focus on core concepts and key techniques.
- Exams will be much easier than homework.
- Homework content will appear in the exams, so the effective weight of homework is higher than its nominal percentage (20%).

Tentative Weekly Plan

Week (Dates)	Content
1 (Jan 05/07)	introduction, basic concepts (projection, optimality condition, convexity, smoothness)
2 (Jan 12/14)	gradient descent and conjugate gradient method
3 (Jan 19/21)	heavy ball and Nesterov's accelerated gradient descent
4 (Jan 26/28)	proximal operator, proximal gradient descent
5 (Feb 02/04)	subgradient methods and mirror descent
6 (Mar 02/04)	stochastic gradient descent
7 (Mar 09/11)	stochastic gradient descent with momentum, adaptive (stochastic) gradient methods.
8 (Mar 16/18)	midterm exam
9 (Mar 23/25)	duality, ALM, ADMM
10 (Mar 30/Apr 01)	low-rank optimization (including conditional gradient methods)
11 (Apr 08/13)	empirical risk minimization (including variance reduction)
12 (Apr 15/20)	minimax optimization (including extragradient methods)
13 (Apr 22/27)	sparse optimization (including coordinate descent)
14 (Apr 29/May 6)	black-box optimization with zeroth-order methods.

Outline

Introduction

Convexity

Optimality Conditions

Projection

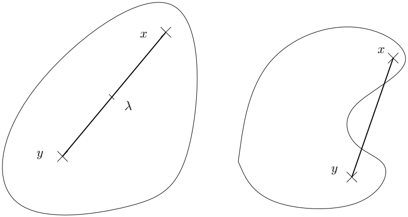
Smoothness

Convex Sets

Definition: The set C is convex if and only if, for all $x, y \in C$ and all $\lambda \in [0, 1]$, we have

$$\lambda x + (1 - \lambda) y \in C$$

Geometrically, the set C contains the line segment $[x, y]$.



Convex set

Non-convex set

Convex sets in ML:

1. Probability simplex: $\Delta = \{\alpha \in \mathbb{R}^K \mid \alpha \geq 0,\ \mathbf{1}^\top \alpha = 1\}$,
   e.g., Large Language Model (LLM); Mixture-of-Experts (MoE).

2. Norm ball: $\{w \in \mathbb{R}^d \mid \|w\|_p \leq R\}$,
   e.g., Adversarial training; Dictionary learning.

3. PSD cone: $\{K \in \mathbb{S}^n \mid K \succeq 0\}$,
   e.g., Kernel learning; Clustering.

Convex Sets

Useful properties:

1. Intersection: if C_i is convex for $i = 1, \ldots, n$, then $\cap_{i=1}^n C_i$ is convex.

2. Affine image: if C is convex, then $\{Ax + b : \ x \in C\}$ is convex.

3. Affine preimage: if C is convex, then $\{x : \ Ax + b \in C\}$ is convex.

4. Minkowski sum: if C, D are convex, then $C + D$ is convex, where we define $C + D := \{c + d : \ c \in C, \ d \in D\}$.

What is the use of these properties?

- They let us verify convexity of complicated feasible regions by building them from simple pieces.

- Constraint sets are usually intersections or images of basic convex sets: halfspaces, norm bounds, simplex constraints, etc.

Convex Functions

Definition: A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is convex if and only if, for all x, y and all $\lambda \in [0, 1]$, we have

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y).$$

Convex functions in ML:

1. Ridge regression:

$$\min_{w \in \mathbb{R}^d} \ \frac{1}{2n} \|Xw - y\|_2^2 + \frac{\lambda}{2} \|w\|_2^2.$$

2. Logistic regression:

$$\min_{w \in \mathbb{R}^d} \ \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i x_i^\top w)\right) + \frac{\lambda}{2} \|w\|_2^2.$$

3. Support Vector Machine:

$$\min_{w,b} \ \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{n} \max\{0,\ 1 - y_i(w^\top x_i + b)\}.$$

Convex Functions

Useful properties:

1. Nonnegative sum: if f_i convex and $\alpha_i \geq 0$, then $\sum_i \alpha_i f_i$ is convex.
2. Affine composition: if f convex, then $x \mapsto f(Ax + b)$ is convex.
3. Pointwise maximum: if f_i convex, then $\max_i f_i$ is convex.
4. Expectation: if $\ell(\theta; z)$ is convex in θ, then $\mathbb{E}_z[\ell(\theta; z)]$ is convex.

