Optimization for Machine Learning

(Introduction)

Suvrit Sra
Massachusetts Institute of Technology

PKU Summer School on Data Science (July 2017)

Course materials

- *http://suvrit.de/teaching.html*
- Some references:
  - *Introductory lectures on convex optimization* – Nesterov
  - *Convex optimization* – Boyd & Vandenberghe
  - *Nonlinear programming* – Bertsekas
  - *Convex Analysis* – Rockafellar
  - *Fundamentals of convex analysis* – Urruty, Lemaréchal
  - *Lectures on modern convex optimization* – Nemirovski
  - *Optimization for Machine Learning* – Sra, Nowozin, Wright
  - *Theory of Convex Optimization for Machine Learning* – Bubeck
  - *NIPS 2016 Optimization Tutorial* – Bach, Sra
- Some related courses:
  - EE227A, Spring 2013, (Sra, UC Berkeley)
  - 10-801, Spring 2014 (Sra, CMU)
  - EE364a,b (Boyd, Stanford)
  - EE236b,c (Vandenberghe, UCLA)
- Venues: NIPS, ICML, UAI, AISTATS, SIOPT, Math. Prog.

Lecture Plan

– Introduction (3 lectures)
– Problems and algorithms (5 lectures)
– Non-convex optimization, perspectives (2 lectures)

Introduction

Supervised machine learning

- **Data**: n observations $(x_i, y_i)_{i=1}^n \in \mathcal{X} \times \mathcal{Y}$
- **Prediction function**: $h(x, \theta) \in \mathbb{R}$ parameterized by $\theta \in \mathbb{R}^d$

Introduction

Supervised machine learning

- **Data**: n observations $(x_i, y_i)_{i=1}^n \in \mathcal{X} \times \mathcal{Y}$
- **Prediction function**: $h(x, \theta) \in \mathbb{R}$ parameterized by $\theta \in \mathbb{R}^d$
- **Motivating examples**:
  - **Linear predictions**: $h(x, \theta) = \theta^\top \Phi(x)$ using features $\Phi(x)$
  - **Neural networks**: $h(x, \theta) = \theta_m^\top \sigma(\theta_{m-1}^\top \sigma(\cdots \theta_2^\top \sigma(\theta_1^\top x))$
- Estimating θ parameters is an optimization problem

Introduction

Supervised machine learning

- **Data**: n observations $(x_i, y_i)_{i=1}^n \in \mathcal{X} \times \mathcal{Y}$
- **Prediction function**: $h(x, \theta) \in \mathbb{R}$ parameterized by $\theta \in \mathbb{R}^d$
- **Motivating examples**:
  - **Linear predictions**: $h(x, \theta) = \theta^\top \Phi(x)$ using features $\Phi(x)$
  - **Neural networks**: $h(x, \theta) = \theta_m^\top \sigma(\theta_{m-1}^\top \sigma(\cdots \theta_2^\top \sigma(\theta_1^\top x))$
- Estimating θ parameters is an optimization problem

Unsupervised and other ML setups

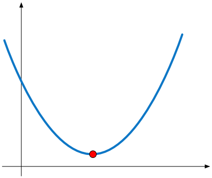
- Different formulations, but ultimately optimization at heart

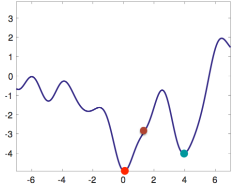
The Problem!

$$\min_{\theta \in \mathcal{S}} \quad f(\theta)$$

The Problem!

$$\min_{\theta \in \mathcal{S}} \quad f(\theta)$$

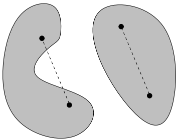




Convex analysis

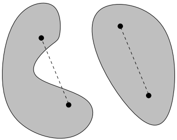
Convex sets

Def. Set $C \subset \mathbb{R}^n$ called **convex**, if for any $x, y \in C$, the line-segment $\lambda x + (1-\lambda)y$, where $\lambda \in [0,1]$, also lies in C.



Convex sets

Def. Set $C \subset \mathbb{R}^n$ called **convex**, if for any $x, y \in C$, the line-segment $\lambda x + (1-\lambda)y$, where $\lambda \in [0,1]$, also lies in C.



Combinations of points

- **Convex**: $\lambda_1 x + \lambda_2 y \in C$, where $\lambda_1, \lambda_2 \geq 0$ and $\lambda_1 + \lambda_2 = 1$.
- **Linear:** if restrictions on λ_1, λ_2 are dropped
- **Conic:** if restriction $\lambda_1 + \lambda_2 = 1$ is dropped

Different restrictions lead to different "algebra"

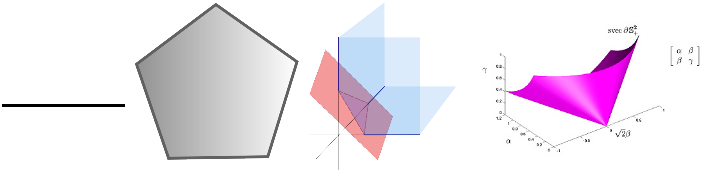
Recognizing / constructing convex sets

Theorem. (Intersection).
Let C_1, C_2 be convex sets. Then, $C_1 \cap C_2$ is also convex.

Proof.

- $\rightarrow$ If $C_1 \cap C_2 = \emptyset$, then true vacuously.
- $\rightarrow$ Let $x, y \in C_1 \cap C_2$. Then, $x, y \in C_1$ and $x, y \in C_2$.
- $\rightarrow$ But C_1, C_2 are convex, hence $\theta x + (1-\theta)y \in C_1$, and also in C_2. Thus, $\theta x + (1-\theta)y \in C_1 \cap C_2$.
- $\rightarrow$ Inductively follows that $\bigcap_{i=1}^{m} C_i$ is also convex.

Convex sets

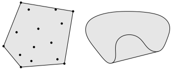


(psdcone image from convexoptimization.com, Dattorro)

Convex sets

♡ Let $x_1, x_2, \ldots, x_m \in \mathbb{R}^n$. Their **convex hull** is

$$\mathrm{co}(x_1, \ldots, x_m) := \left\{ \sum_i \theta_i x_i \mid \theta_i \geq 0, \sum_i \theta_i = 1 \right\}.$$

Example: 

♡ Let $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m$. The set $\{x \mid Ax = b\}$ is convex (it is an *affine space* over subspace of solutions of $Ax = 0$).

♡ *halfspace* $\{x \mid a^T x \leq b\}$.

♡ *polyhedron* $\{x \mid Ax \leq b, Cx = d\}$.

♡ *ellipsoid* $\{x \mid (x - x_0)^T A(x - x_0) \leq 1\}$, ($A$: semidefinite)

♡ *convex cone* $x \in \mathcal{K} \implies \alpha x \in \mathcal{K}$ for $\alpha \geq 0$ (and $\mathcal{K}$ convex)

Exercise: Verify that these sets are convex.

Challenge 1

Let $A, B \in \mathbb{R}^{n\times n}$ be symmetric. Prove that

$$R(A,B) := \left\{(x^TAx, x^TBx) \mid x^Tx = 1\right\}$$

is a compact convex set for $n \geq 3$.

Convex functions

Def. A function $f : \mathbb{R}^d \to \mathbb{R}$ is **convex** if and only if its *epigraph* $\{(x,t) \subseteq \mathbb{R}^{d+1} \mid x \in \mathbb{R}^d, t \in \mathbb{R}, f(x) \leq t\}$ is a convex set.

Convex functions

Def. A function $f : \mathbb{R}^d \to \mathbb{R}$ is **convex** if and only if its *epigraph* $\{(x,t) \subseteq \mathbb{R}^{d+1} \mid x \in \mathbb{R}^d, t \in \mathbb{R}, f(x) \le t\}$ is a convex set.

Def. Function $f : I \to \mathbb{R}$ on interval I called **midpoint convex** if

$$f\left(\frac{x+y}{2}\right) \le \frac{f(x)+f(y)}{2}, \qquad \text{whenever } x, y \in I.$$

Read: f of AM is less than or equal to AM of f.

Convex functions

Def. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called **convex** if its domain $\mathrm{dom}(f)$ is a convex set and for any $x, y \in \mathrm{dom}(f)$ and $\lambda \geq 0$,

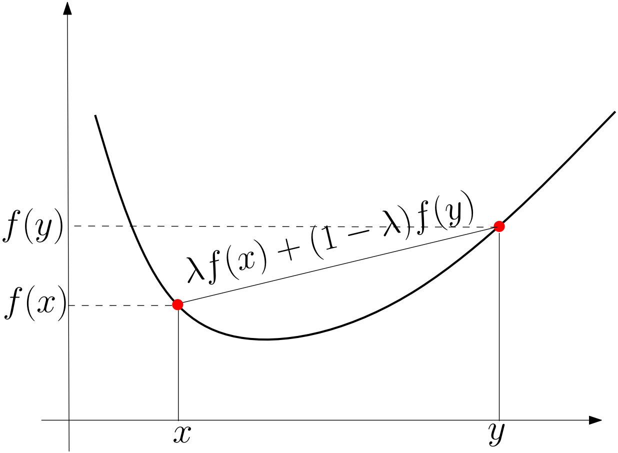
$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y).$$

These functions also known as **Jensen convex**; named after J.L.W.V. Jensen (after his influential 1905 paper).

Theorem. (J.L.W.V. Jensen). Let $f : I \to \mathbb{R}$ be continuous. Then, f is convex *if and only if* it is midpoint convex.

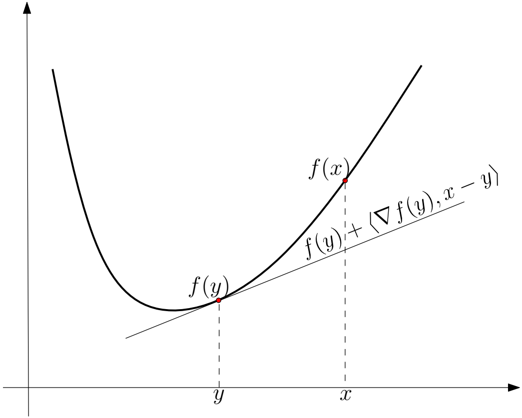
Exercise: Prove Jensen's theorem.

