Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-556 (Fall 2014)

License Information for Mathematics of Data Slides

- This lecture
  1. Learning as an optimization problem
  2. Basic concepts in convex analysis
  3. Three important classes of convex functions
- Next lecture
  1. Optimality conditions
  2. Unconstrained convex minimization
  3. Convergence and convergence rate characterization of methods for unconstrained minimization

Recommended reading

- V. N. Vapnik, “An overview of statistical learning theory,” *IEEE Trans. Inf. Theory*, vol. 10, no. 5, pp. 988–999, Sep. 1999.
- *Chapter 5 in A. W. van der Vaart, *Asymptotic Statistics*, Cambridge Univ. Press, 1998.
- Chapter 2 & 3 in Boyd, Stephen, and Lieven Vandenberghe, *Convex optimization*, Cambridge Univ. Press, 2009.
- Appendices A & B in Bertsekas, Dimitris, *Nonlinear Programming*, Athena Scientific, 1999.
- Chapter 4 in Nesterov, Yurii, *Introductory Lectures on Convex Optimization: A Basic Course*, Vol. 87, Springer, 2004.

Motivation

Motivation

This lecture explains how convex optimization problems naturally arise in data analytics* and feature important properties useful for efficiently obtaining numerical solutions with provable certificates of quality.

- Several important data models lead to convex optimization problems whose solutions have guarantees.
- Convex analysis offer key structures that will help us construct efficient numerical solution methods.

*discovery and communication of meaningful patterns and information in data.

Learning as an optimization problem

Problem

Information in data can be elusive. When we want to extract information form data, we typically have to solve an optimization problem of the following form:

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x} \in \mathcal{X}} \{F(\mathbf{x})\}$$

with some constraints $\mathcal{X} \subseteq \mathbb{R}^p$.

Remark

The seemingly simple optimization formulation above, of course, has applications well beyond learning in many diverse disciplines.

Example 1: Least-squares estimation

Problem

Let $\mathbf{x}^\natural \in \mathbb{R}^p$. Let $\mathbf{A} \in \mathbb{R}^{n\times p}$ with full column rank. How do we estimate $\mathbf{x}^\natural$ given $\mathbf{A}$ and

$$\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w},$$

where $\mathbf{w}$ denotes some unknown noise (either random or deterministic)?

Solution (Least-squares estimator)

$$\hat{\mathbf{x}}_{LS} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{F(\mathbf{x})\right\},$$

with

$$F(\mathbf{x}) := \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 .$$

Example 2: Maximum-likelihood estimation with the linear model

Problem (Gaussian linear model)

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ be an unknown vector and $\mathbf{A} \in \mathbb{R}^{n \times p}$ be a matrix with full column rank. How do we estimate $\mathbf{x}^\natural$ given $\mathbf{A}$ and

$$\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w},$$

where $\mathbf{w}$ is a sample of a Gaussian random vector $\sim \mathcal{N}(\mathbf{0}, \mathbf{I})$?

Solution (Maximum-likelihood estimator)

The maximum-likelihood (ML) estimator in the Gaussian linear model is given by

$$\hat{\mathbf{x}}_{ML} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ F(\mathbf{x}) \right\},$$

with

$$F(\mathbf{x}) := -\sum_{i=1}^{n} \ln \left\{ \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(b_i - \langle \mathbf{a}_i, \mathbf{x} \rangle \right)^2 \right] \right\},$$

where b_i is the i-th entry of $\mathbf{b}$, and $\mathbf{a}_i$ is the i-th row of $\mathbf{A}$.

We may observe that $\hat{\mathbf{x}}_{\mathrm{LS}}$ is equivalent to the $\hat{\mathbf{x}}_{\mathrm{ML}}$ given above.

Example 3: ML estimation in general

Problem (General estimation problem)

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ an unknown vector. Let b_i be a sample of a random variable B_i with unknown probability density function $p_i(b_i; \mathbf{x}^\natural)$ in $\mathcal{P}_i := \{p_i(b_i; \mathbf{x}) : \mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^p\}$. How do we estimate $\mathbf{x}^\natural$ given $\mathcal{P}_1, \ldots, \mathcal{P}_n$ and $b_1, \ldots, b_n$?

Remark

This formulation is essentially equivalent to the formulation $b_i = f^\natural(\mathbf{a}_i) + w_i$ in Lecture 0. Let w_i be the realization of a random variable W with zero mean. Define $B_i := f^\natural(\mathbf{a}_i) + W$ with $f^\natural(\mathbf{a}_i) := \mathbb{E}[B_i]$, and let $p_i(b_i; \mathbf{x}^\natural)$ denote the probability density at b_i given $\mathbf{a}_i$ and $\mathbf{x}^\natural$. Then we obtain the formulation above.

Solution (ML estimator for the general estimation problem)

$$\hat{\mathbf{x}}_{ML} \in \arg\min_{\mathbf{x} \in \mathcal{X}} \{F(\mathbf{x})\},$$

with

$$F(\mathbf{x}) := -\frac{1}{n} \sum_{i=1}^{n} \ln[p_i(b_i; \mathbf{x})].$$

Real application: Poisson imaging

Problem (Poisson observations)

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ be an unknown vector. Let $b_1, \ldots, b_n$ be samples of independent random variables $B_1, \ldots, B_n$, and each B_i is Poisson distributed with parameter $\left\langle \mathbf{a}_i, \mathbf{x}^\natural \right\rangle$, where the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_i$ are given. How do we estimate $\mathbf{x}^\natural$ given $\mathbf{a}_1, \ldots, \mathbf{a}_n$ and the measurement outcomes $b_1, \ldots, b_n$?

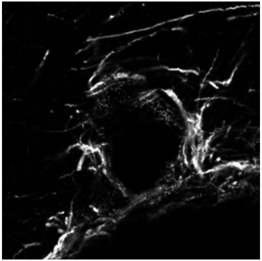
Solution (ML estimator)

We may consider the ML estimator

$$\hat{\mathbf{x}}_{ML} \in \arg\min_{\mathbf{x} \in \mathbb{R}^p} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[\langle \mathbf{a}_i, \mathbf{x} \rangle - b_i \ln\left(\langle \mathbf{a}_i, \mathbf{x} \rangle \right) \right] \right\}.$$

Remark

In confocal imaging, the linear vectors $\mathbf{a}_i$ can be used to capture the lens effects, including blur and (spatial) low-pass filtering (due to the so-called numerical aperture of the lens).



Confocal imaging

Example 4: M-estimators

Problem (General estimation problem)

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ *an unknown vector. Let* b_i *be a sample of a random variable* B_i *with unknown probability density function* $p_i(b_i; \mathbf{x}^\natural)$ *in* $\mathcal{P}_i := \{p_i(b_i; \mathbf{x}) : \mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^p\}$*. How do we estimate* $\mathbf{x}^\natural$ *given* $\mathcal{P}_1, \ldots, \mathcal{P}_n$ *and* $b_1, \ldots, b_n$*?*

Solution (M-estimator)

In general we can replace the negative log-likelihoods by any appropriate functions f_i*, and obtain an* M*-estimator*

$$\hat{\mathbf{x}}_M \in \arg\min_{x \in \mathcal{X}} \{F(\mathbf{x})\},$$

with

$$F(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}; b_i).$$

- When f_i are chosen to be the negative log-likelihoods, the M-estimator is equivalent to the maximum-likelihood estimator.
- The term "M-estimator" denotes "maximum-likelihood-type estimator," as it is a generalization of ML estimators [1].

A machine learning application: Graphical model learning

Problem (Graphical model selection)

Let $\mathbf{x}$ *be a random vector with zero mean and positive-definite covariance matrix* $\mathbf{\Sigma}^\natural$*. How do we estimate* $\mathbf{\Theta}^\natural := {\mathbf{\Sigma}^\natural}^{-1}$ *given independent samples* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *of the random vector* $\mathbf{x}$*?*

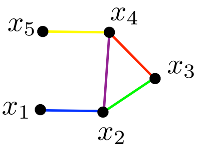
Solution (M-estimator)

We may consider the M*-estimator*

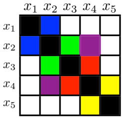
$$\widehat{\mathbf{\Theta}}_M \in \arg\min_{\mathbf{\Theta} \in \mathbb{S}^p_{++}} \left\{ \mathrm{Tr}\left(\widehat{\mathbf{\Sigma}}\mathbf{\Theta}\right) - \log\det\left(\mathbf{\Theta}\right) \right\},$$

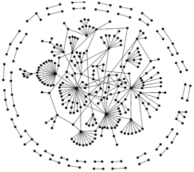
where $\widehat{\mathbf{\Sigma}}$ *is the empirical covariance, i.e.,* $\widehat{\mathbf{\Sigma}} := (1/n)\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T$*.*

This is equivalent to the ML estimator only when $\mathbf{x}$ is Gaussian distributed.



$\Theta =$



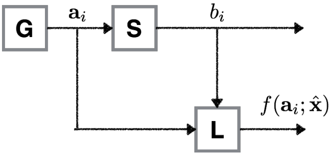


*Example 5: Statistical learning and empirical risk minimization principle

Statistical learning problem [2]

A statistical learning problem consists of three elements.

1. A *generator* that produces samples $\mathbf{a}_i \in \mathbb{R}^p$ of a random variable $\mathbf{a}$ with an unknown probability distribution $\mathbb{P}_{\mathbf{a}}$.
2. A *supervisor* that for each $\mathbf{a}_i \in \mathbb{R}$, generates a sample b_i of a random variable B with an unknown conditional probability distribution $\mathbb{P}_{B|\mathbf{a}}$.
3. A *learning machine* that can respond as any function f of $\mathbf{a}_i$ in the set $\{f_{\mathbf{x}}(\mathbf{a}_i) : \mathbf{x} \in \mathcal{X}\}$ with some fixed $\mathcal{X} \subseteq \mathbb{R}^p$.



*Example 5: Statistical learning and empirical risk minimization principle

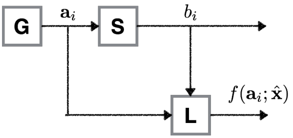
Goal

Choose an $\hat{\mathbf{x}} \in \mathcal{X}$ such that the *risk* $R(\mathbf{x}) := \mathbb{E}\left[\mathcal{L}(B, f_{\mathbf{x}}(\mathbf{a}))\right]$ is minimized for a given loss function $\mathcal{L}$, where the expectation is taken with respect to the joint distribution of $\mathbf{a}$ and B.

Empirical risk minimization (ERM) principle [2]

The risk $R(\mathbf{x})$ is not tractable since we do not know $\mathbb{P}_{\mathbf{a}}$ and $\mathbb{P}_{B|\mathbf{a}}$. But given samples $(\mathbf{a}_i, b_i)$, we can minimize the *empirical risk* $\hat{R}(\mathbf{x})$ as an approximation of $R(\mathbf{x})$, i.e., we can choose

$$\hat{\mathbf{x}}_{\text{ERM}} \in \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \hat{R}(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(b_i, f_{\mathbf{x}}(\mathbf{a}_i)) \right\}.$$



A machine learning application: Pattern classification

Pattern classification by separating hyperplanes

The samples the *generator* produce are given by $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^p$. The samples the *supervisor* generates are given by $b_1, \ldots, b_n \in \{\pm 1\}$. The set of functions that the *learning machine* can implement is given by $\{f_{\mathbf{x}}(\mathbf{a}) := \mathrm{sign}\left(\langle \mathbf{x}, \mathbf{a} \rangle\right) : \mathbf{x} \in \mathcal{X}\}$ with some fixed set $\mathcal{X} \subseteq \mathbb{R}^p$. The *loss function* $\mathcal{L}$ is defined as

$$\mathcal{L}(b_i, f_{\mathbf{x}}(\mathbf{a}_i)) := (b_i - f_{\mathbf{x}}(\mathbf{a}_i))^2 \,.$$

Applying the ERM principle

The corresponding $\hat{\mathbf{x}}_{\mathsf{ERM}}$ is given by

$$\hat{\mathbf{x}}_{\mathsf{ERM}} \in \arg\min_{\mathbf{x} \in \mathcal{X}} \left\{ \hat{R}(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^{n} (b_i - f_{\mathbf{x}}(\mathbf{a}_i))^2 : \mathbf{x} \in \mathcal{X} \right\}.$$

- This stylized method does not apply well in a lot of applications, but it inspires advanced pattern classification algorithms such as the neural network and the support vector machine [2].

Checking the fidelity

Now that we have an estimator $\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathcal{X}} \{F(\mathbf{x})\}$, we need to address two key questions:

1. Is the formulation reasonable?
2. What is the role of the data size?

Standard approach to checking the fidelity

Standard approach

1. Specify a performance criterion $\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}^\natural)$ that should be small if $\hat{\mathbf{x}} = \mathbf{x}^\natural$.
2. Show that $\mathcal{L}$ is actually *small in some sense* when *some condition* is satisfied.

Example

Take the ℓ_2-error $\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}^\natural) := \left\|\hat{\mathbf{x}} - \mathbf{x}^\natural\right\|_2^2$ as an example. Then we may verify the fidelity via one of the following ways, where ε denotes a small enough number:

1. $\mathbb{E}\left[\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}^\natural))\right] \leq \varepsilon$ (expected error),
2. $\mathbb{P}\left(\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}^\natural) > t\right) \leq \varepsilon$ for any $t > 0$ (consistency),
3. $\sqrt{n}(\hat{\mathbf{x}} - \mathbf{x}^\natural)$ converges in distribution to $\mathcal{N}(0, \mathbf{I})$ (asymptotic normality),
4. $\sqrt{n}(\hat{\mathbf{x}} - \mathbf{x}^\natural)$ converges in distribution to $\mathcal{N}(0, \mathbf{I})$ in a local neighborhood (local asymptotic normality).

if *some condition* is satisfied. Such conditions typically revolve around the data size.

Approach 1: Expected error

Gaussian linear model

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ and let $\mathbf{A} \in \mathbb{R}^{n\times p}$. The samples are given by $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ is a sample of a Gaussian random vector $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$.

What is the performance of the ML estimator

$$\hat{\mathbf{x}}_{\mathsf{ML}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 \right\}?$$

Theorem (Performance of the LS estimator [3])

If $\mathbf{A}$ is a matrix of independent and identically distributed (i.i.d.) standard Gaussian distributed entries, and if $n > p+1$, then

$$\mathbb{E}\left[\left\|\hat{\mathbf{x}}_{ML} - \mathbf{x}^\natural\right\|_2^2\right] = \frac{p}{n-p-1}\sigma^2 \to 0 \text{ as } \frac{n}{p} \to \infty.$$

Approach 2: Consistency

Covariance estimation

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be samples of a sub-Gaussian random vector with zero mean and some unknown positive-definite covariance matrix $\mathbf{\Sigma}^\natural \in \mathbb{R}^{p \times p}$. (Sub-Gaussian random variables will be defined in recitation.)

What is the performance of the M-estimator $\widehat{\mathbf{\Sigma}} := \widehat{\mathbf{\Theta}}^{-1}$, where

$$\widehat{\mathbf{\Theta}}_{\mathsf{ML}} \in \arg\min_{\mathbf{\Theta} \in \mathbb{S}^p_{++}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[-\log\det\left(\mathbf{\Theta}\right) + \mathbf{x}_i^T \mathbf{\Theta} \mathbf{x}_i \right] \right\}?$$

- If $\mathbf{y} = f(\mathbf{x})$, then $\hat{\mathbf{y}}_{\mathsf{ML}} = f(\hat{\mathbf{x}}_{\mathsf{ML}})$. This is called the *functional invariance* property of ML estimators.

Theorem (Performance of the ML estimator [4])

Suppose that the diagonal elements of $\mathbf{\Sigma}^\natural$ are bounded above by $\kappa > 0$, and each $X_i / \sqrt{\left(\Sigma^\natural\right)_{i,i}}$ is sub-Gaussian with parameter c. Then

$$\mathbb{P}\left(\left\{ \left| \left(\widehat{\mathbf{\Sigma}}_{ML}\right)_{i,j} - \left(\mathbf{\Sigma}^\natural\right)_{i,j} \right| > t \right\} \right) \leq 4 \exp\left[-\frac{nt^2}{128\left(1 + 4c^2\right)\kappa^2} \right] \to 0 \text{ as } n \to \infty$$

for all $t \in \left(0, 8\kappa\left(1 + 4c^2\right)\right)$.

