Optimization for Machine Learning

(Lecture 1)

Suvrit Sra

Massachusetts Institute of Technology

MPI-IS Tübingen

Machine Learning Summer School, June 2017

Course materials

- My website (Teaching)
- 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
  - *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
– Recap of convexity, sets, functions
– Recap of duality, optimality, problems
– First-order optimization algorithms and techniques
– Large-scale optimization (SGD and friends)
– Directions in non-convex optimization

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

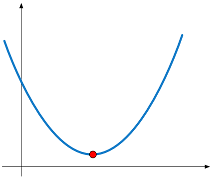
- Different formulations, but ultimately optimization at heart

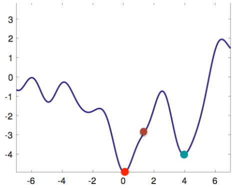
The Problem!

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

The Problem!

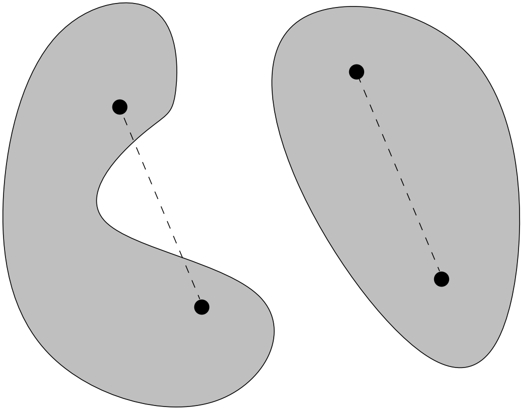
$$\min_{\theta \in \mathcal{S}} \quad f(\theta)$$





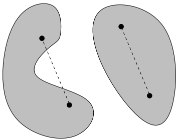
Convex analysis

Convex sets



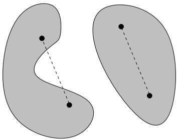
Convex sets

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"

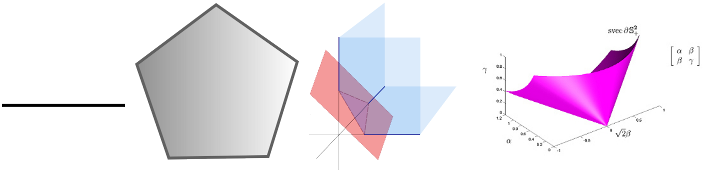
Recognizing / constructing convex sets

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

Proof.

→ If $C_1 \cap C_2 = \emptyset$, then true vacuously.
→ Let $x, y \in C_1 \cap C_2$. Then, $x, y \in C_1$ and $x, y \in C_2$.
→ But C_1, C_2 are convex, hence $\theta x + (1-\theta)y \in C_1$, and also in C_2. Thus, $\theta x + (1-\theta)y \in C_1 \cap C_2$.
→ 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

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

- ♡ 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* $\{(x,t) \subseteq \mathbb{R}^{d+1} \mid x \in \mathbb{R}^d, t \in \mathbb{R}, f(x) \leq t\}$ is a convex set.

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).

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.

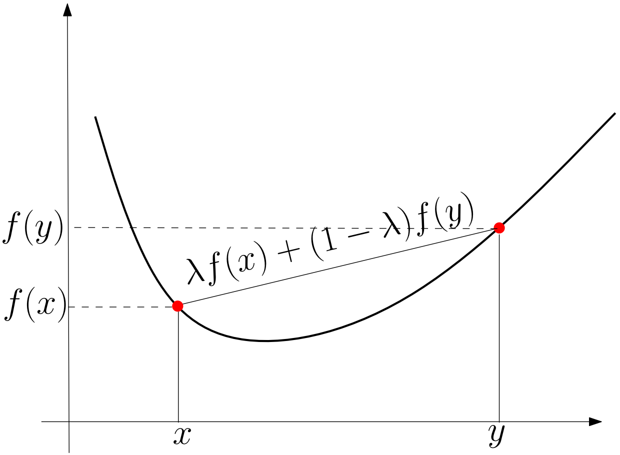
Def. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called **convex** if its domain $\mathrm{dom}(f)$ is a convex set and for any $x, y \in \mathrm{dom}(f)$ and $\lambda \ge 0$,

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

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

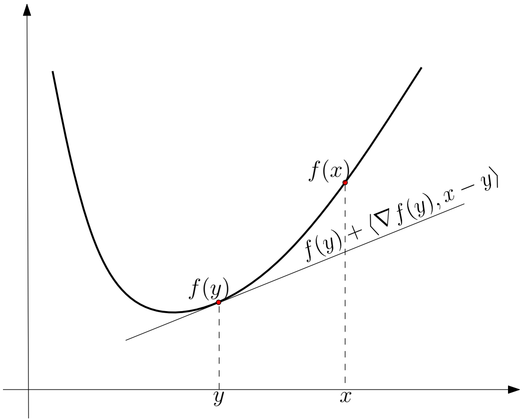
Exercise: Why are we focusing on these functions?

Convex functions: Jensen's inequality



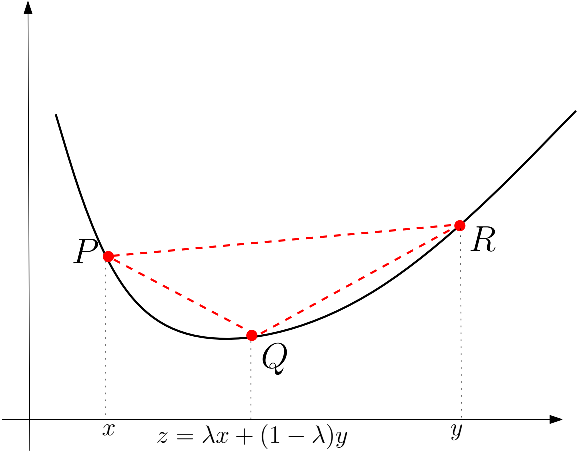
$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y)$$

Convex functions: affine lower bounds



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

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

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

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!

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 $\text{range}(f) \subseteq I_2$. If f and g are convex, and g is increasing, then $g \circ f$ is convex on I_1

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 $\mathrm{range}(f) \subseteq I_2$. 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 $\lambda \in (0, 1)$.

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

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

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

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

- Do not miss out on 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!

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!

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.

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: 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 $A \in \mathbb{R}^{m \times n}$. $\|A\|_F := \sqrt{\sum_{ij} |a_{ij}|^2}$

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

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

Exercise: Verify that above functions are actually norms!

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.

Fenchel conjugate

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

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

Fenchel conjugate

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

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

Note: f^* is pointwise (over x) sup of linear functions of z. Hence, it is always convex (even if f is not convex).