What is the use of these properties?

- They let us verify convexity of complicated objectives by building them from simple convex functions (sum, expectation, max, affine composition).
- Many ML losses are exactly such compositions: ERM is an average of per-sample losses; regularization adds a convex penalty; hinge loss is a max; softmax cross-entropy uses log-sum-exp.

Epigraph Characterization

Epigraph: $\mathrm{epi}(f) := \{(x, \alpha) \in \mathbb{R}^d \times \mathbb{R} : f(x) \leq \alpha\}$.

Epigraph Characterization

A function f is convex if and only if $\mathrm{epi}(f)$ is convex.

Proof sketch:

- " $\Longrightarrow$ ": take any $(x_1, \alpha_1), (x_2, \alpha_2) \in \mathrm{epi}(f)$ so $f(x_1) \leq \alpha_1$, $f(x_2) \leq \alpha_2$; use convexity of f to show
$$f(\lambda x_1 + (1-\lambda)x_2) \ \leq \ \lambda\alpha_1 + (1-\lambda)\alpha_2,$$
hence $\lambda(x_1, \alpha_1) + (1-\lambda)(x_2, \alpha_2) \in \mathrm{epi}(f)$.

- " $\Longleftarrow$ ": for any x_1, x_2, note $(x_1, f(x_1)), (x_2, f(x_2)) \in \mathrm{epi}(f)$; convexity of $\mathrm{epi}(f)$ implies
$$(\lambda x_1 + (1-\lambda)x_2,\ \lambda f(x_1) + (1-\lambda)f(x_2)) \in \mathrm{epi}(f),$$
which means $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$.

First-Order Characterization

Gradient Inequality

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable. Then

$$f \text{ is convex } \iff f(y) \geq f(x) + \nabla f(x)^{\mathrm{T}}(y - x), \quad \forall x, y.$$

Proof sketch:

- " $\implies$ ": apply convexity to x and $x + t(y - x)$, divide by t, and take $t \downarrow 0$. This shows the tangent hyperplane at x is a global under-estimator.
- " $\impliedby$ ": use the inequality twice (swap x, y) along the line segment $z_\lambda = \lambda x + (1 - \lambda)y$ to recover Jensen's inequality.

Convex Optimization

Optimization problem:

$$\begin{aligned} \min_{x\in\mathbb{R}^d} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \le 0, \quad i = 1, \dots, m; \\ & h_j(x) = 0, \quad j = 1, \dots, p. \end{aligned}$$

Jargon:

- Decision variable: x (e.g., parameters, weights, latent variables).
- Objective: f (e.g., training loss).
- Constraints: $g_i(x) \le 0$ (inequality), $h_j(x) = 0$ (equality).
- Feasible point & region: $x \in \mathcal{X}$, where
  $\mathcal{X} := \{x \in \mathbb{R}^d : g_i(x) \le 0,\ i = 1, \dots, m;\ h_j(x) = 0,\ j = 1, \dots, p\}$.
- Convex optimization: both f and $\mathcal{X}$ are convex.

Outline

Introduction

Convexity

Optimality Conditions

Projection

Smoothness

Optimality Conditions: Unconstrained

Solution concepts: Suppose f is differentiable.

- Global minimum: $f(x^\star) \leq f(x)$ for all $x \in \mathbb{R}^d$.
- Local minimum: $f(x^\star) \leq f(x)$ in a neighborhood of $x^\star$.
- Critical point: $\nabla f(x^\star) = 0$.

Optimality conditions: Suppose f is differentiable.

- First-order necessary: if $x^\star$ is a local minimum, then $\nabla f(x^\star) = 0$.
- Second-order necessary: if $f \in C^2$ and $x^\star$ is a local minimum, then Hessian matrix $\nabla^2 f(x^\star) \succeq 0$.
- Second-order sufficient: if $\nabla f(x^\star) = 0$ and $\nabla^2 f(x^\star) \succ 0$, then $x^\star$ is a strict local minimum.

Optimality Conditions: Constrained

Constrained problem: $\min_{x \in \mathcal{X}} f(x)$, where $\mathcal{X}$ is a closed convex set.

Variational Characterization

If f is differentiable and convex, then $x^\star$ is optimal if and only if

$$\langle \nabla f(x^\star), x - x^\star \rangle \geq 0, \qquad \forall x \in \mathcal{X}.$$

Geometric meaning: At $x^\star$, there is no feasible descent direction.