We will actually prove this result in a later recitation.

*Approach 3: Asymptotic normality

Logistic regression

Let $\mathbf{x}^\natural \in \mathbb{R}^p$, and let $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{R}^p$. Let $b_1, \ldots, b_n$ be samples of independent random variables $B_1, \ldots, B_n$. Each random variable B_i takes values in $\{-1, 1\}$ and follows $\mathbb{P}\left(\{B_i = 1\}\right) := \ell_i(\mathbf{x}^\natural) = \left[1 + \exp\left(-\left\langle \mathbf{a}_i, \mathbf{x}^\natural \right\rangle\right)\right]^{-1}$ (i.e., the logistics loss).

What is the performance of the ML estimator

$$\hat{\mathbf{x}}_{\mathsf{ML}} \in \arg\min_{\mathbf{x} \in \mathbb{R}^p} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ln\left[\mathbb{I}_{\{B_i = 1\}} \ell_i(\mathbf{x}) + \mathbb{I}_{\{B_i = 0\}}\left(1 - \ell_i(\mathbf{x})\right)\right] := -\frac{1}{n} f_n(\mathbf{x}) \right\}?$$

*Approach 3: Asymptotic normality

Theorem (Performance of the ML estimator [5] (*also valid for generalized linear models))

The random variable $\mathbf{J}(\mathbf{x}^\natural)^{-1/2}\left(\hat{\mathbf{x}}_{ML} - \mathbf{x}^\natural\right)$ *converges in distribution to* $\mathcal{N}(\mathbf{0}, \mathbf{I})$ *if* $\lambda_{\min}(\mathbf{J}(\mathbf{x}^\natural)) \to \infty$ *and*

$$\max_{\mathbf{x}\in\mathbb{R}^p} \left\{ \left\| \mathbf{J}(\mathbf{x}^\natural)^{-1/2}\mathbf{J}(\mathbf{x})\mathbf{J}(\mathbf{x}^\natural)^{-1/2} - \mathbf{I} \right\|_{2\to 2} : \left\| \mathbf{J}(\mathbf{x}^\natural)^{1/2}\left(\mathbf{x} - \mathbf{x}^\natural\right) \right\|_2 \leq \delta \right\} \to 0 \quad (1)$$

for all $\delta > 0$ *as* $n \to \infty$, *where* $\mathbf{J}(\mathbf{x}) := -\mathbb{E}\left[\nabla^2 f_n(\mathbf{x})\right]$ *is the Fisher information matrix.*

Roughly speaking, assuming that p is fixed, we have the following observations.

1. The technical condition (1) means that $\mathbf{J}(\mathbf{x}) \sim \mathbf{J}(\mathbf{x}^\natural)$ for all $\mathbf{x}$ in a neighborhood $N_{\mathbf{x}^\natural}(\delta)$ of $\mathbf{x}^\natural$, and $N_{\mathbf{x}^\natural}(\delta)$ becomes larger with increasing n.
2. $\left\| \mathbf{J}(\mathbf{x}^\natural)^{-1/2}\left(\hat{\mathbf{x}}_{\mathsf{ML}} - \mathbf{x}^\natural\right) \right\|_2^2 \sim \mathrm{Tr}\left(\mathbf{I}\right) = p$, which means that $\left\| \hat{\mathbf{x}}_{\mathsf{ML}} - \mathbf{x}^\natural \right\|_2^2$ decreases at the rate $\lambda_{\min}(\mathbf{J}(\mathbf{x}^\natural))^{-1} \to 0$ asymptotically.

*Approach 4: Local asymptotic normality

In general, the asymptotic normality does not hold even in the independent identically distributed (i.i.d.) case, but we may have the *local asymptotic normality (LAN)*.

ML estimation with i.i.d. samples

Let $b_1, \ldots, b_n$ be independent samples of a random variable B, whose probability density function is known to be in the set $\{p_{\mathbf{x}}(b) : \mathbf{x} \in \mathcal{X}\}$ with some $\mathcal{X} \subseteq \mathbb{R}^p$.

What is the performance of the ML estimator

$$\hat{\mathbf{x}}_{\text{ML}} \in \arg\min_{\mathbf{x} \in \mathcal{X}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \ln \left[p_{\mathbf{x}}(b_i) \right] \right\}?$$

*Approach 4: Local asymptotic normality

Theorem (Performance of the ML estimator (cf. [6, 7] for details))

Under some technical conditions, the random variable $\sqrt{n}\,\mathbf{J}^{-1/2}\left(\hat{\mathbf{x}}_{ML} - \mathbf{x}^{\natural}\right)$ *converges in distribution to* $\mathcal{N}(\mathbf{0}, \mathbf{I})$*, where* $\mathbf{J}$ *is the Fisher information matrix associated with one sample, i.e.,*

$$\mathbf{J} := -\mathbb{E}\left[\nabla^2_{\mathbf{x}} \ln\left[p_{\mathbf{x}}(B)\right]\right]\Big|_{\mathbf{x}=\mathbf{x}^{\natural}}.$$

Roughly speaking, assuming that p is fixed, we can observe that

- $\left\|\sqrt{n}\,\mathbf{J}^{-1/2}\left(\hat{\mathbf{x}}_{\mathsf{ML}} - \mathbf{x}^{\natural}\right)\right\|_2^2 \sim \mathrm{Tr}\left(\mathbf{I}\right) = p$,
- $\left\|\hat{\mathbf{x}}_{\mathsf{ML}} - \mathbf{x}^{\natural}\right\|_2^2 = \mathcal{O}(1/n)$.

Example: ML estimation for quantum tomography

Problem (Quantum tomography)

A quantum system of q qubits can be characterized by a density operator, i.e., a Hermitian positive semidefinite $\mathbf{X}^\natural \in \mathbb{C}^{p \times p}$ with $p = 2^q$. Let $\{\mathbf{A}_1, \ldots, \mathbf{A}_m\} \subseteq \mathbb{C}^{p \times p}$ be a probability operator-valued measure, i.e., a set of Hermitian positive semidefinite matrices summing to $\mathbf{I}$. Let $b_1, \ldots, b_n$ be samples of independent random variables $B_1, \ldots, B_n$, with probability distribution

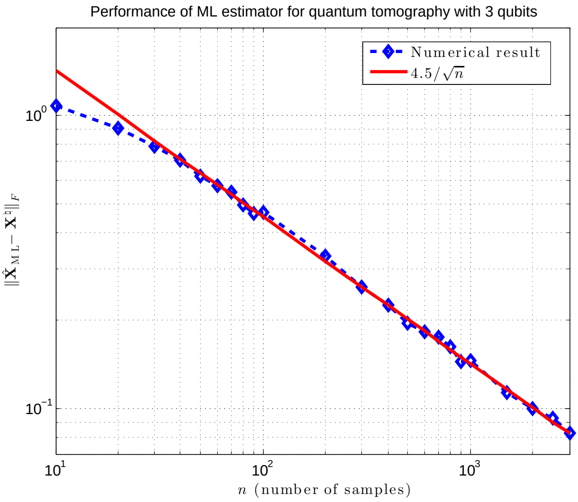
$$\mathbb{P}\left(\{b_i = k\}\right) = \operatorname{Tr}\left(\mathbf{A}_k \mathbf{X}^\natural\right), \quad k = 1, \ldots, m$$

How do we estimate $\mathbf{X}^\natural$ given $\{\mathbf{A}_1, \ldots, \mathbf{A}_m\}$ and $b_1, \ldots, b_n$?

ML approach

$$\hat{\mathbf{X}}_{\mathsf{ML}} \in \arg\min_{\mathbf{X} \in \mathbb{C}^{p \times p}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{m} \mathbb{I}_{\{b_i = k\}} \ln\left[\operatorname{Tr}\left(\mathbf{A}_k \mathbf{X}\right)\right] : \mathbf{X} = \mathbf{X}^H, \mathbf{X} \succeq \mathbf{0} \right\}.$$

Example: ML estimation for quantum tomography



Caveat Emptor

The ML estimator does not always yield the optimal performance. We show a simple yet very powerful example below.

Problem

Let $\mathbf{b}$ *be a sample of a Gaussian random vector* $\mathbf{b} \sim \mathcal{N}(\mathbf{x}^\natural, \mathbf{I})$ *with some* $\mathbf{x}^\natural \in \mathbb{R}^p$*. How do we estimate* $\mathbf{x}^\natural$ *given* $\mathbf{b}$*?*

ML approach

The ML estimator is given by $\hat{\mathbf{x}}_{\text{ML}} := \mathbf{b}$.

James-Stein estimator [8]

The James-Stein estimator is given by

$$\hat{\mathbf{x}}_{\text{JS}} := \left(1 - \frac{p-2}{\|\mathbf{b}\|_2^2}\right)_+ \mathbf{b},$$

for all $p \geq 3$, where $(a)_+ = \max(a, 0)$.

Observation: The James-Stein estimator *shrinks* $\mathbf{b}$ towards the origin.

Caveat Emptor

Theorem (Performance comparison: ML vs. James-Stein [8])

For all $\mathbf{x}^\natural \in \mathbb{R}^p$ *with* $p \geq 3$*, we have*

$$\mathbb{E}\left[\left\|\hat{\mathbf{x}}_{JS} - \mathbf{x}^\natural\right\|_2^2\right] < \mathbb{E}\left[\left\|\hat{\mathbf{x}}_{ML} - \mathbf{x}^\natural\right\|_2^2\right].$$

Performance of the ML estimator is uniformly dominated by the performance of the James-Stein estimator [8].

Important take home message

The ML approach is not always the best.

Caveat Emptor

Theorem (Performance comparison: ML vs. James-Stein [8])

For all $\mathbf{x}^\natural \in \mathbb{R}^p$ *with* $p \geq 3$, *we have*

$$\mathbb{E}\left[\left\|\hat{\mathbf{x}}_{JS} - \mathbf{x}^\natural\right\|_2^2\right] < \mathbb{E}\left[\left\|\hat{\mathbf{x}}_{ML} - \mathbf{x}^\natural\right\|_2^2\right].$$

Performance of the ML estimator is uniformly dominated by the performance of the James-Stein estimator [8].

Important take home message

The ML approach is not always the best.

Remark

The James-Stein estimator inspires the study of *shrinkage estimators* and the use of *oracle inequalities*, which play important roles in contemporary statistics and machine learning [9].

*Minimax performance

In previous slides we focused how good an estimator is. Now we would like to derive a *fundamental limitation* on the statistical performance, posed by the statistical model.

Definition (Minimax risk)

For a given loss function $\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}^\natural)$ and the associated risk function $R(\hat{\mathbf{x}}, \mathbf{x}^\natural) := \mathbb{E}\left[\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}^\natural)\right]$, the minimax risk is defined as

$$R_{minmax} := \min_{\hat{\mathbf{x}}} \max_{\mathbf{x}^\natural \in \mathcal{X}} \left\{ R(\hat{\mathbf{x}}, \mathbf{x}^\natural) \right\},$$

where $\mathcal{X}$ denotes the parameter space.

A game theoretic interpretation:

- Consider a statistician playing a game with Nature.
- Nature is malicious, i.e., Nature prefers *high* risk, while the statistician prefers *low* risk.
- Nature chooses an $\mathbf{x}^\natural \in \mathcal{X}$, and the statistician designs an estimator $\hat{\mathbf{x}}$.
- The best the statistician can choose is the *minimax strategy*, i.e., the estimator $\hat{\mathbf{x}}_{\text{minmax}}$ such that it minimizes the worst-case risk.
- The resulting worst-case risk is the *minimax risk*.

*An information theoretic approach

We choose $R(\hat{\mathbf{x}}, \mathbf{x}^\natural) := \left\| \hat{\mathbf{x}} - \mathbf{x}^\natural \right\|_2$ to illustrate the idea. Generalizations can be found in [10, 11].

There are two key concepts.

*First step: transformation to a multiple hypothesis testing problem

Let $\mathcal{X}_{\text{finite}}$ be a finite subset of the original parameter space $\mathcal{X}$. Then we have

$$R_{\text{minmax}} := \min_{\hat{\mathbf{x}}} \max_{\mathbf{x}^\natural \in \mathcal{X}} \left\{ R(\hat{\mathbf{x}}, \mathbf{x}^\natural) \right\} \geq \min_{\hat{\mathbf{x}} \in \mathcal{X}_{\text{finite}}} \max_{\mathbf{x} \in \mathcal{X}_{\text{finite}}} \left\{ R(\hat{\mathbf{x}}, \mathbf{x}^\natural) \right\},$$

*Second step: randomizing the problem

Let $\mathbb{P}$ be a probability distribution on $\mathcal{X}_{\text{finite}}$, and suppose that $\mathbf{x}^\natural$ is selected randomly following $\mathbb{P}$. Then we have

$$\min_{\hat{\mathbf{x}} \in \mathcal{X}_{\text{finite}}} \max_{\mathbf{x} \in \mathcal{X}_{\text{finite}}} \left\{ R(\hat{\mathbf{x}}, \mathbf{x}^\natural) \right\} \geq \min_{\hat{\mathbf{x}} \in \mathcal{X}_{\text{finite}}} \left\{ \mathbb{E}_{\mathbb{P}} \left[R(\hat{\mathbf{x}}, \mathbf{x}^\natural) \right] \right\}.$$

*An information theoretic approach contd.

Suppose we choose the subset $\mathcal{X}_{\text{finite}}$ such that for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}_{\text{finite}}$, $\mathbf{x} \neq \mathbf{y}$,

$$\|\mathbf{x} - \mathbf{y}\|_2 \geq d_{\min}$$

with some $d_{\min} > 0$. Then we have

$$R_{\text{minmax}} \geq \min_{\hat{\mathbf{x}} \in \mathcal{X}_{\text{finite}}} \left\{ \mathbb{E}_{\mathbb{P}} \left[R(\hat{\mathbf{x}}, \mathbf{x}^\natural) \right] \right\} \geq \frac{1}{2} d_{\min} \mathbb{P} \left(\hat{\mathbf{x}} \neq x^\natural \right).$$

What remains is to bound the probability of error, $\mathbb{P}\left(\hat{\mathbf{x}} \neq x^\natural\right)$.

*An information theoretic approach contd.

A very useful tool from information theory is Fano's inequality.

Theorem (Fano's inequality)

Let X and Y be two random variables taking values in the same finite set $\mathcal{X}$. Then

$$H(X|Y) \leq h(\mathbb{P}(X \neq Y)) + \mathbb{P}(X \neq Y) \ln(|\mathcal{X}| - 1),$$

where $H(X|Y)$ denotes the conditional entropy of X given Y, defined as

$$H(X|Y) := \mathbb{E}_{X,Y}[-\ln(\mathbb{P}(X|Y))],$$

and

$$h(x) := -x \ln x - (1-x)\ln(1-x) \leq \ln 2$$

for any $x \in [0,1]$.

Applying Fano's inequality to our problem with some simplifications, we obtain the following fundamental limit.

Corollary

$$\mathbb{P}\left(\hat{\mathbf{x}} \neq \mathbf{x}^\natural\right) \geq \frac{1}{|\mathcal{X}_{\textit{finite}}|}\left(H(\mathbf{x}^\natural|\hat{\mathbf{x}}) - \ln 2\right).$$

*An information theoretic approach contd.

