Basic information

- http://suvrit.de/mit/optml++
- Current intersections with
  - 6.883 (Jegelka); 6.S978 (Mădry); 18.657 (Rigollet)
  - 9.520 (Poggio); 6.255 (Parrilo); CS229R (Nelson; Harvard)
- Key differences
  - focus (convex, nonconvex, beyond Euclidean)
  - applications, software
  - “bonus” material related to convexity, optimization!
- Typical venues
  - NIPS, ICML, AISTATS, SODA
  - KDD, WWW, CIKM, CVPR, ICLR
  - SIOPT, MathProg, OMS, JOTA
  - SIGMOD, VLDB
  - SISC, SIMAX, IMA JNM,
  - ...
Outline

- Recap on convexity
- Recap on duality, optimality
Convex analysis
Def. Set \( C \subseteq \mathbb{R}^n \) 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 \( C \).

Observations

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

(PSD cone image from convexoptimization.com, Dattorro)
Convex sets

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

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

♥ 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

$$R(A, B) := \left\{ (x^T Ax, x^T Bx) \mid x^T x = 1 \right\}$$

is a compact convex set for $n \geq 3$. 
**Def.** Function \( f : I \rightarrow \mathbb{R} \) on interval \( I \) called **midpoint convex** if

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

**Read:** \( f \) of AM is less than or equal to AM of \( f \).
**Convex functions**

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

$$f \left( \frac{x+y}{2} \right) \leq \frac{f(x) + f(y)}{2},$$

whenever $x, y \in I$.

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

**Think:** What is we use other means, e.g., GM-AM, GM-GM?
Convex functions

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

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

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

**Think:** What is we use other means, e.g., GM-AM, GM-GM?

**Def.** A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is called **convex** if its domain $\text{dom}(f)$ is a convex set and for any $x, y \in \text{dom}(f)$ 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 $f : I \rightarrow \mathbb{R}$ be continuous. Then, $f$ is convex *if and only if* it is midpoint convex.

► Extends to $f : \mathcal{X} \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$; 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 \( PQ \leq \text{slope} \, PR \leq \text{slope} \, QR \)
Theorem: The pointwise sup 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

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

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

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

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

**Proof.** Let $u, v \in \text{dom} f$. Since $f(u) = \inf_{y \in \mathcal{Y}} L(u, y)$, for each $\epsilon > 0$, there is a $y_1 \in \mathcal{Y}$, s.t. $f(u) + \epsilon/2$ is not the infimum. Thus, $L(u, y_1) \leq f(u) + \epsilon/2$.

Similarly, there is $y_2 \in \mathcal{Y}$, such that $L(v, y_2) \leq f(v) + \epsilon/2$.

Now we prove that $f(\lambda u + (1-\lambda)v) \leq \lambda f(u) + (1-\lambda)f(v)$ directly.

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

Since $\epsilon > 0$ is arbitrary, claim follows.
**Theorem** Let $\mathcal{Y}$ be a nonempty convex set. Suppose $L(x, y)$ is convex in $(x, y)$, then,

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

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

**Proof.** Let $u, v \in \text{dom } f$. 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.

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

Since $\epsilon > 0$ is arbitrary, claim follows.
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 $f$ to be a pointwise max of convex functions

♠ By showing $f: \text{dom } (f) \rightarrow \mathbb{R}$ 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 $\{t | x + tv \in \text{dom } (f)\}$).

♠ Exercises (Ch. 3) in Boyd & Vandenberghe

♠ Even more ways exist (may discuss)

Suvrit Sra (MIT) Optimization for ML and beyond: OPTML++
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 $f$ to be a pointwise max of convex functions

Exercises (Ch. 3) in Boyd & Vandenberghe

Even more ways exist (may discuss)
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 $f$ to be a pointwise max of convex functions
♠ By showing $f : \text{dom}(f) \rightarrow \mathbb{R}$ 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 $\{ t \mid x + tv \in \text{dom}(f) \}$).
♠ Exercises (Ch. 3) in Boyd & Vandenberghe
♠ Even more ways exist (may discuss)
Operations preserving convexity

**Pointwise maximum:** \( f(x) = \sup_{y \in Y} f(y; x) \)

**Conic combination:** Let \( a_1, \ldots, a_n \geq 0 \); let \( f_1, \ldots, f_n \) be convex functions. Then, \( f(x) := \sum_i a_i f_i(x) \) is convex.

*Remark:* The set of all convex functions is a *convex cone*.

**Affine composition:** \( f(x) := g(Ax + b) \), where \( g \) is convex.
Operations preserving convexity

**Theorem** Let $f : l_1 \to \mathbb{R}$ and $g : l_2 \to \mathbb{R}$, where range($f$) $\subseteq l_2$. If $f$ and $g$ are convex, and $g$ is increasing, then $g \circ f$ is convex on $l_1$.

**Proof.** Let $x, y \in l_1$, and let $\lambda \in (0, 1)$.

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$$
Theorem Let $f : I_1 \rightarrow \mathbb{R}$ and $g : I_2 \rightarrow \mathbb{R}$, where \( \text{range}(f) \subseteq I_2 \). If $f$ and $g$ are convex, and $g$ is increasing, then $g \circ f$ is convex on $I_1$.

