Mathematical Preliminaries

I-Hsiang Wang

Department of Electrical Engineering
National Taiwan University

ihwang@ntu.edu.tw

September 5, 2023

Let's review some of the mathematical tools that will be used in this course:

1. Probability and random processes
   - Stochastic modeling is essential in information theory.
   - Random variables, random processes, convergences, and limit theorems.

2. Optimization
   - Pursuit of optimality is a central theme of information theory.
   - Many information measures to be introduced are convex or concave.
   - Convex optimization, Lagrange duality, and Karush-Kuhn-Tucker (KKT) conditions.

Notations

Deterministic: lower case. Random: upper case.

Category	Font	Notation
Scalar/Symbol	normal italic serif	$x, y, z; X, Y, Z$
Vector	boldface italic serif	$\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}; \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$
Matrix	boldface upright serif	$\mathbf{a}, \mathbf{g}, \mathbf{w}; \mathbf{A}, \mathbf{G}, \mathbf{W}$
Set/Event	caligraphic	$\mathcal{A}, \mathcal{B}, \mathcal{E}$
Collection of sets/events	script	$\mathscr{C}, \mathscr{F}$
Probability laws	upper case sans serif	$\mathsf{P}, \mathsf{Q}, \mathsf{P}_{X,Y}, \mathsf{P}_{X\|Y}$
Probability mass/density functions	lower case sans serif	$\mathsf{p}_{X,Y}, \mathsf{f}_{X\|Y}$
Expectation, variance	sans serif	$\mathsf{E}, \mathsf{Var}, \mathsf{Cov}$
Specific distributions	normal upright serif	$\mathrm{Ber}, \mathrm{Unif}, \mathrm{N}$
Special functions/quantities	normal upright serif	$\mathrm{H}(X), \mathrm{R}, \mathrm{C}$

1 Probability and Random Processes

2 Optimization

Probability and random variable

A probability space $(\Omega, \mathscr{F}, P)$ comprises three parts:

- Sample space Ω: collection of all sample points.
- Sigma field $\mathscr{F}$: collection of all events (subset of sample points).
- Probability measure P: assigning probability of an event.

A random variable

$$X : \Omega \to \mathcal{X}$$

maps[1] a sample point to an outcome in an *alphabet* $\mathcal{X}$. Correspondingly, there is a sigma field and probability measure induced over $\mathcal{X}$.

We do not go into measure theoretic details unless necessary in this course. In fact, most treatment is about **discrete** probability and random variables.

For the discrete case, the underlying sigma field is usually irrelevant since the collection of all subsets suffice.

[1] This map has to be measurable.

Discrete probability and random variable

For a discrete random variable X taking values in a countable alphabet $\mathcal{X}$, its probability law is completely described by the probability mass function (PMF) $\mathsf{p}_X : \mathcal{X} \to [0,1]$ satisfying

$$\sum_{x \in \mathcal{X}} \mathsf{p}_X(x) = 1.$$

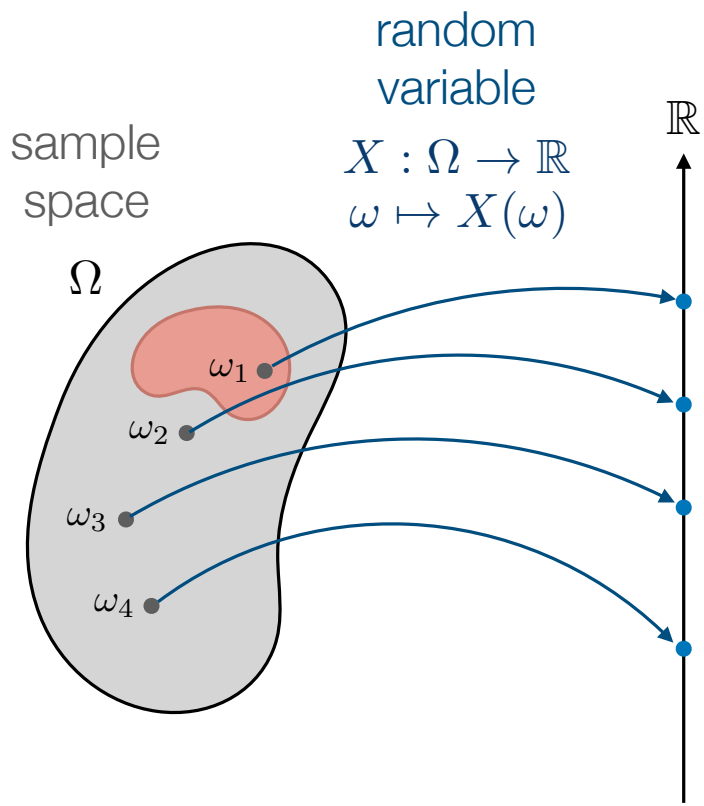
If the cardinality of the alphabet $\mathcal{X}$ is finite:

$$|\mathcal{X}| =: d < \infty.$$

then a discrete probability can be geometrically viewed as an d-dimensional vector $\boldsymbol{p}$ in a **probability simplex**

$$\mathcal{P}_d := \left\{ \boldsymbol{x} \in \mathbb{R}^d \,\middle|\, \sum_{i=1}^{d} x_i = 1,\ x_i \geq 0 \ \forall\, i \right\}.$$