Example. $+\infty$ and $-\infty$ conjugate to each other.

Fenchel conjugate

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

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

Note: f^* is pointwise (over x) sup of linear functions of z. Hence, it is always convex (even if f is not convex).

Example. $+\infty$ and $-\infty$ conjugate to each other.

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.

Proof. $f^*(z) = \sup_x z^T x - \|x\|$. If $\|z\|_* > 1$, by defn. of the dual norm, $\exists u$ such that $\|u\| \le 1$ and $u^T z > 1$. Now select $x = \alpha u$ and let $\alpha \to \infty$. Then, $z^T x - \|x\| = \alpha(z^T u - \|u\|) \to \infty$. If $\|z\|_* \le 1$, then $z^T x \le \|x\| \|z\|_*$, which implies the sup must be zero.

Fenchel conjugate: examples

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

Example. $f(x) = \max(0, 1 - x)$. Verify: $\mathrm{dom}\, f^* = [-1, 0]$, and on this domain, $f^*(z) = z$.

Example. $f(x) = \mathbb{1}_{\mathcal{X}}(x)$: $f^*(z) = \sup_{x \in \mathcal{X}} \langle x,\, z \rangle$ (aka support func)

Example. 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^*?

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).

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 $f_i : \mathbb{R}^n \to \mathbb{R}$ $(0 \le i \le m)$. 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.

- 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

Let $\mathcal{X}$ be **feasible set** and p^* the **optimal value**

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

Convex optimization

Let $\mathcal{X}$ be **feasible set** and p^* the **optimal value**

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

- 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 $f(x^*) \leq f(x)$ 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)

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)

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

If f is convex, then $\nabla f(x^*) = 0$ **sufficient** for global optimality.
(This property makes convex optimization special!)

Optimality – constrained

♠ For every $x, y \in \operatorname{dom} f$, we have $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.

Optimality – constrained

- For every $x, y \in \operatorname{dom} f$, we have $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.
- Thus, x^* is optimal **if** and only if

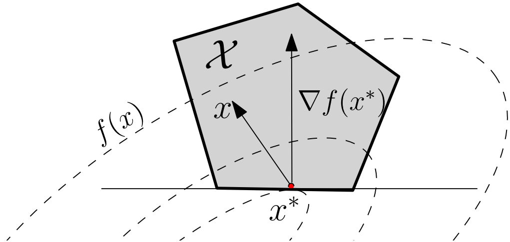
$$\langle \nabla f(x^*), y - x^* \rangle \geq 0, \qquad \text{for all } \; y \in \mathcal{X}.$$

Optimality – constrained

- For every $x, y \in \operatorname{dom} f$, we have $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.
- Thus, x^* is optimal **if** and only if

$$\langle \nabla f(x^*), y - x^* \rangle \geq 0, \qquad \text{for all } y \in \mathcal{X}.$$

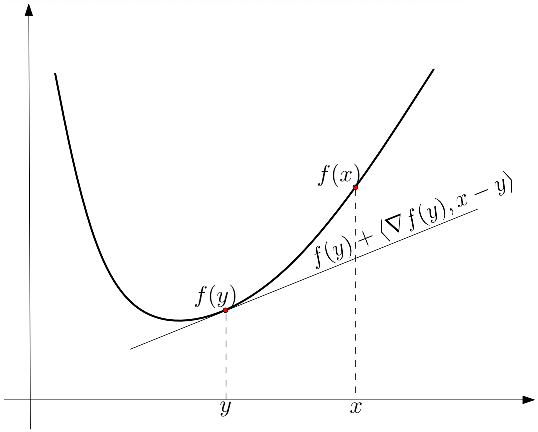
- If $\mathcal{X} = \mathbb{R}^n$, this reduces to $\nabla f(x^*) = 0$



- If $\nabla f(x^*) \neq 0$, it defines supporting hyperplane to $\mathcal{X}$ at x^*

Optimization: via subgradients

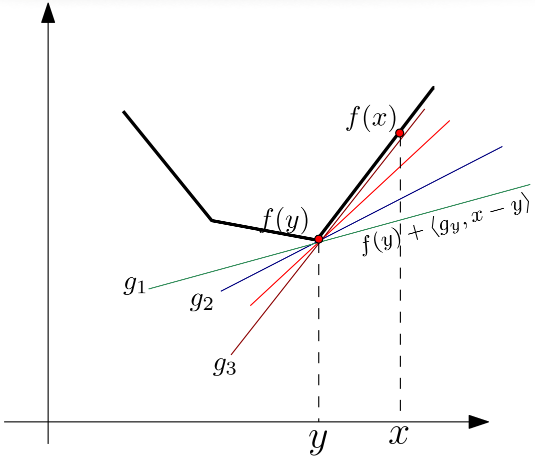
Subgradients: global underestimators



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

Hence $\nabla f(y) = 0$ implies that y is global min.

Subgradients: global underestimators



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

If one of the $g = 0$, then y a global min.

Subgradients – basic facts

- f is convex, differentiable: $\nabla f(y)$ the **unique** subgradient at y
- A vector g is a subgradient at a point y if and only if $f(y) + \langle g, x - y \rangle$ is **globally** smaller than $f(x)$.
- Usually, **one** subgradient costs approx. as much as $f(x)$

Subgradients – basic facts

- f is convex, differentiable: $\nabla f(y)$ the **unique** subgradient at y
- A vector g is a subgradient at a point y if and only if $f(y) + \langle g, x - y \rangle$ is **globally** smaller than $f(x)$.
- Usually, **one** subgradient costs approx. as much as $f(x)$
- Determining all subgradients at a given point — difficult.
- Subgradient calculus—major achievement in convex analysis
- Fenchel-Young inequality: $f(x) + f^*(s) \geq \langle s, x \rangle$ (tight at a subgradient)

Example: computing subgradients

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

Simple way to obtain some $g \in \partial f(x)$:

Example: computing subgradients

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

Simple way to obtain some $g \in \partial f(x)$:

- Pick any y^* for which $h(x, y^*) = f(x)$
- Pick any subgradient $g \in \partial h(x, y^*)$
- This $g \in \partial f(x)$

Proof:

$$\begin{aligned}
h(z, y^*) &\geq h(x, y^*) + g^T(z - x) \\
h(z, y^*) &\geq f(x) + g^T(z - x) \\
f(z) &\geq h(z, y) \qquad \text{(because of sup)} \\
f(z) &\geq f(x) + g^T(z - x).
\end{aligned}$$

Computing subgradients

Several other simple rules can be proved; see Boyd's lecture notes (or my EE227A lecture slides)

- Subgradient from max
- Subgradient from expectation
- Subgradient of composition