Proof sketch of " $\Longleftarrow$ ":

- A direction d is feasible at $x^\star$ if $x^\star + td \in \mathcal{X}$ for all small $t > 0$.
- Plug $x = x^\star + td$ into the inequality:
$$\langle \nabla f(x^\star), (x^\star + td) - x^\star \rangle = t \langle \nabla f(x^\star), d \rangle \geq 0.$$
- Divide by $t > 0$ to get $\langle \nabla f(x^\star), d \rangle \geq 0$, i.e., the directional derivative $f'(x^\star; d) = \nabla f(x^\star)^\top d$ is nonnegative for every feasible d.

Convex Optimization

Why do we care convex optimization in ML?

- No spurious local minima: every local minimum is a global minimum.
- Verifiable optimality: first-order conditions become sufficient.
- Algorithmic guarantees: convergence behaviors are clearest in the convex case (will see in later lectures).
- Many classical ML problems are convex (ridge, logistic, SVM).

Stationary point global optimum

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable. Then

$$\nabla f(x^\star) = 0 \iff x^\star \in \arg\min_{x \in \mathbb{R}^d} f(x).$$

Outline

Introduction

Convexity

Optimality Conditions

Projection

Smoothness

Projection: Basics

Projection: the projection of a point x onto a set $\mathcal{X}$ is

$$\Pi_{\mathcal{X}}(x) := \arg\min_{y \in \mathcal{X}} \tfrac{1}{2}\|y - x\|^2.$$

i.e., a point $y \in \mathcal{X}$ such that

$$\|y - x\| \;\leq\; \|z - x\|, \quad \forall z \in \mathcal{X}.$$

Caveat: in general, a projection may

- not exist: e.g., $D^2 = \{z \in \mathbb{R}^2 : \; \|z\| < 1\}$ with $x \in \mathbb{R}^2 \setminus D^2$.
- exist but not unique: e.g., $S^1 = \{z \in \mathbb{R}^2 : \; \|z\| = 1\}$ with $x = 0$.

Lession:

- Existence: $\mathcal{X}$ needs to be closed.
- Uniqueness: $\mathcal{X}$ needs to be convex.

Projection: Existence

Existence of Projection

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be nonempty and closed. Then for every $x \in \mathbb{R}^d$, the projection exists, i.e., the optimization problem below has a solution:

$$\Pi_{\mathcal{X}}(x) := \arg\min_{y \in \mathcal{X}} \ \tfrac{1}{2}\|y - x\|^2.$$

Proof sketch: Continuous functions attain minima on compact sets.

- Pick any $x_0 \in \mathcal{X}$ and set $r := \|x - x_0\|$.
- Any minimizer must lie in $\mathcal{X}' := \mathcal{X} \cap \overline{B_r(x)}$.
- $\mathcal{X}'$ is closed and bounded, hence compact.
- Continuous map $y \mapsto \|y - x\|^2$ attains its minimum on $\mathcal{X}'$.

Projection: Variational Characterization

Variational Characterization of Projection

Let $\mathcal{X}$ be closed and convex. Then

$$y = \Pi_{\mathcal{X}}(x) \iff \langle z - y, x - y \rangle \leq 0, \quad \forall z \in \mathcal{X}.$$

Proof sketch:

- By definition, $y = \Pi_{\mathcal{X}}(x)$ solves the convex problem
$$\min_{u \in \mathcal{X}} f(u), \text{ where } f(u) := \frac{1}{2}\|u - x\|^2.$$
- Since f is differentiable and convex and $\mathcal{X}$ is closed convex, the constrained optimality condition gives
$$\langle \nabla f(y), z - y \rangle \geq 0, \quad \forall z \in \mathcal{X}.$$
- Compute $\nabla f(y) = y - x$, hence
$$\langle y - x, z - y \rangle \geq 0 \iff \langle z - y, x - y \rangle \leq 0, \quad \forall z \in X.$$

Projection: Uniqueness on Closed Convex Sets

Uniqueness of Projection

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be nonempty, closed and convex. Then for every $x \in \mathbb{R}^d$, the projection $\Pi_{\mathcal{X}}(x)$ is unique.

Proof sketch: apply the variational characterization twice.

- Let $y_1, y_2 \in \Pi_{\mathcal{X}}(x)$. Using the variational characterization with $z = y_2$ and $z = y_1$ gives
$$\langle y_2 - y_1,\ x - y_1 \rangle \leq 0, \quad \langle y_1 - y_2,\ x - y_2 \rangle \leq 0.$$
- Adding yields
$$\|y_1 - y_2\|^2 = \langle x_1 - x_2,\ (y - x_2) - (y - x_1) \rangle \leq 0,$$
hence $y_1 = y_2$.