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

\[
\begin{align*}
    f(\lambda x + (1 - \lambda)y) & \leq \lambda f(x) + (1 - \lambda)f(y) \\
    g(f(\lambda x + (1 - \lambda)y)) & \leq g(\lambda f(x) + (1 - \lambda)f(y))
\end{align*}
\]
**Theorem** Let $f : I_1 \to \mathbb{R}$ and $g : I_2 \to \mathbb{R}$, where $\text{range}(f) \subseteq I_2$. If $f$ and $g$ are convex, and $g$ is increasing, then $g \circ f$ is convex on $I_1$.

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

\[
\begin{align*}
    f(\lambda x + (1 - \lambda)y) & \leq \lambda f(x) + (1 - \lambda)f(y) \\
    g(f(\lambda x + (1 - \lambda)y)) & \leq g(\lambda f(x) + (1 - \lambda)f(y)) \\
                                & \leq \lambda g(f(x)) + (1 - \lambda)g(f(y)).
\end{align*}
\]
Convex functions – Indicator

Let \( 1_X \) be the *indicator function* for \( \mathcal{X} \) defined as:

\[
1_X(x) := \begin{cases} 
0 & \text{if } x \in \mathcal{X}, \\
\infty & \text{otherwise}.
\end{cases}
\]

Note: \( 1_X(x) \) is convex if and only if \( \mathcal{X} \) is convex.
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

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

Note: because $\|x - y\|$ is jointly convex in $(x, y)$, the function $\text{dist}(x, \mathcal{X})$ is a convex function of $x$. 
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 ($\ell_2$-norm): $\|x\|_2 = (\sum_i x_i^2)^{1/2}$

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

Example ($\ell_\infty$-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

$$\|A\| := \sup_{\|x\|_2 \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sigma_{\text{max}}(A).$$
Def. The Fenchel conjugate of a function $f$ is

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

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

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

**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

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

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 $f^*(z) = \mathbb{1}_{\|\cdot\|_* \leq 1}(z)$. 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 $\|u\| \leq 1$ 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 $\|z\|_* \leq 1$, then $z^T x \leq \|x\| \|z\|_*$, which implies the sup must be zero.
Fenchel conjugate

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

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

Example \( f(x) = 1_X(x) \): \( f^*(z) = \sup_{x \in X} \langle x, z \rangle \) (aka support func)
Challenge 2

Consider the following functions on strictly positive variables:

\[
\begin{align*}
    h_1(x) & := \frac{1}{x} \\
    h_2(x, y) & := \frac{1}{x} + \frac{1}{y} - \frac{1}{x+y} \\
    h_3(x, y, z) & := \frac{1}{x} + \frac{1}{y} + \frac{1}{z} - \frac{1}{x+y} - \frac{1}{y+z} - \frac{1}{x+z} + \frac{1}{x+y+z}
\end{align*}
\]

♥ 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

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

Arose in studying *expected random broadcast time* in an unreliable star network (I. Affleck, 1994):

\[
h_n(x) = E_n[T(x)] := \sum_{\sigma \in S_n} \left( \prod_{i=1}^{n} \frac{x_{\sigma(i)}}{\sum_{j=i}^{n} x_{\sigma(j)}} \right) \left( \sum_{i=1}^{n} \frac{1}{\sum_{j=i}^{n} x_{\sigma(j)}} \right)
\]
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

\[ f(y) \] are subgradients at \( y \)

\[ g_1, g_2, g_3 \] 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

\[ f(x) := \max(f_1(x), f_2(x)); \text{ both } f_1, f_2 \text{ convex, differentiable} \]
Subgradients – example

\[ f(x) := \max(f_1(x), f_2(x)) \]; both \( f_1, f_2 \) convex, differentiable
$f(x) := \max(f_1(x), f_2(x))$; 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
Subgradients – example

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

\[ f_1(x) \]

\[ f_2(x) \]

\[ f(x) \]

\[ y \]

\[ \star f_1(x) > f_2(x): \text{ unique subgradient of } f \text{ is } f'_1(x) \]
Subgradients – example

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

\[ f_1(x) > f_2(x): \text{ unique subgradient of } f \text{ is } f_1'(x) \]
\[ f_1(x) < f_2(x): \text{ unique subgradient of } f \text{ is } f_2'(x) \]
Subgradients – example

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

\[ \star f_1(x) > f_2(x): \] unique subgradient of \( f \) is \( f_1'(x) \)

\[ \star f_1(x) < f_2(x): \] unique subgradient of \( f \) is \( f_2'(x) \)

\[ \star f_1(y) = f_2(y): \] 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 dom $f$, then $\partial f(x)$ nonempty.
**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 \( x \in \text{relative interior of dom } f \), then \( \partial f(x) \) nonempty
- ♣ If \( f \) differentiable at \( x \), then \( \partial f(x) = \{\nabla f(x)\} \)
**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

\[ f(x) = |x| \]
Subdifferential – example

\[ f(x) = |x| \]
Subdifferential – example

\[ f(x) = |x| \]

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

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

Proof.

$$
\|z\|_2 \geq \|x\|_2 + \langle g, z - x \rangle
$$
Example $f(x) = \|x\|_2$. Then,

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

Proof.