Theorem ([11])

If there exists a finite subset $\mathcal{X}_{finite}$ *of the parameter space* $\mathcal{X}$ *such that for any* $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}_{finite}$, $\mathbf{x}_1 \neq \mathbf{x}_2$,

$$\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \geq d_{\min}$$

with some $d_{\min} > 0$ *and*[1]

$$D(\mathbb{P}_{\mathbf{x}_1} \| \mathbb{P}_{\mathbf{x}_2}) := \int \ln\left(\frac{d\mathbb{P}_{\mathbf{x}_1}}{d\mathbb{P}_{\mathbf{x}_2}}\right) d\mathbb{P}_{\mathbf{x}_1} \leq r$$

with some $r > 0$, *where* $\mathbb{P}_{\mathbf{x}}$ *denotes the probability distribution of the observations when* $\mathbf{x}^\natural = \mathbf{x}$ *for any* $\mathbf{x} \in \mathcal{X}_{finite}$. *Then*

$$R_{minmax} \geq \frac{d_{\min}}{2}\left(1 - \frac{r + \ln 2}{\ln |\mathcal{X}_{finite}|}\right).$$

Proof.

Combine the results in previous slides, and take $\mathbb{P}_{\text{finite}}$ to be the uniform distribution on $\mathcal{X}_{\text{finite}}$. □

[1] The function $D(\mathbb{P}\|\mathbb{Q})$ is called the Kullback-Leibler divergence or the relative entropy between probability distributions $\mathbb{P}$ and $\mathbb{Q}$.

*Example

Problem (Gaussian linear regression on the ℓ_1-ball)

Let $\mathbf{A} \in \mathbb{R}^{n\times p}$ and let $\mathbf{x}^\natural \in \mathbb{R}^p$. Define $\mathbf{y} := \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ with some $\sigma > 0$. It is known that $\mathbf{x}^\natural \in \mathcal{X} := \left\{\mathbf{x} : \|\mathbf{x}\|_1 \leq R\right\}$. What is the minimax risk R_{minmax} with respect to $R(\hat{\mathbf{x}}, \mathbf{x}^\natural) := \mathbb{E}\left[\left\|\hat{\mathbf{x}} - \mathbf{x}^\natural\right\|_2\right]$?

Theorem ([12])

Suppose the ℓ_2-norm of each column of $\mathbf{A}$ is less than or equal to $\sqrt{n}$ and some technical conditions are satisfied. Then with high probability,

$$R_{\text{minmax}} \geq c\sigma R\sqrt{\frac{\ln p}{n}}$$

with some $c > 0$.

Bound the minimax risk from above

- The worst-case risk of any explicitly given estimator is an upper bound of R_{minmax}.
- If the upper bound equals Θ(lower bound), then Θ(lower bound) is the *optimal minimax rate*. For example, the result of the theorem above is optimal [12].

Practical Issues

Take the ℓ_2 loss, i.e., $\mathcal{L}(\hat{\mathbf{x}}, \mathbf{x}^\natural) := \left\|\hat{\mathbf{x}} - \mathbf{x}^\natural\right\|_2^2$, as an example. Is evaluating $\left\|\hat{\mathbf{x}} - \mathbf{x}^\natural\right\|_2^2$ enough for evaluating the performance of an $\hat{\mathbf{x}}$?

Practical Issues

No, because in general we can only *numerically approximate* the solution of

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \{F(\mathbf{x})\}.$$

Implementation

How do we *numerically approximate* $\hat{\mathbf{x}}$?

Practical performance

Denote the numerical approximation by $\mathbf{x}^\star_\epsilon$. The practical performance is determined by

$$\left\|\mathbf{x}^\star_\epsilon - \mathbf{x}^\natural\right\|_2 \leq \underbrace{\|\mathbf{x}^\star_\epsilon - \hat{\mathbf{x}}\|_2}_{\text{approximation error}} + \underbrace{\left\|\hat{\mathbf{x}} - \mathbf{x}^\natural\right\|_2}_{\text{statistical error}}.$$

How do we evaluate $\|\mathbf{x}^\star_\epsilon - \hat{\mathbf{x}}\|_2^2$?

- The ϵ-approximation solution, $\mathbf{x}^\star_\epsilon$, will be defined rigorously in the later lectures.

Practical issues

How do we *numerically approximate* $\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \{F(\mathbf{x})\}$ for a given F?

General idea of an optimization algorithm

Guess a solution, and then *refine* it based on *oracle information*.

Repeat the procedure until the result is *good enough*.

How do we evaluate the approximation error $\|\mathbf{x}^\star_\epsilon - \hat{\mathbf{x}}\|_2$?

General concept about the approximation error

It depends on the *characteristics* of the function F and the chosen numerical *optimization algorithm*.

Need for convex analysis

General idea of an optimization algorithm

Guess a solution, and then *refine* it based on *oracle information*.

Repeat the procedure until the result is *good enough*.

General concept about the approximation error

It depends on the *characteristics* of the function F and the chosen numerical *optimization algorithm*.

Role of convexity

Convex optimization provides a key framework in obtaining numerical approximations at well-understood computational costs.

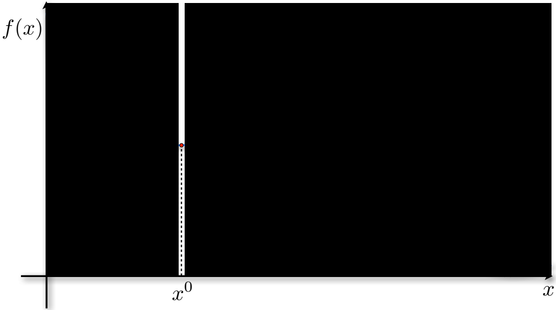
To precisely understand these ideas, we need to understand basics of *convex analysis*.

Challenges for an iterative optimization algorithm

Problem

Find the minimum $x^\star$ of $f(x)$, given starting point x^0 based on only local information.

- **Fog of war**

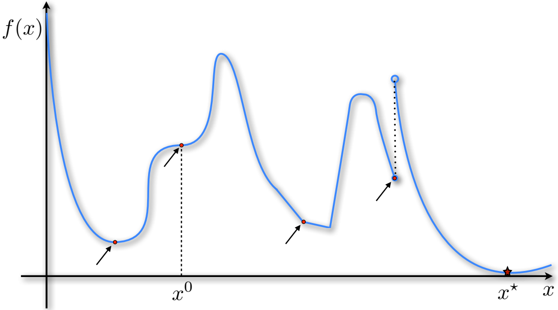


Challenges for an iterative optimization algorithm

Problem

Find the minimum $x^\star$ of $f(x)$, given starting point x^0 based on only local information.

- Fog of war, non-differentiability, discontinuities, local minima, stationary points...



Basics of functions

Definition (Function)

A function f with domain $\mathcal{Q} \subseteq \mathbb{R}^p$ and codomain $\mathcal{U} \subseteq \mathbb{R}$ is denoted as:

$$f \ : \ \mathcal{Q} \to \mathcal{U}.$$

The domain $\mathcal{Q}$ represents the set of values in $\mathbb{R}^p$ on which f is defined and is denoted as $\mathrm{dom}(f) \equiv \mathcal{Q} = \{\mathbf{x} : -\infty < f(\mathbf{x}) < +\infty\}$. The codomain $\mathcal{U}$ is the set of function values of f for any input in $\mathcal{Q}$.

Continuity in functions

Definition (Continuity)

Let $f : \mathcal{Q} \to \mathbb{R}$ where $\mathcal{Q} \subseteq \mathbb{R}^p$. Then, f is a continuous function over its domain $\mathcal{Q}$ if and only if

$$\lim_{\mathbf{x} \to \mathbf{y}} f(\mathbf{x}) = f(\mathbf{y}), \quad \forall \mathbf{y} \in \mathcal{Q},$$

i.e., the limit of f—as $\mathbf{x}$ approaches $\mathbf{y}$—exists and is equal to $f(\mathbf{y})$.

Definition (Class of continuous functions)

We denote the class of continuous functions f over the domain $\mathcal{Q}$ as $f \in \mathcal{C}(\mathcal{Q})$.

Definition (Lipschitz continuity)

Let $f : \mathcal{Q} \to \mathbb{R}$ where $\mathcal{Q} \subseteq \mathbb{R}^p$. Then, f is called Lipschitz continuous if there exists a constant value $K \geq 0$ such that:

$$|f(\mathbf{y}) - f(\mathbf{x})| \leq K \|\mathbf{y} - \mathbf{x}\|_2, \quad \forall \mathbf{x},\ \mathbf{y} \in \mathcal{Q}.$$

- "Small" changes in the input result into "small" changes in the function values.

Continuity in functions

Definition (Continuity)

Let $f : \mathcal{Q} \to \mathbb{R}$ where $\mathcal{Q} \subseteq \mathbb{R}^p$. Then, f is a continuous function over its domain $\mathcal{Q}$ if and only if

$$\lim_{\mathbf{x} \to \mathbf{y}} f(\mathbf{x}) = f(\mathbf{y}), \quad \forall \mathbf{y} \in \mathcal{Q},$$

i.e., the limit of f—as $\mathbf{x}$ approaches $\mathbf{y}$—exists and is equal to $f(\mathbf{y})$.

Definition (Class of continuous functions)

We denote the class of continuous functions f over the domain $\mathcal{Q}$ as $f \in \mathcal{C}(\mathcal{Q})$.

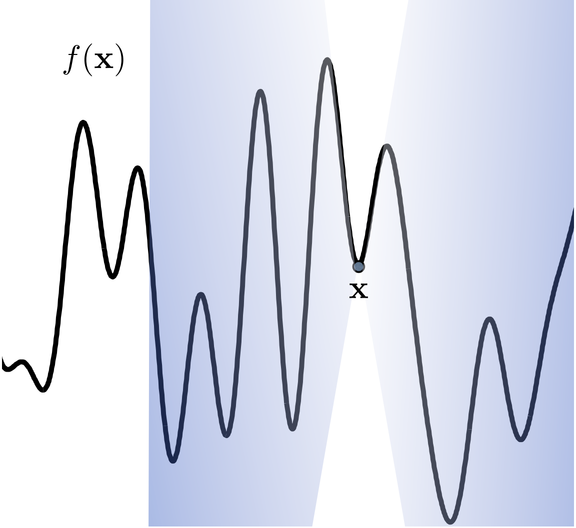
Definition (Lipschitz continuity)

Let $f : \mathcal{Q} \to \mathbb{R}$ where $\mathcal{Q} \subseteq \mathbb{R}^p$. Then, f is called Lipschitz continuous if there exists a constant value $K \geq 0$ such that:

$$|f(\mathbf{y}) - f(\mathbf{x})| \leq K\|\mathbf{y} - \mathbf{x}\|_2, \quad \forall \mathbf{x},\ \mathbf{y} \in \mathcal{Q}.$$

- "Small" changes in the input result into "small" changes in the function values.

Continuity in functions



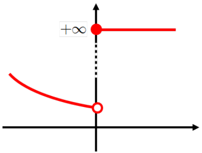
Lower semi-continuity

Definition

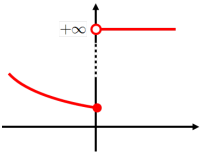
A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is lower semi-continuous (l.s.c.) if

$$\liminf_{\mathbf{x}\to\mathbf{y}} f(\mathbf{x}) \geq f(\mathbf{y}), \quad \text{for any } \mathbf{y} \in \text{dom}(f).$$

$$f(x) = \begin{cases} e^{-x}, & \text{if } x < 0 \\ +\infty, & \text{if } x \geq 0 \end{cases}$$



$$f(x) = \begin{cases} e^{-x}, & \text{if } x \leq 0 \\ +\infty, & \text{if } x > 0 \end{cases}$$



Unless stated otherwise, we only consider l.s.c. functions.

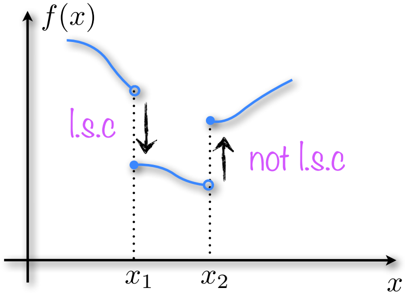
Lower semi-continuity

Definition

A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is lower semi-continuous (l.s.c.) if

$$\liminf_{\mathbf{x} \to \mathbf{y}} f(\mathbf{x}) \geq f(\mathbf{y}), \quad \text{for any } \mathbf{y} \in \mathrm{dom}(f).$$

- **Rule of thumb:** a lower semi-continuous function *only jumps down.*



Differentiability in functions

- We use $\nabla f(\mathbf{x})$ to denote the *gradient* of f at $\mathbf{x} \in \mathbb{R}^p$ such that:

$$\nabla f(\mathbf{x}) = \sum_{i=1}^{p} \frac{\partial f}{\partial x_i}\mathbf{e}_i = \frac{\partial f}{\partial x_1}\mathbf{e}_1 + \cdots + \frac{\partial f}{\partial x_p}\mathbf{e}_p$$

Example: $f(\mathbf{x}) = \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2$

$\nabla f(\mathbf{x}) = -2\mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x})$.

Definition (Differentiability)

Let $f \in \mathcal{C}(\mathcal{Q})$ where $\mathcal{Q} \subseteq \mathbb{R}^p$. Then, f is a k-times continuously differentiable on $\mathcal{Q}$ if and only if $\nabla^k f(\mathbf{x})$ exists $\forall \mathbf{x} \in \mathcal{Q}$.

Definition (Class of differentiable functions)

We denote the class of k-times continuously differentiable functions f on $\mathcal{Q}$ as $f \in \mathcal{C}^k(\mathcal{Q})$.

- In the special case of $k = 2$, we dub $\nabla^2 f(\mathbf{x})$ the **Hessian** of $f(\mathbf{x})$.
- We have $\mathcal{C}^q(\mathcal{Q}) \subseteq \mathcal{C}^k(\mathcal{Q})$ where $q \leq k$. That is, a twice differentiable function is at least differentiable once.
- For complex cases $\mathbb{C}$, we refer to the Matrix Cookbook online.

Differentiability in functions

- Some examples:

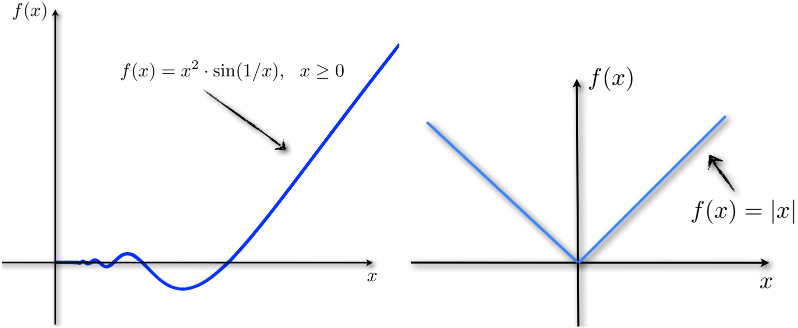


Figure: **(Left panel)** ∞-times continuously differentiable function in $\mathbb{R}$. **(Right panel)** Non-differentiable $f(x) = |x|$ in $\mathbb{R}$.

Stationary points of differentiable functions

Definition (Stationary point)

A point $\bar{\mathbf{x}}$ is called a stationary point of a twice differentiable function $f(\mathbf{x})$ if

$$\nabla f(\bar{\mathbf{x}}) = \mathbf{0}.$$

Definition (Local minima, maxima, and saddle points)

Let $\bar{\mathbf{x}}$ be a stationary point of a twice differentiable function $f(\mathbf{x})$.

- If $\nabla^2 f(\bar{\mathbf{x}}) \succ 0$, then the point $\bar{\mathbf{x}}$ is called a local minimum.
- If $\nabla^2 f(\bar{\mathbf{x}}) \prec 0$, then the point $\bar{\mathbf{x}}$ is called a local maximum.
- If $\nabla^2 f(\bar{\mathbf{x}}) = 0$, then the point $\bar{\mathbf{x}}$ can be a saddle point depending on the sign change.

Stationary points of smooth functions contd.