Subdifferential*

Subdifferential

Def. The set of all subgradients at y denoted by $\partial f(y)$. This set is called **subdifferential** of f at y

Subdifferential

Def. The set of all subgradients at y denoted by $\partial f(y)$. This set is called **subdifferential** of f at y

If f is convex, $\partial f(x)$ is nice:

♣ If $x \in$ relative interior of $\operatorname{dom} f$, then $\partial f(x)$ nonempty

Subdifferential

Def. The set of all subgradients at y denoted by $\partial f(y)$. This set is called **subdifferential** of f at y

If f is convex, $\partial f(x)$ is nice:

- ♣ If $x \in$ relative interior of $\mathrm{dom} f$, then $\partial f(x)$ nonempty
- ♣ If f differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$

Subdifferential

Def. The set of all subgradients at y denoted by $\partial f(y)$. This set is called **subdifferential** of f at y

If f is convex, $\partial f(x)$ is nice:

- ♣ If $x \in$ relative interior of $\operatorname{dom} f$, then $\partial f(x)$ nonempty
- ♣ If f differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$
- ♣ If $\partial f(x) = \{g\}$, then f is differentiable and $g = \nabla f(x)$

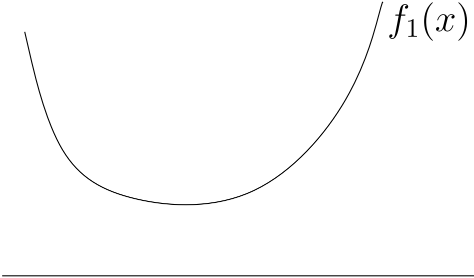
Exercise: What is $\partial f(x)$ for the *ReLU* function: $\max(0, x)$?

Subdifferential – example

$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable

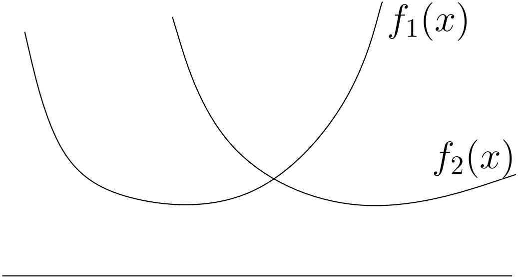
Subdifferential – example

$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable



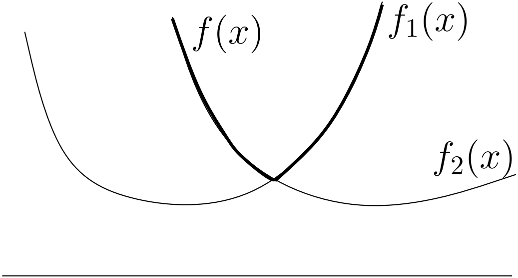
Subdifferential – example

$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable



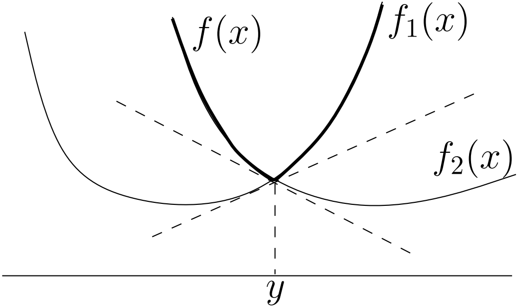
Subdifferential – example

$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable



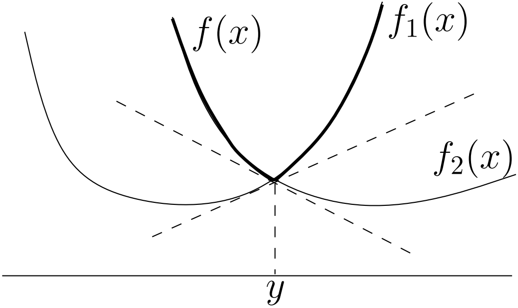
Subdifferential – example

$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable



Subdifferential – example

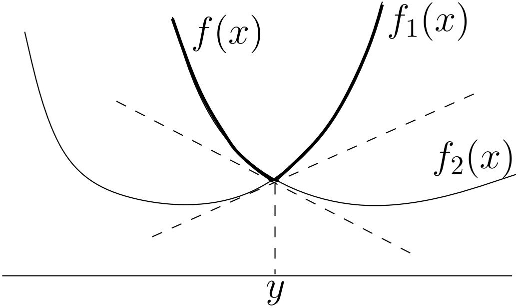
$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable



⋆ $f_1(x) > f_2(x)$: unique subgradient of f is $f_1'(x)$

Subdifferential – example

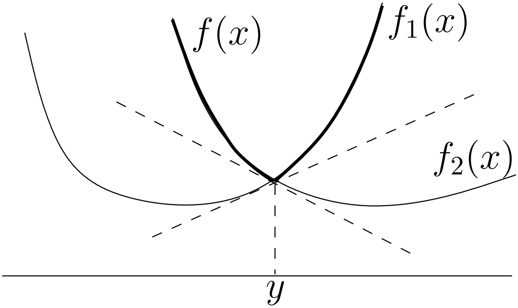
$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable



- ⋆ $f_1(x) > f_2(x)$: unique subgradient of f is $f_1'(x)$
- ⋆ $f_1(x) < f_2(x)$: unique subgradient of f is $f_2'(x)$

Subdifferential – example

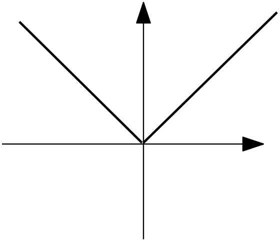
$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable



- ⋆ $f_1(x) > f_2(x)$: unique subgradient of f is $f_1'(x)$
- ⋆ $f_1(x) < f_2(x)$: unique subgradient of f is $f_2'(x)$
- ⋆ $f_1(y) = f_2(y)$: subgradients, the segment $[f_1'(y), f_2'(y)]$ (imagine all supporting lines turning about point y)