Convex functions: Jensen's inequality



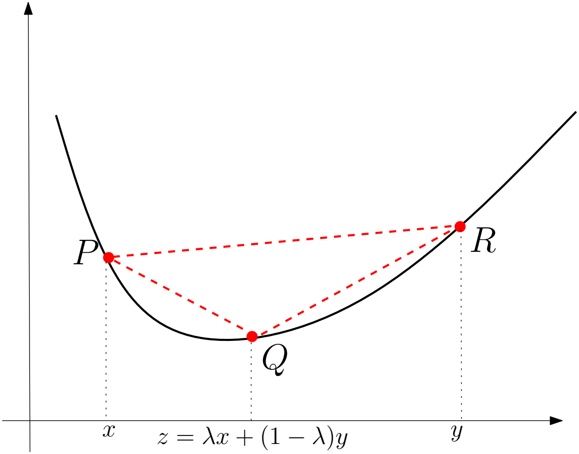
$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$

Convex functions: via gradients



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

Convex functions: increasing slopes



slope PQ $\leq$ slope PR $\leq$ slope QR

Recognizing convex functions

- ♠ If f is continuous and midpoint convex, then it is convex.
- ♠ If f is differentiable, then f is convex *if and only if* $\operatorname{dom} f$ is convex and $f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$ for all $x, y \in \operatorname{dom} f$.
- ♠ If f is twice differentiable, then f is convex *if and only if* $\operatorname{dom} f$ is convex and $\nabla^2 f(x) \succeq 0$ at every $x \in \operatorname{dom} f$.

Recognizing convex functions

- If f is continuous and midpoint convex, then it is convex.
- If f is differentiable, then f is convex *if and only if* $\operatorname{dom} f$ is convex and $f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$ for all $x, y \in \operatorname{dom} f$.
- If f is twice differentiable, then f is convex *if and only if* $\operatorname{dom} f$ is convex and $\nabla^2 f(x) \succeq 0$ at every $x \in \operatorname{dom} f$.
- By showing $f : \operatorname{dom}(f) \to \mathbb{R}$ is convex *if and only if* its restriction to **any** line that intersects $\operatorname{dom}(f)$ is convex. That is, for any $x \in \operatorname{dom}(f)$ and any v, the function $g(t) = f(x + tv)$ is convex (on its domain $\{t \mid x + tv \in \operatorname{dom}(f)\}$).

Recognizing convex functions

- If f is continuous and midpoint convex, then it is convex.
- If f is differentiable, then f is convex *if and only if* $\operatorname{dom} f$ is convex and $f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$ for all $x, y \in \operatorname{dom} f$.
- If f is twice differentiable, then f is convex *if and only if* $\operatorname{dom} f$ is convex and $\nabla^2 f(x) \succeq 0$ at every $x \in \operatorname{dom} f$.
- By showing $f : \operatorname{dom}(f) \to \mathbb{R}$ is convex *if and only if* its restriction to **any** line that intersects $\operatorname{dom}(f)$ is convex. That is, for any $x \in \operatorname{dom}(f)$ and any v, the function $g(t) = f(x + tv)$ is convex (on its domain $\{t \mid x + tv \in \operatorname{dom}(f)\}$).
- By showing f to be a pointwise max of convex functions
- See exercises (Ch. 3) in Boyd & Vandenberghe for more!

Example: Quadratic

Let $f(x) = x^T Ax + b^T x + c$, where $A \succeq 0$, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$.

Example: Quadratic

Let $f(x) = x^T Ax + b^T x + c$, where $A \succeq 0$, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$.

What is: $\nabla^2 f(x)$?

Example: Quadratic

Let $f(x) = x^T Ax + b^T x + c$, where $A \succeq 0$, $b \in \mathbb{R}^n$, and $c \in \mathbb{R}$.

What is: $\nabla^2 f(x)$?

$\nabla f(x) = 2Ax + b$, $\nabla^2 f(x) = A \succeq 0$, hence f is convex.

Examples

Exercise: Prove the convexity of the following functions in **at least** two different ways

1. $f(x, y) = x^2/y$ for $y > 0$ on $\mathbb{R} \times \mathbb{R}_{++}$
2. $f(x) = \log(1 + e^{\sum_i a_i x_i})$ on $\mathbb{R}^n$ ($a_i \in \mathbb{R}$ for $1 \leq i \leq n$).
3. Using 2 show that

$$\det(X + Y)^{1/n} \geq \det(X)^{1/n} + \det(Y)^{1/n}$$

for $X, Y \in \mathbb{S}^n_{++}$ (i.e., positive definite matrices).

4. **Challenge:** $f(X) = X^{-1}$ on positive definite matrices. (*This question is about convexity/concavity over matrices, so we have to replace the $\leq$ by the Löwner order $\preceq$*).

Operations preserving convexity

Example. Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. Let $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^m$. Prove that $g(x) = f(Ax + b)$ is convex.

Exercise: Verify!

Operations preserving convexity

Example. Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. Let $A \in \mathbb{R}^{m\times n}$, and $b \in \mathbb{R}^m$. Prove that $g(x) = f(Ax + b)$ is convex.

Exercise: Verify!

Theorem. Let $f : I_1 \to \mathbb{R}$ and $g : I_2 \to \mathbb{R}$, where range$(f) \subseteq I_2$. If f and g are convex, and g is increasing, then $g \circ f$ is convex on I_1

Operations preserving convexity

Example. Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. Let $A \in \mathbb{R}^{m\times n}$, and $b \in \mathbb{R}^m$. Prove that $g(x) = f(Ax + b)$ is convex.

Exercise: Verify!

Theorem. Let $f : I_1 \to \mathbb{R}$ and $g : I_2 \to \mathbb{R}$, where $\text{range}(f) \subseteq I_2$. If f and g are convex, and g is increasing, then $g \circ f$ is convex on I_1

Proof. Let $x, y \in I_1$, and let $\lambda \in (0, 1)$.

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &\leq \lambda f(x) + (1-\lambda)f(y) \\ g(f(\lambda x + (1-\lambda)y)) &\leq g\big(\lambda f(x) + (1-\lambda)f(y)\big) \\ &\leq \lambda g\big(f(x)\big) + (1-\lambda)g\big(f(y)\big). \end{aligned}$$

- Check out several other important examples in BV!

Constructing convex functions: sup

Example. The *pointwise maximum* of a family of convex functions is convex. That is, if $f(x; y)$ is a convex function of x for every y in an arbitrary "index set" $\mathcal{Y}$, then

$$f(x) := \sup_{y \in \mathcal{Y}} f(x; y)$$

is a convex function of x.

Exercise: Verify!

Example. The ℓ_∞-norm $\|x\|_\infty := \max_{1 \le i \le n} |x_i|$

Exercise: Prove that $|x|$ is a convex function.

Constructing convex functions: joint inf

Theorem. Let $\mathcal{Y}$ be a nonempty convex set. Suppose $L(x, y)$ is convex in **both** (x, y), then,

$$f(x) := \inf_{y \in \mathcal{Y}} \quad L(x, y)$$

is a convex function of x, provided $f(x) > -\infty$.

Constructing convex functions: joint inf

Theorem. Let $\mathcal{Y}$ be a nonempty convex set. Suppose $L(x, y)$ is convex in **both** (x, y), then,

$$f(x) := \inf_{y \in \mathcal{Y}} \quad L(x, y)$$

is a convex function of x, provided $f(x) > -\infty$.

Proof. Let $u, v \in \operatorname{dom} f$. Since $f(u) = \inf_y L(u, y)$, for each $\epsilon > 0$, there is a $y_1 \in \mathcal{Y}$, s.t. $f(u) + \frac{\epsilon}{2}$ is not the infimum. Thus, $L(u, y_1) \leq f(u) + \frac{\epsilon}{2}$.
Similarly, there is $y_2 \in \mathcal{Y}$, such that $L(v, y_2) \leq f(v) + \frac{\epsilon}{2}$.
Now we prove that $f(\lambda u + (1 - \lambda)v) \leq \lambda f(u) + (1 - \lambda)f(v)$ directly.

$$\begin{aligned} f(\lambda u + (1 - \lambda)v) &= \inf_{y \in \mathcal{Y}} L(\lambda u + (1 - \lambda)v, y) \\ &\leq L(\lambda u + (1 - \lambda)v, \lambda y_1 + (1 - \lambda)y_2) \\ &\leq \lambda L(u, y_1) + (1 - \lambda)L(v, y_2) \\ &\leq \lambda f(u) + (1 - \lambda)f(v) + \epsilon. \end{aligned}$$

Since $\epsilon > 0$ is arbitrary, claim follows.

Example: Schur complement

Let A, B, C be matrices such that $C \succ 0$, and let

$$Z := \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \succeq 0,$$

then the **Schur complement** $A - BC^{-1}B^T \succeq 0$.

Example: Schur complement

Let A, B, C be matrices such that $C \succ 0$, and let

$$Z := \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \succeq 0,$$

then the **Schur complement** $A - BC^{-1}B^T \succeq 0$.
Proof. $L(x,y) = [x,y]^T Z [x,y]$ is convex in (x,y) since $Z \succeq 0$

Example: Schur complement

Let A, B, C be matrices such that $C \succ 0$, and let

$$Z := \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \succeq 0,$$

then the **Schur complement** $A - BC^{-1}B^T \succeq 0$.

Proof. $L(x, y) = [x, y]^T Z [x, y]$ is convex in (x, y) since $Z \succeq 0$

Observe that $f(x) = \inf_y L(x, y) = x^T(A - BC^{-1}B^T)x$ is convex.

(We skipped ahead and solved $\nabla_y L(x, y) = 0$ to minimize).

Exercise: Verify the above example!

Convex functions – Indicator

Let $\mathbb{1}_{\mathcal{X}}$ be the *indicator function* for $\mathcal{X}$ defined as:

$$\mathbb{1}_{\mathcal{X}}(x) := \begin{cases} 0 & \text{if } x \in \mathcal{X}, \\ \infty & \text{otherwise.} \end{cases}$$

Note: $\mathbb{1}_{\mathcal{X}}(x)$ is convex **if and only if** $\mathcal{X}$ is convex.

- Also called "extended value" convex function.