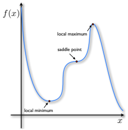
Intuition

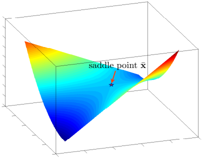
Recall Taylor's theorem for the function f around $\bar{\mathbf{x}}$ for all $\mathbf{y}$ that satisfy $\|\mathbf{y} - \bar{\mathbf{x}}\|_2 \leq r$ in a local region with radius r as follows

$$f(\mathbf{y}) = f(\bar{\mathbf{x}}) + \langle \nabla f(\bar{\mathbf{x}}), \mathbf{y} - \bar{\mathbf{x}} \rangle + \frac{1}{2}(\mathbf{y} - \bar{\mathbf{x}})^T \nabla^2 f(\mathbf{z})(\mathbf{y} - \bar{\mathbf{x}}),$$

where $\mathbf{z}$ is a point between $\bar{\mathbf{x}}$ and $\mathbf{y}$. When $r \to 0$, the second-order term becomes $\nabla^2 f(\mathbf{z}) \to \nabla^2 f(\bar{\mathbf{x}})$. Since $\nabla f(\bar{\mathbf{x}}) = 0$, Taylor's theorem leads to

- $f(\mathbf{y}) > f(\bar{\mathbf{x}})$ when $\nabla^2 f(\bar{\mathbf{x}}) \succ 0$. Hence, the point $\bar{\mathbf{x}}$ is a local minimum.
- $f(\mathbf{y}) < f(\bar{\mathbf{x}})$ when $\nabla^2 f(\bar{\mathbf{x}}) \prec 0$. Hence, the point $\bar{\mathbf{x}}$ is a local maximum.
- $f(\mathbf{y}) \gtrless f(\bar{\mathbf{x}})$ when $\nabla^2 f(\bar{\mathbf{x}}) = 0$. Hence, the point $\bar{\mathbf{x}}$ can be a saddle point (i.e., $f(x) = x^3$ at $\bar{x} = 0$), a local minima (i.e., $f(x) = x^4$ at $\bar{x} = 0$) or a local maxima (i.e., $f(x) = -x^4$ at $\bar{x} = 0$).





Convexity

Definition

A function $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ is called convex on its domain $\mathcal{Q}$ if and only if for any $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{Q}$ and $\alpha \in [0, 1]$ we have:

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2).$$

- If $-f(\mathbf{x})$ is convex, then $f(\mathbf{x})$ is called concave.

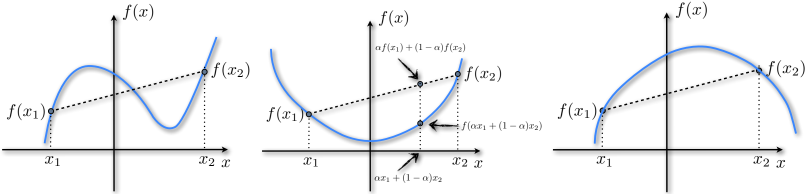


Figure: (**Left**) Non-convex (**Middle**) Convex (**Right**) Concave

Convexity

Definition

A function $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ is called convex on its domain $\mathcal{Q}$ if and only if for any $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{Q}$ and $\alpha \in [0, 1]$ we have:

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2).$$

- Additional terms that you will encounter in the literature

Definition (Proper)

A convex function f is called proper if its domain satisfies $\mathrm{dom}(f) \neq \emptyset$ and, $f(\mathbf{x}) > -\infty,\ \forall x \in \mathrm{dom}(f)$.

Definition (Extended real-valued convex functions)

We define the extended real-valued convex functions f as

$$f(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{if } \mathbf{x} \in \mathrm{dom}(f) \\ +\infty & \text{if otherwise} \end{cases}$$

To denote this concept, we use $f : \mathrm{dom}(f) \to \mathbb{R} \cup \{+\infty\}$. (Note how l.s.c. might be useful)

Convexity

Definition

A function $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ is called convex on its domain $\mathcal{Q}$ if and only if for any $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{Q}$ and $\alpha \in [0, 1]$ we have:

$$f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2).$$

Example

Function	Example	Attributes
ℓ_p vector norms, $p \geq 1$	$\\|\mathbf{x}\\|_2, \\|\mathbf{x}\\|_1, \\|\mathbf{x}\\|_\infty$	convex
ℓ_p matrix norms, $p \geq 1$	$\\|\mathbf{X}\\|_* = \sum_{i=1}^{\text{rank}(\mathbf{X})} \sigma_i$	convex
Square root function	$\sqrt{x}$	concave, nondecreasing
Maximum of functions	$\max\{x_1, \ldots, x_n\}$	convex, nondecreasing
Minimum of functions	$\min\{x_1, \ldots, x_n\}$	concave, nondecreasing
Logarithmic functions	$\log(\det(\mathbf{X}))$	concave, assumes $\mathbf{X} \succ 0$
Affine/linear functions	$\sum_{i=1}^n X_{ii}$	both convex and concave
Eigenvalue functions	$\lambda_{max}(\mathbf{X})$	convex, assumes $\mathbf{X} = \mathbf{X}^T$

Strict convexity

Definition

A function $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ is called *strictly convex* on its domain $\mathcal{Q}$ if and only if for any $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{Q}$ and $\alpha \in [0, 1]$ we have:

$$f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) < \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2).$$

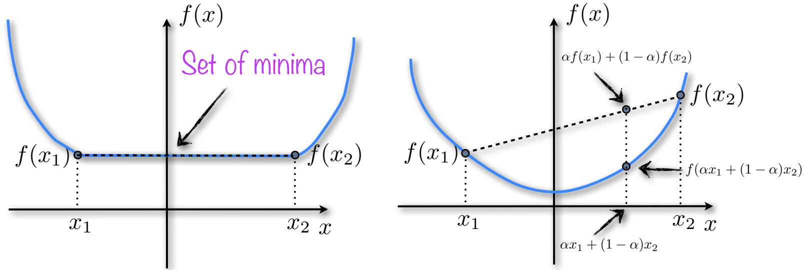


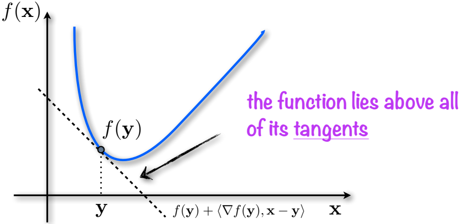
Figure: (**Left panel**) Convex function. (**Right panel**) Strictly convex function.

Revisiting: Alternative definitions of function convexity II

Definition

A function $f \in \mathcal{C}^1(\mathcal{Q})$ is called convex on its domain if for any $\mathbf{x},\ \mathbf{y} \in \mathcal{Q}$:

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}),\ \mathbf{x} - \mathbf{y} \rangle.$$



Definition

A function $f \in \mathcal{C}^1(\mathcal{Q})$ is called convex on its domain if for any $\mathbf{x},\ \mathbf{y} \in \mathcal{Q}$:

$$\langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}),\ \mathbf{y} - \mathbf{x} \rangle \geq 0.$$

*That is, if its gradient is a monotone operator (cf., Lecture 8).

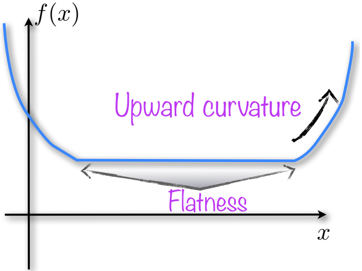
Revisiting: Alternative definitions of function convexity III

Definition

A function $f \in \mathcal{C}^2(\mathbb{R}^p)$ is called convex on its domain if for any $\mathbf{x},\ \mathbf{y} \in \mathbb{R}^p$:

$$\nabla^2 f(\mathbf{x}) \succeq 0.$$

- Geometrical interpretation: the graph of f has zero or positive (upward) curvature.
- However, this does not exclude flatness of f.



What about some "ill-posed" cases...?

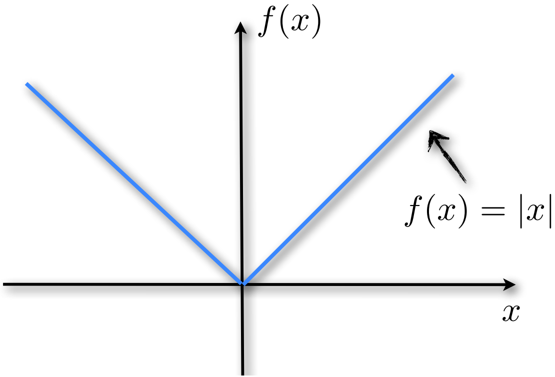


Figure: Non-differentiable at the origin

Subdifferentials and (sub)gradients in convex functions

- Subdifferential: generalizes ∇ to *nondifferentiable functions*

Definition

Let $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ be a convex function. The subdifferential of f at a point $\mathbf{x} \in \mathcal{Q}$ is defined by the set:

$$\partial f(\mathbf{x}) = \{\mathbf{v} \in \mathbb{R}^p \ : \ f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{v}, \ \mathbf{y} - \mathbf{x} \rangle \text{ for all } \mathbf{y} \in \mathcal{Q}\} \,.$$

Each element $\mathbf{v}$ of $\partial f(\mathbf{x})$ is called *subgradient* of f at $\mathbf{x}$.

Definition

Let $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ be a differentiable convex function. Then, the subdifferential of f at a point $\mathbf{x} \in \mathcal{Q}$ contains only the gradient, i.e., $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.

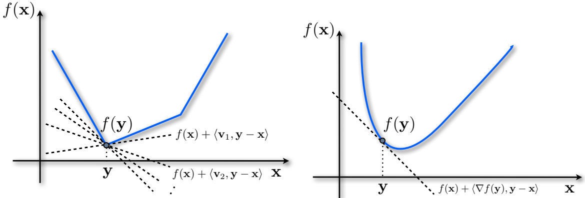


Figure: (**Left**) Non-differentiability at point $\mathbf{y}$. (**Right**) Gradient as a subdifferential with a singleton entry.

Subdifferentials and (sub)gradients in convex functions

Example

- $f(\mathbf{x}) = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \quad \longrightarrow \quad \nabla f(\mathbf{x}) = -2\mathbf{A}^T(\mathbf{y} - \mathbf{A}\mathbf{x})$.
- $f(\mathbf{X}) = -\log\det(\mathbf{X}) \quad \longrightarrow \quad \nabla f(\mathbf{X}) = \mathbf{X}^{-1}$
- $f(x) = |x| \quad \longrightarrow \quad \partial|x| = \{\mathrm{sgn}(x)\}$, if $x \neq 0$, but $[-1, 1]$, if $x = 0$.

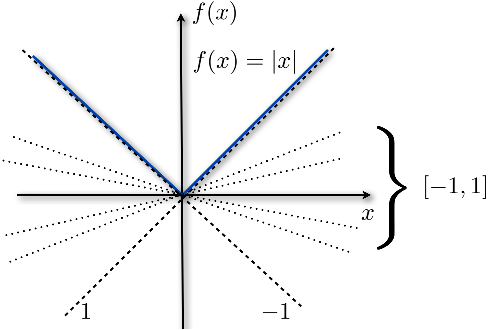
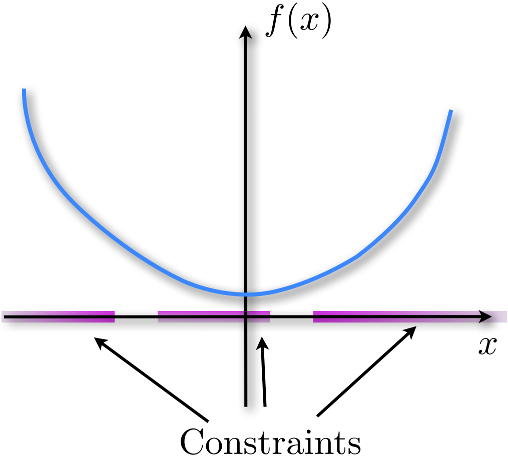


Figure: Subdifferential of $f(x) = |x|$ in $\mathbb{R}$.

Is convexity of f enough for an iterative optimization algorithm?



Convexity over sets

Definition

- $\mathcal{Q} \subseteq \mathbb{R}^p$ is a convex set if $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{Q} \Rightarrow \forall \alpha \in [0,1],\quad \alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2 \in \mathcal{Q}$.
- $\mathcal{Q} \subseteq \mathbb{R}^p$ is a *strictly* convex set if
$$\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{Q} \Longrightarrow \forall \alpha \in (0,1),\quad \alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2 \in \text{interior}(\mathcal{Q}).$$

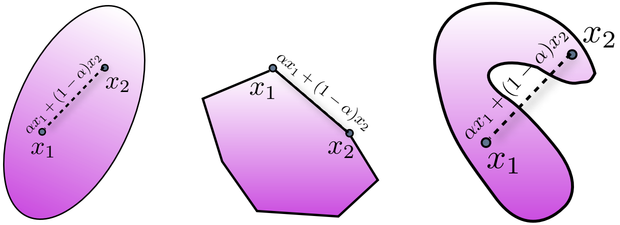


Figure: (**Left**) Strictly convex (**Middle**) Convex (**Right**) Non-convex

Convexity over sets

Definition

- $\mathcal{Q} \subseteq \mathbb{R}^p$ is a convex set if $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{Q} \Rightarrow \forall \alpha \in [0,1],\quad \alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2 \in \mathcal{Q}$.
- $\mathcal{Q} \subseteq \mathbb{R}^p$ is a *strictly* convex set if
$$\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{Q} \Longrightarrow \forall \alpha \in (0,1),\quad \alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2 \in \text{interior}(\mathcal{Q}).$$

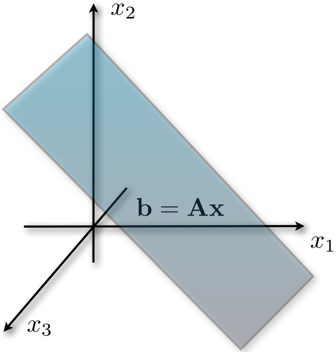


Figure: A linear set of equations $\mathbf{b} = \mathbf{A}\mathbf{x}$ defines an affine (thus convex) set.

Convexity over sets

Definition

- $\mathcal{Q} \subseteq \mathbb{R}^p$ is a convex set if $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{Q} \Rightarrow \forall \alpha \in [0,1],\quad \alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2 \in \mathcal{Q}$.
- $\mathcal{Q} \subseteq \mathbb{R}^p$ is a *strictly* convex set if

$$\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{Q} \Longrightarrow \forall \alpha \in (0,1),\quad \alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2 \in \text{interior}(\mathcal{Q}).$$

Why is this also important/useful?

- convex sets $<>$ convex optimization constraints

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & f_0(\mathbf{x}) \\ \text{subject to} & \text{constraints} \end{array}$$

Some basic notions on sets I

Definition (Closed set)

A set is called *closed* if it contains all its limit points.

Definition (Closure of a set)

Let $\mathcal{Q} \subseteq \mathbb{R}^p$ be a given open set, i.e., the limit points on the boundaries of $\mathcal{Q}$ do not belong into $\mathcal{Q}$. Then, the closure of $\mathcal{Q}$, denoted as $\mathrm{cl}(\mathcal{Q})$, is the smallest set in $\mathbb{R}^p$ that includes $\mathcal{Q}$ with its boundary points.

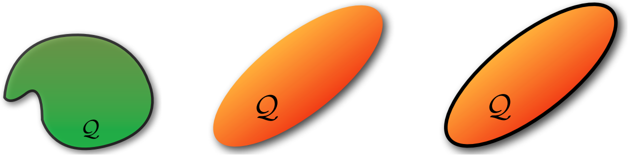


Figure: (**Left panel**) Closed set $\mathcal{Q}$. (**Middle panel**) Open set $\mathcal{Q}$ and its closure $\mathcal{Q}$ (**Right panel**).

Some basic notions on sets II

Definition (Interior)

Let $\mathcal{Q} \subseteq \mathbb{R}^p$. Then, a point $\mathbf{x} \in \mathbb{R}^p$ is an *interior* of $\mathcal{Q}$ if a neighborhood with radius r of $\mathbf{x}$ is also included in $\mathcal{Q}$. That is, there exists $r > 0$, such that $\{\mathbf{y} \ : \ \|\mathbf{y} - \mathbf{x}\|_2 \leq r\} \in \mathcal{Q}$. The set of all interior points is denoted as $\text{int}(\mathcal{Q})$.

