Introduction to large-scale optimization
(Lecture 1)

Suvrit Sra
Massachusetts Institute of Technology

Microsoft Research India
Machine Learning Summer School, June 2015

Course materials

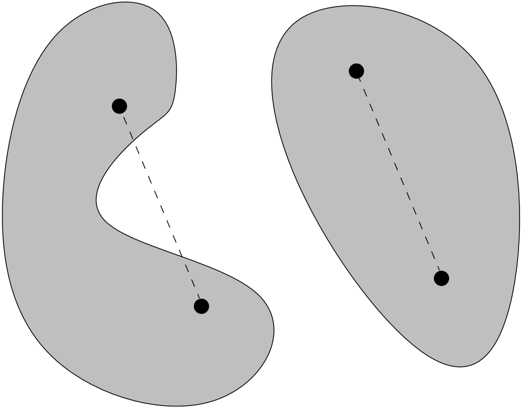
- http://suvrit.de/teach/msr2015/
- Some references:
  - *Introductory lectures on convex optimization* – Nesterov
  - *Convex optimization* – Boyd & Vandenberghe
  - *Nonlinear programming* – Bertsekas
  - *Convex Analysis* – Rockafellar
  - *Fundamentals of convex analysis* – Urruty, Lemaréchal
  - *Lectures on modern convex optimization* – Nemirovski
  - *Optimization for Machine Learning* – Sra, Nowozin, Wright
- Some related courses:
  - EE227A, Spring 2013, (UC Berkeley)
  - 10-801, Spring 2014 (CMU)
  - EE364a,b (Boyd, Stanford)
  - EE236b,c (Vandenberghe, UCLA)
- NIPS, ICML, UAI, AISTATS, SIOPT, Math. Prog.

Outline

- Recap on convexity
- Recap on duality, optimality
- First-order optimization algorithms
- Proximal methods, operator splitting
- Incremental methods
- High-level view of parallel, distributed
- Some words on nonconvex

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 $\theta x + (1-\theta)y$, where $\theta \in [0,1]$, 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 $\theta x + (1-\theta)y$, where $\theta \in [0,1]$, also lies in C.

Combinations

- Convex: $\theta_1 x + \theta_2 y \in C$, where $\theta_1, \theta_2 \geq 0$ and $\theta_1 + \theta_2 = 1$.
- Linear: if restrictions on θ_1, θ_2 are dropped
- Conic: if restriction $\theta_1 + \theta_2 = 1$ is dropped

Convex sets

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

Proof.

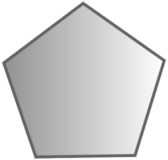
→ If $C_1 \cap C_2 = \emptyset$, then true vacuously.

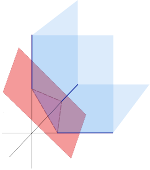
→ Let $x, y \in C_1 \cap C_2$. Then, $x, y \in C_1$ and $x, y \in C_2$.

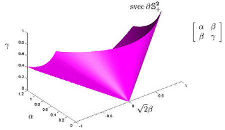
→ But C_1, C_2 are convex, hence $\theta x + (1-\theta)y \in C_1$, and also in C_2. Thus, $\theta x + (1-\theta)y \in C_1 \cap C_2$.

→ Inductively follows that $\cap_{i=1}^m C_i$ 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. 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. Function $f : I \to \mathbb{R}$ on interval I called **midpoint convex** if

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

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

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 $\theta \geq 0$

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

Convex functions

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

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

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

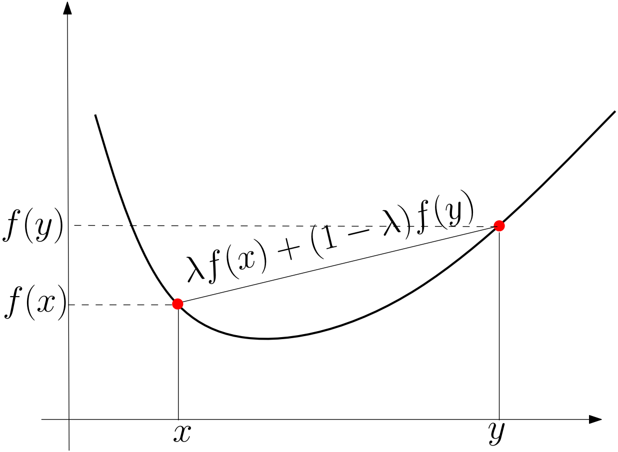
Def. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called **convex** if its domain $\mathrm{dom}(f)$ is a convex set and for any $x, y \in \mathrm{dom}(f)$ and $\theta \geq 0$

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

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.