Convex functions – norms

Let $\Omega : \mathbb{R}^d \to \mathbb{R}$ be a function that satisfies

1. $\Omega(x) \geq 0$, and $\Omega(x) = 0$ if and only if $x = 0$ (definiteness)
2. $\Omega(\lambda x) = |\lambda|\Omega(x)$ for any $\lambda \in \mathbb{R}$ (positive homogeneity)
3. $\Omega(x + y) \leq \Omega(x) + \Omega(y)$ (subadditivity)

Such function called *norms*—usually denoted $\|x\|$.

Theorem. Norms are convex.

Convex functions – norms

Let $\Omega : \mathbb{R}^d \to \mathbb{R}$ be a function that satisfies

1. $\Omega(x) \geq 0$, and $\Omega(x) = 0$ if and only if $x = 0$ (definiteness)
2. $\Omega(\lambda x) = |\lambda|\Omega(x)$ for any $\lambda \in \mathbb{R}$ (positive homogeneity)
3. $\Omega(x + y) \leq \Omega(x) + \Omega(y)$ (subadditivity)

Such function called *norms*—usually denoted $\|x\|$.

Theorem. Norms are convex.

Often used in "regularized" ML problems

$$\min_{\theta} \quad f(\theta) + \mu\Omega(\theta).$$

Norms and distances

Example. Let $\mathcal{X}$ be a convex set. Let $x \in \mathbb{R}^n$ be some point. The distance of x to the set $\mathcal{X}$ is defined as

$$\operatorname{dist}(x, \mathcal{X}) := \inf_{y \in \mathcal{X}} \quad \|x - y\|.$$

Exercise: Prove the above claim.
(*Hint:* argue that $\|x - y\|$ is jointly convex in (x, y))

Norms: important examples

Example. (ℓ_2-norm): $\|x\|_2 = \left(\sum_i x_i^2\right)^{1/2}$

Example. (ℓ_p-norm): Let $p \geq 1$. $\|x\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$

Example. (ℓ_∞-norm): $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$

Example. (Frobenius-norm): Let $A \in \mathbb{R}^{m \times n}$. $\|A\|_F := \sqrt{\sum_{ij} |a_{ij}|^2}$

Mixed norms

Def. Let $x \in \mathbb{R}^{n_1+n_2+\cdots+n_G}$ be a vector partitioned into subvectors $x_j \in \mathbb{R}^{n_j}$, $1 \le j \le G$. Let $\boldsymbol{p} := (p_0, p_1, p_2, \ldots, p_G)$, where $p_j \ge 1$. Consider the vector $\xi := (\|x_1\|_{p_1}, \cdots, \|x_G\|_{p_G})$. Then, we define the **mixed-norm** of x as

$$\|x\|_{\boldsymbol{p}} := \|\xi\|_{p_0}.$$

Mixed norms

Def. Let $x \in \mathbb{R}^{n_1+n_2+\cdots+n_G}$ be a vector partitioned into subvectors $x_j \in \mathbb{R}^{n_j}$, $1 \le j \le G$. Let $\boldsymbol{p} := (p_0, p_1, p_2, \ldots, p_G)$, where $p_j \ge 1$. Consider the vector $\xi := (\|x_1\|_{p_1}, \cdots, \|x_G\|_{p_G})$. Then, we define the **mixed-norm** of x as

$$\|x\|_{\boldsymbol{p}} := \|\xi\|_{p_0}.$$

Example. $\ell_{1,q}$-norm: Let x be as above.

$$\|x\|_{1,q} := \sum_{i=1}^{G} \|x_i\|_q.$$

This norm is popular in machine learning, statistics.

Matrix Norms

Induced norm

Let $A \in \mathbb{R}^{m \times n}$, and let $\|\cdot\|$ be any vector norm. We define an **induced matrix norm** as

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

Matrix Norms

Induced norm

Let $A \in \mathbb{R}^{m\times n}$, and let $\|\cdot\|$ be any vector norm. We define an **induced matrix norm** as

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

Verify that above definition yields a norm.

Matrix Norms

Induced norm

Let $A \in \mathbb{R}^{m\times n}$, and let $\|\cdot\|$ be any vector norm. We define an **induced matrix norm** as

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

Verify that above definition yields a norm.

- Clearly, $\|A\| = 0$ iff $A = 0$ (definiteness)
- $\|\alpha A\| = |\alpha|\ \|A\|$ (homogeneity)
- $\|A + B\| = \sup \frac{\|(A+B)x\|}{\|x\|} \leq \sup \frac{\|Ax\|+\|Bx\|}{\|x\|} \leq \|A\| + \|B\|.$

Operator norm

Example. Let A be any matrix. Then, the **operator norm** of A is

$$\|A\|_2 := \sup_{\|x\|_2 \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.$$

$\|A\|_2 = \sigma_{\max}(A)$, where $\sigma_{\max}$ is the largest singular value of A.

Operator norm

Example. Let A be any matrix. Then, the **operator norm** of A is

$$\|A\|_2 := \sup_{\|x\|_2 \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.$$

$\|A\|_2 = \sigma_{\max}(A)$, where $\sigma_{\max}$ is the largest singular value of A.

- Warning! Generally, largest eigenvalue **not** a norm!

Operator norm

Example. Let A be any matrix. Then, the **operator norm** of A is

$$\|A\|_2 := \sup_{\|x\|_2 \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.$$

$\|A\|_2 = \sigma_{\max}(A)$, where $\sigma_{\max}$ is the largest singular value of A.

- Warning! Generally, largest eigenvalue **not** a norm!
- $\|A\|_1$ and $\|A\|_\infty$—max-abs-column and max-abs-row sums.

Operator norm

Example. Let A be any matrix. Then, the **operator norm** of A is

$$\|A\|_2 := \sup_{\|x\|_2 \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.$$

$\|A\|_2 = \sigma_{\max}(A)$, where $\sigma_{\max}$ is the largest singular value of A.

- Warning! Generally, largest eigenvalue **not** a norm!
- $\|A\|_1$ and $\|A\|_\infty$—max-abs-column and max-abs-row sums.
- $\|A\|_p$ generally NP-Hard to compute for $p \notin \{1, 2, \infty\}$

Operator norm

Example. Let A be any matrix. Then, the **operator norm** of A is

$$\|A\|_2 := \sup_{\|x\|_2 \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.$$

$\|A\|_2 = \sigma_{\max}(A)$, where $\sigma_{\max}$ is the largest singular value of A.

- Warning! Generally, largest eigenvalue **not** a norm!
- $\|A\|_1$ and $\|A\|_\infty$—max-abs-column and max-abs-row sums.
- $\|A\|_p$ generally NP-Hard to compute for $p \notin \{1, 2, \infty\}$
- **Schatten p-norm:** ℓ_p-norm of vector of singular value.

Operator norm

Example. Let A be any matrix. Then, the **operator norm** of A is

$$\|A\|_2 := \sup_{\|x\|_2 \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.$$

$\|A\|_2 = \sigma_{\max}(A)$, where $\sigma_{\max}$ is the largest singular value of A.

- Warning! Generally, largest eigenvalue **not** a norm!
- $\|A\|_1$ and $\|A\|_\infty$—max-abs-column and max-abs-row sums.
- $\|A\|_p$ generally NP-Hard to compute for $p \notin \{1, 2, \infty\}$
- **Schatten p-norm:** ℓ_p-norm of vector of singular value.
- **Exercise:** Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$ be singular values of a matrix $A \in \mathbb{R}^{m \times n}$. Prove that

$$\|A\|_{(k)} := \sum\nolimits_{i=1}^{k} \sigma_i(A),$$

is a norm; $1 \leq k \leq n$.

Proof

Proof. By definition, the largest singular value is defined as

$$\sigma_{\max}(A) := \max_{x:\|x\|_2 \leq 1} \|Ax\|_2.$$

We saw that norms are convex. We also saw that for convex f, $f(Ax)$ is also convex. Thus, $\|Ax\|_2$ is convex.

Since the pointwise max of convex functions (over arbitrary index sets) is convex—here we index over $x \in \mathbb{R}^n$.

———————— ∘ ————————

Thus, $\sigma_{\max}(A)$ is a norm. It is denoted as $\|A\|_2$ or just $\|A\|$ — not to be confused with the Euclidean ℓ_2-norm of a vector!

Dual norms

Def. Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. Its **dual norm** is

$$\|u\|_* := \sup\left\{u^T x \mid \|x\| \leq 1\right\}.$$

Dual norms

Def. Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. Its **dual norm** is

$$\|u\|_* := \sup\left\{u^T x \mid \|x\| \le 1\right\}.$$

Exercise: Verify that we may write $\|u\|_* = \sup_{x\neq 0} \frac{u^T x}{\|x\|}$

Dual norms

Def. Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. Its **dual norm** is

$$\|u\|_* := \sup\left\{u^T x \mid \|x\| \leq 1\right\}.$$

Exercise: Verify that we may write $\|u\|_* = \sup_{x \neq 0} \frac{u^T x}{\|x\|}$

Exercise: Verify that $\|u\|_*$ is a norm.

- $\|u+v\|_* = \sup\left\{(u+v)^T x \mid \|x\| \leq 1\right\}$
- But $\sup\,(A+B) \leq \sup A + \sup B$

Exercise: Let $1/p + 1/q = 1$, where $p, q \geq 1$. Show that $\|\cdot\|_q$ is dual to $\|\cdot\|_p$. In particular, the ℓ_2-norm is self-dual.

Hint: Use *Hölder's inequality:* $u^T v \leq \|u\|_p \|v\|_q$

Challenge 2

Consider the following functions on strictly positive variables:

$$\begin{aligned} h_1(x) &:= \frac{1}{x} \\ h_2(x,y) &:= \frac{1}{x} + \frac{1}{y} - \frac{1}{x+y} \\ h_3(x,y,z) &:= \frac{1}{x} + \frac{1}{y} + \frac{1}{z} - \frac{1}{x+y} - \frac{1}{y+z} - \frac{1}{x+z} + \frac{1}{x+y+z} \end{aligned}$$

- ♡ Prove that $h_n(x) > 0$ (easy)
- ♡ Prove that h_1, h_2, h_3, and in general h_n are convex (hard)
- ♡ Prove that in fact each $1/h_n$ is concave (harder).