\[
\begin{align*}
\|z\|_2 & \geq \|x\|_2 + \langle g, z - x \rangle \\
\|z\|_2 & \geq \langle g, z \rangle
\end{align*}
\]
Example \( f(x) = \|x\|_2 \). Then,

\[
\partial f(x) := \begin{cases} 
\frac{x}{\|x\|_2} & x \neq 0, \\
\{z | \|z\|_2 \leq 1\} & x = 0.
\end{cases}
\]

Proof.

\[
\|z\|_2 \geq \|x\|_2 + \langle g, z - x \rangle
\]

\[
\|z\|_2 \geq \langle g, z \rangle
\]

\[
\implies \|g\|_2 \leq 1.
\]
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 $f(x) := \max_{1 \leq i \leq m} f_i(x)$, then
  $$\partial f(x) = \text{conv} \bigcup \{\partial f_i(x) \mid f_i(x) = f(x)\},$$
  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.
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

\[
f_1(x) := \begin{cases} -2\sqrt{x} & \text{if } x \geq 0, \\ +\infty & \text{if } x < 0, \end{cases}
\]

and

\[
f_2(x) := \begin{cases} +\infty & \text{if } x > 0, \\ -2\sqrt{-x} & \text{if } x \leq 0. \end{cases}
\]

Then, \( f = \max \{f_1, f_2\} = 1_{\{0\}} \), 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, $f = \max \{f_1, f_2\} = 1_{\{0\}}$, 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 $f(x) = \|x\|_\infty$. Then,

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

where $e_i$ is $i$-th canonical basis vector.

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

Then use, \textit{chain rule} and \textit{max rule} and $\partial | \cdot |$.
Example – subgradients

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

Simple way to obtain some \( g \in \partial f(x) \):
Example – subgradients

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

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 \( g \in \partial f(x) \)

\[
\begin{align*}
  h(z, y^*) & \geq h(x, y^*) + g^T(z - x) \\
  h(z, y^*) & \geq f(x) + g^T(z - x)
\end{align*}
\]
Example – subgradients

\[ f(x) := \sup_{y \in Y} h(x, y) \]

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 \( g \in \partial f(x) \)

\[
\begin{align*}
  h(z, y^*) &\geq h(x, y^*) + g^T(z - x) \\
  h(z, y^*) &\geq f(x) + g^T(z - x) \\
  f(z) &\geq h(z, y) \quad \text{(because of sup)} \\
  f(z) &\geq f(x) + g^T(z - x).
\end{align*}
\]
Example

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

\[
f(x) := \max_{1 \leq i \leq n} (a_i^T x + b_i).
\]

(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

\[
f(x) := \max_{1 \leq i \leq n} (a_i^T x + b_i).
\]

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

$$f(x) := \int f(x, u)p(u)du$$
Suppose $f = \mathbb{E}f(x, u)$, 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, $g(x) = \int g(x, u)p(u)du = \mathbf{E}g(x, u) \in \partial f(x)$
Subgradient of composition

Suppose $h : \mathbb{R}^n \rightarrow \mathbb{R}$ cvx and nondecreasing; each $f_i$ cvx

$$f(x) := h(f_1(x), f_2(x), \ldots, f_n(x)).$$
Subgradient of composition

Suppose $h : \mathbb{R}^n \rightarrow \mathbb{R}$ convex and nondecreasing; each $f_i$ convex

$$f(x) := h(f_1(x), f_2(x), \ldots, f_n(x)).$$

To find a vector $g \in \partial f(x)$, we may:

1. For $i = 1$ to $n$, compute $g_i \in \partial f_i(x)$.
2. Compute $u \in \partial h(f_1(x), \ldots, f_n(x))$.
3. Set $g = u_1 g_1 + u_2 g_2 + \cdots + u_n g_n$; this $g \in \partial f(x)$.
4. Compare with $\nabla f(x) = J \nabla h(x)$, where $J$ matrix of $\nabla f_i(x)$.

Exercise: Verify $g \in \partial f(x)$ by showing $f(z) \geq f(x) + g^T(z - x)$.
Subgradient of composition

Suppose $h : \mathbb{R}^n \to \mathbb{R}$ cvx and nondecreasing; each $f_i$ cvx

$$f(x) := h(f_1(x), f_2(x), \ldots, f_n(x)).$$

To find a vector $g \in \partial f(x)$, we may:

- For $i = 1$ to $n$, compute $g_i \in \partial f_i(x)$
Suppose $h : \mathbb{R}^n \to \mathbb{R}$ cvx and nondecreasing; each $f_i$ cvx

\[ f(x) := h(f_1(x), f_2(x), \ldots, f_n(x)). \]

To find a vector $g \in \partial f(x)$, we may:
- For $i = 1$ to $n$, compute $g_i \in \partial f_i(x)$
- Compute $u \in \partial h(f_1(x), \ldots, f_n(x))$
To find a vector $g \in \partial f(x)$, we may:

- For $i = 1$ to $n$, compute $g_i \in \partial f_i(x)$
- Compute $u \in \partial h(f_1(x), \ldots, f_n(x))$
- Set $g = u_1 g_1 + u_2 g_2 + \cdots + u_n g_n$; this $g \in \partial f(x)$
Subgradient of composition

Suppose $h : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex and nondecreasing; each $f_i$ is convex

$$f(x) := h(f_1(x), f_2(x), \ldots, f_n(x)).$$

To find a vector $g \in \partial f(x)$, we may:

- For $i = 1$ to $n$, compute $g_i \in \partial f_i(x)$
- Compute $u \in \partial h(f_1(x), \ldots, f_n(x))$
- Set $g = u_1 g_1 + u_2 g_2 + \cdots + u_n g_n$; this $g \in \partial f(x)$
- Compare with $\nabla f(x) = J \nabla h(x)$, where $J$ matrix of $\nabla f_i(x)$
Subgradient of composition

Suppose \( h : \mathbb{R}^n \to \mathbb{R} \) cvx and nondecreasing; each \( f_i \) cvx

\[
f(x) := h(f_1(x), f_2(x), \ldots, f_n(x)).
\]

To find a vector \( g \in \partial f(x) \), we may:

- For \( i = 1 \) to \( n \), compute \( g_i \in \partial f_i(x) \)
- Compute \( u \in \partial h(f_1(x), \ldots, f_n(x)) \)
- Set \( g = u_1 g_1 + u_2 g_2 + \cdots + u_n g_n \); this \( g \in \partial f(x) \)
- Compare with \( \nabla f(x) = J \nabla h(x) \), where \( J \) matrix of \( \nabla f_i(x) \)

**Exercise:** Verify \( g \in \partial f(x) \) by showing \( f(z) \geq f(x) + g^T(z - x) \)
Optimization
Optimization problems

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

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

Henceforth, we drop condition on domains for brevity.
Optimization problems

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

$$\min \ f_0(x)$$

s.t. $f_i(x) \leq 0, \ 1 \leq i \leq m,$

$x \in \{\text{dom } f_0 \cap \text{dom } f_1 \cdots \cap \text{dom } f_m\}.$

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

$$p^* := \inf \{ f_0(x) \mid x \in \mathcal{X} \}$$
Let $\mathcal{X}$ be the feasible set and $p^*$ the optimal value

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

- If $\mathcal{X}$ is empty, we say the problem is infeasible.
- By convention, we set $p^* = +\infty$ for infeasible problems.
- If $p^* = -\infty$, we say the 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.
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 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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 $x, y \in \text{dom } f$, we have $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.

♠ Thus, $x^* \text{ is optimal if and only if } \langle \nabla f(x^*), y - x^* \rangle \geq 0$, for all $y \in X$.

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

♠ If $\nabla f(x^*) \neq 0$, it defines supporting hyperplane to $X$ at $x^*$. 
Optimality – constrained

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

♠ Thus, $x^*$ is optimal if and only if

$$\langle \nabla f(x^*), y - x^* \rangle \geq 0, \quad \text{for all } y \in X.$$
For every $x, y \in \text{dom } f$, we have $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.

Thus, $x^*$ is optimal if and only if

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

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

\[
\text{argmin } f = \text{zer}(\partial f) := \left\{ x \in \mathbb{R}^n \mid 0 \in \partial f(x) \right\}.
\]
## Optimality – nonsmooth

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

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

**Proof:** $x \in \text{argmin } f$ implies that $f(x) \leq f(y)$ for all $y \in \mathbb{R}^n$. 
**Theorem** (Fermat’s rule): Let $f : \mathbb{R}^n \rightarrow (-\infty, +\infty]$. Then,

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

**Proof:** $x \in \text{argmin } f$ implies that $f(x) \leq f(y)$ for all $y \in \mathbb{R}^n$. Equivalently, $f(y) \geq f(x) + \langle 0, y - x \rangle \quad \forall y,$
Theorem (Fermat’s rule): Let $f : \mathbb{R}^n \to (-\infty, +\infty]$. Then,

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

Proof: $x \in \text{argmin } f$ 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,

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

**Proof:** $x \in \arg\min f$ 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**

\[
\begin{align*}
\min & \quad f(x) & \text{s.t.} & \quad x \in \mathcal{X} \\
\min & \quad f(x) + 1_{\mathcal{X}}(x).
\end{align*}
\]
Minimizing $x$ must satisfy: $0 \in \partial (f + 1 \mathbb{1}_x) (x)$
Optimality – nonsmooth

- Minimizing $x$ must satisfy: $0 \in \partial (f + 1_X)(x)$
- (CQ) Assuming $\text{ri}(\text{dom } f) \cap \text{ri}(X) \neq \emptyset$, $0 \in \partial f(x) + \partial 1_X(x)$
Optimality – nonsmooth

- Minimizing $x$ must satisfy: $0 \in \partial(f + 1_x)(x)$
- (CQ) Assuming $\text{ri}(\text{dom } f) \cap \text{ri}(X) \neq \emptyset$, $0 \in \partial f(x) + \partial 1_X(x)$
- Recall, $g \in \partial 1_X(x)$ iff $1_X(y) \geq 1_X(x) + \langle g, y - x \rangle$ for all $y$. 
Minimizing $x$ must satisfy: $0 \in \partial(f + 1_x)(x)$

(CQ) Assuming $\text{ri}(\text{dom } f) \cap \text{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f(x) + \partial 1_x(x)$

Recall, $g \in \partial 1_x(x)$ iff $1_x(y) \geq 1_x(x) + \langle g, y - x \rangle$ for all $y$.