Example

- The interior of an open set is the set itself.
- The interior of the set $\{\mathbf{x} \ : \ \|\mathbf{y} - \mathbf{x}\|_2 \leq r\}$ is the open set $\{\mathbf{x} \ : \ \|\mathbf{y} - \mathbf{x}\|_2 < r\}$.

Some basic notions on sets II

Definition (Interior)

Let $\mathcal{Q} \subseteq \mathbb{R}^p$. Then, a point $\mathbf{x} \in \mathbb{R}^p$ is an *interior* of $\mathcal{Q}$ if a neighborhood with radius r of $\mathbf{x}$ is also included in $\mathcal{Q}$. That is, there exists $r > 0$, such that $\{\mathbf{y} \ : \ \|\mathbf{y} - \mathbf{x}\|_2 \leq r\} \in \mathcal{Q}$. The set of all interior points is denoted as $\mathrm{int}(\mathcal{Q})$.

Example

- The interior of an open set is the set itself.
- The interior of the set $\{\mathbf{x} \ : \ \|\mathbf{y} - \mathbf{x}\|_2 \leq r\}$ is the open set $\{\mathbf{x} \ : \ \|\mathbf{y} - \mathbf{x}\|_2 < r\}$.

Definition (Relative interior)

Let $\mathcal{Q} \subseteq \mathbb{R}^p$. Then, a point $\mathbf{x} \in \mathbb{R}^p$ is a *relative interior* of $\mathcal{Q}$ if $\mathcal{Q}$ contains the intersection of a neighborhood with radius r around $\mathbf{x}$ with the intersection of all affine sets containing $\mathcal{Q}$, i.e., $\mathrm{aff}(\mathcal{Q})$. The set of all relative interior points is denoted as $\mathrm{relint}(\mathcal{Q})$.

Example

The interior of the affine set $\mathcal{X} = \{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{b}\}$ is empty. However, its relative interior is itself, i.e., $\mathrm{relint}(\mathcal{X}) = \mathcal{X}$.

Convex hull

Definition (Convex hull)

Let $\mathcal{V} \subseteq \mathbb{R}^p$ be a set. The convex hull of $\mathcal{V}$, i.e., $\mathrm{conv}(\mathcal{V})$, is the *smallest* convex set that contains $\mathcal{V}$.

Definition (Convex hull of points)

Let $\mathcal{V} \subseteq \mathbb{R}^p$ be a finite set of points with cardinality $|\mathcal{V}|$. The convex hull of $\mathcal{V}$ is the set of all convex combinations of its points, i.e.,

$$\mathrm{conv}(\mathcal{V}) = \left\{ \sum_{i=1}^{|\mathcal{V}|} \alpha_i \mathbf{x}_i \;:\; \sum_{i=1}^{|\mathcal{V}|} \alpha_i = 1,\ \alpha_i \geq 0, \forall i,\ \mathbf{x}_i \in \mathcal{V} \right\}.$$

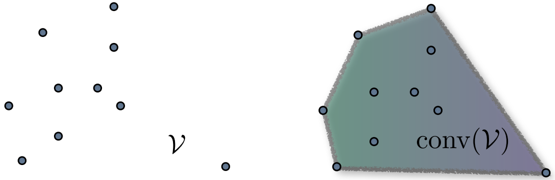


Figure: (**Left**) Discrete set of points $\mathcal{V}$. (**Right**) Convex hull $\mathrm{conv}(\mathcal{V})$.

Properties of convex sets

Lemma (Separating hyperplane theorem)

Let $\mathcal{Q}_1 \subseteq \mathbb{R}^p$ and $\mathcal{Q}_2 \subseteq \mathbb{R}^p$ be two non-empty and disjoint convex sets. Then, there exists at least one hyperplane that separates them, i.e., $\exists \alpha \neq \mathbf{0}$ such that:

$$\boldsymbol{\alpha}^T \mathbf{x}_1 \leq \boldsymbol{\alpha}^T \mathbf{x}_2, \quad \forall \mathbf{x}_1 \in \mathcal{Q}_1,\ \mathbf{x}_2 \in \mathcal{Q}_2$$

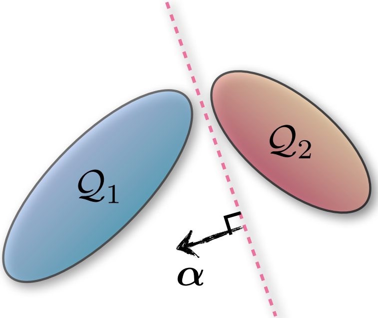


Figure: Illustration of a strictly separating hyperplane of two disjoint convex sets $\mathcal{Q}_1$ and $\mathcal{Q}_2$.

Revisiting: Alternative definition of function convexity I

Definition

The epigraph of a function $f : \mathcal{Q} \to \mathbb{R}$, $\mathcal{Q} \subseteq \mathbb{R}^p$ is the subset of $\mathbb{R}^{p+1}$ given by:

$$\text{epi}(f) = \{(\mathbf{x}, w) : \mathbf{x} \in \mathcal{Q}, w \in \mathbb{R}, f(\mathbf{x}) \leq w\} .$$

Lemma

A function $f : \mathcal{Q} \to \mathbb{R}$ is convex if and only if its epigraph, i.e, the region above its graph, is a convex set.

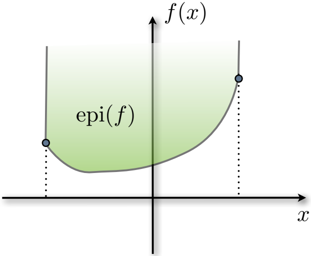


Figure: Epigraph — the region in green above graph $f(\cdot)$.

Cones

Definition (Convex cone)

A subset $\mathcal{K} \subseteq \mathbb{R}^p$ is called a convex cone if and only if for any $\mathbf{x}_1,\ \mathbf{x}_2 \in \mathcal{K}$, the point $\alpha\mathbf{x}_1 + \beta\mathbf{x}_2 \in \mathcal{K}$ for all nonnegative constants $\alpha, \beta \geq 0$.

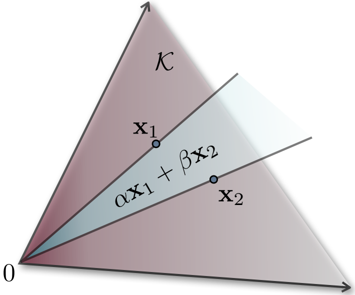


Figure: Illustration of a convex cone $\mathcal{K}$. The depicted cones extend to infinity.

Cones

Definition (Convex cone of an arbitrary set $\mathcal{Q}$)

A subset $\mathcal{K} \subseteq \mathbb{R}^p$ is called a convex cone of a given set $\mathcal{Q}$ if it contains all vectors $\lambda \mathbf{x}$ where $\mathbf{x}$ belongs to $\mathcal{Q}$ and λ is a non-negative scalar.

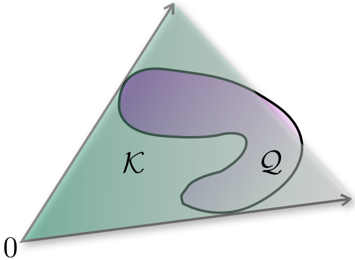


Figure: Illustration of a convex cone $\mathcal{K}$ of an arbitrary set $\mathcal{Q}$. The depicted cones extend till infinity.

Cones

Definition (Normal cone)

Let $\mathcal{Q} \subseteq \mathbb{R}^p$ be an arbitrary convex set in the linear space $\mathbb{R}^p$. The normal cone $\mathcal{N}_{\mathcal{Q}}(\mathbf{x})$ of $\mathcal{Q}$ at a point $\mathbf{x}$ is defined as:

$$\mathcal{N}_{\mathcal{Q}}(\mathbf{x}) = \text{cone}\left\{\mathbf{s} \ : \ \langle \mathbf{s}, \mathbf{y} - \mathbf{x} \rangle \leq 0, \ \forall \mathbf{y} \in \mathcal{Q}\right\},$$

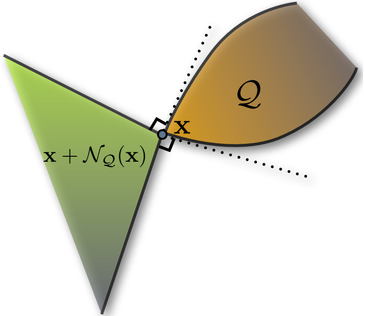


Figure: Illustration of the normal cone $\mathcal{N}_{\mathcal{Q}}(\mathbf{x})$ at a point $\mathbf{x}$ of an convex set $\mathcal{Q}$. The depicted normal cone extends till infinity.

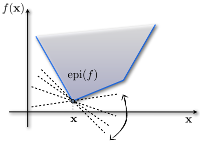
Revisiting: Subdifferential through epi(f) and normal cones

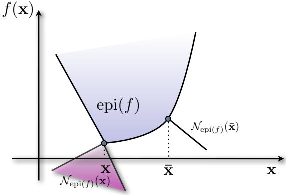
Definition (Subdifferential)

Let $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$ be a convex function. The subdifferential of f at a point $\mathbf{x} \in \mathcal{Q}$ is defined by the set:

$$\partial f(\mathbf{x}) = \left\{ \mathbf{v} \in \mathbb{R}^p \ : \ f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{v}, \ \mathbf{y} - \mathbf{x} \rangle \text{ for all } \mathbf{y} \in \mathcal{Q} \right\}.$$

Each element $\mathbf{v}$ of $\partial f(\mathbf{x})$ is called *subgradient* of f at $\mathbf{x}$.





Subdifferentials and normal cones

With some abuse on the notation, the set $\partial f(\mathbf{x})$ is related to the normal cone $\mathcal{N}_{\mathrm{epi}(f)}(\mathbf{x})$ of the epi(f) at a point $(\mathbf{x}, f(\mathbf{x}))$ as follows

$$[\partial f(\mathbf{x})^T \quad -\mathbf{1}]^T \subseteq \mathcal{N}_{\mathrm{epi}(f)}(\mathbf{x}),$$

where $\mathbf{1}$ is the vector of ones.

Cones

Definition (Dual cone)

Let $\mathcal{Q} \subseteq \mathbb{R}^p$ be an arbitrary subset in the linear space $\mathbb{R}^p$, and let $\mathcal{K}$ be its convex cone. The dual cone $\mathcal{K}^\star$ of $\mathcal{Q}$ is defined as:

$$\mathcal{K}^\star = \{\mathbf{y} \in \mathbb{R}^p \ : \ \langle \mathbf{y}, \mathbf{x} \rangle \geq 0, \ \forall \mathbf{x} \in \mathcal{K}\}.$$

- $\mathcal{K}^\star$ is always a *convex* cone, even if $\mathcal{Q}$ is not a convex set.

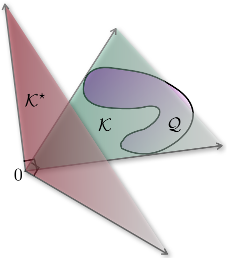


Figure: Illustration of a dual cone $\mathcal{K}^\star$ for subset $\mathcal{Q}$. The depicted cones extend to infinity.

Cones

Definition (Dual cone)

Let $\mathcal{Q} \subseteq \mathbb{R}^p$ be an arbitrary subset in the linear space $\mathbb{R}^p$, and let $\mathcal{K}$ be its convex cone. The dual cone $\mathcal{K}^\star$ of $\mathcal{Q}$ is defined as:

$$\mathcal{K}^\star = \{\mathbf{y} \in \mathbb{R}^p \ : \ \langle \mathbf{y}, \mathbf{x} \rangle \geq 0, \ \forall \mathbf{x} \in \mathcal{K}\}.$$

- $\mathcal{K}^\star$ is always a *convex* cone, even if $\mathcal{Q}$ is not a convex set.

Definition (Self-dual cone)

A cone $\mathcal{K}$ is *self-dual* if its dual cone (relative to inner product) is equal to $\mathcal{K}$.

- Examples: nonnegative orthant, cone of positive semidefinite matrices, etc.

Cones

Definition (Polar cone)

Let $\mathcal{Q} \subseteq \mathbb{R}^p$ be an arbitrary subset in the linear space $\mathbb{R}^p$, and let $\mathcal{K}$ be its convex cone. The polar cone $\mathcal{K}^\circ$ of $\mathcal{Q}$ is defined as:

$$\mathcal{K}^\circ = \{\mathbf{y} \in \mathbb{R}^p \ : \ \langle \mathbf{y}, \mathbf{x} \rangle \leq 0, \ \forall \mathbf{x} \in \mathcal{K}\}\,.$$

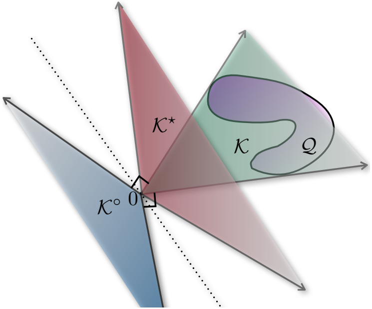


Figure: Illustration of a polar cone $\mathcal{K}^\circ$ for subset $\mathcal{Q}$.

Cones

Definition (Cone of descent directions)

Let $f : \mathcal{Q} \to \mathbb{R}$, $\mathcal{Q} \subseteq \mathbb{R}^p$ be a given function. Then, the cone of descent directions $\mathcal{D}(f, \mathbf{x})$ for f at a point $\mathbf{x} \in \mathcal{Q}$ is given by

$$\mathcal{D}(f, \mathbf{x}) = \text{cone}\left\{\mathbf{d} : \; f(\mathbf{x} + \mathbf{d}) \le f(\mathbf{x}) \; \text{ such that } \; \mathbf{x} + \mathbf{d} \in \mathcal{Q}\right\}.$$

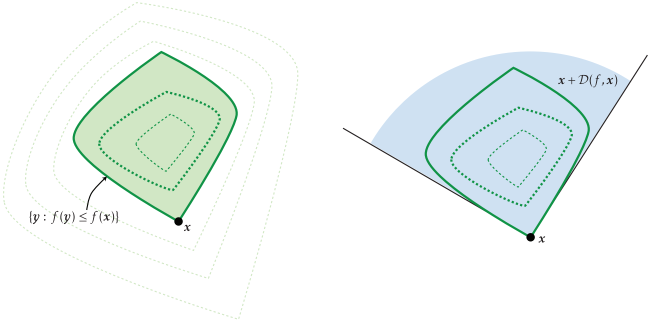


Figure: Illustration of a descent cone $\mathcal{D}(f, \mathbf{x})$ for a toy example.

- This lecture

  1. Learning as an **optimization problem**
  2. Basic concepts in convex analysis
  3. Three important classes of convex functions

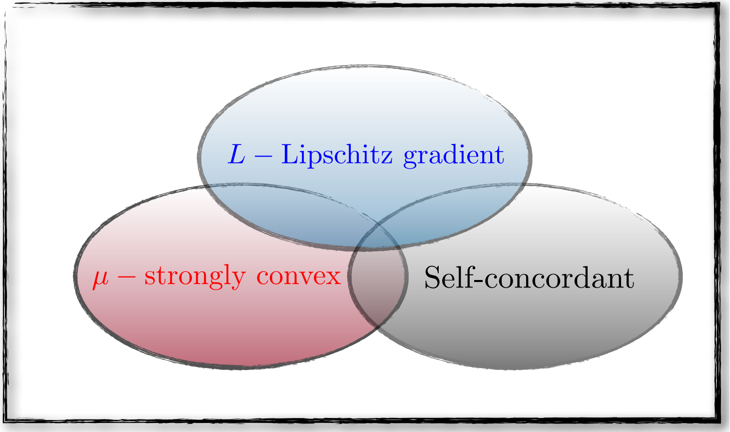
- Next lecture

  1. Optimality conditions
  2. Unconstrained convex minimization
  3. Convergence and convergence rate characterization of methods for unconstrained minimization

Classes of convex functions

Definition

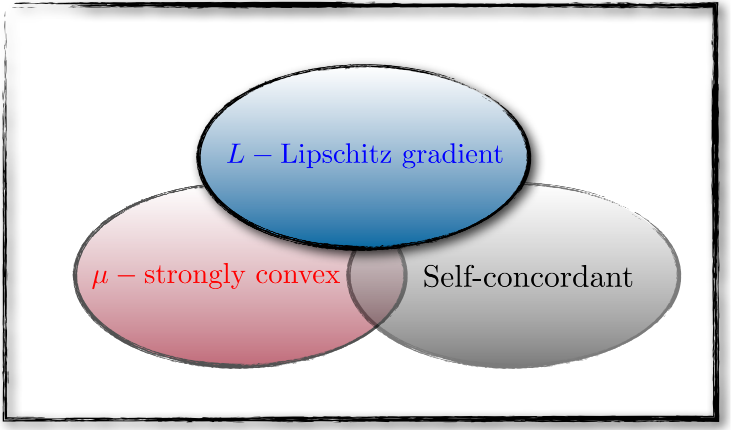
We use $\mathcal{F}$ to denote the class of convex functions f. The domain of f will be apparent from the context.