Fenchel conjugate

Def. The **Fenchel conjugate** of a function f is

$$f^*(z) := \sup_{x \in \operatorname{dom} f} \quad x^T z - f(x).$$

Exercise: Why is f^* convex? What if $f(x)$ is nonconvex?

Example. Let $f(x) = \|x\|$. We have $f^*(z) = \mathbb{1}_{\|\cdot\|_* \leq 1}(z)$. That is, conjugate of norm is the indicator function of dual norm ball.

Fenchel conjugate

Def. The **Fenchel conjugate** of a function f is

$$f^*(z) := \sup_{x \in \mathrm{dom} f} \quad x^T z - f(x).$$

Exercise: Why is f^* convex? What if $f(x)$ is nonconvex?

Example. Let $f(x) = \|x\|$. We have $f^*(z) = \mathbb{1}_{\|\cdot\|_* \leq 1}(z)$. That is, conjugate of norm is the indicator function of dual norm ball.

- Consider two cases: (i) $\|z\|_* > 1$; (ii) $\|z\|_* \leq 1$

Fenchel conjugate

Def. The **Fenchel conjugate** of a function f is

$$f^*(z) := \sup_{x \in \mathrm{dom} f} \quad x^T z - f(x).$$

Exercise: Why is f^* convex? What if $f(x)$ is nonconvex?

Example. Let $f(x) = \|x\|$. We have $f^*(z) = \mathbb{1}_{\|\cdot\|_* \leq 1}(z)$. That is, conjugate of norm is the indicator function of dual norm ball.

- Consider two cases: (i) $\|z\|_* > 1$; (ii) $\|z\|_* \leq 1$
- Case (i), by definition of dual norm (sup over $z^T u$) there is a u s.t. $\|u\| \leq 1$ and $z^T u > 1$

Fenchel conjugate

Def. The **Fenchel conjugate** of a function f is

$$f^*(z) := \sup_{x \in \mathrm{dom} f} \quad x^T z - f(x).$$

Exercise: Why is f^* convex? What if $f(x)$ is nonconvex?

Example. Let $f(x) = \|x\|$. We have $f^*(z) = \mathbb{1}_{\|\cdot\|_* \leq 1}(z)$. That is, conjugate of norm is the indicator function of dual norm ball.

- Consider two cases: (i) $\|z\|_* > 1$; (ii) $\|z\|_* \leq 1$
- Case (i), by definition of dual norm (sup over $z^T u$) there is a u s.t. $\|u\| \leq 1$ and $z^T u > 1$
- $f^*(z) = \sup_x x^T z - f(x)$. Rewrite $x = \alpha u$, and let $\alpha \to \infty$

Fenchel conjugate

Def. The **Fenchel conjugate** of a function f is

$$f^*(z) := \sup_{x \in \mathrm{dom} f} \quad x^T z - f(x).$$

Exercise: Why is f^* convex? What if $f(x)$ is nonconvex?

Example. Let $f(x) = \|x\|$. We have $f^*(z) = \mathbb{1}_{\|\cdot\|_* \leq 1}(z)$. That is, conjugate of norm is the indicator function of dual norm ball.

- Consider two cases: (i) $\|z\|_* > 1$; (ii) $\|z\|_* \leq 1$
- Case (i), by definition of dual norm (sup over $z^T u$) there is a u s.t. $\|u\| \leq 1$ and $z^T u > 1$
- $f^*(z) = \sup_x x^T z - f(x)$. Rewrite $x = \alpha u$, and let $\alpha \to \infty$
- Then, $z^T x - \|x\| = \alpha z^T u - \|\alpha u\| = \alpha(z^T u - \|u\|)$;

Fenchel conjugate

Def. The **Fenchel conjugate** of a function f is

$$f^*(z) := \sup_{x \in \mathrm{dom} f} \quad x^T z - f(x).$$

Exercise: Why is f^* convex? What if $f(x)$ is nonconvex?

Example. Let $f(x) = \|x\|$. We have $f^*(z) = \mathbb{1}_{\|\cdot\|_* \leq 1}(z)$. That is, conjugate of norm is the indicator function of dual norm ball.

- Consider two cases: (i) $\|z\|_* > 1$; (ii) $\|z\|_* \leq 1$
- Case (i), by definition of dual norm (sup over $z^T u$) there is a u s.t. $\|u\| \leq 1$ and $z^T u > 1$
- $f^*(z) = \sup_x x^T z - f(x)$. Rewrite $x = \alpha u$, and let $\alpha \to \infty$
- Then, $z^T x - \|x\| = \alpha z^T u - \|\alpha u\| = \alpha(z^T u - \|u\|); \to \infty$

Fenchel conjugate

Def. The **Fenchel conjugate** of a function f is

$$f^*(z) := \sup_{x \in \mathrm{dom} f} \quad x^T z - f(x).$$

Exercise: Why is f^* convex? What if $f(x)$ is nonconvex?

Example. Let $f(x) = \|x\|$. We have $f^*(z) = \mathbb{1}_{\|\cdot\|_* \le 1}(z)$. That is, conjugate of norm is the indicator function of dual norm ball.

- Consider two cases: (i) $\|z\|_* > 1$; (ii) $\|z\|_* \le 1$
- Case (i), by definition of dual norm (sup over $z^T u$) there is a u s.t. $\|u\| \le 1$ and $z^T u > 1$
- $f^*(z) = \sup_x x^T z - f(x)$. Rewrite $x = \alpha u$, and let $\alpha \to \infty$
- Then, $z^T x - \|x\| = \alpha z^T u - \|\alpha u\| = \alpha(z^T u - \|u\|); \to \infty$
- Case (ii): Since $z^T x \le \|x\| \|z\|_*$,

Fenchel conjugate

Def. The **Fenchel conjugate** of a function f is

$$f^*(z) := \sup_{x \in \mathrm{dom} f} \quad x^T z - f(x).$$

Exercise: Why is f^* convex? What if $f(x)$ is nonconvex?

Example. Let $f(x) = \|x\|$. We have $f^*(z) = \mathbb{1}_{\|\cdot\|_* \le 1}(z)$. That is, conjugate of norm is the indicator function of dual norm ball.

- Consider two cases: (i) $\|z\|_* > 1$; (ii) $\|z\|_* \le 1$
- Case (i), by definition of dual norm (sup over $z^T u$) there is a u s.t. $\|u\| \le 1$ and $z^T u > 1$
- $f^*(z) = \sup_x x^T z - f(x)$. Rewrite $x = \alpha u$, and let $\alpha \to \infty$
- Then, $z^T x - \|x\| = \alpha z^T u - \|\alpha u\| = \alpha(z^T u - \|u\|); \to \infty$
- Case (ii): Since $z^T x \le \|x\| \|z\|_*$, $\quad x^T z - \|x\| \le \|x\|(\|z\|_* - 1) \le 0$.

Fenchel conjugate

Def. The **Fenchel conjugate** of a function f is

$$f^*(z) := \sup_{x \in \mathrm{dom} f} \quad x^T z - f(x).$$

Exercise: Why is f^* convex? What if $f(x)$ is nonconvex?

Example. Let $f(x) = \|x\|$. We have $f^*(z) = \mathbb{1}_{\|\cdot\|_* \leq 1}(z)$. That is, conjugate of norm is the indicator function of dual norm ball.

- Consider two cases: (i) $\|z\|_* > 1$; (ii) $\|z\|_* \leq 1$
- Case (i), by definition of dual norm (sup over $z^T u$) there is a u s.t. $\|u\| \leq 1$ and $z^T u > 1$
- $f^*(z) = \sup_x x^T z - f(x)$. Rewrite $x = \alpha u$, and let $\alpha \to \infty$
- Then, $z^T x - \|x\| = \alpha z^T u - \|\alpha u\| = \alpha(z^T u - \|u\|)$; $\to \infty$
- Case (ii): Since $z^T x \leq \|x\| \|z\|_*$, $\;x^T z - \|x\| \leq \|x\|(\|z\|_* - 1) \leq 0$.
- $x = 0$ maximizes $\|x\|(\|z\|_* - 1)$, hence $f(z) = 0$.

Fenchel conjugate

Def. The **Fenchel conjugate** of a function f is

$$f^*(z) := \sup_{x \in \mathrm{dom} f} \quad x^T z - f(x).$$

Exercise: Why is f^* convex? What if $f(x)$ is nonconvex?

Example. Let $f(x) = \|x\|$. We have $f^*(z) = \mathbb{1}_{\|\cdot\|_* \le 1}(z)$. That is, conjugate of norm is the indicator function of dual norm ball.

- Consider two cases: (i) $\|z\|_* > 1$; (ii) $\|z\|_* \le 1$
- Case (i), by definition of dual norm (sup over $z^T u$) there is a u s.t. $\|u\| \le 1$ and $z^T u > 1$
- $f^*(z) = \sup_x x^T z - f(x)$. Rewrite $x = \alpha u$, and let $\alpha \to \infty$
- Then, $z^T x - \|x\| = \alpha z^T u - \|\alpha u\| = \alpha(z^T u - \|u\|); \to \infty$
- Case (ii): Since $z^T x \le \|x\| \|z\|_*$, $x^T z - \|x\| \le \|x\|(\|z\|_* - 1) \le 0$.
- $x = 0$ maximizes $\|x\|(\|z\|_* - 1)$, hence $f(z) = 0$.
- Thus, $f(z) = +\infty$ if (i), and 0 if (ii), as desired.

Fenchel conjugate

Example. $f(x) = ax + b$; then,

$$f^*(z) = \sup_x zx - (ax + b)$$

Fenchel conjugate

Example. $f(x) = ax + b$; then,

$$\begin{aligned} f^*(z) &= \sup_x zx - (ax + b) \\ &= \infty, \quad \text{if } (z - a) \neq 0. \end{aligned}$$

Fenchel conjugate

Example. $f(x) = ax + b$; then,

$$\begin{aligned} f^*(z) &= \sup_x zx - (ax + b) \\ &= \infty, \quad \text{if } (z - a) \neq 0. \end{aligned}$$

Thus, $\text{dom}\, f^* = \{a\}$, and $f^*(a) = -b$.