So $g \in \partial 1_x(x)$ means $x \in \mathcal{X}$ and $0 \geq \langle g, y - x \rangle \ \forall y \in \mathcal{X}$. 
Optimality – nonsmooth

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

\[
\mathcal{N}_\mathcal{X}(x) := \{ g \in \mathbb{R}^n \mid 0 \geq \langle g, y - x \rangle \ \forall y \in \mathcal{X} \}
\]
Minimizing $x$ must satisfy: $0 \in \partial (f + 1_X)(x)$

(CQ) Assuming $\text{ri} (\text{dom } f) \cap \text{ri}(\mathcal{X}) \neq \emptyset$, $0 \in \partial f(x) + \partial 1_X(x)$

Recall, $g \in \partial 1_X(x)$ iff $1_X(y) \geq 1_X(x) + \langle g, y - x \rangle$ for all $y$.

So $g \in \partial 1_X(x)$ means $x \in \mathcal{X}$ and $0 \geq \langle g, y - x \rangle \forall y \in \mathcal{X}$.

Normal cone:

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

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

$\diamond$ If $f$ is diff., we get $0 \in \nabla f(x^*) + \mathcal{N}_X(x^*)$
Optimality – nonsmooth

- Minimizing \( x \) must satisfy: \( 0 \in \partial (f + 1_X)(x) \)
- (CQ) Assuming \( \text{ri}(\text{dom } f) \cap \text{ri}(\mathcal{X}) \neq \emptyset \), \( 0 \in \partial f(x) + \partial 1_X(x) \)
- Recall, \( g \in \partial 1_X(x) \) iff \( 1_X(y) \geq 1_X(x) + \langle g, y - x \rangle \) for all \( y \).
- So \( g \in \partial 1_X(x) \) means \( x \in \mathcal{X} \) and \( 0 \geq \langle g, y - x \rangle \) \( \forall y \in \mathcal{X} \).
- Normal cone:
  \[
  \mathcal{N}_\mathcal{X}(x) := \{ g \in \mathbb{R}^n | 0 \geq \langle g, y - x \rangle \ \forall y \in \mathcal{X} \}
  \]

**Application.** \( \min f(x) \) s.t. \( x \in \mathcal{X} \):
- \( \diamond \) If \( f \) is diff., we get \( 0 \in \nabla f(x^*) + \mathcal{N}_\mathcal{X}(x^*) \)
- \( \diamond \) \( -\nabla f(x^*) \in \mathcal{N}_\mathcal{X}(x^*) \iff \langle \nabla f(x^*), y - x^* \rangle \geq 0 \) for all \( y \in \mathcal{X} \).
Optimality – projection operator

\[ P_{\mathcal{X}}(y) := \arg\min_{x \in \mathcal{X}} \|x - y\|^2 \]

(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:**
  \[ \|P_{\mathcal{X}}(x) - P_{\mathcal{X}}(y)\|^2 \leq \|x - y\|^2 \text{ for all } x, y \in \mathbb{R}^n. \]
Optimality – projection operator

\[ P_\mathcal{X}(y) := \arg\min_{x \in \mathcal{X}} \| x - y \|^2 \]

(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:

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

Proof: Exercise!
Duality
Primal problem

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

$$\min \quad f_0(x)$$

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

(P)

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

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

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

♠ Variables $\lambda \in \mathbb{R}^m$ called **Lagrange multipliers**
To the primal problem, associate **Lagrangian** $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, 

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

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

♠ Suppose $x$ is feasible, and $\lambda \geq 0$. Then, we get the lower-bound:

$$f_0(x) \geq L(x, \lambda) \quad \forall x \in \mathcal{X}, \lambda \in \mathbb{R}_+^m.$$
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) \quad \forall x \in \mathcal{X}, \lambda \in \mathbb{R}_+.\]

♠ Lagrangian helps write problem in unconstrained form
Def. We define the **Lagrangian dual** as

\[ g(\lambda) := \inf_x \mathcal{L}(x, \lambda). \]
Lagrange dual function

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

\[ g(\lambda) := \inf_x \mathcal{L}(x, \lambda). \]

**Observations:**

- \( g \) is pointwise inf of affine functions of \( \lambda \)
- Thus, \( g \) is concave; it may take value \(-\infty\)
Def. We define the **Lagrangian dual** as

\[ g(\lambda) := \inf_x \mathcal{L}(x, \lambda). \]

**Observations:**

- \( g \) is pointwise inf of affine functions of \( \lambda \)
- 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

\[ \forall \lambda \in \mathbb{R}_+^m \quad p^* \geq g(\lambda). \]
Lagrange dual problem

\[
\sup_{\lambda} g(\lambda) \quad \text{s.t.} \quad \lambda \geq 0.
\]
Lagrange dual problem

\[
\sup_{\lambda} g(\lambda) \quad \text{s.t. } \lambda \geq 0.
\]

- **dual feasible**: if \( \lambda \geq 0 \) and \( g(\lambda) > -\infty \)
- **dual optimal**: \( \lambda^* \) if sup is achieved
- Lagrange dual is always concave, regardless of original
**Def.** Denote **dual optimal value** by $d^*$, i.e.,

$$d^* := \sup_{\lambda \geq 0} g(\lambda).$$
Def. Denote dual optimal value by $d^*$, i.e.,

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

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

Weak duality
Weak duality

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

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

**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!

“Easy” necessary and sufficient conditions: **unknown**
Zero duality gap: nonconvex example

**Trust region subproblem (TRS)**

\[
\min x^T Ax + 2b^T x \quad x^T x \leq 1.
\]

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 \( D = \{(x, y) \mid y > 0\} \).

