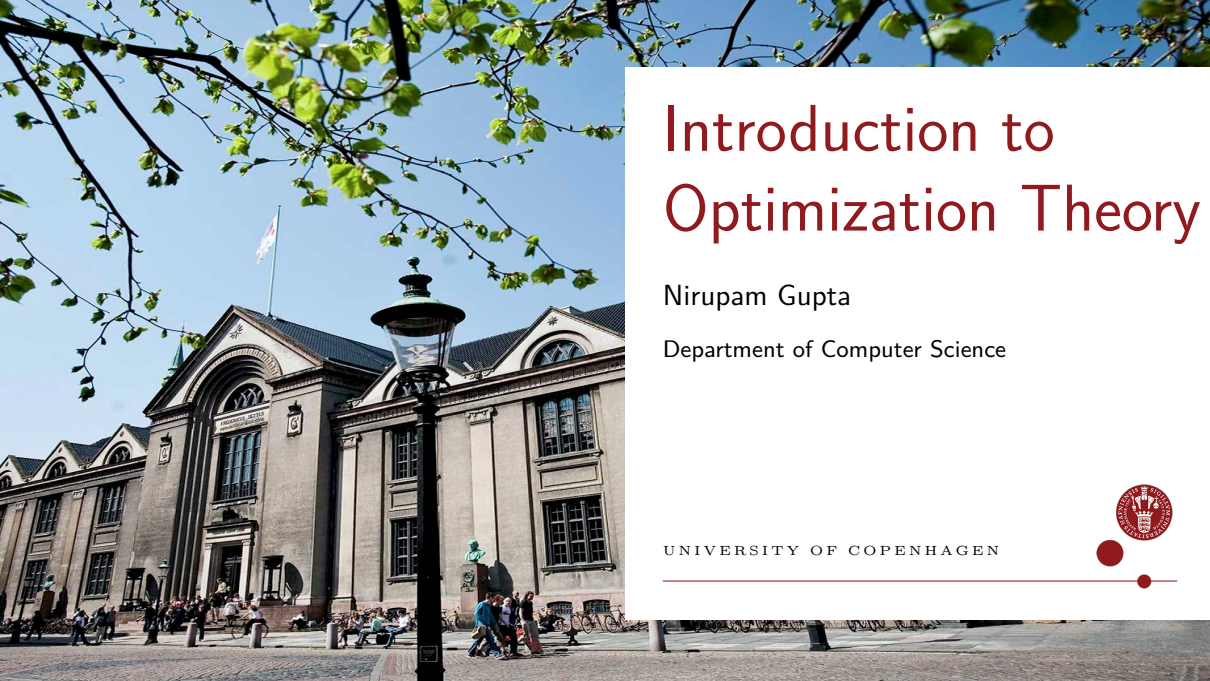


Introduction to Optimization Theory

Nirupam Gupta

Department of Computer Science

UNIVERSITY OF COPENHAGEN

What is optimization?

Mathematical optimization

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

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

Mathematical optimization

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

- $w \in \mathbb{R}^d$ represents scalar **variables** $w_1, \ldots, w_d \in \mathbb{R}$, to be computed.
- $f : \mathbb{R}^d \to \mathbb{R}$ is the **objective function**, to be minimized.
- $\{f_1, \ldots, f_p\}$ with $f_i : \mathbb{R}^d \to \mathbb{R}$ are the **inequality constraints**.
- $\{g_1, \ldots, g_q\}$ with $g_i : \mathbb{R}^d \to \mathbb{R}$ 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) $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^m \times \{-1, 1\}$.
- Linear **hypothesis class** $\mathcal{H} = (h : x \mapsto \mathsf{sign}\,(w^\intercal x + w_o) \mid (w, w_o) \in \mathbb{R}^m \times \mathbb{R})$.
- For each sample (x_i, y_i), **define loss** by $f_i(w, w_o) := \mathbb{1}[h(x_i) \neq y_i]$ where $\mathbb{1}[\cdot]$ is the indicator function

Solve the following **optimization problem**:

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

The inequality constraint ensures sparsity in the weights when $k < m$.