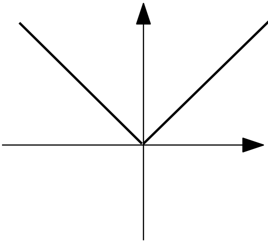
Subdifferential for abs value

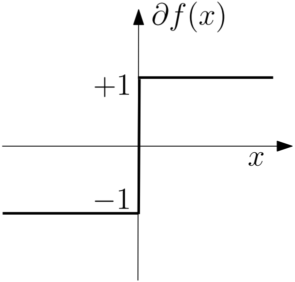
$$f(x) = |x|$$



Subdifferential for abs value

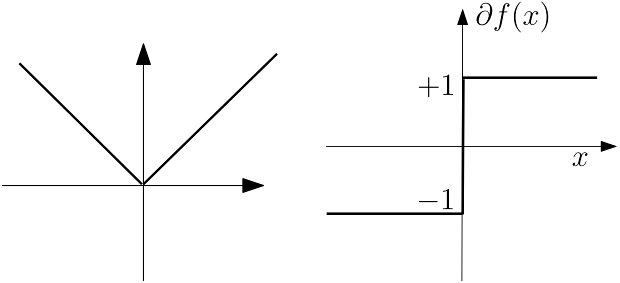
$$f(x) = |x|$$





Subdifferential for abs value

$$f(x) = |x|$$



$$\partial|x| = \begin{cases} -1 & x < 0, \\ +1 & x > 0, \\ [-1, 1] & x = 0. \end{cases}$$

Subdifferential for Euclidean norm

Example. $f(x) = \|x\|_2$. Then,

$$\partial f(x) := \begin{cases} x/\|x\|_2 & x \neq 0, \\ \{z \mid \|z\|_2 \leq 1\} & x = 0. \end{cases}$$

Subdifferential for Euclidean norm

Example. $f(x) = \|x\|_2$. Then,

$$\partial f(x) := \begin{cases} x/\|x\|_2 & x \neq 0, \\ \{z \mid \|z\|_2 \leq 1\} & x = 0. \end{cases}$$

Proof.

$$\begin{aligned} \|z\|_2 &\geq \|x\|_2 + \langle g,\, z - x\rangle \\ \|z\|_2 &\geq \langle g,\, z\rangle \\ &\implies \|g\|_2 \leq 1. \end{aligned}$$

Example: difficulties

Example. A convex function need not be subdifferentiable everywhere. Let

$$f(x) := \begin{cases} -(1 - \|x\|_2^2)^{1/2} & \text{if } \|x\|_2 \leq 1, \\ +\infty & \text{otherwise.} \end{cases}$$

f diff. for all x with $\|x\|_2 < 1$, but $\partial f(x) = \emptyset$ whenever $\|x\|_2 \geq 1$.

Subdifferential calculus

- Finding one subgradient within $\partial f(x)$
- Determining entire subdifferential $\partial f(x)$ at a point x
- Do we have the chain rule?

Subdifferential calculus

- If f is differentiable, $\partial f(x) = \{\nabla f(x)\}$
- **Scaling** $\alpha > 0$, $\partial(\alpha f)(x) = \alpha\partial f(x) = \{\alpha g \mid g \in \partial f(x)\}$
- **Addition***: $\partial(f+k)(x) = \partial f(x) + \partial k(x)$ (set addition)
- **Chain rule***: Let $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $f : \mathbb{R}^m \to \mathbb{R}$, and $h : \mathbb{R}^n \to \mathbb{R}$ be given by $h(x) = f(Ax+b)$. Then,

$$\partial h(x) = A^T \partial f(Ax+b).$$

- **Chain rule***: $h(x) = f \circ k$, where $k : X \to Y$ is diff.

$$\partial h(x) = \partial f(k(x)) \circ Dk(x) = [Dk(x)]^T \partial f(k(x))$$

- **Max function***: If $f(x) := \max_{1\le i\le m} f_i(x)$, then

$$\partial f(x) = \operatorname{conv} \bigcup \{\partial f_i(x) \mid f_i(x) = f(x)\},$$

  convex hull over subdifferentials of "active" functions at x
- **Conjugation:** $z \in \partial f(x)$ if and only if $x \in \partial f^*(z)$

* — can fail to hold without precise assumptions.

Example: breakdown

It can happen that $\partial(f_1 + f_2) \neq \partial f_1 + \partial f_2$

Example. Define f_1 and f_2 by

$$f_1(x) := \begin{cases} -2\sqrt{x} & \text{if } x \geq 0, \\ +\infty & \text{if } x < 0, \end{cases} \quad \text{and} \quad f_2(x) := \begin{cases} +\infty & \text{if } x > 0, \\ -2\sqrt{-x} & \text{if } x \leq 0. \end{cases}$$

Then, $f = \max\{f_1, f_2\} = \mathbb{1}_{\{0\}}$, whereby $\partial f(0) = \mathbb{R}$
But $\partial f_1(0) = \partial f_2(0) = \emptyset$.

However, $\partial f_1(x) + \partial f_2(x) \subset \partial(f_1 + f_2)(x)$ always holds.

Subdifferential – example

Example. $f(x) = \|x\|_\infty$. Then,

$$\partial f(0) = \operatorname{conv}\{\pm e_1, \ldots, \pm e_n\},$$

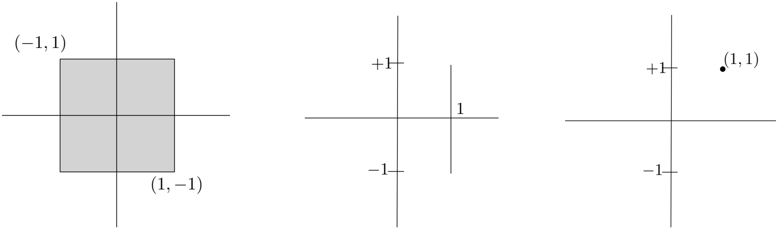
where e_i is i-th canonical basis vector.

To prove, notice that $f(x) = \max_{1 \le i \le n} \{|e_i^T x|\}$

Then use, *chain rule* and *max rule* and $\partial|\cdot|$

Subdifferential - example (Boyd)

Example. Let $f(x) = \max\{s^T x \mid s_i \in \{-1, 1\}\}$ (2^n members)



∂f at $x = (0, 0)$ $\quad$ ∂f at $x = (1, 0)$ $\quad$ ∂f at $x = (1, 1)$

Optimality via subdifferentials

Theorem. (Fermat's rule): Let $f : \mathbb{R}^n \to (-\infty, +\infty]$. Then,

$$\operatorname{argmin} f = \operatorname{zer}(\partial f) := \{x \in \mathbb{R}^n \mid 0 \in \partial f(x)\}.$$

Optimality via subdifferentials

Theorem. (Fermat's rule): Let $f : \mathbb{R}^n \to (-\infty, +\infty]$. Then,

$$\operatorname{argmin} f = \operatorname{zer}(\partial f) := \{x \in \mathbb{R}^n \mid 0 \in \partial f(x)\}.$$

Proof: $x \in \operatorname{argmin} f$ implies that $f(x) \leq f(y)$ for all $y \in \mathbb{R}^n$.