Fenchel conjugate

Example. $f(x) = ax + b$; then,

$$\begin{aligned} f^*(z) &= \sup_x zx - (ax + b) \\ &= \infty, \quad \text{if } (z - a) \neq 0. \end{aligned}$$

Thus, $\operatorname{dom} f^* = \{a\}$, and $f^*(a) = -b$.

Example. Let $a \geq 0$, and set $f(x) = -\sqrt{a^2 - x^2}$ if $|x| \leq a$, and $+\infty$ otherwise. Then, $f^*(z) = a\sqrt{1 + z^2}$.

Fenchel conjugate

Example. $f(x) = ax + b$; then,

$$\begin{aligned} f^*(z) &= \sup_x zx - (ax + b) \\ &= \infty, \quad \text{if } (z - a) \neq 0. \end{aligned}$$

Thus, $\operatorname{dom} f^* = \{a\}$, and $f^*(a) = -b$.

Example. Let $a \geq 0$, and set $f(x) = -\sqrt{a^2 - x^2}$ if $|x| \leq a$, and $+\infty$ otherwise. Then, $f^*(z) = a\sqrt{1 + z^2}$.

Example. $f(x) = \frac{1}{2}x^T Ax$, where $A \succ 0$. Then, $f^*(z) = \frac{1}{2}z^T A^{-1} z$.

Fenchel conjugate – exercises

Exercise: If $f(x) = \max(0, 1-x)$ (hinge loss) then $\mathrm{dom} f^*$ is $[-1, 0]$, and within this domain, $f^*(z) = z$.

If $f^{**} = f$, we say f is a closed convex function.

Exercise: Suppose $f(x) = (\sum_i |x_i|^{1/2})^2$. What is f^{**}?

Exercise: Suppose $f(x) = x^T Ax + b^T x$ but $A \succeq 0$; what is f^*?

Exercise: For which functions is $f^* = f$?

Optimization

Optimization problems

Let $f_i : \mathbb{R}^n \to \mathbb{R}$ $(0 \le i \le m)$. Generic **nonlinear program**

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.}\ & f_i(x) \le 0, \quad 1 \le i \le m, \\ & x \in \{\operatorname{dom} f_0 \cap \operatorname{dom} f_1 \cdots \cap \operatorname{dom} f_m\}\,. \end{aligned}$$

Henceforth, we drop condition on domains for brevity.

Optimization problems

Let $f_i : \mathbb{R}^n \to \mathbb{R}$ $(0 \leq i \leq m)$. Generic **nonlinear program**

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.}\ & f_i(x) \leq 0, \quad 1 \leq i \leq m, \\ & x \in \{\operatorname{dom} f_0 \cap \operatorname{dom} f_1 \cdots \cap \operatorname{dom} f_m\}\,. \end{aligned}$$

Henceforth, we drop condition on domains for brevity.

- If f_i are **differentiable** — smooth optimization
- If any f_i is **non-differentiable** — nonsmooth optimization
- If all f_i are **convex** — convex optimization
- If $m = 0$, i.e., only f_0 is there — **unconstrained** minimization

Convex optimization problems

Standard form

$$
\begin{aligned}
\min \quad & f_0(x) \\
\text{s.t.} \quad & f_i(x) \leq 0, \quad 1 \leq i \leq m, \\
& Ax = b.
\end{aligned}
$$

Convex optimization problems

Standard form

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, \quad 1 \leq i \leq m, \\ & Ax = b. \end{aligned}$$

Some observations

- All f_i are convex

Convex optimization problems

Standard form

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, \quad 1 \leq i \leq m, \\ & Ax = b. \end{aligned}$$

Some observations

- All f_i are convex
- Direction of inequality $f_i(x) \leq 0$ **crucial**

Convex optimization problems

Standard form

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, \quad 1 \leq i \leq m, \\ & Ax = b. \end{aligned}$$

Some observations

- All f_i are convex
- Direction of inequality $f_i(x) \leq 0$ **crucial**
- The only equality constraints we allow are affine

Convex optimization problems

Standard form

$$\begin{aligned} \min \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, \quad 1 \leq i \leq m, \\ & Ax = b. \end{aligned}$$

Some observations

- All f_i are convex
- Direction of inequality $f_i(x) \leq 0$ **crucial**
- The only equality constraints we allow are affine
- This ensures, set of feasible solutions is also **convex**

Convex optimization problems

Def. We denote by $\mathcal{X}$ the **feasible set**

$$\mathcal{X} := \{x \in \mathbb{R}^n \mid f_i(x) \leq 0,\ 1 \leq i \leq m,\ Ax = b\}.$$

Convex optimization problems

Def. We denote by $\mathcal{X}$ the **feasible set**

$$\mathcal{X} := \{x \in \mathbb{R}^n \mid f_i(x) \leq 0,\ 1 \leq i \leq m,\ Ax = b\}.$$

Def. We denote by p^* the **optimal value** of the problem.

$$p^* := \inf \{f_0(x) \mid x \in \mathcal{X}\}$$

Convex optimization problems

Def. We denote by $\mathcal{X}$ the **feasible set**

$$\mathcal{X} := \{x \in \mathbb{R}^n \mid f_i(x) \le 0,\ 1 \le i \le m,\ Ax = b\}.$$

Def. We denote by p^* the **optimal value** of the problem.

$$p^* := \inf \{f_0(x) \mid x \in \mathcal{X}\}$$

- If $\mathcal{X}$ is empty, we say problem is **infeasible**

Convex optimization problems

Def. We denote by $\mathcal{X}$ the **feasible set**

$$\mathcal{X} := \{x \in \mathbb{R}^n \mid f_i(x) \leq 0,\ 1 \leq i \leq m,\ Ax = b\}.$$

Def. We denote by p^* the **optimal value** of the problem.

$$p^* := \inf \{f_0(x) \mid x \in \mathcal{X}\}$$

- If $\mathcal{X}$ is empty, we say problem is **infeasible**
- By **convention**, we set $p^* = +\infty$ for infeasible problems

Convex optimization problems

Def. We denote by $\mathcal{X}$ the **feasible set**

$$\mathcal{X} := \{x \in \mathbb{R}^n \mid f_i(x) \leq 0,\ 1 \leq i \leq m,\ Ax = b\}.$$

Def. We denote by p^* the **optimal value** of the problem.

$$p^* := \inf \{f_0(x) \mid x \in \mathcal{X}\}$$

- If $\mathcal{X}$ is empty, we say problem is **infeasible**
- By **convention**, we set $p^* = +\infty$ for infeasible problems
- If $p^* = -\infty$, we say problem is **unbounded below**.

Convex optimization problems

Def. We denote by $\mathcal{X}$ the **feasible set**

$$\mathcal{X} := \{x \in \mathbb{R}^n \mid f_i(x) \leq 0,\ 1 \leq i \leq m,\ Ax = b\}.$$

Def. We denote by p^* the **optimal value** of the problem.

$$p^* := \inf \{f_0(x) \mid x \in \mathcal{X}\}$$

- If $\mathcal{X}$ is empty, we say problem is **infeasible**
- By **convention**, we set $p^* = +\infty$ for infeasible problems
- If $p^* = -\infty$, we say problem is **unbounded below**.
- Example, $\min x$ on $\mathbb{R}$, or $\min -\log x$ on $\mathbb{R}_{++}$

Convex optimization problems

Def. We denote by $\mathcal{X}$ the **feasible set**

$$\mathcal{X} := \{x \in \mathbb{R}^n \mid f_i(x) \leq 0,\ 1 \leq i \leq m,\ Ax = b\}.$$

Def. We denote by p^* the **optimal value** of the problem.

$$p^* := \inf \{f_0(x) \mid x \in \mathcal{X}\}$$

- If $\mathcal{X}$ is empty, we say problem is **infeasible**
- By **convention**, we set $p^* = +\infty$ for infeasible problems
- If $p^* = -\infty$, we say problem is **unbounded below**.
- Example, $\min x$ on $\mathbb{R}$, or $\min -\log x$ on $\mathbb{R}_{++}$
- Sometimes **minimum doesn't exist** (as $x \to \pm\infty$)

Convex optimization problems

Def. We denote by $\mathcal{X}$ the **feasible set**

$\mathcal{X} := \{x \in \mathbb{R}^n \mid f_i(x) \le 0,\ 1 \le i \le m,\ Ax = b\}.$

Def. We denote by p^* the **optimal value** of the problem.

$$p^* := \inf \{f_0(x) \mid x \in \mathcal{X}\}$$

- If $\mathcal{X}$ is empty, we say problem is **infeasible**
- By **convention**, we set $p^* = +\infty$ for infeasible problems
- If $p^* = -\infty$, we say problem is **unbounded below**.
- Example, $\min x$ on $\mathbb{R}$, or $\min -\log x$ on $\mathbb{R}_{++}$
- Sometimes **minimum doesn't exist** (as $x \to \pm\infty$)
- Say $f_0(x) = 0$, problem is called **convex feasibility**

Optimality

Def. A point $x^* \in \mathcal{X}$ is **locally optimal** if $f(x^*) \le f(x)$ for all x in a **neighborhood** of x^*. **Global** if $f(x^*) \le f(x)$ for **all** $x \in \mathcal{X}$.

Theorem. For convex problems, local $\implies$ global!

Exercise: Prove this theorem (*Hint:* try contradiction)

Optimality

Def. A point $x^* \in \mathcal{X}$ is **locally optimal** if $f(x^*) \leq f(x)$ for all x in a **neighborhood** of x^*. **Global** if $f(x^*) \leq f(x)$ for **all** $x \in \mathcal{X}$.

Theorem. For convex problems, local $\implies$ global!

Exercise: Prove this theorem (*Hint:* try contradiction)

- Let x^* be a local minimizer of $f(x)$ on $\mathcal{X}$ that is not global

Optimality

Def. A point $x^* \in \mathcal{X}$ is **locally optimal** if $f(x^*) \leq f(x)$ for all x in a **neighborhood** of x^*. **Global** if $f(x^*) \leq f(x)$ for **all** $x \in \mathcal{X}$.

Theorem. For convex problems, local $\implies$ global!

