



Introduction to Optimization Theory

Nirupam Gupta

Department of Computer Science

UNIVERSITY OF COPENHAGEN





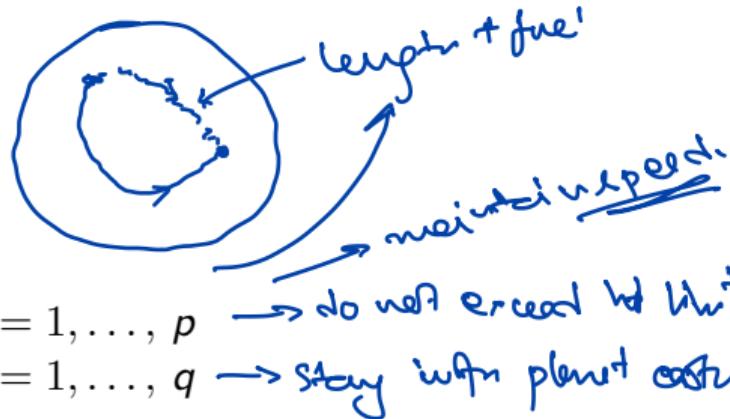
What is optimization?

Mathematical optimization

Minimize $f(w)$

Subject to $f_i(w) \leq 0 \quad i = 1, \dots, p$ → do not exceed w limit.

$g_i(w) = 0 \quad i = 1, \dots, q$ → stay w/in plant const

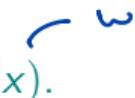


- $w \in \mathbb{R}^d$ represents scalar **variables** $w_1, \dots, w_d \in \mathbb{R}$, to be computed. ✓
- $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is the **objective function**, to be minimized. ✓
- $\{f_1, \dots, f_p\}$ with $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ are the **inequality constraints**. ✓
- $\{g_1, \dots, g_q\}$ with $g_i : \mathbb{R}^d \rightarrow \mathbb{R}$ are the **equality constraints**. ✓

Mathematical optimization

$$\begin{array}{ll}\text{Minimize} & f(w) \\ \text{Subject to} & f_i(w) \leq 0 \quad i = 1, \dots, p \\ & g_i(w) = 0 \quad i = 1, \dots, q\end{array}$$

- $w \in \mathbb{R}^d$ represents scalar **variables** $w_1, \dots, w_d \in \mathbb{R}$, to be computed.
- $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is the **objective function**, to be minimized.
- $\{f_1, \dots, f_p\}$ with $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ are the **inequality constraints**.
- $\{g_1, \dots, g_q\}$ with $g_i : \mathbb{R}^d \rightarrow \mathbb{R}$ are the **equality constraints**.

Minimizing $f(x)$ is equivalent to maximizing $-f(x)$. 

Optimization in machine learning

Solving **empirical risk minimization** (ERM) for designing classifiers.

Example: **Linear Classifier**

- n samples (input-output pairs) $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^m \times \{-1, 1\}$.
- Linear **hypothesis class** $\mathcal{H} = \{h : x \mapsto \text{sign}(w^\top x + w_0) \mid (w, w_0) \in \mathbb{R}^m \times \mathbb{R}\}$.
- For each sample (x_i, y_i) , **define loss** by $f_i(w, w_0) := \mathbb{1}[h(x_i) \neq y_i]$ where $\mathbb{1}[\cdot]$ is the indicator function

Solve the following **optimization problem**:

$$\begin{array}{ll} \text{Minimize} & f(w, w_0) := \frac{1}{n} \sum_{i=1}^n f_i(w, w_0) \\ \text{Subject to} & \|(w, w_0)\|_0 - k \leq 0 \end{array} \quad \leftarrow \text{analoge verallgemeinerung} \quad (\text{ERM})$$

The inequality constraint ensures sparsity in the weights when $k < m$. \leftarrow keep model sparse.

Optimization in machine learning (cont'd)

$\|\cdot\|_0$ - ~~\mathbb{R}^2~~ no. of non-zero entries.

Training artificial neural networks

- n samples (input-output pairs) $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^m \times \{-1, +1\}$.
- Hypothesis $h(x) : \mathbb{R}^m \rightarrow \mathbb{R}^2$ is typically non-linear, determined by tunable parameters $w \in \mathbb{R}^d$.
- For each sample (x_i, y_i) , define loss by cross-entropy function

$$f_i(w) := \mathbb{1}[y = +1] \frac{e^{[h(x)]_1}}{e^{[h(x)]_1} + e^{[h(x)]_2}} + \mathbb{1}[y = -1] \frac{e^{[h(x)]_2}}{e^{[h(x)]_1} + e^{[h(x)]_2}}$$

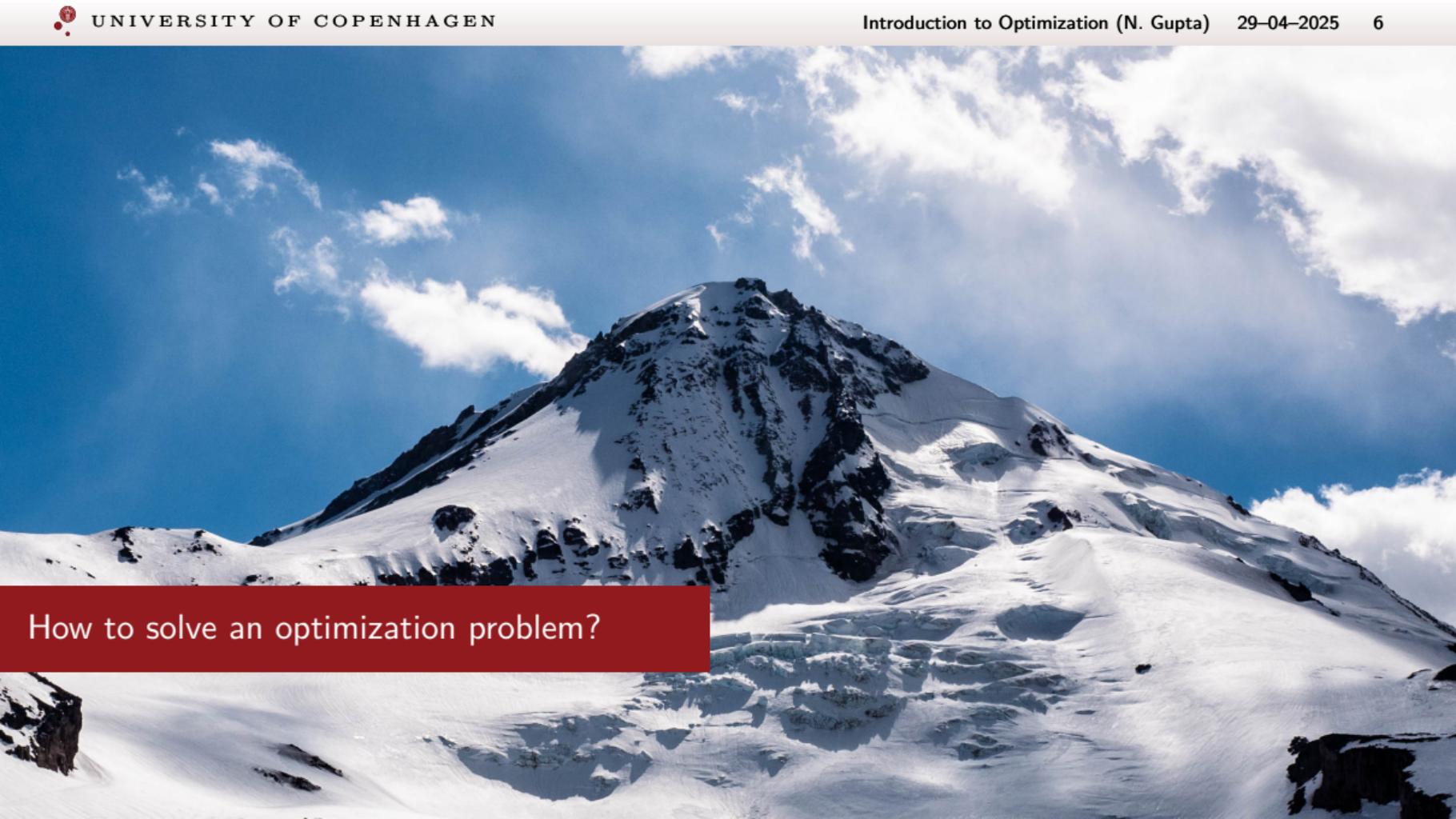
where $[\cdot]_k$ denotes the k -th coordinate value of the vector.