Real-valued random variable



$$(\Omega, \mathscr{F}, P) \xrightarrow{X} (\mathbb{R}, \mathscr{B}, \mathsf{F}_X)$$

Probability space

$$(\Omega, \mathscr{F}, P)$$

sample space — sigma field — probability measure

event: subset of sample points

sigma field: collection of all events

probability measure: assign the probability of an event

random variable: maps a sample point to a numeric outcome

cumulative distribution function:

$$\mathsf{F}_X(x) \triangleq P\{X(\omega) \le x\}, \ x \in \mathbb{R}.$$

(distribution)

Continuous random variable with density

For a *continuous* random variable $X : \Omega \to \mathbb{R}$, its **probability density function** (PDF, or *density* in short) is defined as

$$\mathsf{f}_X(x) := \frac{\mathrm{d}}{\mathrm{d}x}\mathsf{F}_X(x)$$

if the derivative exists.

Accepting the notion of Dirac delta functions, one may also talk about the "density" of a discrete random variable.

Although there are other real-valued random variables that are neither discrete nor continuous, we mainly focus on these two kinds as well as their mixtures throughout this course unless specified otherwise.

Expectation

Consider $g : \mathcal{X} \to \mathbb{R}$, measuring some performance of $x \in \mathcal{X}$ numerically.

The *expected* performance over a discrete (continuous) random variable $X \sim \mathsf{P}$ is just the **expectation** of $g(X)$:

$$\begin{aligned} \mathsf{E}_{X\sim\mathsf{P}}\left[g(X)\right] &:= \sum_{x\in\mathcal{X}} \mathsf{P}(x)g(x) && \text{(discrete random variable)} \\ &:= \int_{x\in\mathbb{R}} \mathsf{P}(x)g(x)\,\mathrm{d}x && \text{(continuous random variable)} \end{aligned}$$

where $\mathsf{P}(\cdot)$ is understood either as the PMF or PDF of X.

More generally $g(X)$ can be a vector in $\mathbb{R}^k$ or a matrix $\mathbb{R}^{k\times m}$, and the expectation can be defined similarly.

Note: the composition $(g \circ X)$ gives a discrete/continuous real-valued random variable/vector/matrix.

Multiple discrete random variables

Multiple discrete random variables $(X_1, ..., X_n) \in \mathcal{X}^n$ can be viewed as a discrete random variable.

The probability law is determined by an d^n-dimensional vector in the probability simplex $\mathcal{P}_{d^n}$.

Usually, we call it the *joint* probability mass function of the random variables, with the following notation:

$$(X_1, ..., X_n) \equiv \boldsymbol{X} \equiv X^n \sim \mathsf{P}_{X_1,...,X_n}.$$

$(X_1, ..., X_n)$ can be viewed as a random vector $\boldsymbol{X}$, or a random sequence X^n.

Marginalization is the procedure of computing the (joint) probability law of an ordered subset of random variables $(X_{i_1}, ..., X_{i_k})$, $i_1, ..., i_k \in \{1, ..., n\}$ from the joint law $\mathsf{P}_{X_1,...,X_n}$. (Recall: marginal vs. joint)

Conditional probability

For two jointly distributed discrete random variables X and Y, the conditional probability law of Y given X is

$$\mathsf{P}_{Y|X}(y|x) := \frac{\mathsf{P}_{X,Y}(x,y)}{\mathsf{P}_X(x)}, \quad \forall\, x \in \underbrace{\mathsf{supp}\{\mathsf{P}_X\} := \{x \in \mathcal{X} \,|\, \mathsf{P}_X(x) \neq 0\}}_{\text{the } \textit{support} \text{ of the PMF } \mathsf{P}_X}.$$

Clearly, the conditional PMF of discrete random variables Y given X with finite supports can be equivalently represented as a *stochastic matrix* $\mathbf{w}$, where the row corresponding to an $x \in \mathsf{supp}\{\mathsf{P}_X\}$ represents the probability law of Y given $X = x$

Independence

Random variables $X_1, ..., X_n$ are independent (notation: $X_1, ..., X_n : \perp\!\!\!\perp$) iff

$$\mathsf{P}_{X_1,...,X_n}(x_1, ..., x_n)] = \prod_{i=1}^{n} \mathsf{P}_{X_i}(x_i), \quad \forall (x_1, ..., x_n).$$