Classes of convex functions

Definition

We use $\mathcal{F}$ to denote the class of convex functions f. The domain of f will be apparent from the context.



L-Lipschitz gradient class of functions

Definition (L-Lipschitz gradient functions)

Let $f : \mathcal{Q} \to \mathbb{R}$ be a differentiable convex function, i.e., $f \in \mathcal{F}^1(\mathcal{Q})$. Then, f is a Lipschitz gradient function if and only if

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{Q}.$$

Here, $L > 0$ is known as the Lipschitz constant.

L-Lipschitz gradient class of functions

Definition (L-Lipschitz gradient functions)

Let $f : \mathcal{Q} \to \mathbb{R}$ be a differentiable convex function, i.e., $f \in \mathcal{F}^1(\mathcal{Q})$. Then, f is a Lipschitz gradient function if and only if

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{Q}.$$

Here, $L > 0$ is known as the Lipschitz constant.

Definition (L-Lipschitz gradient functions in a Banach space)

Let $f : \mathcal{Q} \to \mathbb{R}$ be a differentiable convex function, i.e., $f \in \mathcal{F}^1(\mathcal{Q})$. Then, f is a Lipschitz gradient function if and only if

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^* \leq L\|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{Q}.$$

Here, $L > 0$ is known as the Lipschitz constant.

Definition (L-Lipschitz gradient convex functions)

Let $f : \mathcal{Q} \to \mathbb{R}$ be a differentiable convex function, i.e., $f \in \mathcal{F}^1(\mathcal{Q})$. Then, f is a L-Lipschitz gradient function if and only if the following function is convex

$$h(\mathbf{x}) = \frac{L}{2}\|\mathbf{x}\|_2^2 - f(\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{Q}.$$

L-Lipschitz gradient class of functions

Definition (L-Lipschitz gradient functions)

Let $f : \mathcal{Q} \to \mathbb{R}$ be a differentiable convex function, i.e., $f \in \mathcal{F}^1(\mathcal{Q})$. Then, f is a Lipschitz gradient function if and only if

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{Q}.$$

Here, $L > 0$ is known as the Lipschitz constant.

Definition (Class of 2-nd order Lipschitz functions)

We denote the class of twice continuously differentiable functions f on $\mathcal{Q}$, where their 2^{nd} derivative is Lipschitz continuous, i.e.,

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\|_{2\to 2} \leq L\|\mathbf{x} - \mathbf{y}\|_2, \quad \forall \mathbf{x}, \mathbf{y} \in Q,$$

as $f \in \mathcal{F}_L^{2,2}(\mathcal{Q})$.

- In the sequel, we will use the notation $\mathcal{F}_L^{l,m}$ to denote convex functions that are l-times differentiable with m-th order Lipschitz property.

L-Lipschitz gradient class of functions

Definition (L-Lipschitz gradient functions)

Let $f : \mathcal{Q} \to \mathbb{R}$ be a differentiable convex function, i.e., $f \in \mathcal{F}^1(\mathcal{Q})$. Then, f is a Lipschitz gradient function if and only if

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{Q}.$$

Here, $L > 0$ is known as the Lipschitz constant.

Example (Underdetermined least squares)

Consider an *underdetermined* linear system of equations $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$ where $\mathbf{A} \in \mathbb{R}^{n\times p}$ and $\mathbf{x}^\natural$ is unknown. Let $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2$. Then, f is a L-Lipschitz convex function, i.e., $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^p)$ where:

$$\begin{aligned}\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\|_2 &= \|\mathbf{A}^T\mathbf{A}\left(\mathbf{x}_1 - \mathbf{x}_2\right)\|_2 \\ &\leq \|\mathbf{A}^T\mathbf{A}\|_{2\to 2}\|\mathbf{x}_1 - \mathbf{x}_2\|_2,\end{aligned}$$

for all $\mathbf{x}_1, \mathbf{x}_2$. That is, $L = \sigma^2_{\max}\left(\mathbf{A}\right)$. Also, (**SPOILER ALERT**) $\sigma^2_{\min}\left(\mathbf{A}\right) = 0$.

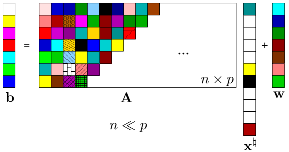


Figure: Compressive sensing.

L-Lipschitz gradient class of functions

Definition (L-Lipschitz gradient functions)

Let $f : \mathcal{Q} \to \mathbb{R}$ be a differentiable convex function, i.e., $f \in \mathcal{F}^1(\mathcal{Q})$. Then, f is a Lipschitz gradient function if and only if

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{Q}.$$

Here, $L > 0$ is known as the Lipschitz constant.

Example (Linear functions)

Consider any linear function $f(\mathbf{x}) = \mathbf{c}^T\mathbf{x} + \beta$. Then, f is a 0-Lipschitz convex function, i.e., $f \in \mathcal{F}_0^{1,1}(\mathbb{R}^p)$ since $\nabla f(\mathbf{x}) = \mathbf{c}$, $\forall \mathbf{x}$ and thus

$$\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\|_2 = 0 \cdot \|\mathbf{x}_1 - \mathbf{x}_2\|_2,$$

for all $\mathbf{x}_1, \mathbf{x}_2$.

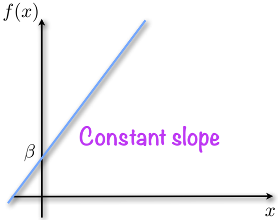


Figure: Linear function have $L = 0$ Lipschitz constant.

L-Lipschitz gradient class of functions

Definition (L-Lipschitz gradient functions)

Let $f : \mathcal{Q} \to \mathbb{R}$ be a differentiable convex function, i.e., $f \in \mathcal{F}^1(\mathcal{Q})$. Then, f is a Lipschitz gradient function if and only if

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{Q}.$$

Here, $L > 0$ is known as the Lipschitz constant.

Example (Underdetermined least squares)

Consider an *underdetermined* linear system of equations $\mathbf{b} = \mathbf{A}\mathbf{x}^{\natural} + \mathbf{w}$ where $\mathbf{A} \in \mathbb{R}^{n \times p}$ and $\mathbf{x}^{\natural}$ is unknown. Let $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2$. Using operator norm properties, we have

$$\begin{aligned} \|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\|_1 &= \|\mathbf{A}^T\mathbf{A}\left(\mathbf{x}_1 - \mathbf{x}_2\right)\|_1 \\ &\leq \|\mathbf{A}^T\mathbf{A}\|_{\infty\to 1}\|\mathbf{x}_1 - \mathbf{x}_2\|_\infty \\ &\text{(derivation on board)} \end{aligned}$$

for all $\mathbf{x}_1, \mathbf{x}_2$.

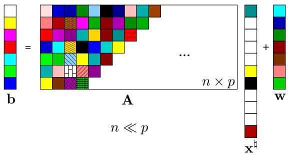


Figure: Compressive sensing.

Properties of L-Lipschitz functions

Lemma

Let $f : \mathcal{Q} \to \mathbb{R},\ \mathcal{Q} \subseteq \mathbb{R}^p$ be a twice differentiable convex function, i.e., $f \in \mathcal{F}^2(\mathcal{Q})$. Then, f is a Lipschitz gradient function if and only if

$$0 \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}, \quad \forall \mathbf{x} \in \mathbb{R}^p.$$

Properties of L-Lipschitz functions

Lemma

Let $f : \mathcal{Q} \to \mathbb{R}$, $\mathcal{Q} \subseteq \mathbb{R}^p$ be a twice differentiable convex function, i.e., $f \in \mathcal{F}^2(\mathcal{Q})$. Then, f is a Lipschitz gradient function if and only if

$$0 \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}, \quad \forall \mathbf{x} \in \mathbb{R}^p.$$

Proof.

($\Longrightarrow$) Key ingredient: *Taylor's theorem*. Then for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, we have:

$$\nabla f(\mathbf{y}) = \nabla f(\mathbf{x}) + \int_0^1 \nabla^2 f(\mathbf{x} + \tau(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x})\, d\tau$$

By the Cauchy-Schwartz and Jensen inequalities, we further have $(1/r + 1/q = 1)$:

$$\begin{aligned} \|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_r &\leq \left\| \int_0^1 \nabla^2 f(\mathbf{x} + \tau(\mathbf{y} - \mathbf{x}))\, d\tau \right\|_{q \to r} \cdot \|\mathbf{y} - \mathbf{x}\|_q \\ &\leq \int_0^1 \left\| \nabla^2 f(\mathbf{x} + \tau(\mathbf{y} - \mathbf{x})) \right\|_{2 \to 2} d\tau \cdot \|\mathbf{y} - \mathbf{x}\|_2 [q = r = 2] \\ &\leq L\|\mathbf{y} - \mathbf{x}\|_2 \end{aligned}$$

$\square$

Properties of L-Lipschitz functions

Lemma

Let $f : \mathcal{Q} \to \mathbb{R}$, $\mathcal{Q} \subseteq \mathbb{R}^p$ be a twice differentiable convex function, i.e., $f \in \mathcal{F}^2(\mathcal{Q})$. Then, f is a Lipschitz gradient function if and only if

$$0 \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}, \quad \forall \mathbf{x} \in \mathbb{R}^p.$$

Example (Positive semi-definite quadratic functions)

Consider any quadratic function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{\Phi}\mathbf{x} + \mathbf{c}^T\mathbf{x} + \beta$ where $\mathbf{\Phi} \succeq 0$. Then, f is a L-Lipschitz convex function, i.e., $f \in \mathcal{F}_L^{2,1}(\mathbb{R}^p)$ with $L = \|\mathbf{\Phi}\|_{2\to 2}$ since

$$\nabla f(\mathbf{x}) = \mathbf{\Phi}\mathbf{x} + \mathbf{c} \quad \text{and} \quad \nabla^2 f(\mathbf{x}) = \mathbf{\Phi}.$$

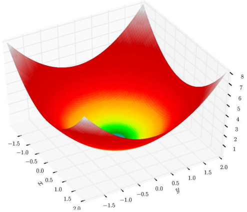


Figure: Quadratic function with $\mathbf{\Phi} \succeq 0$ has $L = \|\mathbf{\Phi}\|_{2\to 2}$ Lipschitz constant.

Additional properties of L-Lipschitz functions

Lemma

Let $f \in \mathcal{F}_L^{1,1}(\mathcal{Q})$. Then, we have:

$$|f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| \leq \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|_2^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{Q}.$$

Proof.

By the Taylor's theorem:

$$f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 \langle \nabla f(\mathbf{x} + \tau(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \, d\tau.$$

Therefore,

$$\begin{aligned} |f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| &\leq \int_0^1 \|\nabla f(\mathbf{x} + \tau(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})\|^* \cdot \|\mathbf{y} - \mathbf{x}\| \, d\tau \\ &\leq L\|\mathbf{y} - \mathbf{x}\|_2^2 \int_0^1 \tau \, d\tau = \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 \end{aligned}$$

□

Geometric illustration of lower/upper Lipschitz bounds

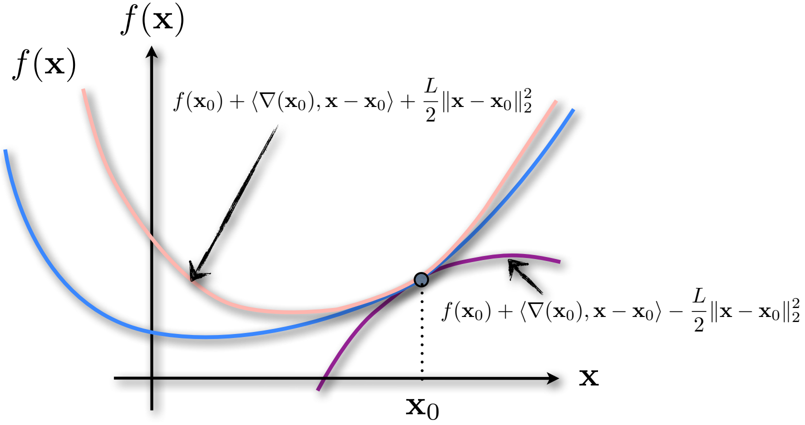
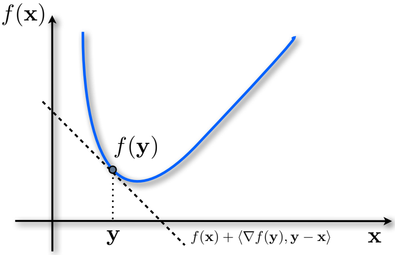


Figure: The function f is located between the lower quadratic $f(\mathbf{x}_0) + \langle \nabla(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle - \frac{L}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2$ and the upper quadratic $f(\mathbf{x}_0) + \langle \nabla(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2$.

Lipschitz continuity and Taylor series

- Let $f \in \mathcal{F}_L^2(\mathbb{R}^p)$ with gradient $\nabla f(\mathbf{x})$ and Hessian $\nabla^2 f(\mathbf{x})$.
- First-order Taylor approximation of f at $\mathbf{y}$:

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$$



- Convex functions: 1^{st}-order Taylor approximation is a global lower surrogate.

Lipschitz continuity and Taylor series approximation

- Let $f \in \mathcal{F}_L^2(\mathbb{R}^p)$ with gradient $\nabla f(\mathbf{x})$ and Hessian $\nabla^2 f(\mathbf{x})$.
- Second-order Taylor approximation of f at $\mathbf{y}$: there exists $\alpha \in [0, 1]$ such that

$$f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2} \langle \nabla^2 f(\mathbf{x} + \alpha(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

- By convexity and L-Lipschitz gradient assumption (Hessian is globally bounded):

$$\mathbf{0} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}$$

- Thus:

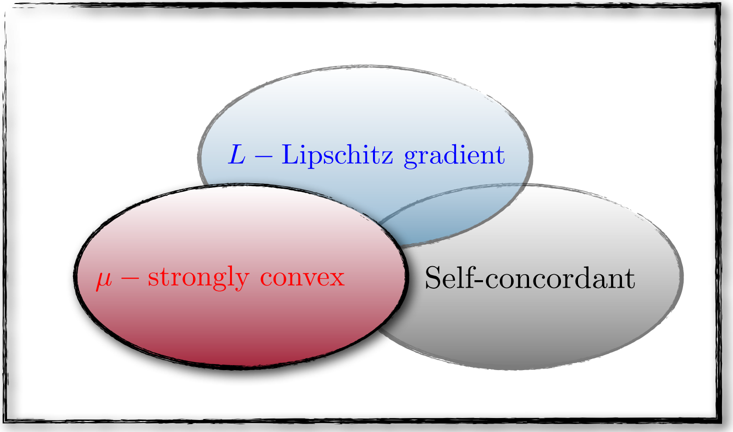
$$\begin{aligned} f(\mathbf{y}) &= f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2} \langle \nabla^2 f(\mathbf{x} + \alpha(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \\ &\leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2} \| \nabla^2 f(\mathbf{x} + \alpha(\mathbf{y} - \mathbf{x})) \|_{2 \to 2} \| \mathbf{y} - \mathbf{x} \|_2^2 \\ &\leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \| \mathbf{y} - \mathbf{x} \|_2^2 \end{aligned}$$

- Convex functions with L-Lipschitz gradient. We can use **2^{st}-order Taylor approximation to obtain a global upper surrogate.**

Classes of convex functions

Definition

We use $\mathcal{F}$ to denote the class of convex functions f. The domain of f will be apparent from the context.