Solve the following **optimization problem**:

$$\begin{array}{ll} \text{Minimize} & f(w) := \frac{1}{n} \sum_{i=1}^n f_i(w) \\ \text{Subject to} & \|w\|_0 - k \leq 0 \end{array}$$

Inequality constraint ensures sparsity of the network when $k < m$.

(ERM)

The background of the slide features a majestic, snow-capped mountain peak against a backdrop of a clear blue sky with wispy white clouds. The mountain's surface is a mix of bright white snow and dark, rocky outcrops. In the foreground, there's a solid red rectangular block containing the text.

How to solve an optimization problem?

Solving an optimization problem

$$\underset{w \in \mathbb{R}^d}{\text{Minimize}} \quad f(w)$$

$$\begin{aligned} \text{Subject to} \quad & f_i(w) \leq 0 \quad i = 1, \dots, p \\ & g_i(w) = 0 \quad i = 1, \dots, q \end{aligned}$$

In general, it is *practically impossible* to solve the above problem.

$$\begin{aligned} w \in \{0, 1\}^d & \quad f: w \mapsto \{0, 1\} \\ \# f's & \quad 2^{2^d} \\ & \quad \nearrow 2^d \\ S \subset \{0, 1\}^d & \\ f(w) = 0, \quad w \in S & \\ f(w) = 1, \quad w \notin S & \end{aligned}$$

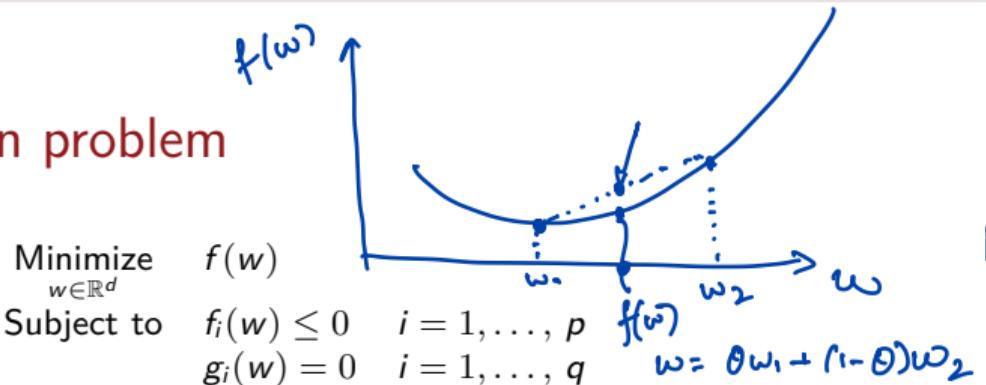
Solving an optimization problem

$$\begin{aligned} & \underset{w \in \mathbb{R}^d}{\text{Minimize}} \quad f(w) \\ & \text{Subject to} \quad f_i(w) \leq 0 \quad i = 1, \dots, p \\ & \quad g_i(w) = 0 \quad i = 1, \dots, q \end{aligned}$$

In general, it is *practically impossible* to solve the above problem.

However, many optimization problems can be solved efficiently, e.g., **convex optimization** problems.

Solving an optimization problem



In general, it is *practically impossible* to solve the above problem.

However, many optimization problems can be solved efficiently, e.g., **convex optimization** problems.

- **Objective function** $f(w)$ and **inequality constraints** $f_1(w), \dots, f_p(w)$ are **convex** functions -

$$f(\theta w_1 + (1 - \theta) w_2) \leq \theta f(w_1) + (1 - \theta) f(w_2), \quad \text{for all } w_1, w_2 \in \mathbb{R}^d \text{ and } \theta \in [0, 1].$$

Jensen's Ineq.

$$\frac{f(w_1) + f(w_2)}{2}$$

$$g(w) ; \quad g(\alpha w + b) = \alpha g(w) + b.$$

Solving an optimization problem

w is feasible if $w \in S_{\text{feas}}$.

S_{pos} is a convex set.

$$\left[\begin{array}{l} \text{Minimize}_{w \in \mathbb{R}^d} f(w) \\ \text{Subject to } \begin{array}{ll} f_i(w) \leq 0 & i = 1, \dots, p \\ g_i(w) = 0 & i = 1, \dots, q \end{array} \end{array} \right] \text{ convex.}$$

In general, it is *practically impossible* to solve the above problem.

However, many optimization problems can be solved efficiently, e.g., **convex optimization** problems.

- **Objective function** $f(w)$ and **inequality constraints** $f_1(w), \dots, f_p(w)$ are **convex** functions -

$$f(\theta w_1 + (1 - \theta) w_2) \leq \theta f(w_1) + (1 - \theta) f(w_2), \quad \text{for all } w_1, w_2 \in \mathbb{R}^d \text{ and } \theta \in [0, 1].$$

- **Equality constraints** $g_1(w), \dots, g_q(w)$ are affine functions:

$$g_i(w) := a_i^T w + b_i, \quad \text{where } (a_i, b_i) \in \mathbb{R}^d \times \mathbb{R}.$$

For $u, v \in \mathbb{R}^d$, we define their **inner product** by $\langle u, v \rangle = u^T v$.



Convex optimization

Convex set

A set S is *convex* if the line segment between any two points in S lies in S , i.e., if for any $x_1, x_2 \in S$ and $\theta \in [0, 1]$ we have $\theta x_1 + (1 - \theta)x_2 \in S$.