Optimization in machine learning (cont'd)

Training **artificial neural networks**

- n **samples** (input-output pairs) $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^m \times \{-1, +1\}$.
- Hypothesis $h(x) : \mathbb{R}^m \to \mathbb{R}^d$ 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**:

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

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

How to solve an optimization problem?

Solving an optimization problem

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

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

Solving an optimization problem

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

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

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

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** $f_1(w), \ldots, f_p(w)$ are **convex** functions -

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

Solving an optimization problem

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

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** $f_1(w), \ldots, f_p(w)$ 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** $g_1(w), \ldots, g_q(w)$ are **affine** functions:

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

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

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

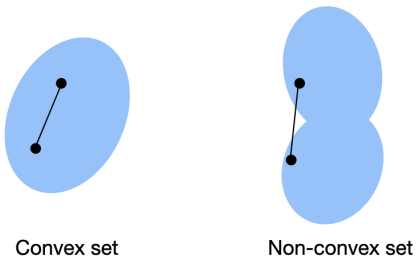


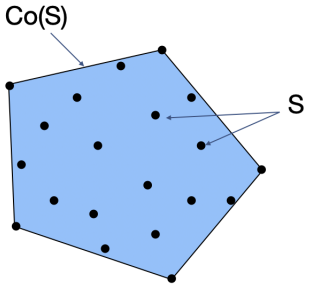
Figure: Difference between convex and nonconvex sets.

Convex Hull

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

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

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



Separating hyperplane theorem

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. $w^\intercal x + b \geq 0,\ \forall x \in C$ and $w^\intercal x + b \leq 0,\ \forall x \in D$.

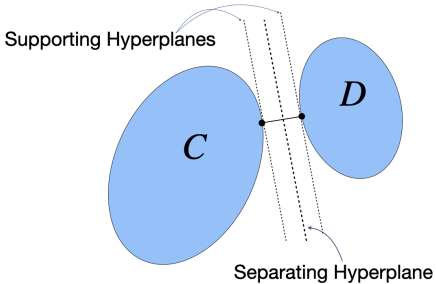


Figure: Separating and supporting hyperplanes for two disjoint *closed* convex sets.

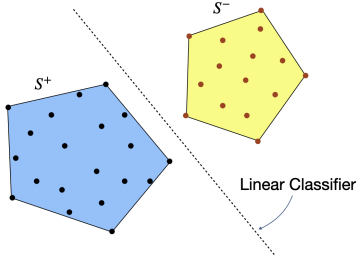
Refer to Section of 2.5 of Convex Optimization, book by Boyd and Vandenberghe, for a proof.

Linear separability

Consider a set of data points $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ 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 $\mathrm{Co}(S^+) \cap \mathrm{Co}(S^-) = \emptyset$. Show that $\exists$ a hyperplane $f(x) := w^\intercal x + b$ 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., $D \coloneqq \left(\frac{x}{t} \,\middle|\, (x, t) \in C\right)$ 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., $C \coloneqq \left((x, t) \,\middle|\, \frac{x}{t} \in D\right)$ is a convex set.

Convex function

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

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

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

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

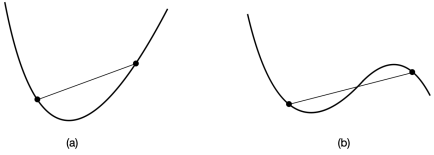


Figure: Difference between (a) convex and (b) nonconvex functions.

Exercise. The *sublevel set* $S_t := (x \in \mathsf{dom}(f) \mid f(x) \leq t)$ is a convex set for all $t > \inf f(x)$.

Minimizing nonconvex functions is difficult

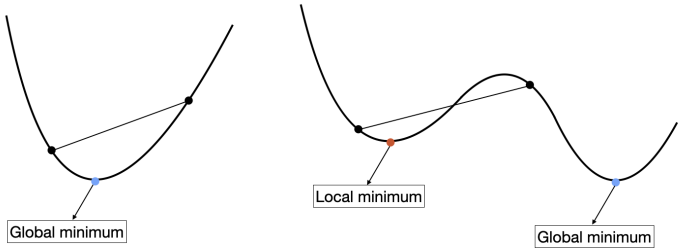


Figure: Inconsistencies between global and local minima under nonconvexity.