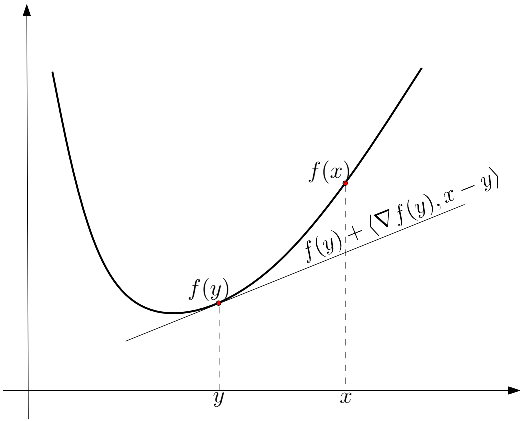
- Extends to $f : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$; useful for proving convexity.

Convex functions



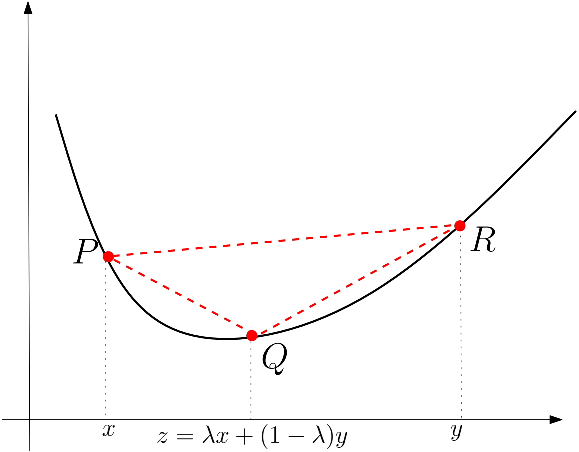
$$f(\lambda x + (1 - \lambda) y) \leq \lambda f(x) + (1 - \lambda) f(y)$$

Convex functions



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

Convex functions



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

Convex functions

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 some "index set" $\mathcal{Y}$, then

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

is a convex function of x (set $\mathcal{Y}$ is arbitrary).

Convex functions

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 some "index set" $\mathcal{Y}$, then

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

is a convex function of x (set $\mathcal{Y}$ is arbitrary).

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 above examples.

Convex functions

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

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

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

Convex functions

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

Convex functions – distance

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

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

Note: because $\|x - y\|$ is jointly convex in (x, y), the function $\operatorname{dist}(x, \mathcal{Y})$ is a convex function of x.

Convex functions – norms

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function that satisfies

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

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

Theorem Norms are convex.

Some norms

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

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

Fenchel conjugate

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

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

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

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: $\operatorname{dom} f^* = [-1, 0]$, 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)

Challenge 2

Consider the following functions on strictly positive variables:

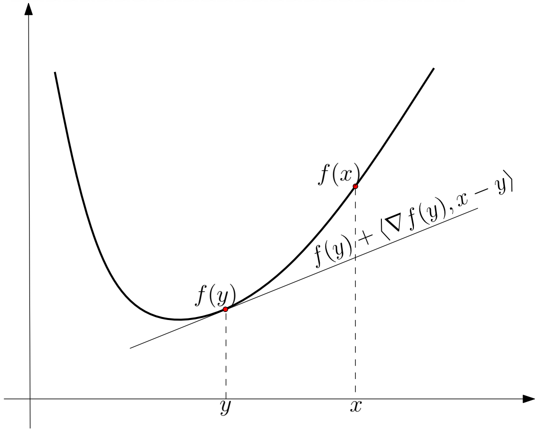
$$\begin{aligned}
h_1(x) &:= \frac{1}{x} \\
h_2(x, y) &:= \frac{1}{x} + \frac{1}{y} - \frac{1}{x+y} \\
h_3(x, y, z) &:= \frac{1}{x} + \frac{1}{y} + \frac{1}{z} - \frac{1}{x+y} - \frac{1}{y+z} - \frac{1}{x+z} + \frac{1}{x+y+z}
\end{aligned}$$

- ♡ Prove that h_1, h_2, h_3, and in general h_n are convex!
- ♡ Prove that in fact each $1/h_n$ is concave