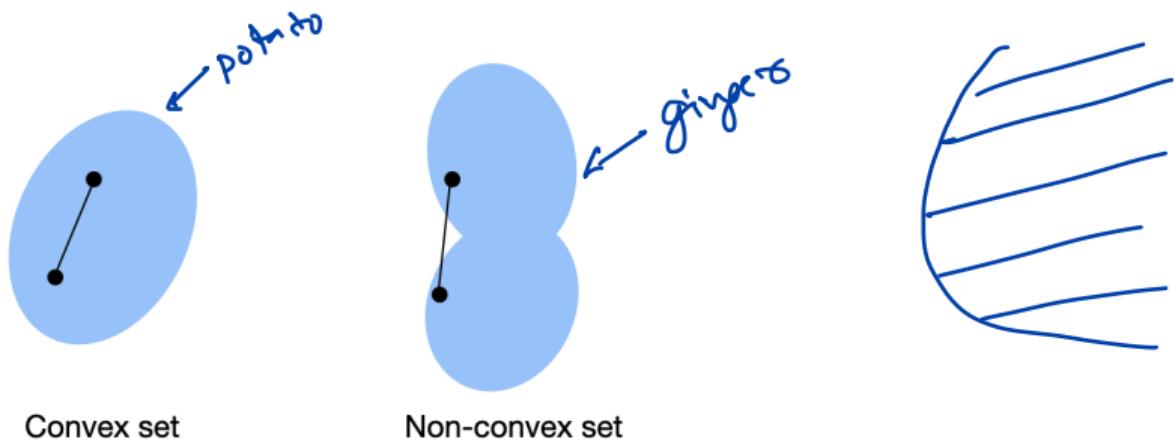


Figure: Difference between convex and nonconvex sets.

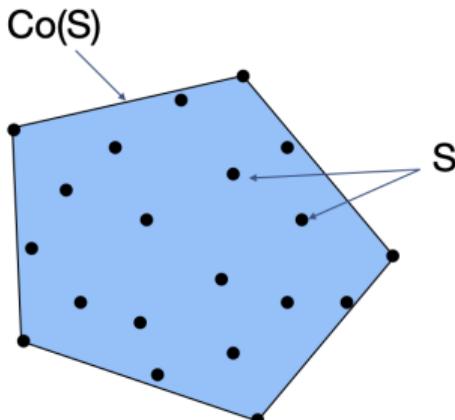
Convex Hull



For a set S , its *convex hull*, denoted by $\text{Co}(S)$ is defined to be the smallest convex set containing S .

Suppose that $S = (x_1, \dots, x_n) \subset \mathbb{R}^m$. Then,

$$\text{Co}(S) = \left(x = \sum_{i=1}^n \theta_i x_i \mid \theta_i \geq 0 \text{ for all } i, \sum_{i=1}^n \theta_i = 1 \right).$$



Separating hyperplane theorem [very important].

Separating hyperplane theorem. Consider 2 disjoint convex sets C and D in \mathbb{R}^m , i.e., $C \cap D = \emptyset$. There exists $(w, b) \in \mathbb{R}^m \times \mathbb{R}$ with $w \neq \mathbf{0}$ s.t. $w^T x + b \geq 0, \forall x \in C$ and $w^T x + b \leq 0, \forall x \in D$.

strict ineq. need

Supporting Hyperplanes

not be true in general: Ex:

$$C = \{x \in \mathbb{R}^2 \mid x_1 < 0\} \cup (0, 0)$$

$$D = \{x \in \mathbb{R}^2 \mid x_2 > 0\} \cup (0, 1)$$

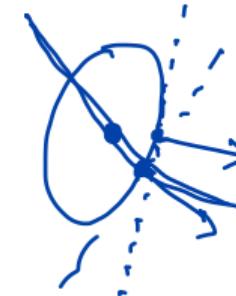
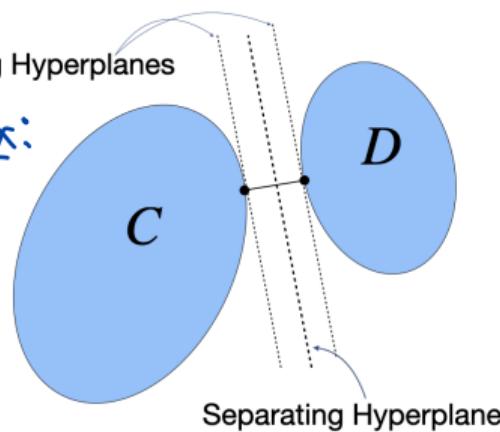


Figure: Separating and supporting hyperplanes for two disjoint closed convex sets.

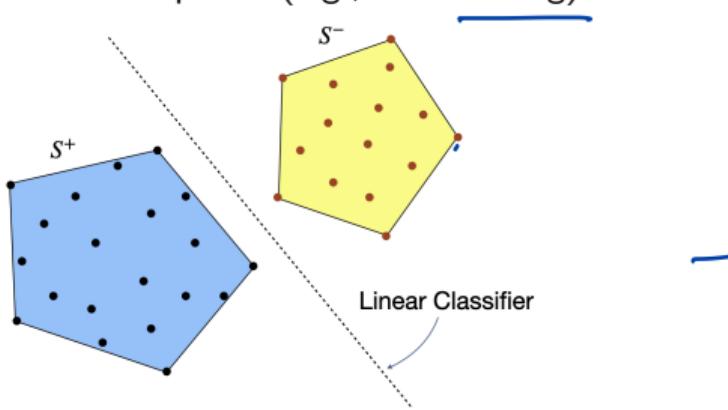
[Refer to Section of 2.5 of Convex Optimization, book by Boyd and Vandenberghe, for a proof.]

Linear separability

Consider a set of data points $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ such that $x_i \in \mathbb{R}^m$ and $y_i \in \{-1, +1\}$ for all i . Define $S^+ = \{x \mid (x, y) \in S, y = +1\}$ and $S^- = \{x \mid (x, y) \in S, y = -1\}$.

Exercise: Suppose that $\text{Co}(S^+) \cap \text{Co}(S^-) = \emptyset$. Show that \exists a hyperplane $f(x) := w^\top x + b$ such that $f(x) > 0, \forall x \in S^+$ and $f(x) < 0, \forall x \in S^-$.

To compute the classifier, we only need *boundary* points of the respective convex hulls, which can be quite small compared to the original number of data points (e.g., see this [blog](#)).



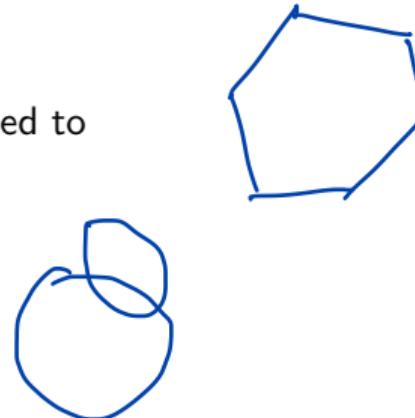
Examples of convex set

Polyhedron. Let $A \in \mathbb{R}^{p \times m}$ and $C \in \mathbb{R}^{q \times m}$. A polyhedron is defined to

$$(x \in \mathbb{R}^m \mid Ax \preceq b, Cx = d)$$

where \preceq is coordinate-wise inequality.

Intersection. Intersection of convex sets is convex.



Perspective mapping. Let $C \subset \mathbb{R}^m \times \mathbb{R}_{++}$ be a convex set. Then the image of C under the perspective mapping $P : (x, t) \mapsto \frac{x}{t}$ is convex, i.e., $D := \left(\frac{x}{t} \mid (x, t) \in C \right)$ is a convex set.

The inverse image of a convex set $D \subset \mathbb{R}^m$ under the perspective mapping $P : (x, t) \mapsto \frac{x}{t}$ is also convex, i.e., $C := \left((x, t) \mid \frac{x}{t} \in D \right)$ is a convex set.