Optimality via subdifferentials

Theorem. (Fermat's rule): Let $f : \mathbb{R}^n \to (-\infty, +\infty]$. Then,

$$\operatorname{argmin} f = \operatorname{zer}(\partial f) := \{x \in \mathbb{R}^n \mid 0 \in \partial f(x)\}.$$

Proof: $x \in \operatorname{argmin} f$ implies that $f(x) \leq f(y)$ for all $y \in \mathbb{R}^n$.
Equivalently, $f(y) \geq f(x) + \langle 0, y - x \rangle \quad \forall y$,

Optimality via subdifferentials

Theorem. (Fermat's rule): Let $f : \mathbb{R}^n \to (-\infty, +\infty]$. Then,

$$\operatorname{argmin} f = \operatorname{zer}(\partial f) := \{x \in \mathbb{R}^n \mid 0 \in \partial f(x)\}.$$

Proof: $x \in \operatorname{argmin} f$ implies that $f(x) \le f(y)$ for all $y \in \mathbb{R}^n$.
Equivalently, $f(y) \ge f(x) + \langle 0, y - x \rangle \quad \forall y, \leftrightarrow 0 \in \partial f(x)$.

Optimality via subdifferentials

Theorem. (Fermat's rule): Let $f : \mathbb{R}^n \to (-\infty, +\infty]$. Then,

$$\operatorname{argmin} f = \operatorname{zer}(\partial f) := \{x \in \mathbb{R}^n \mid 0 \in \partial f(x)\}.$$

Proof: $x \in \operatorname{argmin} f$ implies that $f(x) \leq f(y)$ for all $y \in \mathbb{R}^n$.
Equivalently, $f(y) \geq f(x) + \langle 0, y - x \rangle \quad \forall y, \leftrightarrow 0 \in \partial f(x)$.

Nonsmooth optimality

$$\begin{aligned} &\min && f(x) \quad \text{s.t. } x \in \mathcal{X} \\ &\min && f(x) + \mathbb{1}_{\mathcal{X}}(x). \end{aligned}$$

Optimality via subdifferentials: application

- Minimizing x must satisfy: $0 \in \partial(f + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming $\mathrm{ri}(\mathrm{dom} f) \cap \mathrm{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f(x) + \partial \mathbb{1}_X(x)$
- Recall, $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ iff $\mathbb{1}_{\mathcal{X}}(y) \geq \mathbb{1}_{\mathcal{X}}(x) + \langle g, y - x \rangle$ for all y.
- So $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ means $x \in \mathcal{X}$ and $0 \geq \langle g, y - x \rangle \; \forall y \in \mathcal{X}$.
- **Normal cone:**

$$\mathcal{N}_{\mathcal{X}}(x) := \{ g \in \mathbb{R}^n \mid 0 \geq \langle g, y - x \rangle \quad \forall y \in \mathcal{X} \}$$

Application. $\min f(x) \quad \text{s.t.} \;\; x \in \mathcal{X}$:

◇ If f is diff., we get $0 \in \nabla f(x^*) + \mathcal{N}_{\mathcal{X}}(x^*)$

Optimality via subdifferentials: application

- Minimizing x must satisfy: $0 \in \partial(f + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming $\mathrm{ri}(\mathrm{dom} f) \cap \mathrm{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f(x) + \partial \mathbb{1}_X(x)$
- Recall, $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ iff $\mathbb{1}_{\mathcal{X}}(y) \geq \mathbb{1}_{\mathcal{X}}(x) + \langle g, y - x \rangle$ for all y.
- So $g \in \partial \mathbb{1}_{\mathcal{X}}(x)$ means $x \in \mathcal{X}$ and $0 \geq \langle g, y - x \rangle$ $\forall y \in \mathcal{X}$.
- **Normal cone:**

$$\mathcal{N}_{\mathcal{X}}(x) := \{ g \in \mathbb{R}^n \mid 0 \geq \langle g, y - x \rangle \quad \forall y \in \mathcal{X} \}$$

Application. $\min f(x) \quad \text{s.t. } x \in \mathcal{X}$:

♢ If f is diff., we get $0 \in \nabla f(x^*) + \mathcal{N}_{\mathcal{X}}(x^*)$

♢ $-\nabla f(x^*) \in \mathcal{N}_{\mathcal{X}}(x^*) \iff \langle \nabla f(x^*), y - x^* \rangle \geq 0$ for all $y \in \mathcal{X}$.

Duality

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

Primal problem

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

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

Def. Domain: The set $\mathcal{D} := \{\operatorname{dom} f_0 \cap \operatorname{dom} f_1 \cdots \cap \operatorname{dom} f_m\}$

- We call (P) the **primal problem**
- The variable x is the **primal variable**
- We will attach to (P) a **dual problem**
- In our initial derivation: no restriction to convexity.

Lagrangian

To the primal problem, associate **Lagrangian** $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$,

$$\mathcal{L}(x, \lambda) := f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x).$$

♠ Variables $\lambda \in \mathbb{R}^m$ called **Lagrange multipliers**

Lagrangian

To the primal problem, associate **Lagrangian** $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$,

$$\mathcal{L}(x, \lambda) := f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x).$$

- Variables $\lambda \in \mathbb{R}^m$ called **Lagrange multipliers**
- Suppose x is feasible, and $\lambda \geq 0$. Then, we get the lower-bound:

$$f_0(x) \geq \mathcal{L}(x, \lambda) \qquad \forall x \in \mathcal{X}, \;\; \lambda \in \mathbb{R}^m_+.$$

Lagrangian

To the primal problem, associate **Lagrangian** $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$,

$$\mathcal{L}(x, \lambda) := f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x).$$

- ♠ Variables $\lambda \in \mathbb{R}^m$ called **Lagrange multipliers**
- ♠ Suppose x is feasible, and $\lambda \geq 0$. Then, we get the lower-bound:

$$f_0(x) \geq \mathcal{L}(x, \lambda) \qquad \forall x \in \mathcal{X}, \;\; \lambda \in \mathbb{R}^m_+.$$

- ♠ Lagrangian helps write problem in **unconstrained form**

Lagrange dual function

Def. We define the **Lagrangian dual** as

$$g(\lambda) := \inf_x \quad \mathcal{L}(x, \lambda).$$

Lagrange dual function

Def. We define the **Lagrangian dual** as

$$g(\lambda) := \inf_x \quad \mathcal{L}(x, \lambda).$$

Observations:

- g is pointwise inf of affine functions of λ
- Thus, g is concave; it may take value $-\infty$

Lagrange dual function

Def. We define the **Lagrangian dual** as

$$g(\lambda) := \inf_x \quad \mathcal{L}(x, \lambda).$$

Observations:

- g is pointwise inf of affine functions of λ
- Thus, g is concave; it may take value $-\infty$
- Recall: $f_0(x) \geq \mathcal{L}(x, \lambda) \quad \forall x \in \mathcal{X}, \lambda \geq 0$; thus
- $\forall x \in \mathcal{X}, \quad f_0(x) \geq \inf_{x'} \mathcal{L}(x', \lambda) =: g(\lambda)$
- Now minimize over x on lhs, to obtain

$$\forall\, \lambda \in \mathbb{R}^m_+ \qquad p^* \geq g(\lambda).$$

Lagrange dual problem

$$\sup_{\lambda} g(\lambda) \qquad \text{s.t. } \lambda \geq 0.$$

Lagrange dual problem

$$\sup_{\lambda} g(\lambda) \qquad \text{s.t. } \lambda \geq 0.$$

- **dual feasible:** if $\lambda \geq 0$ and $g(\lambda) > -\infty$
- **dual optimal:** λ^* if sup is achieved
- Lagrange dual is **always concave**, regardless of original

Weak duality

Def. Denote **dual optimal value** by d^*, i.e.,

$$d^* := \sup_{\lambda \geq 0} \quad g(\lambda).$$

Weak duality

Def. Denote **dual optimal value** by d^*, i.e.,

$$d^* := \sup_{\lambda \geq 0} \quad g(\lambda).$$

Theorem. (Weak-duality): For problem (P), we have $p^* \geq d^*$.

Weak duality

Def. Denote **dual optimal value** by d^*, i.e.,

$$d^* := \sup_{\lambda \geq 0} \quad g(\lambda).$$

Theorem. (Weak-duality): For problem (P), we have $p^* \geq d^*$.

Proof: We showed that for all $\lambda \in \mathbb{R}^m_+$, $p^* \geq g(\lambda)$.
Thus, it follows that $p^* \geq \sup g(\lambda) = d^*$.

Duality gap

$$p^* - d^* \geq 0$$

Duality gap

$$p^* - d^* \geq 0$$

Strong duality if duality gap is zero: $p^* = d^*$

Notice: both p^* and d^* may be $+\infty$

Duality gap

$$p^* - d^* \geq 0$$

Strong duality if duality gap is zero: $p^* = d^*$

Notice: both p^* and d^* may be $+\infty$

Several **sufficient** conditions known, especially for convex optimization.

"Easy" necessary and sufficient conditions: **unknown**

Example: Slater's sufficient conditions

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

Example: Slater's sufficient conditions

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

Constraint qualification: There exists $x \in \operatorname{ri} \mathcal{D}$ s.t.

$$f_i(x) < 0, \qquad Ax = b.$$

That is, there is a **strictly feasible** point.

Theorem. Let the primal problem be convex. If there is a feasible point such that is strictly feasible for the non-affine constraints (and merely feasible for affine, linear ones), then strong duality holds. Moreover, the dual optimal is attained (i.e., $d^* > -\infty$).

Reading: Read BV §5.3.2 for a proof.

Example: failure of strong duality

$$\min_{x,y} e^{-x} \quad x^2/y \leq 0,$$

over the domain $\mathcal{D} = \{(x, y) \mid y > 0\}$.

Example: failure of strong duality

$$\min_{x,y} e^{-x} \quad x^2/y \leq 0,$$

over the domain $\mathcal{D} = \{(x, y) \mid y > 0\}$.
Clearly, only feasible $x = 0$. So $p^* = 1$

Example: failure of strong duality

$$\min_{x,y} e^{-x} \quad x^2/y \leq 0,$$

over the domain $\mathcal{D} = \{(x, y) \mid y > 0\}$.
Clearly, only feasible $x = 0$. So $p^* = 1$

$$\mathcal{L}(x, y, \lambda) = e^{-x} + \lambda x^2/y,$$

so dual function is

$$g(\lambda) = \inf_{x,y>0} e^{-x} + \lambda x^2 y = \begin{cases} 0 & \lambda \geq 0 \\ -\infty & \lambda < 0. \end{cases}$$

Example: failure of strong duality

$$\min_{x,y} e^{-x} \quad x^2/y \leq 0,$$

over the domain $\mathcal{D} = \{(x, y) \mid y > 0\}$.
Clearly, only feasible $x = 0$. So $p^* = 1$

$$\mathcal{L}(x, y, \lambda) = e^{-x} + \lambda x^2/y,$$

so dual function is

$$g(\lambda) = \inf_{x,y>0} e^{-x} + \lambda x^2 y = \begin{cases} 0 & \lambda \geq 0 \\ -\infty & \lambda < 0. \end{cases}$$

Dual problem

$$d^* = \max_{\lambda} 0 \qquad \text{s.t. } \lambda \geq 0.$$

Thus, $d^* = 0$, and gap is $p^* - d^* = 1$.
Here, we had no strictly feasible solution.

Zero duality gap: nonconvex example

Trust region subproblem (TRS)

$$\min \quad x^T A x + 2b^T x \qquad x^T x \leq 1.$$

A is symmetric but not necessarily semidefinite!

Theorem. TRS always has zero duality gap.

Remark: Above theorem extremely important result; part of a family of related results on strong duality for certain quadratic nonconvex problems.

Example: dual for Support Vector Machine

$$\begin{aligned} \min_{x,\xi} \quad & \tfrac{1}{2}\|x\|_2^2 + C\sum\nolimits_i \xi_i \\ \text{s.t.} \quad & Ax \geq 1 - \xi, \quad \xi \geq 0. \end{aligned}$$

Example: dual for Support Vector Machine

$$\begin{aligned} \min_{x,\xi} \quad & \tfrac{1}{2}\|x\|_2^2 + C\sum\nolimits_i \xi_i \\ \text{s.t.} \quad & Ax \geq 1 - \xi, \quad \xi \geq 0. \end{aligned}$$

$$L(x,\xi,\lambda,\nu) = \tfrac{1}{2}\|x\|_2^2 + C1^T\xi - \lambda^T(Ax - 1 + \xi) - \nu^T\xi$$

Example: dual for Support Vector Machine

$$\begin{aligned} \min_{x,\xi} \quad & \tfrac{1}{2}\|x\|_2^2 + C\sum\nolimits_i \xi_i \\ \text{s.t.} \quad & Ax \geq 1 - \xi, \quad \xi \geq 0. \end{aligned}$$

$$L(x, \xi, \lambda, \nu) = \tfrac{1}{2}\|x\|_2^2 + C1^T\xi - \lambda^T(Ax - 1 + \xi) - \nu^T\xi$$

$$\begin{aligned} g(\lambda, \nu) \quad &:= \quad \inf L(x, \xi, \lambda, \nu) \\ &= \quad \begin{cases} \lambda^T 1 - \tfrac{1}{2}\|A^T\lambda\|_2^2 & \lambda + \nu = C\mathbf{1} \\ +\infty & \text{otherwise} \end{cases} \\ d^* \quad &= \quad \max_{\lambda \geq 0, \nu \geq 0} \quad g(\lambda, \nu) \end{aligned}$$

Exercise: Using $\nu \geq 0$, eliminate ν from above problem.