$\boxed{\nabla^2 h_n(x) \succeq 0 \text{ is not recommended}}$

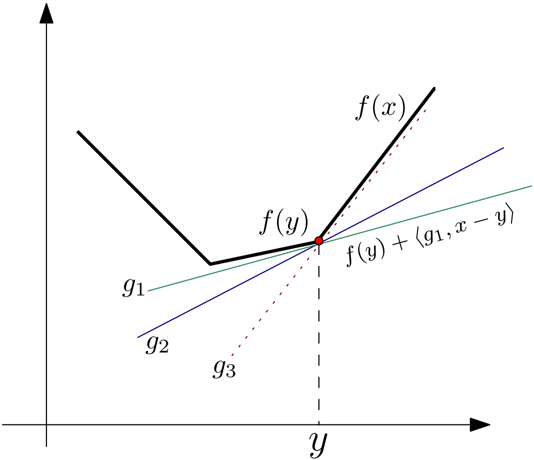
Subgradients

Subgradients: global underestimators



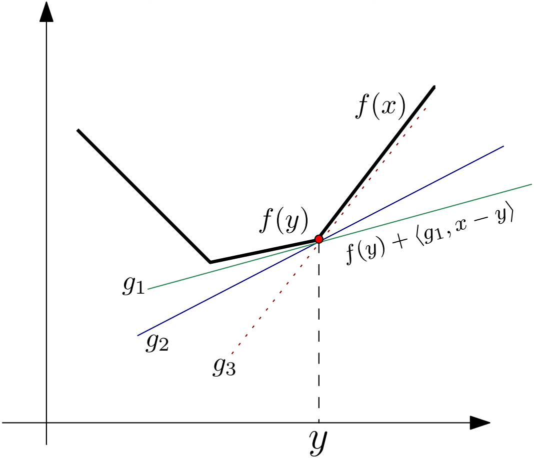
$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$$

Subgradients: global underestimators



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

Subgradients: global underestimators



g_1, g_2, g_3 are subgradients at y

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

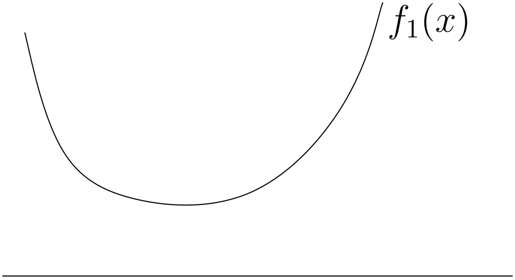
- f is convex, differentiable: $\nabla f(y)$ the **unique** subgradient at y
- A vector g is a subgradient at a point y if and only if $f(y) + \langle g, x - y \rangle$ is **globally** smaller than $f(x)$.
- Usually, **one** subgradient costs approx. as much as $f(x)$
- Determining all subgradients at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

Subgradients – example

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

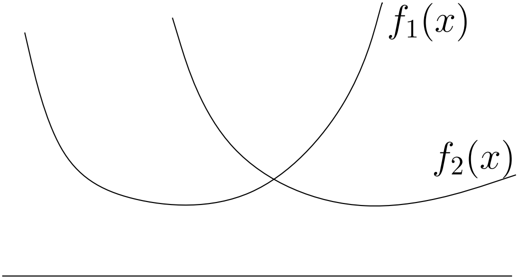
Subgradients – example

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



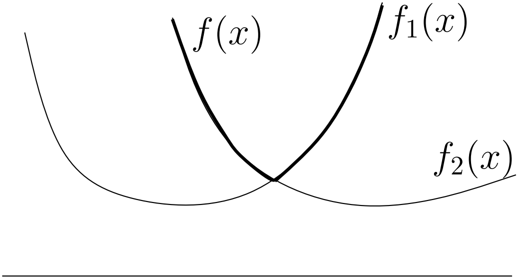
Subgradients – example

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



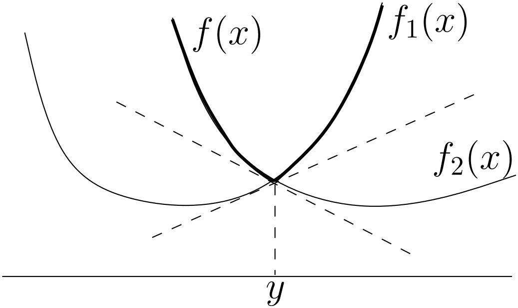
Subgradients – example

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



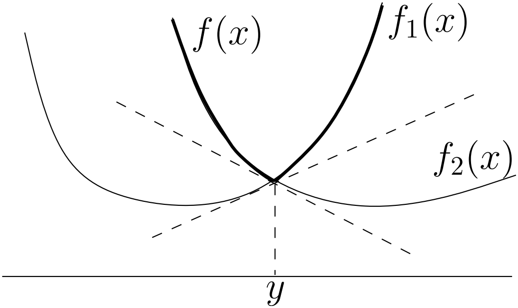
Subgradients – example

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



Subgradients – example

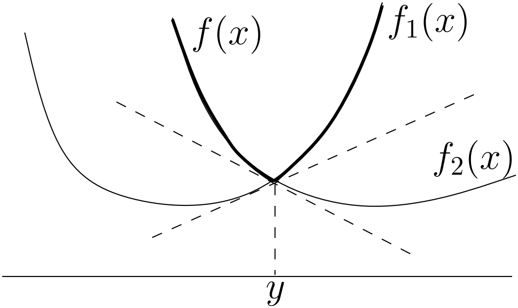
$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable



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

Subgradients – example

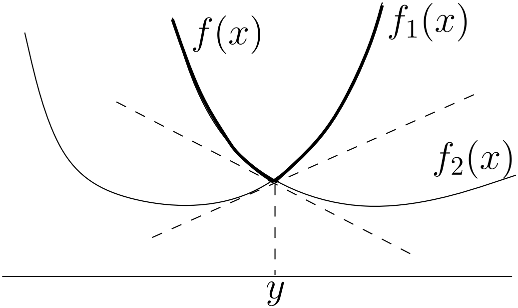
$f(x) := \max(f_1(x), f_2(x))$; both f_1, f_2 convex, differentiable