↳ KL-Divergence



Convex function

Convex function: $f : \mathbb{R}^m \rightarrow \mathbb{R}$ is a convex function if the domain of f (or $\text{dom}(f)$) is a convex set and

$$f(\theta x_1 + (1 - \theta)x_2) \leq \theta f(x_1) + (1 - \theta)f(x_2), \quad \text{for all } x_1, x_2 \in \text{dom}(f) \text{ and } \theta \in [0, 1].$$

The above is commonly referred to as Jensen's inequality.

Function $f(x)$ is convex if and only if $-f(x)$ is concave.

$$\forall x_1 \in S_t, x_2 \in S_t, \exists \theta \in [0, 1]$$

$$\theta x_1 + (1 - \theta)x_2 \in S_t.$$

$$f(\theta x_1 + (1 - \theta)x_2)$$

$$\leq \theta f(x_1) + (1 - \theta)f(x_2) \stackrel{(a)}{\leq} \theta t + (1 - \theta)t = t \stackrel{(b)}{=}$$

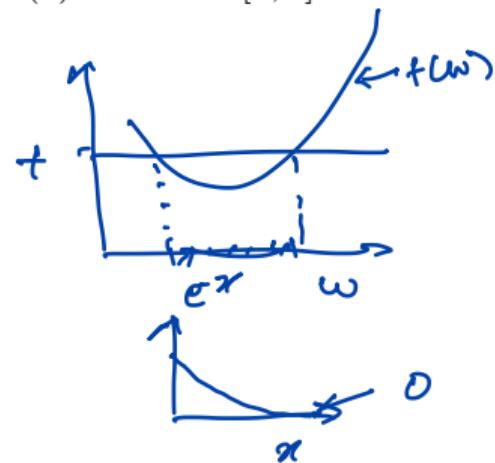
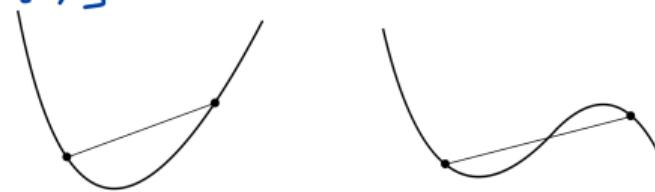


Figure: Difference between (a) convex and (b) nonconvex functions.

Exercise. The sublevel set $S_t := \{x \in \text{dom}(f) | f(x) \leq t\}$ is a convex set for all $t \geq \inf f(x)$. *[minimum]*

Minimizing nonconvex functions is difficult

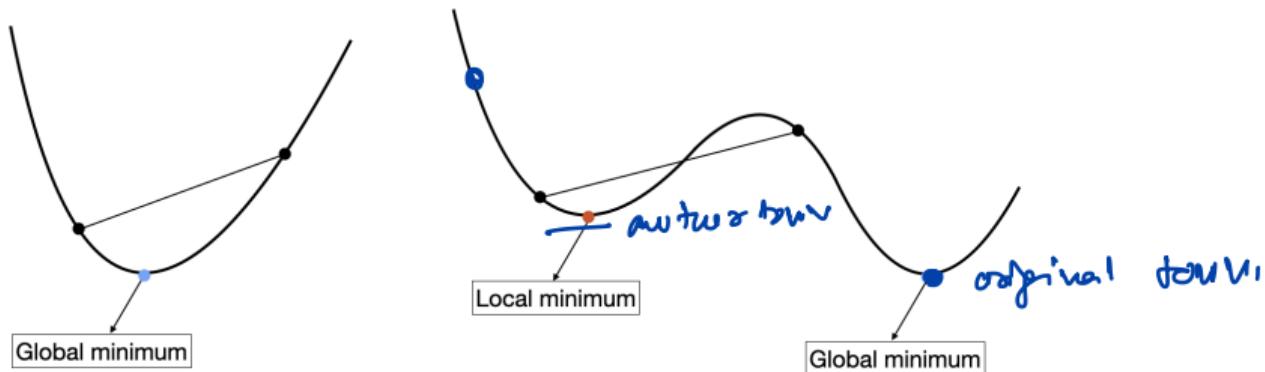


Figure: Inconsistencies between global and local minima under nonconvexity.



Examples of convex functions

Power. For $x \in \mathbb{R}$, $f(x) = x^a$ is convex for all $a \geq 1$ or $a \leq 0$. What about when $\underline{a \in [0, 1]}$?

Norm. $f(x) = \|x\|$, where $\|\cdot\|$ denotes a norm (e.g., ℓ_2 (or Euclidean) norm: $\|x\| := \sqrt{\sum_{i=1}^m x_i^2}$)

Exponent. For $x \in \mathbb{R}$, $f(x) = a^x$ is convex for all $a > 1$.
 $(a \leq 1 \text{ ?})$

- $\|x + y\| \leq \|x\| + \|y\|$
- $\|\lambda x\| = |\lambda| \|x\|$
- $\|x\| = 0 \Leftrightarrow x = 0$

Positive linear combination. If f_1, f_2 are convex functions then for all $\theta_1, \theta_2 \in \mathbb{R}_+$,
 $f(x) = \underline{\theta_1 f_1(x) + \theta_2 f_2(x)}$ is convex.

Composition. Let f_1, f_2 be convex functions. Is $f := f_2 \circ f_1$ convex?

Examples of convex functions

Power. For $x \in \mathbb{R}$, $f(x) = x^a$ is convex for all $a \geq 1$ or $a \leq 0$. What about when $a \in [0, 1]$?

Norm. $f(x) = \|x\|$, where $\|\cdot\|$ denotes a norm (e.g., ℓ_2 (or Euclidean) norm: $\|x\| := \sqrt{\sum_{i=1}^m x_i^2}$).

Exponent. For $x \in \mathbb{R}$, $f(x) = a^x$ is convex for all $a > 1$.

Positive linear combination. If f_1, f_2 are convex functions then for all $\theta_1, \theta_2 \in \mathbb{R}_+$,
 $f(x) = \theta_1 f_1(x) + \theta_2 f_2(x)$ is convex.

Composition. Let f_1, f_2 be convex functions. Is $f := f_2 \circ f_1$ convex?

For $x \in \mathbb{R}$, consider convex functions $f_1(x) = \frac{1}{\sqrt{x}}$ and $f_2(x) = \frac{1}{\sqrt{x}}$. However, $f_2(f_1(x)) = x^{\frac{1}{4}}$ is concave.



Examples of convex functions

Power. For $x \in \mathbb{R}$, $f(x) = x^a$ is convex for all $a \geq 1$ or $a \leq 0$. What about when $a \in [0, 1]$?

Norm. $f(x) = \|x\|$, where $\|\cdot\|$ denotes a norm (e.g., ℓ_2 (or Euclidean) norm: $\|x\| := \sqrt{\sum_{i=1}^m x_i^2}$).

Exponent. For $x \in \mathbb{R}$, $f(x) = a^x$ is convex for all $a > 1$.

Positive linear combination. If f_1, f_2 are convex functions then for all $\theta_1, \theta_2 \in \mathbb{R}_+$,
 $f(x) = \theta_1 f_1(x) + \theta_2 f_2(x)$ is convex.

Affine composition. Let f and g be convex and affine functions, resp. Then, $f \circ g$ is convex.



Examples of convex functions

Power. For $x \in \mathbb{R}$, $f(x) = x^a$ is convex for all $a \geq 1$ or $a \leq 0$. What about when $a \in [0, 1]$?

Norm. $f(x) = \|x\|$, where $\|\cdot\|$ denotes a norm (e.g., ℓ_2 (or Euclidean) norm: $\|x\| := \sqrt{\sum_{i=1}^m x_i^2}$).