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 $a \in [0, 1]$?

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 $\theta_1, \theta_2 \in \mathbb{R}_+$, $f(x) = \theta_1 f_1(x) + \theta_2 f_1(x)$ 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 $\theta_1, \theta_2 \in \mathbb{R}_+$, $f(x) = \theta_1 f_1(x) + \theta_2 f_1(x)$ 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 $f : \mathbb{R}^m \to \mathbb{R}$ is convex *if and only if* the domain of f (or $\mathbf{dom}(f)$) is a convex set and

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

First-order convexity condition

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

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

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

First-order convexity condition

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

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

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

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

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

$$f(y) \geq f(x) + \nabla f(x)^\intercal (y - x) + \mu \left\| y - x \right\|^2 .$$

Second-order convexity condition

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

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

Second-order convexity condition

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

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

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

Second-order convexity condition

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

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

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

Example (quadratic function). $f(x) := x^\intercal P x + Qx$ 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

$$\begin{array}{lll} \underset{w \in \mathbb{R}^d}{\text{Minimize}} & f(w) & \\ \text{Subject to} & f_i(w) \leq 0 & i = 1, \ldots, p \\ & a_i^\mathsf{T} w + b_i = 0 & i = 1, \ldots, q \end{array} \tag{CVX}$$

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

Coming back to convex optimization problem

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

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

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

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

Then, solving (CVX) reduces to finding:

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

Epigraph formulation

Original optimization problem:

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

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

$$\begin{array}{lll} \underset{w \in \mathbb{R}^d \ , \ t \in \mathbb{R}}{\text{Minimize}} & t & \\ \text{Subject to} & f(w) - t \leq 0 & \\ & f_i(w) \leq 0 & i = 1, \ldots, p \\ & g_i(w) = 0 & i = 1, \ldots, q \end{array} \tag{EPI}$$

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

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)

$$\begin{array}{lll} \underset{w \in \mathbb{R}^d}{\text{Minimize}} & f(w) \coloneqq a^\mathsf{T} w + b & \\ \text{Subject to} & Gw \preceq c & ;\ G \in \mathbb{R}^{p \times d},\ c \in \mathbb{R}^p \\ & Hw = d & ;\ H \in \mathbb{R}^{q \times d},\ d \in \mathbb{R}^q \end{array}$$

Convex optimization with affine objective and constraint functions.

The feasibility set is a polyhedron.

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

Quadratic programming (QP)

Convex optimization with quadratic objective and affine constraint functions.

$$\begin{array}{lll} \underset{w \in \mathbb{R}^d}{\text{Minimize}} & f(w) \coloneqq w^\intercal P w + Qx & \\ \text{Subject to} & Gw \preceq c & ;\ G \in \mathbb{R}^{p \times d},\ c \in \mathbb{R}^p \\ & Hw = d & ;\ H \in \mathbb{R}^{q \times d},\ d \in \mathbb{R}^q \end{array}$$

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

Quadratic programming (QP)

Convex optimization with quadratic objective and affine constraint functions.

$$\begin{array}{lll} \underset{w \in \mathbb{R}^d}{\text{Minimize}} & f(w) \coloneqq w^\mathsf{T} P w + Qx & \\ \text{Subject to} & Gw \preceq c & ;\ G \in \mathbb{R}^{p \times d},\ c \in \mathbb{R}^p \\ & Hw = d & ;\ H \in \mathbb{R}^{q \times d},\ d \in \mathbb{R}^q \end{array}$$

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.

Returning to classification: linear classifier

Consider a set of data points $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ 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 $\mathsf{sign}(w^\mathsf{T} x + b) > 0$ for all $x \in S^+$ and $\mathsf{sign}(w^\mathsf{T} x + b) < 0$ for all $x \in S^-$ reduces to the following LP:

$$\begin{array}{ll} \underset{w \in \mathbb{R}^m, \ b \in \mathbb{R}}{\text{Minimize}} & 1 \\ \text{Subject to} & y_i(w^\mathsf{T} x_i + b) > 0 \quad ; i = 1, \ldots, n \end{array}$$

We will revisit the above when talking about support vector machines (SVMs).