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

\[ \min_{x, y} e^{-x} \quad \frac{x^2}{y} \leq 0, \]

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

Clearly, only feasible \( x = 0 \). So \( p^* = 1 \)

\[ \mathcal{L}(x, y, \lambda) = e^{-x} + \lambda \frac{x^2}{y}, \]

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**

\[ d^* = \max_{\lambda} 0 \quad \text{s.t. } \lambda \geq 0. \]

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 \( D = \{(x, y) \mid y > 0\} \).

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

\[
d^* = \max_{\lambda} 0 \quad \text{s.t. } \lambda \geq 0.
\]

Thus, \( d^* = 0 \), and gap is \( p^* - d^* = 1 \).

Here, we had no strictly feasible solution.
Support vector machine

\[
\begin{align*}
\min_{x, \xi} & \quad \frac{1}{2} \|x\|_2^2 + C \sum_i \xi_i \\
\text{s.t.} & \quad Ax \geq 1 - \xi, \quad \xi \geq 0.
\end{align*}
\]
Support vector machine

$$\min_{x, \xi} \frac{1}{2} \|x\|_2^2 + C \sum_i \xi_i$$

s.t. \quad Ax \geq 1 - \xi, \quad \xi \geq 0.

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

g(\lambda, \nu) := \inf L(x, \xi, \lambda, \nu) = \begin{cases} \lambda^T 1 - \frac{1}{2} \|A^T \lambda\|_2^2 + \nu = C + \infty \text{ otherwise} \end{cases}

d^* = \max_{\lambda \geq 0, \nu \geq 0} g(\lambda, \nu)$$

Exercise: Using $\nu \geq 0$, eliminate $\nu$ from above problem.
Support vector machine

\[
\begin{align*}
\min_{x, \xi} & \quad \frac{1}{2} \| x \|_2^2 + C \sum_i \xi_i \\
\text{s.t.} & \quad A x \geq 1 - \xi, \quad \xi \geq 0.
\end{align*}
\]

\[
L(x, \xi, \lambda, \nu) = \frac{1}{2} \| x \|_2^2 + C 1^T \xi - \lambda^T (A x - 1 + \xi) - \nu^T \xi
\]

\[
g(\lambda, \nu) := \inf L(x, \xi, \lambda, \nu)
\]

\[
= \begin{cases} \\
\lambda^T 1 - \frac{1}{2} \| A^T \lambda \|_2^2 & \lambda + \nu = C 1 \\
+\infty & \text{otherwise}
\end{cases}
\]

\[
d^* = \max_{\lambda \geq 0, \nu \geq 0} g(\lambda, \nu)
\]

Exercise: Using \(\nu \geq 0\), eliminate \(\nu\) from above problem.
Regularized optimization

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

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

**Dual problem**

\[
\inf_{u \in \mathcal{Y}} \ f^*(-A^T u) + r^*(u).
\]
Regularized optimization

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

**Dual problem**

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

▶ **Introduce new variable** \( z = Ax \)

\[
\inf_{x \in \mathcal{X}, \ z \in \mathcal{Y}} \ f(x) + r(z), \quad \text{s.t.} \quad z = Ax.
\]
Regularized optimization

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

**Dual problem**

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

- Introduce new variable \( z = Ax \)

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

- 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

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

**Dual problem**

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

- Introduce new variable \( z = Ax \)

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

- 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

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

Dual problem

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

The infimum above can be rearranged as follows

\[
g(y) = \inf_{x \in X} f(x) + y^T A x + \inf_{z \in \mathcal{Y}} r(z) - y^T z
\]
Regularized optimization

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

**Dual problem**

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

The infimum above can be rearranged as follows

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

\[
= -\sup_{x \in \mathcal{X}} \left\{ -x^T A^T y - f(x) \right\} - \sup_{z \in \mathcal{Y}} \left\{ z^T y - r(z) \right\}
\]
Regularized optimization

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

**Dual problem**

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

The infimum above can be rearranged as follows

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

\[
= -\sup_{x \in \mathcal{X}} \left\{ -x^T A^T y - f(x) \right\} - \sup_{z \in \mathcal{Y}} \left\{ z^T y - r(z) \right\}
\]

\[
= -f^*(-A^T y) - r^*(y) \quad \text{s.t.} \quad y \in \mathcal{Y}.
\]
Regularized optimization

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

**Dual problem**

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

The infimum above can be rearranged as follows

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

\[
= -\sup_{x \in \mathcal{X}} \left\{-x^T A^T y - f(x)\right\} - \sup_{z \in \mathcal{Y}} \left\{z^T y - r(z)\right\}
\]

\[
= -f^*(-A^T y) - r^*(y) \quad \text{s.t. } y \in \mathcal{Y}.
\]