Exponent. For $x \in \mathbb{R}$, $f(x) = a^x$ is convex for all $a > 1$.

Positive linear combination. If f_1, f_2 are convex functions then for all $\theta_1, \theta_2 \in \mathbb{R}_+$,
 $f(x) = \theta_1 f_1(x) + \theta_2 f_2(x)$ is convex.

Affine composition. Let f and g be convex and affine functions, resp. Then, $f \circ g$ is convex.

Maximum of convex functions. Let f and g be convex functions. Then, $\max\{f, g\}$ is convex.



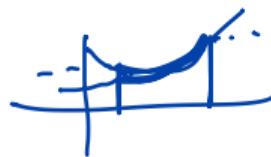
First-order convexity condition

A differentiable function $f : \mathbb{R}^m \rightarrow \mathbb{R}$ is convex if and only if the domain of f (or $\text{dom}(f)$) is a convex set and

$$f(y) \geq f(x) + \nabla f(x)^T(y - x), \quad \forall x, y \in \text{dom}(f)$$



First-order convexity condition



Suppose $\min f(x)$ exists.

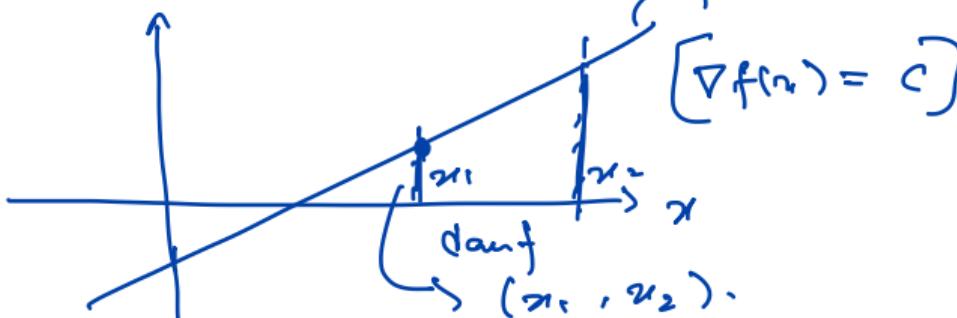
A differentiable function $f : \mathbb{R}^m \rightarrow \mathbb{R}$ is convex if and only if the domain of f (or $\text{dom}(f)$) is a convex set and

$$f(y) \geq f(x) + \nabla f(x)^T(y - x), \quad \forall x, y \in \underline{\text{dom}(f)}$$

If $\nabla f(x) = \mathbf{0}$ then $f(y) \geq f(x)$. That is, if $\nabla f(x) = \mathbf{0}$ then $x \in \underline{\arg \min f(x)}$.

$\arg \inf f(x)$

$\underline{\text{dom}} f$



* If $\text{dom}(f) = \mathbb{R}^n$ then $\nabla f(x) = \mathbf{0}$ is necessary for $x \in \arg \min f(x)$.

First-order convexity condition

A differentiable function $f : \mathbb{R}^m \rightarrow \mathbb{R}$ is convex if and only if the domain of f (or $\text{dom}(f)$) is a convex set and

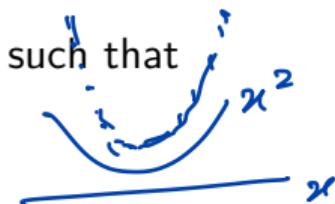
$$f(y) \geq f(x) + \nabla f(x)^T(y - x), \quad \forall x, y \in \text{dom}(f)$$

If $\nabla f(x) = \mathbf{0}$ then $f(y) \geq f(x)$. That is, if $\nabla f(x) = \mathbf{0}$ then $x \in \arg \min f(x)$.

Strict convexity. Function f is strictly convex if $f(y) > f(x) + \nabla f(x)^T(y - x)$ for all $x \neq y \in \text{dom}(f)$.

Strong convexity. Function f is strongly convex if there exists $\mu > 0$ such that

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \underline{\mu \|y - x\|^2}.$$



Second-order convexity condition

If $f : \mathbb{R}^m \rightarrow \mathbb{R}$ is twice differentiable then it is convex if and only if the domain of f (or $\text{dom}(f)$) is convex and its Hessian is positive semidefinite, i.e.,

(Important)

$$\nabla^2 f(x) \succeq 0, \quad \forall x \in \text{dom}(f).$$

$$\geq \begin{pmatrix} \text{smallest} \\ \text{eig.}(M) \end{pmatrix} \|x\|^2.$$

$M \in \mathbb{R}^{m \times m}$ is
positive semidefinite if

$$[x^T M x \geq 0] \quad \forall x \in \mathbb{R}^m.$$

$x \in \mathbb{R}^m$.

$$x^T (Mx) \geq 0$$

- If $M \in \mathbb{R}^{m \times m}$ is not symmetric

we can still have $x^T M x \geq 0 \quad \forall x$

Consider $M = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2$

$$x_1^2 + x_2^2.$$

$$\underline{x^T M x} = [x_1 \ x_2] \begin{bmatrix} x_2 \\ -x_1 \end{bmatrix} = 0. \quad \checkmark$$

(You can also consider $M = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$)

$$= x^T M^T x = (Mx)^T x$$

Second-order convexity condition

If $f : \mathbb{R}^m \rightarrow \mathbb{R}$ is twice differentiable then it is convex *if and only if* the domain of f ($\text{dom}(f)$) is convex and its Hessian is positive semidefinite, i.e.,

$$\nabla^2 f(x) \succeq 0, \quad \forall x \in \text{dom}(f).$$

Strict convexity. Function f is strictly convex if $\nabla^2 f(x) \succ 0, \quad \forall x \in \text{dom}(f)$.

Second-order convexity condition

If $f : \mathbb{R}^m \rightarrow \mathbb{R}$ is twice differentiable then it is convex *if and only if* the domain of f ($\text{dom}(f)$) is convex and its Hessian is positive semidefinite, i.e.,

$$\nabla^2 f(x) \succeq 0, \quad \forall x \in \text{dom}(f).$$

Strict convexity. Function f is strictly convex if $\nabla^2 f(x) \succ 0, \quad \forall x \in \text{dom}(f)$.

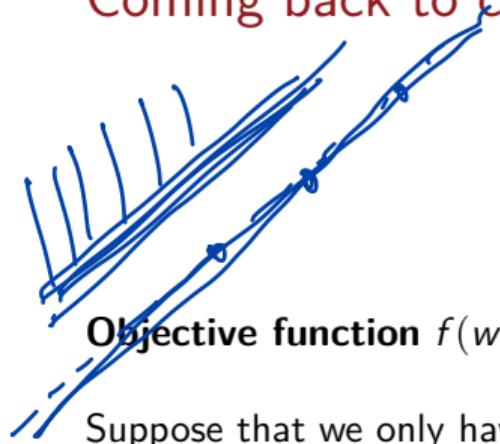
Example (quadratic function). $f(x) := x^\top P x + Qx$ where $P \in \mathbf{S}_{++}^m$ and $Q \in \mathbb{R}^{m \times m}$.
 \mathbf{S}_{++}^m denotes the set of (symmetric) positive definite matrices in $\mathbb{R}^{m \times m}$.