μ-strongly convex functions

Definition

A function $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$, $\mathcal{Q} \subseteq \mathbb{R}^p$ is called μ-*strongly* convex on its domain if and only if for any $\mathbf{x}, \ \mathbf{y} \in \mathcal{Q}$ and $\alpha \in [0, 1]$ we have:

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

The constant μ is called the convexity parameter of function f. We denote the class of k-differentiable μ-strongly functions as $f \in \mathcal{F}_\mu^k(\mathcal{Q})$.

- Strong convexity $\Rightarrow$ strict convexity, BUT strict convexity $\not\Rightarrow$ strong convexity

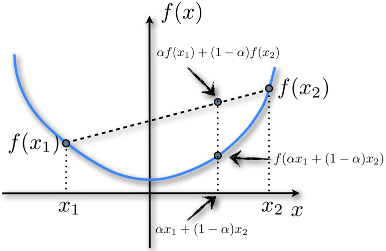


Figure: (**Left**) Convex (**Right**) Strongly convex

μ-strongly convex functions

Definition

A function $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$, $\mathcal{Q} \subseteq \mathbb{R}^p$ is called *μ-strongly* convex on its domain if and only if for any $\mathbf{x}, \ \mathbf{y} \in \mathcal{Q}$ and $\alpha \in [0, 1]$ we have:

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

The constant μ is called the convexity parameter of function f. We denote the class of k-differentiable μ-strongly functions as $f \in \mathcal{F}_\mu^k(\mathcal{Q})$.

- Strong convexity $\Rightarrow$ strict convexity, BUT strict convexity $\not\Rightarrow$ strong convexity

Example

Function $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2 + \|\mathbf{x}\|_1$ is non-differentiable but strongly convex.

Example

Function $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_{1.5}^2 + \|\mathbf{x}\|_1$ is non-differentiable, strictly convex but not strongly convex.

μ-strongly convex functions

Definition

A function $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}, \mathcal{Q} \subseteq \mathbb{R}^p$ is called *μ-strongly* convex on its domain if and only if for any $\mathbf{x},\ \mathbf{y} \in \mathcal{Q}$ and $\alpha \in [0,1]$ we have:

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

The constant μ is called the convexity parameter of function f. We denote the class of k-differentiable μ-strongly functions as $f \in \mathcal{F}_\mu^k(\mathcal{Q})$.

- Strong convexity $\Rightarrow$ strict convexity, BUT strict convexity $\not\Rightarrow$ strong convexity

Definition (Alternative definition)

Let $f : \mathcal{Q} \to \mathbb{R}$ be a convex function, i.e., $f \in \mathcal{F}(\mathcal{Q})$. Then, f is a μ-strongly convex function if and only if the following function is convex

$$h(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|_2^2 \quad \forall \mathbf{x} \in \mathcal{Q}.$$

Properties of μ-strongly convex functions

Lemma

Let $f : \mathcal{Q} \to \mathbb{R}, \mathcal{Q} \subseteq \mathbb{R}^p$ be a twice differentiable convex function, i.e., $f \in \mathcal{F}^2(\mathcal{Q})$. Then, f is μ-strongly convex function if and only if

$$\nabla^2 f(\mathbf{x}) \succeq \mu \mathbf{I}, \quad \forall \mathbf{x} \in \mathbb{R}^p.$$

Properties of μ-strongly convex functions

Lemma

Let $f : \mathcal{Q} \to \mathbb{R}, \mathcal{Q} \subseteq \mathbb{R}^p$ be a twice differentiable convex function, i.e., $f \in \mathcal{F}^2(\mathcal{Q})$. Then, f is μ-strongly convex function if and only if

$$\nabla^2 f(\mathbf{x}) \succeq \mu \mathbf{I}, \quad \forall \mathbf{x} \in \mathbb{R}^p.$$

Example (Toy example)

Consider the quadratic function $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$.
Then, f is a μ-strongly convex since
$\nabla^2 f(\mathbf{x}) = \mathbf{I} \quad \Longrightarrow \quad \mu = 1.$

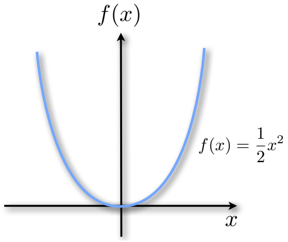


Figure: Toy example for μ-strongly convex functions.

Properties of μ-strongly convex functions

Lemma

Let $f : \mathcal{Q} \to \mathbb{R}, \mathcal{Q} \subseteq \mathbb{R}^p$ be a twice differentiable convex function, i.e., $f \in \mathcal{F}^2(\mathcal{Q})$. Then, f is μ-strongly convex function if and only if

$$\nabla^2 f(\mathbf{x}) \succeq \mu \mathbf{I}, \quad \forall \mathbf{x} \in \mathbb{R}^p.$$

Example (Overdetermined least squares)

Consider an *overdetermined* linear system of equations $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$ where $\mathbf{A} \in \mathbb{R}^{n \times p}$ is a full column-rank matrix and $\mathbf{x}^\natural$ is unknown. Assume that $\mathbf{A}^T\mathbf{A} \succeq \rho\mathbf{I}, \rho > 0$ and let $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2$. Then, f is a μ-strongly convex function, i.e., $f \in \mathcal{F}_\mu^2(\mathbb{R}^p)$ since:

$$\nabla^2 f(\mathbf{x}) = \mathbf{A}^T\mathbf{A} \quad \text{where} \quad \mathbf{A}^T\mathbf{A} \succeq \rho\mathbf{I} =: \mu\mathbf{I}.$$

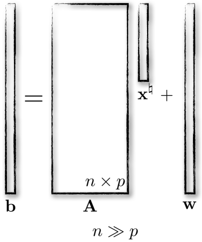


Figure: Overdetermined system of linear equations.

Properties of μ-strongly convex functions

Lemma

Let $f : \mathcal{Q} \to \mathbb{R}, \mathcal{Q} \subseteq \mathbb{R}^p$ be a twice differentiable convex function, i.e., $f \in \mathcal{F}^2(\mathcal{Q})$. Then, f is μ-strongly convex function if and only if

$$\nabla^2 f(\mathbf{x}) \succeq \mu \mathbf{I}, \quad \forall \mathbf{x} \in \mathbb{R}^p.$$

Example (Trivial)

Any linear function $f(\mathbf{x}) = \mathbf{c}^T\mathbf{x} + \beta \in \mathcal{F}^1_\mu(\mathbb{R}^p)$ for $\mu = 0$ since

$$\nabla f(\mathbf{x}) = \mathbf{c} \ \text{ and } \ \nabla^2 f(\mathbf{x}) = \mathbf{0}.$$

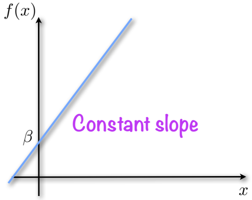


Figure: Counterexample for μ-strongly convex functions.

Properties of μ-strongly convex functions

Lemma

Let $f : \mathcal{Q} \to \mathbb{R}, \mathcal{Q} \subseteq \mathbb{R}^p$ be a twice differentiable convex function, i.e., $f \in \mathcal{F}^2(\mathcal{Q})$. Then, f is μ-strongly convex function if and only if

$$\nabla^2 f(\mathbf{x}) \succeq \mu \mathbf{I}, \quad \forall \mathbf{x} \in \mathbb{R}^p.$$

Lemma

A continuously differentiable function f belongs to $\mathcal{F}^1_\mu(\mathcal{Q})$ if there exists a constant $\mu > 0$ such that for any $\mathbf{x}, \mathbf{y} \in \mathcal{Q}$, we have:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|_2^2$$

Lemma

Let f be continuously differentiable. The following condition, holding for all $\mathbf{x}, \mathbf{y} \in \mathcal{Q} \subseteq \mathbb{R}^p$, is equivalent to inclusion that f is μ-strongly convex function:

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \mu \|\mathbf{x} - \mathbf{y}\|_{\#}^2$$

where $\#$ is the primal norm.

L-Lipschitz, μ-strongly convex functions

Definition

Let $f : \mathcal{Q} \to \mathbb{R}, \mathcal{Q} \subseteq \mathbb{R}^p$ be a continuously differentiable function. Then, f is both μ-strongly and L-Lipschitz convex function if for any $\mathbf{x}, \mathbf{y} \in \mathcal{Q}$, we have:

$$\frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 \leq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \leq \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$$

and

$$\mu\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}$$

for constants $0 < \mu \leq L$. We denote that $f \in \mathcal{F}^{1,1}_{\mu,L}(\mathcal{Q})$.

L-Lipschitz, μ-strongly convex functions

Definition

Let $f : \mathcal{Q} \to \mathbb{R}, \mathcal{Q} \subseteq \mathbb{R}^p$ be a continuously differentiable function. Then, f is both μ-strongly and L-Lipschitz convex function if for any $\mathbf{x}, \mathbf{y} \in \mathcal{Q}$, we have:

$$\frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 \leq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|_2^2$$

and

$$\mu \mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}$$

for constants $0 < \mu \leq L$. We denote that $f \in \mathcal{F}_{\mu,L}^{1,1}(\mathcal{Q})$.

Example

Consider an linear system of equations $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural$ where $\mu\mathbf{I} \preceq \mathbf{A}^T\mathbf{A} \preceq L\mathbf{I}$. Let $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2$. Then, f is both μ-strongly convex and L-Lipschitz continuous gradient function, i.e., $f \in \mathcal{F}_{\mu,L}^{2,1}(\mathbb{R}^p)$ since:

$$\nabla^2 f(\mathbf{x}) = \mathbf{A}^T\mathbf{A} \quad \text{where} \quad \mu\mathbf{I} \preceq \mathbf{A}^T\mathbf{A} \preceq L\mathbf{I}.$$

L-Lipschitz, μ-strongly convex functions

Definition

Let $f : \mathcal{Q} \to \mathbb{R}, \mathcal{Q} \subseteq \mathbb{R}^p$ be a continuously differentiable function. Then, f is both μ-strongly and L-Lipschitz convex function if for any $\mathbf{x}, \mathbf{y} \in \mathcal{Q}$, we have:

$$\frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 \leq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x} \rangle \leq \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$$

and

$$\mu \mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}$$

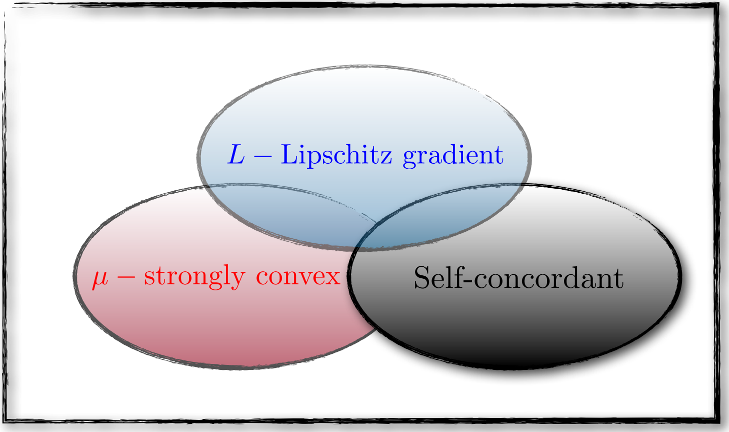
for constants $0 < \mu \leq L$. We denote that $f \in \mathcal{F}^{1,1}_{\mu,L}(\mathcal{Q})$.

- (As will be shown in next sections) μ, L are used in convergence rate characterization of actual algorithmic implementations
- Also used in stopping criteria
- **Unfortunately, μ, L are usually not known a priori...**

Classes of convex functions

Definition

We use $\mathcal{F}$ to denote the class of convex functions f. The domain of f will be apparent from the context.



Self-concordant functions

- Another key structure beyond

$$\mu \mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}$$

- We first explain the concept in the simple 1-dimensional setting...

Definition (Self-concordant functions in 1-dimension)

A convex function $\varphi : \mathbb{R} \to \mathbb{R}$ is self-concordant if

$$|\varphi'''(t)| \leq 2\varphi''(t)^{3/2}, \quad \forall t \in R.$$

- Motivation
  1. Conceptually, self-concordance definition provides a complete convergence analysis for algorithmic solutions (e.g., Newton method) *without knowing such constants*.
  2. Self-concordance leads to convergence analysis which is *affine invariant*; i.e., does not depend on the coordinate basis selected.

Example

Linear and quadratic functions are self-concordant: their 3^{rd} derivative is by definition zero.

Self-concordant functions

- Self-concordance provides a way to control the 3^{rd} derivative of a function.

Lemma

CLAIM: *Let* $\tilde{\varphi}(t) = \varphi(\alpha t + \beta)$ *where* $\alpha \neq 0$. *Then,* $\tilde{\varphi}$ *is self-concordant iff* φ *is.*

Self-concordant functions

- Self-concordance provides a way to control the 3^{rd} derivative of a function.

Lemma

CLAIM: *Let* $\tilde{\varphi}(t) = \varphi(\alpha t + \beta)$ *where* $\alpha \neq 0$. *Then,* $\tilde{\varphi}$ *is self-concordant iff* φ *is.*

Proof.

To see this, observe that:

$$\tilde{\varphi}''(t) = \alpha^2 \varphi''(\alpha t + \beta), \quad \tilde{\varphi}'''(t) = \alpha^3 \varphi'''(\alpha t + \beta).$$

Then, by definition of the self-concordance,

$$|\tilde{\varphi}'''(t)| \leq 2\tilde{\varphi}(t)^{3/2} \Longrightarrow |\alpha^3 \varphi'''(\alpha t + \beta)| \leq 2 \left(\alpha^2 \varphi(\alpha t + \beta)\right)^{3/2}$$

□

Affine invariance!!!

Self-concordant functions in higher dimensions

Definition (Self-concordant functions)

A convex function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be self-concordant with parameter $M \geq 0$, if $|\varphi'''(t)| \leq M\varphi''(t)^{3/2}$, where $\varphi(t) := f(\mathbf{x} + t\mathbf{v})$ for all $t \in \mathbb{R}$, $\mathbf{x} \in \mathrm{dom} f$ and $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{x} + t\mathbf{v} \in \mathrm{dom} f$. When $M = 2$, the function f is said to be a *standard* self-concordant.

Self-concordant functions in higher dimensions

Definition (Self-concordant functions)

A convex function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be self-concordant with parameter $M \geq 0$, if $|\varphi'''(t)| \leq M\varphi''(t)^{3/2}$, where $\varphi(t) := f(\mathbf{x} + t\mathbf{v})$ for all $t \in \mathbb{R}$, $\mathbf{x} \in \mathrm{dom} f$ and $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{x} + t\mathbf{v} \in \mathrm{dom} f$. When $M = 2$, the function f is said to be a *standard* self-concordant.

Example

The function $f(x) = -\log x$ is self-concordant. To see this, observe:

$$f''(x) = 1/x^2, \quad f'''(x) = -2/x^3.$$

Thus:

$$\frac{|f'''(x)|}{2f''(x)^{3/2}} = \frac{2/x^3}{2(1/x^2)^{3/2}} = 1$$

Self-concordant functions in higher dimensions

Definition (Self-concordant functions)

A convex function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be self-concordant with parameter $M \geq 0$, if $|\varphi'''(t)| \leq M\varphi''(t)^{3/2}$, where $\varphi(t) := f(\mathbf{x} + t\mathbf{v})$ for all $t \in \mathbb{R}$, $\mathbf{x} \in \text{dom}f$ and $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{x} + t\mathbf{v} \in \text{dom}f$. When $M = 2$, the function f is said to be a *standard* self-concordant.