Projection: Basic Inequalities

From now on, we will assume constraints $\mathcal{X}$ are always closed and convex, unless stated otherwise.

Firm Nonexpansiveness

For any $x_1, x_2 \in \mathbb{R}^d$,

$$\|\Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2)\|^2 \leq \langle \Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2), x_1 - x_2 \rangle.$$

Corollary

For any $x_1, x_2 \in \mathbb{R}^d$,

1. Monotonicity:

$$\langle \Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2),\ x_1 - x_2 \rangle \geq 0.$$

2. Nonexpansiveness (1-Lipschitz):

$$\|\Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2)\| \leq \|x_1 - x_2\|.$$

Outline

Introduction

Convexity

Optimality Conditions

Projection

Smoothness

Smoothness: Lipschitz Continuity

L-Lipschitz: A function $f : \mathcal{X} \to \mathbb{R}$ is L-Lipschitz if

$$|f(x) - f(y)| \leq L\|x - y\|, \quad \forall\, x, y \in \mathcal{X}.$$

Interpretation: f cannot change too drastically. L measures how drastic f can change.

Lipschitz $\implies$ bounded gradients

If $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz and differentiable, then

$$\|\nabla f(x)\| \leq L, \quad \forall\, x \in \mathbb{R}^n.$$

Proof sketch:

- Use directional derivative: $\nabla f(x)^\top v = \lim_{t\to 0} \frac{f(x+tv)-f(x)}{t}$.
- Apply Lipschitz definition to get $|\nabla f(x)^\top v| \leq L\|v\|$.
- Take supremum over $v \neq 0$ to obtain $\|\nabla f(x)\| \leq L$.

Smoothness: Lipschitz Gradient

ℓ-smooth: A differentiable function $f : X \to \mathbb{R}$ is ℓ-smooth if

$$\|\nabla f(x) - \nabla f(y)\| \leq \ell \|x - y\|, \quad \forall x, y \in \mathcal{X}.$$

Interpretation: ∇f cannot change too fast (bounded curvature). ℓ measures how drastic ∇f can change.

Quadratic Taylor Bound

If $f : \mathbb{R}^n \to \mathbb{R}$ is ℓ-smooth, then for all $x, y \in \mathbb{R}^n$,

$$|f(y) - f(x) - \nabla f(x)^{\mathrm{T}}(y - x)| \leq \frac{\ell}{2}\|y - x\|^2.$$

Proof sketch:

- Let $g(t) = f(x + t(y - x))$ and write $g(1) - g(0) = \int_0^1 g'(s)\, ds$.
- Use $g'(s) = \nabla f(x + s(y - x))^{\mathrm{T}}(y - x)$ and ℓ-smoothness.

Smoothness: Lipschitz Gradient

Refined Gradient Inequality

If $f : \mathbb{R}^n \to \mathbb{R}$ is ℓ-smooth and convex, then for all $x, y \in \mathbb{R}^n$,

$$f(x) - f(y) \leq \nabla f(x)^{\mathrm{T}}(x - y) - \frac{1}{2\ell}\|\nabla f(x) - \nabla f(y)\|^2.$$

Proof sketch: Insert $f(z)$

- Use convexity to relate $f(x) - f(z)$ to $\nabla f(x)^{\mathrm{T}}(x - z)$.
- Use the quadratic Taylor bound to estimate $f(z) - f(y)$.
- Choose a specific $z := y - \frac{1}{\ell}(\nabla f(y) - \nabla f(x))$.

Co-coercivity of ∇f

If $f : \mathbb{R}^n \to \mathbb{R}$ is ℓ-smooth and convex, then for all $x, y \in \mathbb{R}^n$,

$$(\nabla f(x) - \nabla f(y))^{\mathrm{T}}(x - y) \geq \frac{1}{\ell}\|\nabla f(x) - \nabla f(y)\|^2.$$

Wrap-up and Next Week

Week 1: Key takeaways

- Convex sets & functions: global guarantees.
- Optimality conditions: local/global; unconstrained/constrained.
- Projection: existence & uniqueness; nonexpansiveness.
- Smoothness: quadratic Taylor bound & co-coercivity.

Thank you for your attendance!

Questions?

Next week (Week 2): Gradient Descent & Conjugate Gradient