Coming back to convex optimization problem

$$\begin{aligned} & \underset{w \in \mathbb{R}^d}{\text{Minimize}} \quad f(w) \\ & \text{Subject to} \quad f_i(w) \leq 0 \quad i = 1, \dots, p \\ & \quad a_i^\top w + b_i = 0 \quad i = 1, \dots, q \end{aligned} \tag{CVX}$$

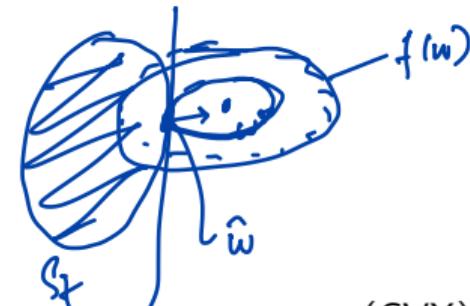
Objective function $f(w)$ and inequality constraints $f_1(w), \dots, f_p(w)$ are convex functions.

Coming back to convex optimization problem



$$\underset{w \in \mathbb{R}^d}{\text{Minimize}} \quad f(w)$$

$$\begin{aligned} \text{Subject to} \quad & f_i(w) \leq 0 \quad i = 1, \dots, p \\ & a_i^T w + b_i = 0 \quad i = 1, \dots, q \end{aligned}$$



Objective function $f(w)$ and inequality constraints $f_1(w), \dots, f_p(w)$ are convex functions.

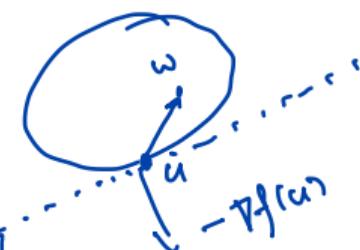
Suppose that we only have inequality constraints. Define the feasible set

$$S_f = \{w \mid f_i(w) \leq 0, i = 1, \dots, p\}.$$

Then, solving (CVX) reduces to finding:

$u \in S_f$ such that $H(w) := \nabla f(u)^T(w - u)$ is a **supporting hyperplane** of S_f (with $H(w) \geq 0, \forall w \in S_f$).

u can be non-unique.



$$\text{epi } f = \{(x, t) \mid f(x) \leq t\}$$

$x \in \mathbb{R}^d, t \in \mathbb{R}$

Epigraph formulation

Original optimization problem:

\hat{w} → does not solve epf

$$\begin{aligned} & \underset{w \in \mathbb{R}^d}{\text{Minimize}} \quad f(w) \\ & \text{Subject to} \quad f_i(w) \leq 0 \quad i = 1, \dots, p \\ & \quad g_i(w) = 0 \quad i = 1, \dots, q \end{aligned} \tag{OPT}$$

The above optimization problem is **reducible** to the following:



$$\begin{aligned} & \underset{w \in \mathbb{R}^d, t \in \mathbb{R}}{\text{Minimize}} \quad t \\ & \text{Subject to} \quad \underline{f(w) - t \leq 0} \\ & \quad \underline{f_i(w) \leq 0} \quad i = 1, \dots, p \\ & \quad g_i(w) = 0 \quad i = 1, \dots, q \end{aligned}$$

$\hat{t} = f(\hat{w})$ ← convex

$f(w^*) \leq t^*$

$$\begin{aligned} & f(\hat{w}) < f(w^*) \\ & f(\hat{w}) \leq t^* \end{aligned}$$

(EPI)

The latter is called the **epigraph form** of the latter optimization problem.

(prove by contradiction)

- Specifically, a solution (w^*, t^*) of (EPI) is a solution and optimal value of the original problem (OPT).



Standard convex optimization problems

Linear programming (LP)

~~algo~~ → standard →

$$\left[\begin{array}{ll} \text{Minimize}_{w \in \mathbb{R}^d} & f(w) := a^T w + b \\ \text{Subject to} & Gw \leq c \quad ; \quad G \in \mathbb{R}^{p \times d}, \quad c \in \mathbb{R}^p \\ & Hw = d \quad ; \quad H \in \mathbb{R}^{q \times d}, \quad d \in \mathbb{R}^q \end{array} \right] \xrightarrow{\text{affine}} a^T w - c \leq 0$$

Convex optimization with affine objective and constraint functions.



The feasibility set is a polyhedron.

Exercise: Minimizing piecewise-linear function $f(w) := \max_{i=1,\dots,p} (a_i^T w + b_i)$ is reducible to linear programming.

minimise $f(w)$.

Epigraph form: $\left[\begin{array}{ll} \text{mr. } t & \text{s.t. } f(w) \leq t \Rightarrow a_i^T w + b_i \leq t \quad \forall i \end{array} \right]$

Quadratic programming (QP)

$$\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geq 0\}$$

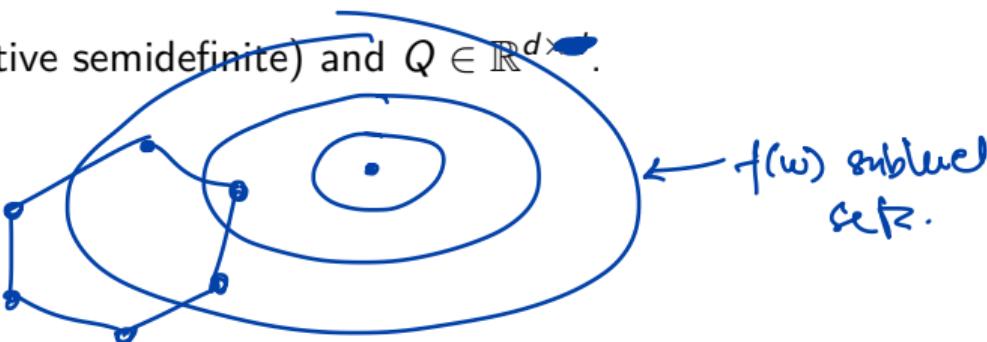
$$\mathbb{R}_{++} = \{x \in \mathbb{R} \mid x > 0\}.$$

Convex optimization with quadratic objective and affine constraint functions.

$$\underset{w \in \mathbb{R}^d}{\text{Minimize}} \quad f(w) := \underline{w^T P w + Q^T w}.$$

$$\begin{array}{ll} \text{Subject to} & Gw \leq c \quad ; \quad G \in \mathbb{R}^{p \times d}, \quad c \in \mathbb{R}^p \\ & Hw = d \quad ; \quad H \in \mathbb{R}^{q \times d}, \quad d \in \mathbb{R}^q \end{array}$$

where $P \in \mathbf{S}_+^d$ (i.e., symmetric positive semidefinite) and $Q \in \mathbb{R}^{d \times 1}$.



Quadratic programming (QP)

Convex optimization with quadratic objective and affine constraint functions.

$$\underset{w \in \mathbb{R}^d}{\text{Minimize}} \quad f(w) := w^T P w + Q^T w$$

$$\begin{array}{ll} \text{Subject to} & Gw \preceq c \\ & Hw = d \end{array} ; \quad \begin{array}{l} G \in \mathbb{R}^{p \times d}, \quad c \in \mathbb{R}^p \\ H \in \mathbb{R}^{q \times d}, \quad d \in \mathbb{R}^q \end{array}$$



where $P \in \mathbf{S}_+^d$ (i.e., symmetric positive semidefinite) and $Q \in \mathbb{R}^{d \times d}$.