Example: norm regularized problems

$$\min \quad f(x) + \|Ax\|$$

Example: norm regularized problems

$$\min \quad f(x) + \|Ax\|$$

Dual problem

$$\min_y \quad f^*(-A^T y) \quad \text{s.t. } \|y\|_* \leq 1.$$

Example: norm regularized problems

$$\min \quad f(x) + \|Ax\|$$

Dual problem

$$\min_y \quad f^*(-A^T y) \quad \text{s.t. } \|y\|_* \le 1.$$

Say $\|\bar{y}\|_* < 1$, such that $A^T\bar{y} \in \text{ri}(\text{dom} f^*)$, then we have strong duality (e.g., for instance $0 \in \text{ri}(\text{dom} f^*)$)

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda\|x\|_1.$$

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda \|x\|_1.$$

$$\|x\|_1 = \max\left\{x^T v \mid \|v\|_\infty \le 1\right\}$$
$$\|x\|_2 = \max\left\{x^T u \mid \|u\|_2 \le 1\right\}.$$

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda\|x\|_1.$$

$$\|x\|_1 = \max\left\{x^T v \mid \|v\|_\infty \leq 1\right\}$$
$$\|x\|_2 = \max\left\{x^T u \mid \|u\|_2 \leq 1\right\}.$$

Saddle-point formulation

$$p^* \quad = \quad \min_x \max_{u,v} \left\{u^T(b - Ax) + v^T x \mid \|u\|_2 \leq 1, \;\; \|v\|_\infty \leq \lambda\right\}$$

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda\|x\|_1.$$

$$\|x\|_1 = \max\left\{x^T v \mid \|v\|_\infty \le 1\right\}$$
$$\|x\|_2 = \max\left\{x^T u \mid \|u\|_2 \le 1\right\}.$$

Saddle-point formulation

$$\begin{aligned} p^* &= \min_x \max_{u,v} \left\{u^T(b - Ax) + v^T x \mid \|u\|_2 \le 1,\ \|v\|_\infty \le \lambda\right\} \\ &= \max_{u,v} \min_x \left\{u^T(b - Ax) + x^T v \mid \|u\|_2 \le 1,\ \|v\|_\infty \le \lambda\right\} \end{aligned}$$

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda\|x\|_1.$$

$$\|x\|_1 = \max\left\{x^T v \mid \|v\|_\infty \le 1\right\}$$
$$\|x\|_2 = \max\left\{x^T u \mid \|u\|_2 \le 1\right\}.$$

Saddle-point formulation

$$\begin{aligned} p^* &= \min_x \max_{u,v} \left\{u^T(b - Ax) + v^T x \mid \|u\|_2 \le 1,\ \ \|v\|_\infty \le \lambda\right\} \\ &= \max_{u,v} \min_x \left\{u^T(b - Ax) + x^T v \mid \|u\|_2 \le 1,\ \ \|v\|_\infty \le \lambda\right\} \\ &= \max_{u,v} u^T b \qquad A^T u = v,\ \|u\|_2 \le 1,\ \ \|v\|_\infty \le \lambda \end{aligned}$$

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda\|x\|_1.$$

$$\|x\|_1 = \max\left\{x^T v \mid \|v\|_\infty \leq 1\right\}$$
$$\|x\|_2 = \max\left\{x^T u \mid \|u\|_2 \leq 1\right\}.$$

Saddle-point formulation

$$\begin{aligned}
p^* &= \min_x \max_{u,v} \left\{u^T(b - Ax) + v^T x \mid \|u\|_2 \leq 1,\ \ \|v\|_\infty \leq \lambda\right\} \\
&= \max_{u,v} \min_x \left\{u^T(b - Ax) + x^T v \mid \|u\|_2 \leq 1,\ \ \|v\|_\infty \leq \lambda\right\} \\
&= \max_{u,v} u^T b \qquad A^T u = v,\ \|u\|_2 \leq 1,\ \ \|v\|_\infty \leq \lambda \\
&= \max_u u^T b \qquad \|u\|_2 \leq 1, \quad \|A^T v\|_\infty \leq \lambda.
\end{aligned}$$

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f_0(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f_0(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f_0(x) + \sum_i \lambda_i f_i(x)$

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \le 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f_0(x^*), x - x^* \rangle \ge 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f_0(x) + \sum_i \lambda_i f_i(x)$

Assume strong duality; and both p^* and d^* attained!

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \leq 0, \quad i = 1, \dots, m.$$

- Recall: $\langle \nabla f_0(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f_0(x) + \sum_i \lambda_i f_i(x)$

Assume strong duality; and both p^* and d^* attained!

Thus, there exists a pair (x^*, λ^*) such that

$p^* = f_0(x^*)$

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \le 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f_0(x^*),\, x - x^* \rangle \ge 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f_0(x) + \sum_i \lambda_i f_i(x)$

Assume strong duality; and both p^* and d^* attained!

Thus, there exists a pair (x^*, λ^*) such that

$$p^* = f_0(x^*) = d^* = g(\lambda^*)$$

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f_0(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f_0(x) + \sum_i \lambda_i f_i(x)$

Assume strong duality; and both p^* and d^* attained!

Thus, there exists a pair (x^*, λ^*) such that

$$p^* = f_0(x^*) = d^* = g(\lambda^*) = \min_x \mathcal{L}(x, \lambda^*)$$

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f_0(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f_0(x) + \sum_i \lambda_i f_i(x)$

Assume strong duality; and both p^* and d^* attained!

Thus, there exists a pair (x^*, λ^*) such that

$$p^* = f_0(x^*) = d^* = g(\lambda^*) = \min_x \mathcal{L}(x, \lambda^*) \leq \mathcal{L}(x^*, \lambda^*)$$

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f_0(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f_0(x) + \sum_i \lambda_i f_i(x)$

Assume strong duality; and both p^* and d^* attained!

Thus, there exists a pair (x^*, λ^*) such that

$$p^* = f_0(x^*) = d^* = g(\lambda^*) = \min_x \mathcal{L}(x, \lambda^*) \leq \mathcal{L}(x^*, \lambda^*) \leq f_0(x^*) = p^*$$

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f_0(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f_0(x) + \sum_i \lambda_i f_i(x)$

Assume strong duality; and both p^* and d^* attained!