Duality

Lagrangian function

Primal optimization problem:

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

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

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

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

Lagrange dual function

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

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

Lagrange dual function

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

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

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

Lagrange dual function

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

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

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,

$$p^* \geq \phi(\lambda, \nu).$$

Lagrange dual optimization problem

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

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

Lagrange dual optimization problem

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

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

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

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

Lagrange dual optimization problem

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

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

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:

$$\begin{array}{ll} \underset{w \in \mathbb{R}^d}{\text{Minimize}} & c^\top w \\ \text{Subject to} & Aw = b \quad ; \; A \in \mathbb{R}^{p \times d} \, , \; b \in \mathbb{R}^p \\ & w \succeq 0 \end{array} \tag{Primal LP}$$

Lagrange dual of LP

Primal problem:

$$\begin{array}{ll} \underset{w \in \mathbb{R}^d}{\text{Minimize}} & c^\mathsf{T} w \\ \text{Subject to} & Aw = b \quad ; \; A \in \mathbb{R}^{p \times d} \, , \; b \in \mathbb{R}^p \\ & w \succeq 0 \end{array} \tag{Primal LP}$$

Exercise. The dual problem is

$$\begin{array}{ll} \text{Maximize} & -b^\mathsf{T} \nu \\ \text{Subject to} & A^\mathsf{T} \nu + c \succeq 0 \end{array} \tag{Dual LP}$$

Strong duality

We have **strong duality** when

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

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**.

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

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

Optimality conditions

Let's consider a possibly nonconvex primal problem:

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

Optimality conditions

Let's consider a possibly nonconvex primal problem:

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

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

Optimality conditions

Let's consider a possibly nonconvex primal problem:

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

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

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

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

Optimality conditions

Let's consider a possibly nonconvex primal problem:

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

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

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

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

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,

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

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 $i = 1, \ldots, p$ and $g_i(w^*) = 0$ for $i = 1, \ldots, q$. 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,

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

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

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

KKT optimality conditions

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

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

$$\begin{aligned} f_i(w^*) &\leq 0\,, \quad i = 1, \ldots, p \\ g_i(w^*) &= 0\,, \quad i = 1, \ldots, q \\ \lambda^* &\succeq \mathbf{0} \\ \lambda_i^* f_i(w^*) &= 0\,, \quad i = 1, \ldots, p \\ \nabla_w f(w^*) + \textstyle\sum_{i=1}^p \lambda_i \nabla_w f_i(w^*) + \sum_{i=1}^q \nu_i \nabla_w g_i(w^*) &= \mathbf{0} \end{aligned} \tag{KKT}$$

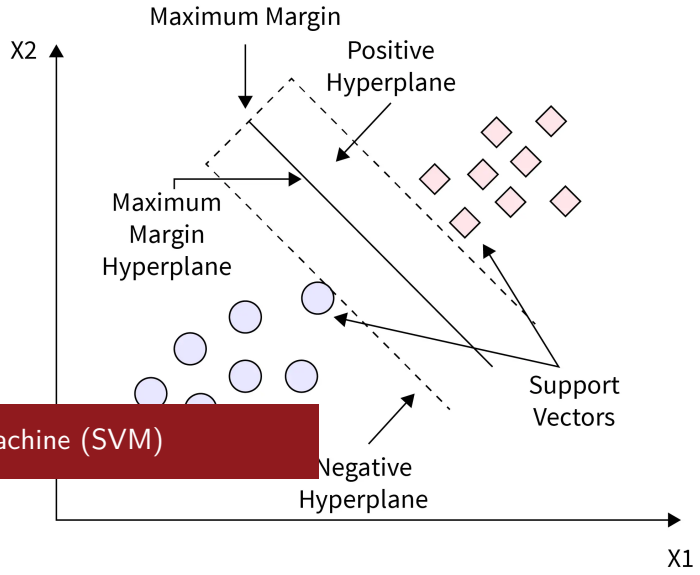
If the **primal** is **convex** then **KKT conditions are sufficient** for optimality.

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.



Support Vector Machine (SVM)