Exercise: Prove this theorem (*Hint:* try contradiction)

- Let x^* be a local minimizer of $f(x)$ on $\mathcal{X}$ that is not global
- Then there is a point $y \in \mathcal{X}$ such that $f(y) < f(x^*)$

Optimality

Def. A point $x^* \in \mathcal{X}$ is **locally optimal** if $f(x^*) \leq f(x)$ for all x in a **neighborhood** of x^*. **Global** if $f(x^*) \leq f(x)$ for **all** $x \in \mathcal{X}$.

Theorem. For convex problems, local $\implies$ global!

Exercise: Prove this theorem (*Hint:* try contradiction)

- Let x^* be a local minimizer of $f(x)$ on $\mathcal{X}$ that is not global
- Then there is a point $y \in \mathcal{X}$ such that $f(y) < f(x^*)$
- $\mathcal{X}$ is cvx., so we have $x_\theta = \theta y + (1-\theta)x^* \in \mathcal{X}$ for $\theta \in (0,1)$

Optimality

Def. A point $x^* \in \mathcal{X}$ is **locally optimal** if $f(x^*) \leq f(x)$ for all x in a **neighborhood** of x^*. **Global** if $f(x^*) \leq f(x)$ for **all** $x \in \mathcal{X}$.

Theorem. For convex problems, local $\implies$ global!

Exercise: Prove this theorem (*Hint:* try contradiction)

- Let x^* be a local minimizer of $f(x)$ on $\mathcal{X}$ that is not global
- Then there is a point $y \in \mathcal{X}$ such that $f(y) < f(x^*)$
- $\mathcal{X}$ is cvx., so we have $x_\theta = \theta y + (1-\theta)x^* \in \mathcal{X}$ for $\theta \in (0,1)$
- Since f is cvx, and $x^*, y \in \operatorname{dom} f$, we have

$$f(x_\theta) - f(x^*) \leq \theta(f(y) - f(x^*)).$$

Optimality

Def. A point $x^* \in \mathcal{X}$ is **locally optimal** if $f(x^*) \leq f(x)$ for all x in a **neighborhood** of x^*. **Global** if $f(x^*) \leq f(x)$ for **all** $x \in \mathcal{X}$.

Theorem. For convex problems, local $\implies$ global!

Exercise: Prove this theorem (*Hint:* try contradiction)

- Let x^* be a local minimizer of $f(x)$ on $\mathcal{X}$ that is not global
- Then there is a point $y \in \mathcal{X}$ such that $f(y) < f(x^*)$
- $\mathcal{X}$ is cvx., so we have $x_\theta = \theta y + (1-\theta)x^* \in \mathcal{X}$ for $\theta \in (0,1)$
- Since f is cvx, and $x^*, y \in \operatorname{dom} f$, we have

$$f(x_\theta) - f(x^*) \leq \theta(f(y) - f(x^*)).$$

- Since x^* is a local minimizer, for small enough $\theta > 0$, lhs ≥ 0.

Optimality

Def. A point $x^* \in \mathcal{X}$ is **locally optimal** if $f(x^*) \leq f(x)$ for all x in a **neighborhood** of x^*. **Global** if $f(x^*) \leq f(x)$ for **all** $x \in \mathcal{X}$.

Theorem. For convex problems, local $\implies$ global!

Exercise: Prove this theorem (*Hint:* try contradiction)

- Let x^* be a local minimizer of $f(x)$ on $\mathcal{X}$ that is not global
- Then there is a point $y \in \mathcal{X}$ such that $f(y) < f(x^*)$
- $\mathcal{X}$ is cvx., so we have $x_\theta = \theta y + (1-\theta)x^* \in \mathcal{X}$ for $\theta \in (0,1)$
- Since f is cvx, and $x^*, y \in \operatorname{dom} f$, we have

$$f(x_\theta) - f(x^*) \leq \theta(f(y) - f(x^*)).$$

- Since x^* is a local minimizer, for small enough $\theta > 0$, lhs ≥ 0.
- But the rhs is negative, which is a contradiction.

First-order optimality conditions

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable in an open set S containing x^*, a local min of f. Then, $\nabla f(x^*) = 0$.

Proof: Consider function $g(t) = f(x^* + td)$, where $d \in \mathbb{R}^n$; $t > 0$.
Since x^* is a local min, for small enough t, $f(x^* + td) \geq f(x^*)$.

First-order optimality conditions

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable in an open set S containing x^*, a local min of f. Then, $\nabla f(x^*) = 0$.

Proof: Consider function $g(t) = f(x^* + td)$, where $d \in \mathbb{R}^n$; $t > 0$. Since x^* is a local min, for small enough t, $f(x^* + td) \geq f(x^*)$.

First-order optimality conditions

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable in an open set S containing x^*, a local min of f. Then, $\nabla f(x^*) = 0$.

Proof: Consider function $g(t) = f(x^* + td)$, where $d \in \mathbb{R}^n$; $t > 0$. Since x^* is a local min, for small enough t, $f(x^* + td) \geq f(x^*)$.

$$0 \leq \quad \frac{f(x^* + td) - f(x^*)}{}$$

First-order optimality conditions

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable in an open set S containing x^*, a local min of f. Then, $\nabla f(x^*) = 0$.

Proof: Consider function $g(t) = f(x^* + td)$, where $d \in \mathbb{R}^n$; $t > 0$. Since x^* is a local min, for small enough t, $f(x^* + td) \geq f(x^*)$.

$$0 \leq \frac{f(x^* + td) - f(x^*)}{t}$$

First-order optimality conditions

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable in an open set S containing x^*, a local min of f. Then, $\nabla f(x^*) = 0$.

Proof: Consider function $g(t) = f(x^* + td)$, where $d \in \mathbb{R}^n$; $t > 0$. Since x^* is a local min, for small enough t, $f(x^* + td) \geq f(x^*)$.

$$0 \leq \lim_{t \downarrow 0} \frac{f(x^* + td) - f(x^*)}{t}$$

First-order optimality conditions

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable in an open set S containing x^*, a local min of f. Then, $\nabla f(x^*) = 0$.

Proof: Consider function $g(t) = f(x^* + td)$, where $d \in \mathbb{R}^n$; $t > 0$. Since x^* is a local min, for small enough t, $f(x^* + td) \geq f(x^*)$.

$$0 \leq \lim_{t \downarrow 0} \frac{f(x^* + td) - f(x^*)}{t} = \frac{dg(0)}{dt}$$

First-order optimality conditions

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable in an open set S containing x^*, a local min of f. Then, $\nabla f(x^*) = 0$.

Proof: Consider function $g(t) = f(x^* + td)$, where $d \in \mathbb{R}^n$; $t > 0$. Since x^* is a local min, for small enough t, $f(x^* + td) \geq f(x^*)$.

$$0 \leq \lim_{t \downarrow 0} \frac{f(x^* + td) - f(x^*)}{t} = \frac{dg(0)}{dt} = \langle \nabla f(x^*), d \rangle.$$

First-order optimality conditions

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable in an open set S containing x^*, a local min of f. Then, $\nabla f(x^*) = 0$.

Proof: Consider function $g(t) = f(x^* + td)$, where $d \in \mathbb{R}^n$; $t > 0$. Since x^* is a local min, for small enough t, $f(x^* + td) \geq f(x^*)$.

$$0 \leq \lim_{t \downarrow 0} \frac{f(x^* + td) - f(x^*)}{t} = \frac{dg(0)}{dt} = \langle \nabla f(x^*), d \rangle.$$

Similarly, using $-d$ it follows that $\langle \nabla f(x^*), d \rangle \leq 0$, so $\langle \nabla f(x^*), d \rangle = 0$ **must hold**. Since d is arbitrary, $\nabla f(x^*) = 0$.

First-order optimality conditions

Theorem. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable in an open set S containing x^*, a local min of f. Then, $\nabla f(x^*) = 0$.

Proof: Consider function $g(t) = f(x^* + td)$, where $d \in \mathbb{R}^n$; $t > 0$. Since x^* is a local min, for small enough t, $f(x^* + td) \geq f(x^*)$.

$$0 \leq \lim_{t \downarrow 0} \frac{f(x^* + td) - f(x^*)}{t} = \frac{dg(0)}{dt} = \langle \nabla f(x^*), d \rangle.$$

Similarly, using $-d$ it follows that $\langle \nabla f(x^*), d \rangle \leq 0$, so $\langle \nabla f(x^*), d \rangle = 0$ **must hold**. Since d is arbitrary, $\nabla f(x^*) = 0$.

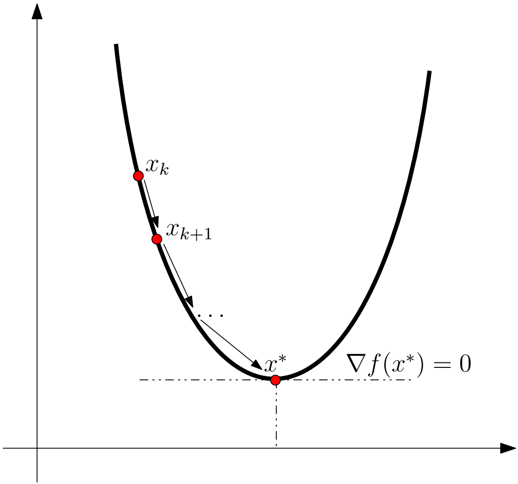
Exercise: Prove that if f is convex, then $\nabla f(x^*) = 0$ is actually **sufficient** for global optimality! For general f this is **not** true. (This property that makes convex optimization special!)