Example. (Euclidean distance between polyhedra)

Let $P_1 = (w \mid A_1 w + b_1 \preceq 0)$ and $P_2 = (w \mid A_2 w + b_2 \preceq 0)$.

Determining the distance between P_1 and P_2 is a quadratic optimization problem.

$$\text{dist}(P_1, P_2) = \min_{x \in P_1, y \in P_2} \|x - y\|$$

Returning to classification: linear classifier

Consider a set of data points $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ such that $x_i \in \mathbb{R}^m$ and $y_i \in \{-1, +1\}$ for all i . Suppose the points are linearly separable.

Define $S^+ = \{x \mid (x, y) \in S, y = +1\}$ and $S^- = \{x \mid (x, y) \in S, y = -1\}$.

Training classifier. Determining $(w, b) \in \mathbb{R}^m \times \mathbb{R}$ such that $\text{sign}(w^T x + b) > 0$ for all $x \in S^+$ and $\text{sign}(w^T x + b) < 0$ for all $x \in S^-$ reduces to the following LP:

$$\begin{array}{ll} \text{Minimize}_{w \in \mathbb{R}^m, b \in \mathbb{R}} & 1 \rightarrow \text{optimizing weight} \\ \text{Subject to} & y_i(w^T x_i + b) > 0 \quad ; i = 1, \dots, n \end{array} \quad \left. \begin{array}{l} \text{s.t.} \\ \text{sign}(w^T x_i + b) \\ = y_i \quad \text{if} \end{array} \right\}$$

We will revisit the above when talking about support vector machines (SVMs). \rightarrow maximizing separating margin.



Duality

Lagrangian function

Primal optimization problem:

$$\underset{w \in \mathbb{R}^d}{\text{Minimize}} \quad f(w)$$

$$\begin{aligned} \text{Subject to} \quad & f_i(w) \leq 0 \quad i = 1, \dots, p \\ & g_i(w) = 0 \quad i = 1, \dots, q \end{aligned}$$

(Primal)

We define the **Lagrangian** $\mathcal{L} : \mathbb{R}^d \times \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}$ of the above problem to be

$$\mathcal{L}(w, \lambda, \nu) = f(w) + \sum_{i=1}^p \underline{\lambda_i} f_i(w) + \sum_{i=1}^q \underline{\nu_i} g_i(w).$$

$\lambda = (\lambda_1, \dots, \lambda_p)$ and $\nu = (\nu_1, \dots, \nu_q)$ are called **dual variables** or **Lagrange multipliers**.

Lagrange dual function

The **Lagrange dual function** $\phi : \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}$ is defined to be

$$\checkmark \quad \phi(\lambda, \nu) := \min_{w \in \mathbb{R}^d} \mathcal{L}(w, \lambda, \nu) = \underbrace{\min_{w \in \mathbb{R}^d}}_{\text{---}} \left(f(w) + \sum_{i=1}^p \lambda_i f_i(w) + \sum_{i=1}^q \nu_i g_i(w) \right).$$

$$\max_{i=1,\dots,n} (\alpha_i^T w + b_i) \rightarrow \text{convex}$$


Lagrange dual function

The **Lagrange dual function** $\phi : \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}$ is defined to be

$$\phi(\lambda, \nu) := \min_{w \in \mathbb{R}^d} \mathcal{L}(w, \lambda, \nu) = \underbrace{\min_{w \in \mathbb{R}^d}}_{-\max(-\dots)} \left(f(w) + \sum_{i=1}^p \lambda_i f_i(w) + \sum_{i=1}^q \nu_i g_i(w) \right).$$

Exercise. Show that $\phi(\lambda, \nu)$ is a concave function. (regardless of f, g).

Lagrange dual function

The **Lagrange dual function** $\phi : \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}$ is defined to be

$$\begin{aligned} & \text{s.t. } f_i(w) \leq 0 \\ & g_j(w) = 0 \end{aligned}$$

$$\underline{\phi(\lambda, \nu)} := \min_{w \in \mathbb{R}^d} \mathcal{L}(w, \lambda, \nu) = \min_{w \in \mathbb{R}^d} \left(f(w) + \sum_{i=1}^p \lambda_i f_i(w) + \sum_{i=1}^q \nu_i g_i(w) \right).$$

Exercise. Show that $\phi(\lambda, \nu)$ is a concave function.

Let p^* be the optimal value of the primal problem (i.e., $p^* = \min f(w)$ subject to the constraints). For all $\lambda \succeq 0$ and ν ,

$$\underline{p^* \geq \phi(\lambda, \nu).} - ?$$

Lagrange dual optimization problem

The **Lagrange dual problem** is defined to be

$$\begin{array}{ll} \text{Maximize} & \phi \\ \text{Subject to} & \underline{\lambda \geq 0} \end{array} \quad (\text{Dual})$$

$$g \longleftrightarrow \phi(\lambda, \nu)$$

Lagrange dual optimization problem

The **Lagrange dual problem** is defined to be

$$\begin{aligned} & \text{Maximize} && g(\lambda, \nu) \\ & \text{Subject to} && \lambda \succeq 0 \end{aligned} \tag{Dual}$$

Suppose that d^* is the optimal value of (Dual) (i.e., $\underline{\max g(\lambda, \nu)}$ s.t. $\lambda \succeq 0$.) Then,

$$p^* \geq d^*.$$

Lagrange dual optimization problem

The **Lagrange dual problem** is defined to be

$$\begin{aligned} & \text{Maximize} && g(\lambda, \nu) \\ & \text{Subject to} && \lambda \succeq 0 \end{aligned} \tag{Dual}$$

Suppose that d^* is the optimal value of (Dual) (i.e., $d^* = \max g(\lambda, \nu)$ s.t. $\lambda \succeq 0$.) Then,

$$p^* \geq d^*.$$

This property is called **weak duality**. It holds true even if the primal problem is nonconvex.

Lagrange dual of LP

Primal problem:

$$\underset{w \in \mathbb{R}^d}{\text{Minimize}} \quad c^T w$$

$$\begin{aligned} \text{Subject to} \quad & Aw = b \quad ; \quad A \in \mathbb{R}^{p \times d}, \quad b \in \mathbb{R}^p \\ & w \succeq 0 \end{aligned} \quad (\text{Primal LP})$$

Lagrange dual of LP

Primal problem:

$$\text{Minimize}_{w \in \mathbb{R}^d} c^T w$$

$$\text{Subject to } Aw = b \quad ; \quad A \in \mathbb{R}^{p \times d}, \quad b \in \mathbb{R}^p$$

$$w \succeq 0$$

$$\begin{aligned} L(w, \lambda, v) &= c^T w + v^T (Aw - b) + (-\lambda^T w) \\ \Phi(\lambda, v) &= \min_w \left\{ (c - \lambda)^T + v^T A \right\} w - v^T b \end{aligned}$$

Exercise. The dual problem is

$$\text{Maximize}_{v} -b^T v$$

$$\text{Subject to } A^T v + c \succeq 0$$

]

(Dual LP)

$$\underline{v^T A + c^T = \lambda^T}$$

(Primal LP)

Strong duality

We have **strong duality** when

$$p^* \geq d^*$$

no duality gap

$$p^* = d^*.$$

Strong duality

We have **strong duality** when

$$p^* = d^*.$$

Strong duality usually (but not always) holds when the **primal problem is convex**.

Strong duality

We have **strong duality** when

$$p^* = d^*.$$

Strong duality usually (but not always) holds when the **primal problem is convex**.