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

Subgradients – 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

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

Subdifferential

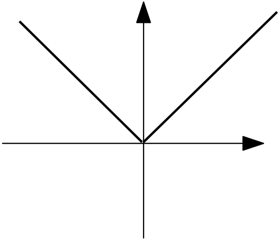
Def. The set of all subgradients at y denoted by $\partial f(y)$. This set is called **subdifferential** of f at y

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

- ♣ If $x \in$ relative interior of 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 $g = \nabla f(x)$

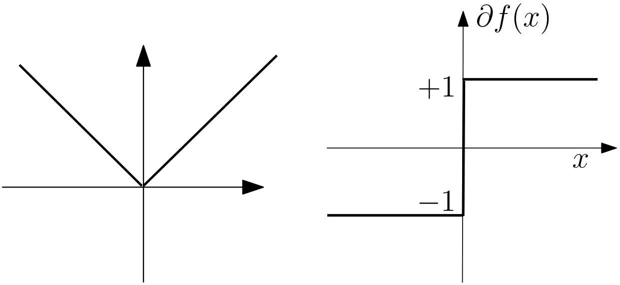
Subdifferential – example

$$\boxed{f(x) = |x|}$$



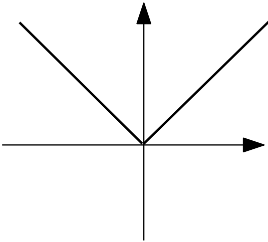
Subdifferential – example

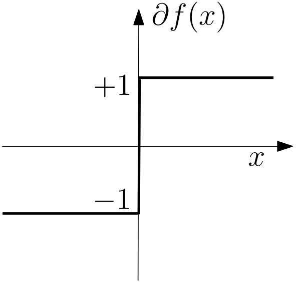
$$\boxed{f(x) = |x|}$$



Subdifferential – example

$$\boxed{f(x) = |x|}$$





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

More examples

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

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

More examples

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.

$$\|z\|_2 \quad \geq \quad \|x\|_2 + \langle g,\, z - x\rangle$$

More examples

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.

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

More examples

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

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

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

  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

$\boxed{\text{It can happen that } \partial(f_1 + f_2) \neq \partial f_1 + \partial f_2}$

Example

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 $\partial f_1(0) = \partial f_2(0) = \emptyset$.

Example

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.

Example

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

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

where e_i is i-th canonical basis vector.

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

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

Example – subgradients

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

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

Example – 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 $g \in \partial f(x)$

$$\begin{aligned} h(z, y^*) &\geq h(x, y^*) + g^T(z - x) \\ h(z, y^*) &\geq f(x) + g^T(z - x) \end{aligned}$$

Example – 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 $g \in \partial f(x)$

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

Example

Suppose $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$. And

$$f(x) := \max_{1 \le i \le n} (a_i^T x + b_i).$$

(This f is a max over a finite number of terms)

Example

Suppose $a_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$. And

$$f(x) := \max_{1 \le i \le n} (a_i^T x + b_i).$$

(This f is a max over a finite number of terms)

- Let $f_k(x) = a_k^T x + b_k$
- Suppose $f(x) = a_k^T x + b_k$ for some index k
- Here $\partial f_k(x) = \{\nabla f_k(x)\}$
- Hence, $a_k \in \partial f(x)$ is a subgradient

Subgradient of expectation

Suppose $f = \mathbf{E}f(x,u)$, where f is convex in x for each u (r.v.)

$$f(x) := \int f(x,u)p(u)du$$

Subgradient of expectation

Suppose $f = \mathbf{E} f(x,u)$, where f is convex in x for each u (r.v.)

$$f(x) := \int f(x,u)p(u)du$$

- For each u choose any $g(x,u) \in \partial_x f(x,u)$

Subgradient of expectation

Suppose $f = \mathbf{E} f(x, u)$, where f is convex in x for each u (r.v.)

$$f(x) := \int f(x, u) p(u) du$$

- For each u choose any $g(x, u) \in \partial_x f(x, u)$
- Then, $g(x) = \int g(x, u) p(u) du = \mathbf{E} g(x, u) \in \partial f(x)$

Optimization

Optimization problems

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

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

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, locally optimal also globally so.

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, locally optimal also globally so.

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

If f is convex, then $\nabla f(x^*) = 0$ is actually **sufficient** for global optimality! For general f this is **not** true.