Thus, there exists a pair (x^*, λ^*) such that

$$p^* = f_0(x^*) = d^* = g(\lambda^*) = \min_x \mathcal{L}(x, \lambda^*) \leq \mathcal{L}(x^*, \lambda^*) \leq f_0(x^*) = p^*$$

- Thus, equalities hold in above chain.

Example: KKT conditions

$$\min \quad f_0(x) \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m.$$

- Recall: $\langle \nabla f_0(x^*), x - x^* \rangle \geq 0$ for all feasible $x \in \mathcal{X}$
- Can we simplify this using Lagrangian?
- $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f_0(x) + \sum_i \lambda_i f_i(x)$

Assume strong duality; and both p^* and d^* attained!

Thus, there exists a pair (x^*, λ^*) such that

$$p^* = f_0(x^*) = d^* = g(\lambda^*) = \min_x \mathcal{L}(x, \lambda^*) \leq \mathcal{L}(x^*, \lambda^*) \leq f_0(x^*) = p^*$$

- Thus, equalities hold in above chain.

$$\boxed{x^* \in \operatorname{argmin}_x \mathcal{L}(x, \lambda^*).}$$

Example: KKT conditions

$$\boxed{x^* \in \operatorname{argmin}_x \mathcal{L}(x, \lambda^*).}$$

If $f_0, f_1, \ldots, f_m$ are differentiable, this implies

Example: KKT conditions

$$x^* \in \operatorname{argmin}_x \mathcal{L}(x, \lambda^*).$$

If $f_0, f_1, \ldots, f_m$ are differentiable, this implies

$$\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} = \nabla f_0(x^*) + \sum_i \lambda_i^* \nabla f_i(x^*) = 0.$$

Example: KKT conditions

$$\boxed{x^* \in \operatorname{argmin}_x \mathcal{L}(x, \lambda^*).}$$

If $f_0, f_1, \ldots, f_m$ are differentiable, this implies

$$\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} = \nabla f_0(x^*) + \sum\nolimits_i \lambda_i^* \nabla f_i(x^*) = 0.$$

Moreover, since $\mathcal{L}(x^*, \lambda^*) = f_0(x^*)$, we also have

Example: KKT conditions

$$x^* \in \operatorname{argmin}_x \mathcal{L}(x, \lambda^*).$$

If $f_0, f_1, \ldots, f_m$ are differentiable, this implies

$$\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} = \nabla f_0(x^*) + \sum_i \lambda_i^* \nabla f_i(x^*) = 0.$$

Moreover, since $\mathcal{L}(x^*, \lambda^*) = f_0(x^*)$, we also have

$$\sum_i \lambda_i^* f_i(x^*) = 0.$$

Example: KKT conditions

$$x^* \in \operatorname{argmin}_x \mathcal{L}(x, \lambda^*).$$

If $f_0, f_1, \ldots, f_m$ are differentiable, this implies

$$\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} = \nabla f_0(x^*) + \sum\nolimits_i \lambda_i^* \nabla f_i(x^*) = 0.$$

Moreover, since $\mathcal{L}(x^*, \lambda^*) = f_0(x^*)$, we also have

$$\sum\nolimits_i \lambda_i^* f_i(x^*) = 0.$$

But $\lambda_i^* \geq 0$ and $f_i(x^*) \leq 0$,

Example: KKT conditions

$$\boxed{x^* \in \operatorname{argmin}_x \mathcal{L}(x, \lambda^*).}$$

If $f_0, f_1, \ldots, f_m$ are differentiable, this implies

$$\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} = \nabla f_0(x^*) + \sum\nolimits_i \lambda_i^* \nabla f_i(x^*) = 0.$$

Moreover, since $\mathcal{L}(x^*, \lambda^*) = f_0(x^*)$, we also have

$$\sum\nolimits_i \lambda_i^* f_i(x^*) = 0.$$

But $\lambda_i^* \geq 0$ and $f_i(x^*) \leq 0$, so **complementary slackness**

$$\lambda_i^* f_i(x^*) = 0, \quad i = 1, \ldots, m.$$

KKT conditions

$$\begin{aligned}
f_i(x^*) &\leq 0, \quad i = 1, \ldots, m && \text{(primal feasibility)} \\
\lambda_i^* &\geq 0, \quad i = 1, \ldots, m && \text{(dual feasibility)} \\
\lambda_i^* f_i(x^*) &= 0, \quad i = 1, \ldots, m && \text{(compl. slackness)} \\
\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} &= 0 && \text{(Lagrangian stationarity)}
\end{aligned}$$

KKT conditions

$$\begin{aligned}
f_i(x^*) &\leq 0, \quad i = 1, \ldots, m && \text{(primal feasibility)} \\
\lambda_i^* &\geq 0, \quad i = 1, \ldots, m && \text{(dual feasibility)} \\
\lambda_i^* f_i(x^*) &= 0, \quad i = 1, \ldots, m && \text{(compl. slackness)} \\
\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} &= 0 && \text{(Lagrangian stationarity)}
\end{aligned}$$

- We showed: if strong duality holds, and (x^*, λ^*) exist, then KKT conditions are **necessary** for pair (x^*, λ^*) to be optimal

KKT conditions

$$
\begin{aligned}
f_i(x^*) &\leq 0, \quad i = 1, \ldots, m && \text{(primal feasibility)} \\
\lambda_i^* &\geq 0, \quad i = 1, \ldots, m && \text{(dual feasibility)} \\
\lambda_i^* f_i(x^*) &= 0, \quad i = 1, \ldots, m && \text{(compl. slackness)} \\
\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} &= 0 && \text{(Lagrangian stationarity)}
\end{aligned}
$$

- We showed: if strong duality holds, and (x^*, λ^*) exist, then KKT conditions are **necessary** for pair (x^*, λ^*) to be optimal
- If problem is convex, then KKT also **sufficient**

KKT conditions

$$
\begin{aligned}
f_i(x^*) &\le 0, \quad i = 1, \ldots, m && \text{(primal feasibility)}\\
\lambda_i^* &\ge 0, \quad i = 1, \ldots, m && \text{(dual feasibility)}\\
\lambda_i^* f_i(x^*) &= 0, \quad i = 1, \ldots, m && \text{(compl. slackness)}\\
\nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} &= 0 && \text{(Lagrangian stationarity)}
\end{aligned}
$$

- We showed: if strong duality holds, and (x^*, λ^*) exist, then KKT conditions are **necessary** for pair (x^*, λ^*) to be optimal
- If problem is convex, then KKT also **sufficient**

Exercise: Prove the above sufficiency of KKT. *Hint:* Use that $\mathcal{L}(x, \lambda^*)$ is convex, and conclude from KKT conditions that $g(\lambda^*) = f_0(x^*)$, so that (x^*, λ^*) optimal primal-dual pair.