Conditions, beyond convexity, that ensure strong duality are called **constraint qualifications**.

Strong duality

We have **strong duality** when

$$p^* = d^*.$$

Strong duality usually (but not always) holds when the **primal problem is convex**.

Conditions, beyond convexity, that ensure strong duality are called **constraint qualifications**.

↓ a camera

Slater's condition: There exists $\underline{w} \in \mathbb{R}^d$ such that

$$\begin{aligned} f_i(\underline{w}) &< 0, \quad i = 1, \dots, p \\ a_i^T \underline{w} + b_i &= 0, \quad i = 1, \dots, q \end{aligned}$$

+ Convex

Optimality conditions

Let's consider a possibly nonconvex primal problem:

$$\underset{w \in \mathbb{R}^d}{\text{Minimize}} \quad f(w)$$

$$\begin{aligned} \text{Subject to} \quad & f_i(w) \leq 0 \quad i = 1, \dots, p \\ & g_i(w) = 0 \quad i = 1, \dots, q \end{aligned}$$

Optimality conditions

Let's consider a possibly nonconvex primal problem:

$$\begin{aligned} & \underset{w \in \mathbb{R}^d}{\text{Minimize}} \quad f(w) \\ & \text{Subject to} \quad f_i(w) \leq 0 \quad i = 1, \dots, p \\ & \qquad \qquad \qquad g_i(w) = 0 \quad i = 1, \dots, q \end{aligned}$$

Suppose that $p^* = d^*$ (strong duality holds true).

Optimality conditions

Let's consider a possibly nonconvex primal problem:

$$\underset{w \in \mathbb{R}^d}{\text{Minimize}} \quad f(w)$$

$$\begin{aligned} \text{Subject to} \quad & f_i(w) \leq 0 \quad i = 1, \dots, p \\ & g_i(w) = 0 \quad i = 1, \dots, q \end{aligned}$$

Suppose that $\underline{p^*} = d^*$ (strong duality holds true).

$$\begin{aligned} p^* &\geq \varphi(\lambda, \nu) \\ p^* &= d^* \end{aligned}$$

$$\lambda^* \geq 0$$

$$f_i(w^*) \leq 0$$

$$\sum_{i=1}^p \lambda_i^* f_i(w^*) = 0$$

Complementary slackness. For any primal solution $\underline{w^*}$ and dual solution (λ^*, ν^*) ,

$$\boxed{\lambda_i^* f_i(w^*) = 0, \quad i = 1, \dots, p.}$$

Optimality conditions

Let's consider a possibly nonconvex primal problem:

$$\begin{aligned} & \underset{w \in \mathbb{R}^d}{\text{Minimize}} \quad f(w) \\ & \text{Subject to} \quad f_i(w) \leq 0 \quad i = 1, \dots, p \\ & \qquad \qquad g_i(w) = 0 \quad i = 1, \dots, q \end{aligned}$$

Suppose that $p^* = d^*$ (strong duality holds true).

Complementary slackness. For any primal solution w^* and dual solution (λ^*, ν^*) ,

$$\lambda_i^* f_i(w^*) = 0, \quad i = 1, \dots, p.$$

Why is it called “complementary slackness”?

Optimality conditions (cont'd)

Suppose the objective and constraint functions are **differentiable** and **strong duality** holds.

Optimality conditions (cont'd)

Suppose the objective and constraint functions are **differentiable** and **strong duality** holds.

Let w^* and (λ^*, ν^*) be the primal and dual optimal solutions, respectively. Then,

Optimality conditions (cont'd)

Suppose the objective and constraint functions are **differentiable** and **strong duality** holds.

Let w^* and (λ^*, ν^*) be the primal and dual optimal solutions, respectively. Then,

$$\nabla_w \mathcal{L}(w^*, \lambda^*, \nu^*) = \underbrace{\nabla_w f(w^*)}_{\text{Primal}} + \sum_{i=1}^p \lambda_i \nabla_w f_i(w^*) + \sum_{i=1}^q \nu_i \nabla_w g_i(w^*) = \mathbf{0}. \quad \checkmark$$

Optimality conditions (cont'd)

Suppose the objective and constraint functions are **differentiable** and **strong duality** holds.

Let w^* and (λ^*, ν^*) be the primal and dual optimal solutions, respectively. Then,

$$\nabla_w \mathcal{L}(w^*, \lambda^*, \nu^*) = \nabla_w f(w^*) + \sum_{i=1}^p \lambda_i \nabla_w f_i(w^*) + \sum_{i=1}^q \nu_i \nabla_w g_i(w^*) = \mathbf{0}.$$

Recall that $f_i(w^*) \leq 0$ for $i = 1, \dots, p$ and $g_i(w^*) = 0$ for $i = 1, \dots, q$. Moreover, $\lambda^* \succeq 0$.

Optimality conditions (cont'd)

Suppose the objective and constraint functions are **differentiable** and **strong duality** holds.

Let w^* and (λ^*, ν^*) be the primal and dual optimal solutions, respectively. Then,

$$\nabla_w \mathcal{L}(w^*, \lambda^*, \nu^*) = \nabla_w f(w^*) + \sum_{i=1}^p \lambda_i \nabla_w f_i(w^*) + \sum_{i=1}^q \nu_i \nabla_w g_i(w^*) = \mathbf{0}. \quad \bullet$$

Recall that $f_i(w^*) \leq 0$ for $i = 1, \dots, p$ and $g_i(w^*) = 0$ for $i = 1, \dots, q$. Moreover, $\lambda^* \succeq 0$.

The above, together with complementary slackness, form the famous **KKT conditions**.
(KKT stands for Karush-Kuhn-Tucker.)

KKT optimality conditions

$$\begin{aligned}
 & \underset{w \in \mathbb{R}^d}{\text{Minimize}} \quad f(w) \\
 & \text{Subject to} \quad f_i(w) \leq 0 \quad i = 1, \dots, p \\
 & \qquad \qquad g_i(w) = 0 \quad i = 1, \dots, q
 \end{aligned} \tag{Primal}$$

KKT necessary conditions for optimality (under differentiability and strong duality):

$$\left[\begin{array}{l}
 \begin{matrix}
 f_i(w^*) & \leq 0, & i = 1, \dots, p \\
 g_i(w^*) & = 0, & i = 1, \dots, q \\
 \lambda^* & \succeq \mathbf{0} \\
 \lambda_i^* f_i(w^*) & = 0, & i = 1, \dots, p
 \end{matrix} \\
 \nabla_w f(w^*) + \sum_{i=1}^p \lambda_i^* \nabla_w f_i(w^*) + \sum_{i=1}^q \nu_i^* \nabla_w g_i(w^*) = \mathbf{0}
 \end{array} \right] \tag{KKT}$$

complementarity
slackness

differentiability.

If the **primal** is **convex** then **KKT conditions** are **sufficient** for optimality.

Further readings

The lecture notes are based on Chapters 1 - 5 of “Convex Optimization” by S. Boyd and L. Vandenberghe

Further readings:

- **Sections 2.2 - 2.4.** Examples and properties of convex set.
- **Sections 3.2 - 3.4.** Operations that preserve convexity and quasi-convex functions.
- **Sections 4.2 - 4.5.** Standard convex optimization reductions and problems.
- **Sections 5.3.** Geometric interpretation of weak and strong duality.