Descent methods

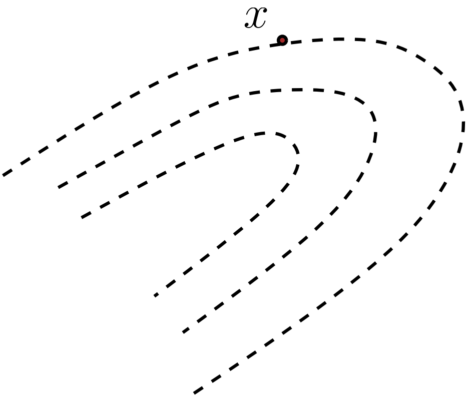
$$\min_x \quad f(x)$$

Descent methods

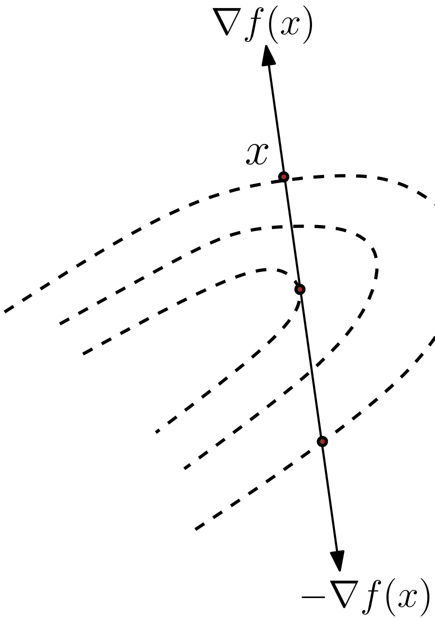
$$\min_x \quad f(x)$$



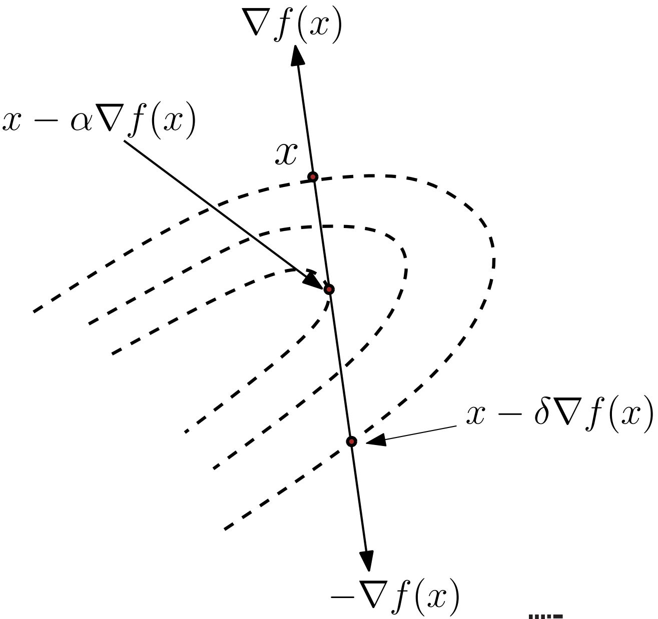
Descent methods



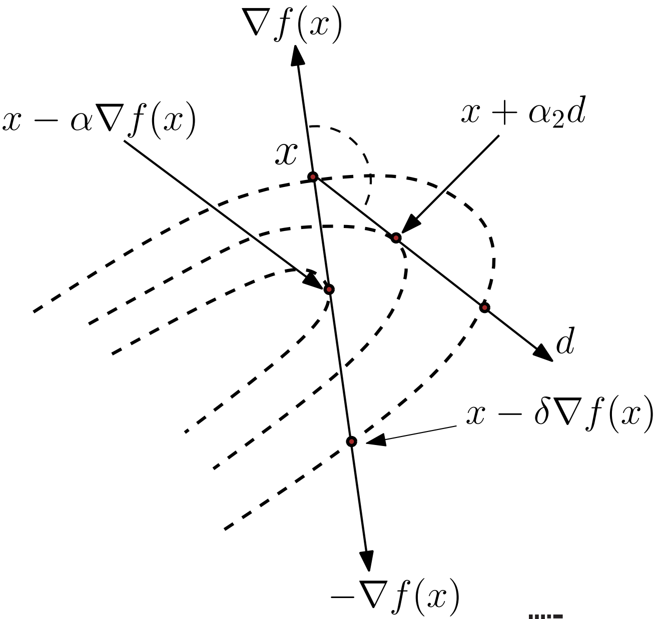
Descent methods



Descent methods



Descent methods



Iterative Algorithm

1. Start with some guess x^0;
2. For each $k = 0, 1, \ldots$
   - "Guess" α_k and d^k
   - $x^{k+1} \leftarrow x^k + \alpha_k d^k$
   - Check when to stop (e.g., if $\nabla f(x^{k+1}) \approx 0$)

(Batch) Gradient methods

$$x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots$$

- **stepsize** $\alpha_k \geq 0$, usually ensures $f(x^{k+1}) < f(x^k)$

(Batch) Gradient methods

$$x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots$$

- **stepsize** $\alpha_k \geq 0$, usually ensures $f(x^{k+1}) < f(x^k)$
- **Descent direction** d^k satisfies

$$\langle \nabla f(x^k), d^k \rangle < 0$$

(Batch) Gradient methods

$$x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots$$

- **stepsize** $\alpha_k \geq 0$, usually ensures $f(x^{k+1}) < f(x^k)$
- **Descent direction** d^k satisfies

$$\langle \nabla f(x^k), d^k \rangle < 0$$

Numerous ways to select α_k and d^k

(Batch) Gradient methods

$$x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots$$

- **stepsize** $\alpha_k \geq 0$, usually ensures $f(x^{k+1}) < f(x^k)$
- **Descent direction** d^k satisfies

$$\langle \nabla f(x^k), d^k \rangle < 0$$

Numerous ways to select α_k and d^k

Usually (batch) methods **seek monotonic descent**

$$f(x^{k+1}) < f(x^k)$$

Gradient methods – direction

$$x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots$$

- Different choices of direction d^k
  - **Scaled gradient:** $d^k = -D^k \nabla f(x^k)$, $D^k \succ 0$
  - **Newton's method:** $(D^k = [\nabla^2 f(x^k)]^{-1})$
  - **Quasi-Newton:** $D^k \approx [\nabla^2 f(x^k)]^{-1}$
  - **Steepest descent:** $D^k = \boldsymbol{I}$
  - **Diagonally scaled:** D^k diagonal with $D^k_{ii} \approx \left(\frac{\partial^2 f(x^k)}{(\partial x_i)^2}\right)^{-1}$
  - **Discretized Newton:** $D^k = [H(x^k)]^{-1}$, H via finite-diff.

Gradient methods – direction

$$x^{k+1} = x^k + \alpha_k d^k, \quad k = 0, 1, \ldots$$

- Different choices of direction d^k
  - **Scaled gradient:** $d^k = -D^k \nabla f(x^k)$, $D^k \succ 0$
  - **Newton's method:** $(D^k = [\nabla^2 f(x^k)]^{-1})$
  - **Quasi-Newton:** $D^k \approx [\nabla^2 f(x^k)]^{-1}$
  - **Steepest descent:** $D^k = \boldsymbol{I}$
  - **Diagonally scaled:** D^k diagonal with $D^k_{ii} \approx \left(\frac{\partial^2 f(x^k)}{(\partial x_i)^2}\right)^{-1}$
  - **Discretized Newton:** $D^k = [H(x^k)]^{-1}$, H via finite-diff.
  - …

Exercise: Verify that $\langle \nabla f(x^k), d^k \rangle < 0$ for above choices

Gradient methods – stepsize

- **Exact:** $\alpha_k := \underset{\alpha \geq 0}{\operatorname{argmin}} f(x^k + \alpha d^k)$

Gradient methods – stepsize

- **Exact:** $\alpha_k := \operatorname{argmin}_{\alpha \geq 0} f(x^k + \alpha d^k)$
- **Limited min:** $\alpha_k = \operatorname{argmin}_{0 \leq \alpha \leq s} f(x^k + \alpha d^k)$

Gradient methods – stepsize

- **Exact:** $\alpha_k := \operatorname{argmin}_{\alpha \geq 0} f(x^k + \alpha d^k)$
- **Limited min:** $\alpha_k = \operatorname{argmin}_{0 \leq \alpha \leq s} f(x^k + \alpha d^k)$
- **Armijo-rule**. Given **fixed** scalars, s, β, σ with $0 < \beta < 1$ and $0 < \sigma < 1$ (chosen experimentally). Set

$$\alpha_k = \beta^{m_k} s,$$

  where we **try** $\beta^m s$ for $m = 0, 1, \ldots$ until **sufficient descent**

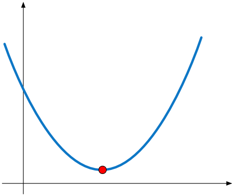
$$f(x^k) - f(x + \beta^m s d^k) \geq -\sigma \beta^m s \langle \nabla f(x^k), d^k \rangle$$

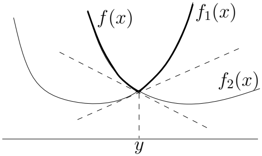
- **Constant:** $\alpha_k = 1/L$ (for suitable value of L)
- **Diminishing:** $\alpha_k \to 0$ but $\sum_k \alpha_k = \infty$.

Convergence

Assumption: Lipschitz continuous gradient; denoted $f \in C_L^1$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$

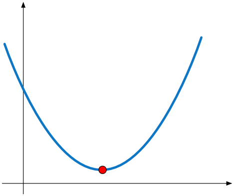


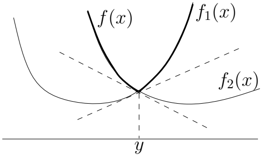


Convergence

Assumption: Lipschitz continuous gradient; denoted $f \in C_L^1$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$





- Gradient vectors of closeby points are close to each other
- Objective function has "bounded curvature"
- Speed at which gradient varies is bounded

Convergence

Assumption: Lipschitz continuous gradient; denoted $f \in C_L^1$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2$$

Lemma (Descent). Let $f \in C_L^1$. Then,

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

Theorem. Let $f \in C_L^1$ be convex, and $\{x^k\}$ is sequence generated as above, with $\alpha_k = 1/L$. Then, $f(x^{k+1}) - f(x^*) = O(1/k)$.

Remark: $f \in C_L^1$ is "good" for nonconvex too, except for $f - f^*$.

Strong convexity (faster convergence)

Assumption: Strong convexity; denote $f \in S^1_{L,\mu}$

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

- A twice diff. $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if

$$\forall x \in \mathbb{R}^d, \text{ eigenvalues}\left[\nabla^2 f(x)\right] \geqslant 0.$$

- A twice diff. $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex if and only if

$$\forall x \in \mathbb{R}^d, \text{ eigenvalues}\left[\nabla^2 f(x)\right] \geqslant \mu.$$

Strong convexity (faster convergence)

Assumption: **Strong convexity**; denote $f \in S^1_{L,\mu}$

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

- A twice diff. $f : \mathbb{R}^d \to \mathbb{R}$ is convex if and only if

$$\forall x \in \mathbb{R}^d, \text{ eigenvalues}\left[\nabla^2 f(x)\right] \geqslant 0.$$

- A twice diff. $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex if and only if

$$\forall x \in \mathbb{R}^d, \text{ eigenvalues}\left[\nabla^2 f(x)\right] \geqslant \mu.$$

Condition number: $\kappa := \frac{L}{\mu} \geq 1$ influences convergence speed.