(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

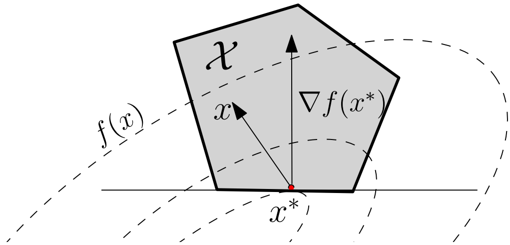
$$\langle \nabla f(x^*), y - x^* \rangle \geq 0, \qquad \text{for all } \; y \in \mathcal{X}.$$

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^*

Optimality – nonsmooth

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 – nonsmooth

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

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

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

Optimality – nonsmooth

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 – nonsmooth

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

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

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

Optimality – nonsmooth

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

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

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

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

Optimality – nonsmooth

- Minimizing x must satisfy: $0 \in \partial(f_0 + \mathbb{1}_{\mathcal{X}})(x)$

Optimality – nonsmooth

- Minimizing x must satisfy: $0 \in \partial(f_0 + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming $\mathrm{ri}(\mathrm{dom}\, f_0) \cap \mathrm{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f_0(x) + \partial \mathbb{1}_X(x)$

Optimality – nonsmooth

- Minimizing x must satisfy: $0 \in \partial(f_0 + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming $\mathrm{ri}(\mathrm{dom}\, f_0) \cap \mathrm{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f_0(x) + \partial \mathbb{1}_X(x)$
- 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.

Optimality – nonsmooth

- Minimizing x must satisfy: $0 \in \partial(f_0 + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming $\mathrm{ri}(\mathrm{dom}\, f_0) \cap \mathrm{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f_0(x) + \partial \mathbb{1}_X(x)$
- 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 $0 \geq \langle g, y - x \rangle\ \forall y \in \mathcal{X}$.

Optimality – nonsmooth

- Minimizing x must satisfy: $0 \in \partial(f_0 + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming ri(dom f_0) $\cap$ ri($\mathcal{X}$) $\neq \emptyset$, $0 \in \partial f_0(x) + \partial \mathbb{1}_X(x)$
- 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:**

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

Optimality – nonsmooth

- Minimizing x must satisfy: $0 \in \partial(f_0 + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming $\mathrm{ri}(\mathrm{dom}\, f_0) \cap \mathrm{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f_0(x) + \partial \mathbb{1}_X(x)$
- 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:**

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

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 – nonsmooth

- Minimizing x must satisfy: $0 \in \partial(f_0 + \mathbb{1}_{\mathcal{X}})(x)$
- (**CQ**) Assuming $\mathrm{ri}(\mathrm{dom}\, f_0) \cap \mathrm{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f_0(x) + \partial \mathbb{1}_X(x)$
- 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 $y \in \mathcal{X}$.

Optimality – projection operator

$$P_{\mathcal{X}}(y) := \underset{x \in \mathcal{X}}{\operatorname{argmin}} \|x - y\|^2$$

(Assume $\mathcal{X}$ is closed and convex, then projection is unique)
Let $\mathcal{X}$ be nonempty, closed and convex.

- Optimality condition: $x^* = P_{\mathcal{X}}(y)$ iff

$$\langle x^* - y, z - x^* \rangle \geq 0 \text{ for all } z \in \mathcal{X}$$

- Projection is **nonexpansive**:

$$\|P_{\mathcal{X}}(x) - P_{\mathcal{X}}(y)\|^2 \leq \|x - y\|^2 \quad \text{for all } x, y \in \mathbb{R}^n.$$

Optimality – projection operator

$$P_{\mathcal{X}}(y) := \operatorname*{argmin}_{x \in \mathcal{X}} \|x - y\|^2$$

(Assume $\mathcal{X}$ is closed and convex, then projection is unique)
Let $\mathcal{X}$ be nonempty, closed and convex.

- Optimality condition: $x^* = P_{\mathcal{X}}(y)$ iff

$$\langle x^* - y, z - x^* \rangle \geq 0 \text{ for all } z \in \mathcal{X}$$

- Projection is **nonexpansive**:

$$\|P_{\mathcal{X}}(x) - P_{\mathcal{X}}(y)\|^2 \;\leq\; \|x - y\|^2 \quad \text{for all } x, y \in \mathbb{R}^n.$$

Proof: **Exercise!**

Duality

Primal problem

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

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

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** $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$,

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

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

Lagrangian

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

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

- 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** $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$,

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

- 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: $f_0(x) \geq \mathcal{L}(x, \lambda) \quad \forall x \in \mathcal{X}$; 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

$$\boxed{p^* - d^* \geq 0}$$

Duality gap

$$\boxed{p^* - d^* \geq 0}$$

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!

"Easy" necessary and sufficient conditions: unknown

Zero duality gap: nonconvex example

Trust region subproblem (TRS)

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

A is symmetric but not necessarily semidefinite!

Theorem TRS always has zero duality gap.

Strong duality – counterexample

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

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

Strong duality – counterexample

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

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

Strong duality – counterexample

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

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

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

Strong duality – counterexample

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

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 $p^* - d^* = 1$.

Strong duality – counterexample

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

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.