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

\[
\inf_{y \in \mathcal{Y}} f^*(-A^T y) + r^*(y).
\]
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. \(\exists x \in \text{ri(dom } f)\) such that \(Ax \in \text{ri(dom } r)\)
2. \(\exists y \in \text{ri(dom } r^*)\) such that \(A^T y \in \text{ri(dom } f^*)\)
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 ‘inf’ attained at some \( x \)
- Condition 2 ensures ‘sup’ attained at some \( y \)
Example: norm regularized problems

$$\min \quad f(x) + \|Ax\|$$
Example: norm regularized problems

\[ \min \ f(x) + \|Ax\| \]

**Dual problem**

\[ \min_y \ f^*(-A^Ty) \quad \text{s.t.} \quad \|y\|_* \leq 1. \]
Example: norm regularized problems

\[ \min f(x) + \|Ax\| \]

**Dual problem**

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

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

$$\min \quad f(x) + g(x)$$
Example: variable splitting

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

Exercise: Fill in the details below

$$\min_{x,z} f(x) + g(z) \ \text{s.t.} \ \ x = z$$
Example: variable splitting

\[ \min f(x) + g(x) \]

**Exercise:** Fill in the details below

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

\[ L(x, z, \nu) = f(x) + g(z) + \nu^T(x - z) \]
Example: variable splitting

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

**Exercise:** Fill in the details below

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

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

$$g(\nu) = \inf_{x,z} L(x, z, \nu)$$
Theorem Let $\phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{-\infty, +\infty\}$ be any function. Then,

$$\sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} \phi(x, y) \leq \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \phi(x, y).$$
**Primal-dual: weak minimax**

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

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

**Proof:**

\[
\forall x, y, \quad \inf_{x' \in \mathcal{X}} \phi(x', y) \leq \phi(x, y)
\]
**Primal-dual: weak minimax**

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

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

**Proof:**

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

$$
\forall x, y, \quad \inf_{x' \in \mathcal{X}} \phi(x', y) \leq \sup_{y' \in \mathcal{Y}} \phi(x, y')
$$
Theorem Let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm \infty\}$ be any function. Then,

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

Proof:

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

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

$$\forall x, \sup_{y \in \mathcal{Y}} \inf_{x' \in \mathcal{X}} \phi(x', y) \leq \sup_{y' \in \mathcal{Y}} \phi(x, y')$$
Theorem Let \( \phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{\pm \infty\} \) be any function. Then,

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

Proof:

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

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

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

\[
\Rightarrow \sup_{y \in \mathcal{Y}} \inf_{x' \in \mathcal{X}} \phi(x', y) \leq \inf_{x \in \mathcal{X}} \sup_{y' \in \mathcal{Y}} \phi(x, y').
\]
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 \(\phi\) be as before. A point \((x^*, y^*)\) is a saddle-point of \(\phi\) (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 \((x^*, y^*)\)

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

**Def.** Let \(\phi\) be as before. A point \((x^*, y^*)\) is a saddle-point of \(\phi\) (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.

\[
x^* \in \arg\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \phi(x, y) \quad \text{and} \quad y^* \in \arg\max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \phi(x, y).
\]
Sufficient conditions for saddle-point

♠ Classes of problems “dual” to each other can be generated by studying classes of functions $\phi$, 

- Function $\phi$ is continuous, and
- It is convex-concave ($\phi(x, \cdot)$ convex for every $y \in Y$, and $\phi(\cdot, y)$ concave for every $x \in X$), and
- Both $X$ and $Y$ are convex; one of them is compact.
Sufficient conditions for saddle-point

♠ Classes of problems “dual” to each other can be generated by studying classes of functions \( \phi \),

♠ More interesting question: Starting from the primal problem over \( \mathcal{X} \), how to introduce a space \( \mathcal{Y} \) and a “useful” function \( \phi \) on \( \mathcal{X} \times \mathcal{Y} \) so that we have a saddle-point?

► Function \( \phi \) 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

\[ p^* := \min_x \|Ax - b\|_2 + \lambda \|x\|_1. \]
Example: Lasso-like problem

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

\[ \| x \|_1 = \max \left\{ x^T v \mid \| v \|_\infty \leq 1 \right\} \]

\[ \| x \|_2 = \max \left\{ x^T u \mid \| u \|_2 \leq 1 \right\}. \]
Example: Lasso-like problem

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

\[ \| x \|_1 = \max \left\{ x^T v \mid \| v \|_\infty \leq 1 \right\} \]

\[ \| x \|_2 = \max \left\{ x^T u \mid \| u \|_2 \leq 1 \right\} \]

**Saddle-point formulation**

\[ p^* = \min_x \max_{u,v} \left\{ u^T (b - Ax) + v^T x \mid \| u \|_2 \leq 1, \| v \|_\infty \leq \lambda \right\} \]
Example: Lasso-like problem

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

\[ \| x \|_1 = \max \left\{ x^T v \mid \| v \|_\infty \leq 1 \right\} \]

\[ \| x \|_2 = \max \left\{ x^T u \mid \| u \|_2 \leq 1 \right\}. \]

Saddle-point formulation

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

\[ = \max_{u,v} \min_x \left\{ u^T (b - Ax) + x^T v \mid \| u \|_2 \leq 1, \| v \|_\infty \leq \lambda \right\} \]
Example: Lasso-like problem

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

\[ \|x\|_1 = \max \left\{ x^T v \mid \|v\|_\infty \leq 1 \right\} \]

\[ \|x\|_2 = \max \left\{ x^T u \mid \|u\|_2 \leq 1 \right\} \]