Setting $\alpha_k = \frac{2}{\mu+L}$ yields linear rate ($\mu > 0$) for gradient descent. That is, $f(x^k) - f(x^*) = O(e^{-k})$.

Strong convexity – linear rate

Theorem. If $f \in S^1_{L,\mu}$, $0 < \alpha < 2/(L+\mu)$, then the gradient method generates a sequence $\{x^k\}$ that satisfies

$$\|x^k - x^*\|_2^2 \leq \left(1 - \frac{2\alpha\mu L}{\mu + L}\right)^k \|x^0 - x^*\|_2.$$

Moreover, if $\alpha = 2/(L+\mu)$ then

$$f(x^k) - f^* \leq \frac{L}{2}\left(\frac{\kappa - 1}{\kappa + 1}\right)^{2k} \|x^0 - x^*\|_2^2,$$

where $\kappa = L/\mu$ is the condition number.

Gradient methods – lower bounds

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k)$$

Theorem. Lower bound I (Nesterov) For any $x^0 \in \mathbb{R}^n$, and $1 \leq k \leq \frac{1}{2}(n-1)$, there is a smooth f, s.t.

$$f(x^k) - f(x^*) \geq \frac{3L\|x^0 - x^*\|_2^2}{32(k+1)^2}$$

Gradient methods – lower bounds

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k)$$

Theorem. Lower bound I (Nesterov) For any $x^0 \in \mathbb{R}^n$, and $1 \leq k \leq \frac{1}{2}(n-1)$, there is a smooth f, s.t.

$$f(x^k) - f(x^*) \geq \frac{3L\|x^0 - x^*\|_2^2}{32(k+1)^2}$$

Theorem. Lower bound II (Nesterov). For class of smooth, strongly convex, i.e., $S^{\infty}_{L,\mu}$ ($\mu > 0$, $\kappa > 1$)

$$f(x^k) - f(x^*) \geq \frac{\mu}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} \|x^0 - x^*\|_2^2.$$

Faster methods*

Optimal gradient methods

♠ We saw efficiency estimates for the gradient method:

$$f \in C_L^1: \qquad f(x^k) - f^* \leq \frac{2L\|x^0 - x^*\|_2^2}{k+4}$$

$$f \in S_{L,\mu}^1: \qquad f(x^k) - f^* \leq \tfrac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^{2k}\|x^0 - x^*\|_2^2.$$

Optimal gradient methods

- We saw efficiency estimates for the gradient method:

$$f \in C_L^1: \qquad f(x^k) - f^* \le \frac{2L\|x^0 - x^*\|_2^2}{k+4}$$

$$f \in S_{L,\mu}^1: \qquad f(x^k) - f^* \le \tfrac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^{2k}\|x^0 - x^*\|_2^2.$$

- We also saw **lower complexity bounds**

$$f \in C_L^1: \qquad f(x^k) - f(x^*) \ge \frac{3L\|x^0 - x^*\|_2^2}{32(k+1)^2}$$

$$fS_{L,\mu}^\infty: \qquad f(x^k) - f(x^*) \ge \frac{\mu}{2}\left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^{2k}\|x^0 - x^*\|_2^2.$$

Optimal gradient methods

- Subgradient method upper and lower bounds

$$f(x^k) - f(x^*) \leq O(1/\sqrt{k})$$
$$f(x^k) - f(x^*) \geq \frac{LD}{2(1+\sqrt{k+1})}.$$

- Composite objective problems: proximal gradient gives same bounds as gradient methods.

Gradient with "momentum"

Polyak's method (aka heavy-ball) for $f \in S^1_{L,\mu}$

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) + \beta_k (x^k - x^{k-1})$$

Gradient with "momentum"

Polyak's method (aka heavy-ball) for $f \in S^1_{L,\mu}$

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) + \beta_k (x^k - x^{k-1})$$

- **Converges** (locally, i.e., for $\|x^0 - x^*\|_2 \leq \epsilon$) as

$$\|x^k - x^*\|_2^2 \leq \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^{2k} \|x^0 - x^*\|_2^2,$$

for $\alpha_k = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ and $\beta_k = \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2$

Nesterov's optimal gradient method

$\min_x f(x)$, where $S^1_{L,\mu}$ with $\mu \geq 0$

Nesterov's optimal gradient method

$$\min_x f(x), \text{ where } S^1_{L,\mu} \text{ with } \mu \geq 0$$

1. Choose $x^0 \in \mathbb{R}^n$, $\alpha_0 \in (0, 1)$
2. Let $y^0 \leftarrow x^0$; set $q = \mu/L$

Nesterov's optimal gradient method

$\min_x f(x)$, where $S^1_{L,\mu}$ with $\mu \geq 0$

1. Choose $x^0 \in \mathbb{R}^n$, $\alpha_0 \in (0, 1)$
2. Let $y^0 \leftarrow x^0$; set $q = \mu/L$
3. k-th iteration ($k \geq 0$):
   a). Compute intermediate update

$$x^{k+1} = y^k - \tfrac{1}{L}\nabla f(y^k)$$

Nesterov's optimal gradient method

$\min_x f(x)$, where $S^1_{L,\mu}$ with $\mu \geq 0$

1. Choose $x^0 \in \mathbb{R}^n$, $\alpha_0 \in (0,1)$
2. Let $y^0 \leftarrow x^0$; set $q = \mu/L$
3. k-th iteration ($k \geq 0$):
   a). Compute intermediate update

$$x^{k+1} = y^k - \tfrac{1}{L}\nabla f(y^k)$$

   b). Compute stepsize α_{k+1} by solving

$$\alpha_{k+1}^2 = (1-\alpha_{k+1})\alpha_k^2 + q\alpha_{k+1}$$

Nesterov's optimal gradient method

$\min_x f(x)$, where $S^1_{L,\mu}$ with $\mu \geq 0$

1. Choose $x^0 \in \mathbb{R}^n$, $\alpha_0 \in (0,1)$
2. Let $y^0 \leftarrow x^0$; set $q = \mu/L$
3. k-th iteration ($k \geq 0$):
   a). Compute intermediate update

$$x^{k+1} = y^k - \tfrac{1}{L}\nabla f(y^k)$$

   b). Compute stepsize α_{k+1} by solving

$$\alpha_{k+1}^2 = (1-\alpha_{k+1})\alpha_k^2 + q\alpha_{k+1}$$

   c). Set $\beta_k = \alpha_k(1-\alpha_k)/(\alpha_k^2 + \alpha_{k+1})$
   d). Update solution estimate

$$y^{k+1} = x^{k+1} + \beta_k(x^{k+1} - x^k)$$

Optimal gradient method – rate

Theorem. Let $\{x^k\}$ be sequence generated by above algorithm. If $\alpha_0 \geq \sqrt{\mu/L}$, then

$$f(x^k) - f(x^*) \leq c_1 \min\left\{\left(1 - \sqrt{\frac{\mu}{L}}\right)^k, \frac{4L}{(2\sqrt{L} + c_2 k)^2}\right\},$$

where constants c_1, c_2 depend on α_0, L, μ.

Strongly convex case – simplification

If $\mu > 0$, select $\alpha_0 = \sqrt{\mu/L}$. The two main steps get simplified:

1. Set $\beta_k = \alpha_k(1-\alpha_k)/(\alpha_k^2 + \alpha_{k+1})$
2. $y^{k+1} = x^{k+1} + \beta_k(x^{k+1} - x^k)$

$$\alpha_k = \sqrt{\frac{\mu}{L}} \qquad \beta_k = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}, \qquad k \geq 0.$$

Strongly convex case – simplification

If $\mu > 0$, select $\alpha_0 = \sqrt{\mu/L}$. The two main steps get simplified:

1. Set $\beta_k = \alpha_k(1-\alpha_k)/(\alpha_k^2 + \alpha_{k+1})$
2. $y^{k+1} = x^{k+1} + \beta_k(x^{k+1} - x^k)$

$$\alpha_k = \sqrt{\frac{\mu}{L}} \qquad \beta_k = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}, \qquad k \geq 0.$$

Optimal method simplifies to

1. Choose $y^0 = x^0 \in \mathbb{R}^n$
2. k-th iteration ($k \geq 0$):

Strongly convex case – simplification

If $\mu > 0$, select $\alpha_0 = \sqrt{\mu/L}$. The two main steps get simplified:

1. Set $\beta_k = \alpha_k(1-\alpha_k)/(\alpha_k^2 + \alpha_{k+1})$
2. $y^{k+1} = x^{k+1} + \beta_k(x^{k+1} - x^k)$

$$\alpha_k = \sqrt{\frac{\mu}{L}} \qquad \beta_k = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}, \qquad k \geq 0.$$

Optimal method simplifies to

1. Choose $y^0 = x^0 \in \mathbb{R}^n$
2. k-th iteration ($k \geq 0$):
   a). $x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$
   b). $y^{k+1} = x^{k+1} + \beta(x^{k+1} - x^k)$

Strongly convex case – simplification

If $\mu > 0$, select $\alpha_0 = \sqrt{\mu/L}$. The two main steps get simplified:

1. Set $\beta_k = \alpha_k(1 - \alpha_k)/(\alpha_k^2 + \alpha_{k+1})$
2. $y^{k+1} = x^{k+1} + \beta_k(x^{k+1} - x^k)$

$$\alpha_k = \sqrt{\frac{\mu}{L}} \qquad \beta_k = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}, \qquad k \geq 0.$$

Optimal method simplifies to

1. Choose $y^0 = x^0 \in \mathbb{R}^n$
2. k-th iteration ($k \geq 0$):
   a). $x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$
   b). $y^{k+1} = x^{k+1} + \beta(x^{k+1} - x^k)$

Notice similarity to Polyak's method!