Support vector machine

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

Support vector machine

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

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

Support vector machine

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

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

$$\begin{aligned} g(\lambda, \nu) \quad &:= \quad \inf L(x, \xi, \lambda, \nu) \\ &= \quad \begin{cases} \lambda^T 1 - \tfrac{1}{2}\|A^T\lambda\|_2^2 & \lambda + \nu = C1 \\ +\infty & \text{otherwise} \end{cases} \\ d^* \quad &= \quad \max_{\lambda \geq 0, \nu \geq 0} \quad g(\lambda, \nu) \end{aligned}$$

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

Regularized optimization

$$\inf_{x \in \mathcal{X}} \quad f(x) + r(Ax) \quad \text{s.t.} \;\; Ax \in \mathcal{Y}.$$

Regularized optimization

$$\inf_{x \in \mathcal{X}} \quad f(x) + r(Ax) \quad \text{s.t.} \;\; Ax \in \mathcal{Y}.$$

Dual problem

$$\inf_{u \in \mathcal{Y}} \quad f^*(-A^T u) + r^*(u).$$

Regularized optimization

$$\inf_{x \in \mathcal{X}} \quad f(x) + r(Ax) \quad \text{s.t.} \;\; Ax \in \mathcal{Y}.$$

Dual problem

$$\inf_{u \in \mathcal{Y}} \quad f^*(-A^T u) + r^*(u).$$

- Introduce new variable $z = Ax$

$$\inf_{x \in \mathcal{X}, z \in \mathcal{Y}} \quad f(x) + r(z), \qquad \text{s.t.} \quad z = Ax.$$

Regularized optimization

$$\inf_{x \in \mathcal{X}} \quad f(x) + r(Ax) \quad \text{s.t. } \; Ax \in \mathcal{Y}.$$

Dual problem

$$\inf_{u \in \mathcal{Y}} \quad f^*(-A^T u) + r^*(u).$$

- Introduce new variable $z = Ax$

$$\inf_{x \in \mathcal{X}, z \in \mathcal{Y}} \quad f(x) + r(z), \qquad \text{s.t.} \quad z = Ax.$$

- The (partial)-Lagrangian is

$$L(x, z; u) := f(x) + r(z) + u^T(Ax - z), \quad x \in \mathcal{X}, z \in \mathcal{Y};$$

Regularized optimization

$$\inf_{x \in \mathcal{X}} \quad f(x) + r(Ax) \quad \text{s.t.} \;\; Ax \in \mathcal{Y}.$$

Dual problem

$$\inf_{u \in \mathcal{Y}} \quad f^*(-A^T u) + r^*(u).$$

- Introduce new variable $z = Ax$

$$\inf_{x \in \mathcal{X}, z \in \mathcal{Y}} \quad f(x) + r(z), \qquad \text{s.t.} \quad z = Ax.$$

- The (partial)-Lagrangian is

$$L(x, z; u) := f(x) + r(z) + u^T(Ax - z), \quad x \in \mathcal{X}, z \in \mathcal{Y};$$

- Associated dual function

$$g(u) := \inf_{x \in \mathcal{X}, z \in \mathcal{Y}} L(x, z; u).$$

Regularized optimization

$$\inf_{x \in \mathcal{X}} \quad f(x) + r(Ax) \quad \text{s.t. } Ax \in \mathcal{Y}.$$

Dual problem

$$\inf_{y \in \mathcal{Y}} \quad f^*(-A^T y) + r^*(y).$$

The infimum above can be rearranged as follows

$$g(y) \quad = \quad \inf_{x \in \mathcal{X}} f(x) + y^T Ax + \inf_{z \in \mathcal{Y}} r(z) - y^T z$$

Regularized optimization

$$\inf_{x \in \mathcal{X}} \quad f(x) + r(Ax) \quad \text{s.t.} \;\; Ax \in \mathcal{Y}.$$

Dual problem

$$\inf_{y \in \mathcal{Y}} \quad f^*(-A^T y) + r^*(y).$$

The infimum above can be rearranged as follows

$$\begin{aligned} g(y) &= \inf_{x \in \mathcal{X}} f(x) + y^T Ax + \inf_{z \in \mathcal{Y}} r(z) - y^T z \\ &= -\sup_{x \in \mathcal{X}} \left\{ -x^T A^T y - f(x) \right\} - \sup_{z \in \mathcal{Y}} \left\{ z^T y - r(z) \right\} \end{aligned}$$

Regularized optimization

$$\inf_{x\in\mathcal{X}} \quad f(x) + r(Ax) \quad \text{s.t. } \; Ax \in \mathcal{Y}.$$

Dual problem

$$\inf_{y\in\mathcal{Y}} \quad f^*(-A^T y) + r^*(y).$$

The infimum above can be rearranged as follows

$$\begin{aligned}
g(y) &= \inf_{x\in\mathcal{X}} f(x) + y^T Ax + \inf_{z\in\mathcal{Y}} r(z) - y^T z \\
&= -\sup_{x\in\mathcal{X}} \left\{-x^T A^T y - f(x)\right\} - \sup_{z\in\mathcal{Y}} \left\{z^T y - r(z)\right\} \\
&= -f^*(-A^T y) - r^*(y) \quad \text{s.t. } y \in \mathcal{Y}.
\end{aligned}$$