Example

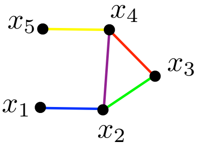
Similarly, the following example functions are self-concordant

1. $f(x) = x \log x - \log x$,
2. $f(x) = \sum_{i=1}^m \log(b_i - \mathbf{a}_i^T\mathbf{x})$ with domain $\text{dom}(f) = \left\{\mathbf{x} \ : \ \mathbf{a}_i^T\mathbf{x} < b_i, i = 1, \ldots, m\right\}$,
3. $f(\mathbf{X}) = -\log\det(\mathbf{X})$ with domain $\text{dom}(f) = \mathbb{S}_n^{++}$,
4. $f(\mathbf{x}) = -\log\left(\mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{q}^T\mathbf{x} + r\right)$ with domain $\text{dom}(f) = \left\{\mathbf{x} \ : \ \mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{q}^T\mathbf{x} + r > 0\right\}$ and $\mathbf{P} \in \mathbb{S}_n^{++}$.

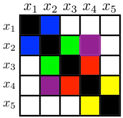
Example: Graphical model learning

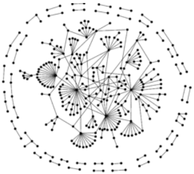
Problem (Graphical model selection)

Given a data set $\mathcal{D} := \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$*, where* $\mathbf{x}_i \in \mathbb{R}^p$ *(*$p \gg N$*) is a Gaussian random variable with sample covariance* $\widehat{\boldsymbol{\Sigma}}$*. Let* $\boldsymbol{\Sigma}$ *be the unknown covariance matrix corresponding to the graphical model of the Gaussian Markov random field. The aim is to learn a matrix* $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$*.*



$\Theta =$

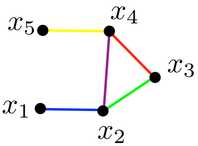




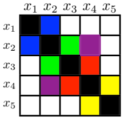
Example: Graphical model learning

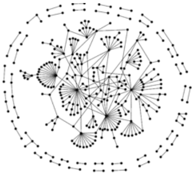
Problem (Graphical model selection)

Given a data set $\mathcal{D} := \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$, *where* $\mathbf{x}_i \in \mathbb{R}^p$ *(*$p \gg N$*) is a Gaussian random variable with sample covariance* $\widehat{\boldsymbol{\Sigma}}$. *Let* $\boldsymbol{\Sigma}$ *be the unknown covariance matrix corresponding to the graphical model of the Gaussian Markov random field. The aim is to learn a matrix* $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$.



$\Theta =$





Optimization formulation

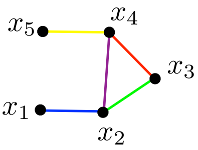
$$\min_{\boldsymbol{\Theta}} \left\{ \underbrace{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}) - \log\det(\boldsymbol{\Theta})}_{f(\boldsymbol{\Theta})} \right\}$$

where $f(\boldsymbol{\Theta})$ forces $\boldsymbol{\Theta}$ to be symmetric and positive definite through the *self-concordant* function $\log\det(\cdot)$.

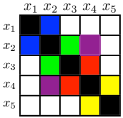
Example: Graphical model learning

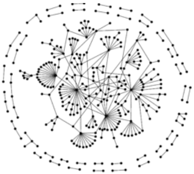
Problem (Graphical model selection)

Given a data set $\mathcal{D} := \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$, where $\mathbf{x}_i \in \mathbb{R}^p$ $(p \gg N)$ is a Gaussian random variable with sample covariance $\widehat{\boldsymbol{\Sigma}}$. Let $\boldsymbol{\Sigma}$ be the unknown covariance matrix corresponding to the graphical model of the Gaussian Markov random field. The aim is to learn a matrix $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$.



$\Theta =$





- $f(\boldsymbol{\Theta}) = \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Theta}) - \log\det(\boldsymbol{\Theta})$ is only *locally* Lipschitz continuous gradient function, restricted on a compact subset of $\mathbb{S}^p_{++}$.
- Observe that, for $\mathbf{X}, \mathbf{Y} \in \mathbb{S}^p_{++}$ where $\alpha\mathbf{I} \preceq \mathbf{X}, \mathbf{Y}, \preceq \beta\mathbf{I}$:

$$\|\nabla f(\mathbf{X}) - \nabla f(\mathbf{Y})\|_F = \|\mathbf{X}^{-1} - \mathbf{Y}^{-1}\|_F \leq \sqrt{p}\|\mathbf{X}^{-1} - \mathbf{Y}^{-1}\|_{2\to 2}$$
$$\leq \frac{\sqrt{p}}{\alpha^2}\|\mathbf{X} - \mathbf{Y}\|_{2\to 2} \leq \frac{\sqrt{p}}{\alpha^2}\|\mathbf{X} - \mathbf{Y}\|_F$$

Some geometric intuition behind self-concordant functions

Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Hessian surrogates	$\mu\mathbb{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbb{I}$	$\mathbf{x} \in \mathrm{dom}(f)$

Global

Some geometric intuition behind self-concordant functions

Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|_2^2$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Hessian surrogates	$\mu\mathbb{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbb{I}$	$\mathbf{x} \in \mathrm{dom}(f)$

Global

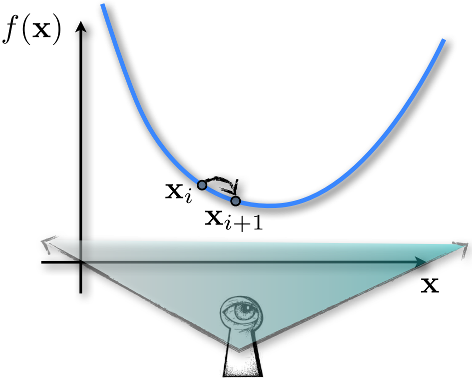


Figure: Global assumptions must hold *a priori* to operate with Lipschitz or μ-strongly convex machinery.

Some geometric intuition behind self-concordant functions

Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \omega\left(\\|\mathbf{y}-\mathbf{x}\\|_{\mathbf{x}}\right)$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \omega_*\left(\\|\mathbf{y}-\mathbf{x}\\|_{\mathbf{x}}\right)$	$\\|\mathbf{y}-\mathbf{x}\\|_{\mathbf{x}} < 1$
Hessian surrogates	$(1-\\|\mathbf{y}-\mathbf{x}\\|_{\mathbf{x}})^2\nabla^2 f(\mathbf{x}) \preceq \nabla^2 f(\mathbf{y}) \preceq (1-\\|\mathbf{y}-\mathbf{x}\\|_{\mathbf{x}})^{-2}\nabla^2 f(\mathbf{x})$	$\\|\mathbf{y}-\mathbf{x}\\|_{\mathbf{x}} < 1$

Local

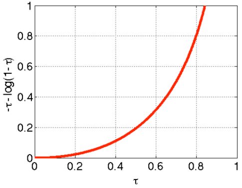
Some geometric intuition behind self-concordant functions

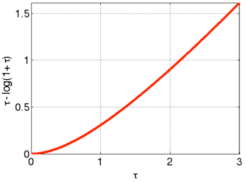
Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \omega\left(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right)$	$\mathbf{x},\mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \omega_*\left(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right)$	$\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < 1$
Hessian surrogates	$(1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^2\nabla^2 f(\mathbf{x}) \preceq \nabla^2 f(\mathbf{y}) \preceq (1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^{-2}\nabla^2 f(\mathbf{x})$	$\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < 1$

Local

Local norm: $\|\mathbf{u}\|_{\mathbf{x}} := \left[\mathbf{u}^T\nabla^2 f(\mathbf{x})\mathbf{u}\right]^{1/2}$

Utility functions: $\omega_*(\tau) = -\tau - \ln(1-\tau),\ \tau \in [0,1)$ $\qquad \omega(\tau) = \tau - \ln(1+\tau),\ \tau \geq 0$





Some geometric intuition behind self-concordant functions

Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \omega\left(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right)$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \omega_*\left(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}\right)$	$\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < 1$
Hessian surrogates	$(1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^2\nabla^2 f(\mathbf{x}) \preceq \nabla^2 f(\mathbf{y}) \preceq (1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^{-2}\nabla^2 f(\mathbf{x})$	$\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < 1$

Local

Definition

For any $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$, we have:

$$\langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle \geq \frac{\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^2}{1+\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}}$$

Some geometric intuition behind self-concordant functions

Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \omega(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \omega_*(\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})$	$\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < 1$
Hessian surrogates	$(1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^2\nabla^2 f(\mathbf{x}) \preceq \nabla^2 f(\mathbf{y}) \preceq (1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}})^{-2}\nabla^2 f(\mathbf{x})$	Local $\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}} < 1$

Definition

Let $\mathbf{x} \in \mathrm{dom}(f)$ and $\|\mathbf{x}-\mathbf{y}\|_{\mathbf{x}} < 1$. Then:

$$\langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x} \rangle \leq \frac{\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}^2}{1-\|\mathbf{y}-\mathbf{x}\|_{\mathbf{x}}}$$

Some geometric intuition behind self-concordant functions

Lower surrogate	$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \omega\left(\lVert\mathbf{y}-\mathbf{x}\rVert_{\mathbf{x}}\right)$	$\mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$
Upper surrogate	$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \omega_*\left(\lVert\mathbf{y}-\mathbf{x}\rVert_{\mathbf{x}}\right)$	$\lVert\mathbf{y}-\mathbf{x}\rVert_{\mathbf{x}} < 1$
Hessian surrogates	$(1-\lVert\mathbf{y}-\mathbf{x}\rVert_{\mathbf{x}})^2\nabla^2 f(\mathbf{x}) \preceq \nabla^2 f(\mathbf{y}) \preceq (1-\lVert\mathbf{y}-\mathbf{x}\rVert_{\mathbf{x}})^{-2}\nabla^2 f(\mathbf{x})$	$\lVert\mathbf{y}-\mathbf{x}\rVert_{\mathbf{x}} < 1$

Local

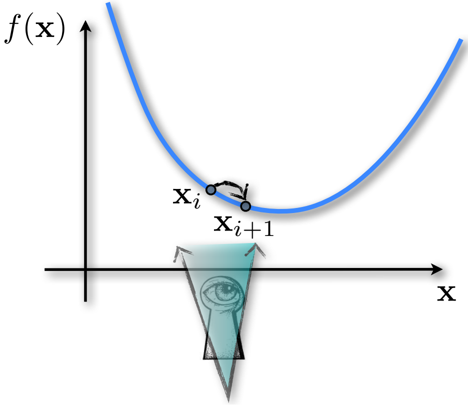
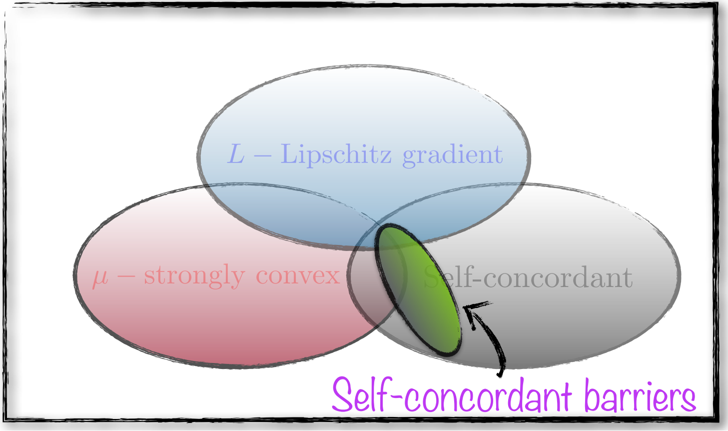


Figure: Only local information is used such to operate with self-concordant machinery.

Classes of convex functions

Definition

We use $\mathcal{F}$ to denote the class of convex functions f. The domain of f will be apparent from the context.



*Self-concordant barriers

- In the problems above, the self-concordant function $f(\cdot)$ appears in the objective function.

Problem

For our discussion, we consider the following constrained optimization problem

$$\begin{aligned} \underset{\mathbf{x}}{\textit{minimize}} \quad & g(\mathbf{x}) \\ \textit{subject to} \quad & \mathbf{x} \in \mathcal{Q} \end{aligned}$$

where $\mathcal{Q}$ is a closed convex set and is endowed with a self-concordant barrier.

- That is, we assume that we know a self-concordant function f such that $\text{dom}(f) \equiv \mathcal{Q}$.

Definition

A standard self-concordant function f is a ν-self-concordant barrier of a given convex set $\mathcal{Q}$ with parameter $\nu > 0$ if

$$\sup_{\mathbf{u} \in \mathbb{R}^p} \left\{ 2\mathbf{u}^T \nabla f(\mathbf{x}) - \mathbf{u}^T \nabla^2 f(\mathbf{x}) \mathbf{u} \right\} \leq \nu, \quad \forall \mathbf{x} \in \text{dom}(f).$$

[2] This material will be covered again in the next lectures

*Self-concordant barriers

- Used in sequential **unconstrained** minimization

Problem

We define the following parametric penalty function:

$$F(t;\ \mathbf{x}) = f(\mathbf{x}) + tg(\mathbf{x})$$

and solve the following sequential problem for increasing values of t*:*

$$\underset{\mathbf{x}\in dom(f)}{\textit{minimize}} \quad F(t;\ \mathbf{x}).$$

- Intuition: we expect $\mathbf{x}^{\natural}(t) \to \mathbf{x}$ (optimal solution) as $t \to \infty$.

Example

- All linear and convex quadratic functions are not self-concordant barriers.
- $f(\mathbf{x}) := -\sum_{i=1}^{p} \log(x_i)$ is an p-self-concordant barrier of the orthogonal cone $\mathbb{R}^p_+$.
- $f(\mathbf{x}, u) = -\log(u^2 - \|\mathbf{x}\|_2^2)$ is a 2-self-concordant barrier of the Lorentz cone $\mathcal{L}_{p+1} := \{(\mathbf{x}, u) \in \mathbb{R}^p \times \mathbb{R}_+ \mid \|\mathbf{x}\|_2 \leq u\}$.
- The semidefinite cone $\mathbb{S}^p_+$ is endowed with an p-self-concordant barrier $f(\mathbf{X}) := -\log\det(\mathbf{X})$.

[3] This material will be covered again in the next lectures

References

[1] P. J. Huber and E. M. Ronchetti, *Robust Statistics*. Hoboken, NJ: John Wiley & Sons, 2009.

[2] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Inf. Theory*, vol. 10, no. 5, pp. 988–999, Sep. 1999.

[3] S. Oymak, C. Thrampoulidis, and B. Hassibi, "The squared-error of generalized LASSO: A precise analysis," 2013, arXiv:1311.0830v2 [cs.IT].

[4] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Electron. J. Stat.*, vol. 5, pp. 935–980, 2011.

[5] L. Fahrmeir and H. Kaufmann, "Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models," *Ann. Stat.*, vol. 13, no. 1, pp. 342–368, 1985.

[6] L. Le Cam, *Asymptotic methods in Statistical Decision Theory*. New York, NY: Springer-Verl., 1986.

[7] A. W. van der Vaart, *Asymptotic Statistics*. Cambridge, UK: Cambridge Univ. Press, 1998.

[8] W. James and C. Stein, "Estimation with quadratic loss," in *Proc. 4th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1. Univ. Calif. Press, 1961, pp. 361–379.

[9] E. J. Candès, "Modern statistical estimation via oracle inequalities," *Acta Numer.*, vol. 15, pp. 257–325, May 2006.

References

[10] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Ann. Stat.*, vol. 27, no. 5, pp. 1564–1599, 1999.

[11] B. Yu, "Assouad, Fano, and Le Cam," in *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, D. Pollard, T. Erik, and G. L. Yang, Eds. New York: Springer, 1997, pp. 423–435.

[12] G. Raskutti, M. J. Wainwright, and B. Yu, "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6976–6994, Oct. 2011.