Saddle-point formulation

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

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

\[ = \max_{u,v} u^T b \quad A^T u = v, \|u\|_2 \leq 1, \|v\|_\infty \leq \lambda \]
Example: Lasso-like problem

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

\[ \| x \|_1 = \max \left\{ x^T v \mid \| v \|_\infty \leq 1 \right\} \]

\[ \| x \|_2 = \max \left\{ x^T u \mid \| u \|_2 \leq 1 \right\}. \]

Saddle-point formulation

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

\[ = \max_{u,v} \min_{x} \left\{ u^T (b - Ax) + x^T v \mid \| u \|_2 \leq 1, \| v \|_\infty \leq \lambda \right\} \]

\[ = \max_{u,v} u^T b \quad A^T u = v, \| u \|_2 \leq 1, \| v \|_\infty \leq \lambda \]

\[ = \max_{u} u^T b \quad \| u \|_2 \leq 1, \| A^T v \|_\infty \leq \lambda. \]
Example: KKT conditions

\[ \min f_0(x) \quad f_i(x) \leq 0, \quad i = 1, \ldots, m. \]
Example: KKT conditions

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

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

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

- 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

\[
\begin{align*}
\min & \quad f_0(x) \quad f_i(x) \leq 0, \quad i = 1, \ldots, m. \\
\end{align*}
\]

- 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!
Example: KKT conditions

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

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

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

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

\[
p^* = f_0(x^*)
\]
Example: KKT conditions

\[
\begin{align*}
\min & \quad f_0(x) \quad f_i(x) \leq 0, \quad i = 1, \ldots, m. \\
\end{align*}
\]

- 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 \((x^*, \lambda^*)\) such that

\[
p^* = f_0(x^*) = d^* = g(\lambda^*)
\]
Example: KKT conditions

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

- 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 \( (x^*, \lambda^*) \) such that

\[
p^* = f_0(x^*) = d^* = g(\lambda^*) = \min_x \mathcal{L}(x, \lambda^*)
\]
Example: KKT conditions

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

- 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 \((x^*, \lambda^*)\) such that

\[
p^* = f_0(x^*) = d^* = g(\lambda^*) = \min_x \mathcal{L}(x, \lambda^*) \leq \mathcal{L}(x^*, \lambda^*)
\]
Example: KKT conditions

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

- 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 \( (x^*, \lambda^*) \) 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

\[
\begin{align*}
\min\; f_0(x) & \quad f_i(x) \leq 0, \quad i = 1, \ldots, m. \\
\end{align*}
\]

- 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 \( (x^*, \lambda^*) \) such that

\[
\begin{align*}
p^* = f_0(x^*) = d^* = g(\lambda^*) &= \min_x \mathcal{L}(x, \lambda^*) \leq \mathcal{L}(x^*, \lambda^*) \leq f_0(x^*) = p^* \\
\end{align*}
\]

- Thus, equalities hold in above chain.
Example: KKT conditions

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

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

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

Thus, there exists a pair \((x^*, \lambda^*)\) 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.

\[
x^* \in \arg\min_x \mathcal{L}(x, \lambda^*).
\]
Example: KKT conditions

\[ x^* \in \arg\min_x \mathcal{L}(x, \lambda^*). \]

If \( f_0, f_1, \ldots, f_m \) are differentiable, this implies
Example: KKT conditions

\[ x^* \in \arg\min_x \mathcal{L}(x, \lambda^*). \]

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

\[ x^* \in \text{argmin}_x \mathcal{L}(x, \lambda^*). \]

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
Example: KKT conditions

\[ x^* \in \text{argmin}_x \mathcal{L}(x, \lambda^*). \]

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

\[ x^* \in \arg\min_x \mathcal{L}(x, \lambda^*). \]

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

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

But \( \lambda_i^* \geq 0 \) and \( f_i(x^*) \leq 0 \),
Example: KKT conditions

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

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

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

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

\[ f_i(x^*) \leq 0, \quad i = 1, \ldots, m \quad \text{(primal feasibility)} \]
\[ \lambda^*_i \geq 0, \quad i = 1, \ldots, m \quad \text{(dual feasibility)} \]
\[ \lambda^*_i f_i(x^*) = 0, \quad i = 1, \ldots, m \quad \text{(compl. slackness)} \]
\[ \nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} = 0 \quad \text{(Lagrangian stationarity)} \]
KKT conditions

\[ f_i(x^*) \leq 0, \quad i = 1, \ldots, m \]  \quad \text{(primal feasibility)}

\[ \lambda^*_i \geq 0, \quad i = 1, \ldots, m \]  \quad \text{(dual feasibility)}

\[ \lambda^*_i f_i(x^*) = 0, \quad i = 1, \ldots, m \]  \quad \text{(compl. slackness)}

\[ \nabla_x \mathcal{L}(x, \lambda^*)|_{x=x^*} = 0 \]  \quad \text{(Lagrangian stationarity)}

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

\[ f_i(x^*) \leq 0, \quad i = 1, \ldots, m \]  
(\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})

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

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

▶ We showed: if strong duality holds, and \((x^*, \lambda^*)\) exist, then KKT conditions are necessary for pair \((x^*, \lambda^*)\) 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 \((x^*, \lambda^*)\) optimal primal-dual pair.