Regularized optimization

$$\inf_{x \in \mathcal{X}} \quad f(x) + r(Ax) \quad \text{s.t. } \; Ax \in \mathcal{Y}.$$

Dual problem

$$\inf_{y \in \mathcal{Y}} \quad f^*(-A^T y) + r^*(y).$$

The infimum above can be rearranged as follows

$$\begin{aligned}
g(y) &= \inf_{x \in \mathcal{X}} f(x) + y^T A x + \inf_{z \in \mathcal{Y}} r(z) - y^T z \\
&= -\sup_{x \in \mathcal{X}} \left\{ -x^T A^T y - f(x) \right\} - \sup_{z \in \mathcal{Y}} \left\{ z^T y - r(z) \right\} \\
&= -f^*(-A^T y) - r^*(y) \quad \text{s.t. } y \in \mathcal{Y}.
\end{aligned}$$

Dual problem computes $\sup_{u \in \mathcal{Y}} g(u)$; so equivalently,

$$\inf_{y \in \mathcal{Y}} \quad f^*(-A^T y) + r^*(y).$$

Regularized optimization

Strong duality

$$\inf_x \{f(x) + r(Ax)\} = \sup_y \left\{-f^*(-A^T y) + r^*(y)\right\}$$

if either of the following conditions holds:

Regularized optimization

Strong duality

$$\inf_x \{f(x) + r(Ax)\} = \sup_y \left\{-f^*(-A^T y) + r^*(y)\right\}$$

if either of the following conditions holds:

1. $\exists x \in \text{ri}(\text{dom}\, f)$ such that $Ax \in \text{ri}(\text{dom}\, r)$
2. $\exists y \in \text{ri}(\text{dom}\, r^*)$ such that $A^T y \in \text{ri}(\text{dom}\, f^*)$

Regularized optimization

Strong duality

$$\inf_x \{f(x) + r(Ax)\} = \sup_y \left\{-f^*(-A^T y) + r^*(y)\right\}$$

if either of the following conditions holds:

1. $\exists x \in \text{ri}(\text{dom}\, f)$ such that $Ax \in \text{ri}(\text{dom}\, r)$
2. $\exists y \in \text{ri}(\text{dom}\, r^*)$ such that $A^T y \in \text{ri}(\text{dom}\, f^*)$

- Condition 1 ensures 'sup' attained at some y
- Condition 2 ensures 'inf' attained at some x

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

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

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

Example: variable splitting

$$\min \quad f(x) + g(x)$$

Example: variable splitting

$$\min \quad f(x) + g(x)$$

Exercise: Fill in the details below

$$\min_{x,z} \quad f(x) + g(z) \quad \text{s.t.} \quad x = z$$

Example: variable splitting

$$\min \quad f(x) + g(x)$$

Exercise: Fill in the details below

$$\min_{x,z} \quad f(x) + g(z) \quad \text{s.t.} \quad x = z$$

$$L(x, z, \nu) = f(x) + g(z) + \nu^T (x - z)$$

Example: variable splitting

$$\min \quad f(x) + g(x)$$

Exercise: Fill in the details below

$$\begin{aligned} \min_{x,z} \quad f(x) + g(z) \quad \text{s.t.} \quad x = z \\ L(x, z, \nu) = f(x) + g(z) + \nu^T(x - z) \\ g(\nu) = \inf_{x,z} L(x, z, \nu) \end{aligned}$$

Primal-dual: weak minimax

Theorem Let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$ be any function. Then,

$$\sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} \phi(x, y) \quad \leq \quad \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \phi(x, y)$$

Primal-dual: weak minimax

Theorem Let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$ be any function. Then,

$$\sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} \phi(x, y) \quad \leq \quad \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \phi(x, y)$$

Proof:

$$\forall x, y, \quad \inf_{x' \in \mathcal{X}} \phi(x', y) \quad \leq \quad \phi(x, y)$$

Primal-dual: weak minimax

Theorem Let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$ be any function. Then,

$$\sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} \phi(x, y) \quad \leq \quad \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \phi(x, y)$$

Proof:

$$\forall x, y, \quad \inf_{x' \in \mathcal{X}} \phi(x', y) \quad \leq \quad \phi(x, y)$$

$$\forall x, y, \quad \inf_{x' \in \mathcal{X}} \phi(x', y) \quad \leq \quad \sup_{y' \in \mathcal{Y}} \phi(x, y')$$

Primal-dual: weak minimax

Theorem Let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$ be any function. Then,

$$\sup_{y\in\mathcal{Y}} \inf_{x\in\mathcal{X}} \phi(x,y) \quad \leq \quad \inf_{x\in\mathcal{X}} \sup_{y\in\mathcal{Y}} \phi(x,y)$$

Proof:

$$\forall x, y, \quad \inf_{x'\in\mathcal{X}} \phi(x',y) \quad \leq \quad \phi(x,y)$$

$$\forall x, y, \quad \inf_{x'\in\mathcal{X}} \phi(x',y) \quad \leq \quad \sup_{y'\in\mathcal{Y}} \phi(x,y')$$

$$\forall x, \quad \sup_{y\in\mathcal{Y}} \inf_{x'\in\mathcal{X}} \phi(x',y) \quad \leq \quad \sup_{y'\in\mathcal{Y}} \phi(x,y')$$

Primal-dual: weak minimax

Theorem Let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$ be any function. Then,

$$\sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} \phi(x, y) \quad \leq \quad \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \phi(x, y)$$

Proof:

$$\begin{aligned}
\forall x, y, \quad & \inf_{x' \in \mathcal{X}} \phi(x', y) && \leq \quad \phi(x, y) \\
\forall x, y, \quad & \inf_{x' \in \mathcal{X}} \phi(x', y) && \leq \quad \sup_{y' \in \mathcal{Y}} \phi(x, y') \\
\forall x, \quad & \sup_{y \in \mathcal{Y}} \inf_{x' \in \mathcal{X}} \phi(x', y) && \leq \quad \sup_{y' \in \mathcal{Y}} \phi(x, y') \\
\Longrightarrow \quad & \sup_{y \in \mathcal{Y}} \inf_{x' \in \mathcal{X}} \phi(x', y) && \leq \quad \inf_{x \in \mathcal{X}} \sup_{y' \in \mathcal{Y}} \phi(x, y').
\end{aligned}$$

Primal-dual: strong minimax

- If "inf sup = sup inf", common value called **saddle-value**
- Value exists if there is a **saddle-point**, i.e., pair (x^*, y^*)

$$\phi(x, y^*) \geq \phi(x^*, y^*) \geq \phi(x^*, y) \quad \text{for all } x \in \mathcal{X}, y \in \mathcal{Y}.$$

Def. Let ϕ be as before. A point (x^*, y^*) is a saddle-point of ϕ (min over $\mathcal{X}$ and max over $\mathcal{Y}$) **iff** the infimum in the expression

$$\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \phi(x, y)$$

is **attained** at x^*, and the supremum in the expression

$$\sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} \phi(x, y)$$

is **attained** at y^*, and these two extrema are equal.

Primal-dual: strong minimax

- If "inf sup = sup inf", common value called **saddle-value**
- Value exists if there is a **saddle-point**, i.e., pair (x^*, y^*)

$$\phi(x, y^*) \geq \phi(x^*, y^*) \geq \phi(x^*, y) \quad \text{for all } x \in \mathcal{X}, y \in \mathcal{Y}.$$

Def. Let ϕ be as before. A point (x^*, y^*) is a saddle-point of ϕ (min over $\mathcal{X}$ and max over $\mathcal{Y}$) **iff** the infimum in the expression

$$\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \phi(x, y)$$

is **attained** at x^*, and the supremum in the expression

$$\sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} \phi(x, y)$$

is **attained** at y^*, and these two extrema are equal.

$$x^* \in \operatorname*{argmin}_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \phi(x, y) \qquad y^* \in \operatorname*{argmax}_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \phi(x, y).$$

Sufficient conditions for saddle-point

- Function ϕ is continuous, and
- It is convex-concave ($\phi(\cdot, y)$ convex for every $y \in \mathcal{Y}$, and $\phi(x, \cdot)$ concave for every $x \in \mathcal{X}$), and
- Both $\mathcal{X}$ and $\mathcal{Y}$ are convex; one of them is compact.

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

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

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

Saddle-point formulation

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

Example: Lasso-like problem

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

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

Saddle-point formulation

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

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

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

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

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

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

$$\boxed{\text{Assume strong duality; and both } p^* \text{ and } d^* \text{ attained!}}$$

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?
- $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f_0(x) + \sum_i \lambda_i f_i(x)$

$\boxed{\text{Assume strong duality; and both } p^* \text{ and } d^* \text{ attained!}}$

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

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

Example: KKT conditions

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

- 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

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

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?
- $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f_0(x) + \sum_i \lambda_i f_i(x)$

$$\boxed{\text{Assume strong duality; and both } p^* \text{ and } d^* \text{ attained!}}$$

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?
- $g(\lambda) = \inf_x \mathcal{L}(x, \lambda) := f_0(x) + \sum_i \lambda_i f_i(x)$

$$\boxed{\text{Assume strong duality; and both } p^* \text{ and } d^* \text{ attained!}}$$

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

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

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

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

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

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

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

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

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 $f_i(x^*) \leq 0$,

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

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

KKT conditions

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

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

KKT conditions

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

- 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

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

- 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**

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.