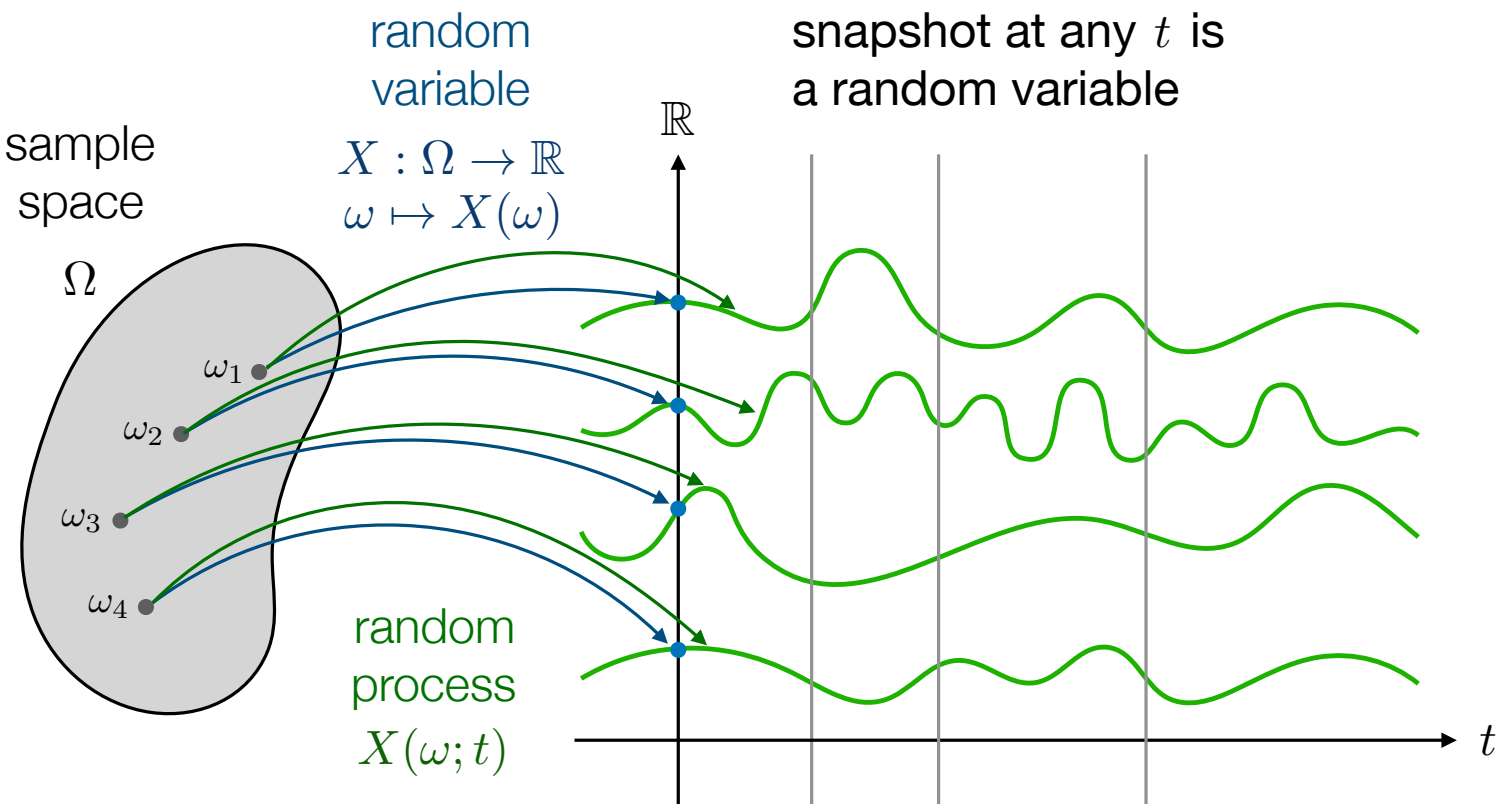
Conditional independence: X is indep. of Z given Y (notation: $X - Y - Z$) iff

$$\mathsf{P}_{Z|Y}(z|y) = \mathsf{P}_{Z|Y,X}(z|y, x), \quad \forall (x, y, z).$$

$X_1 - X_2 - ... - X_n - ...$ form a *Markov chain* iff

$$\mathsf{P}_{X_{n+1}|X^n}(x_{n+1}|x^n) = \mathsf{P}_{X_{n+1}|X_n}(x_{n+1}|x_n), \quad \forall n, \ (x_1, ..., x_n).$$

Random process



When the experiment selects some ω, it maps to a waveform/sequence $X(\omega; t) \equiv X_\omega(t)$

$\Longrightarrow$ random process $\equiv$ random waveform/sequence

Discrete-time random process

The index set $\mathbb{I}$ in a random process $\{X[t] \,|\, t \in \mathbb{I}\}$ can be countable ($\mathbb{I} = \mathbb{N}, \mathbb{Z}$) or uncountable ($\mathbb{I} = \mathbb{R}$).

In this course we will only deal with discrete-time random processes $\{X[t] \,|\, t \in \mathbb{N}\}$, and hence it can be viewed as a random sequence.

Its probability law is characterized by **all** joint PMF's $\mathsf{P}_{X[1],\ldots,X[n]}$, $\forall\, n = 1, 2, \ldots$.

Definition 1 (Stationary Random Process)

A random process $\{X[t]\}$ is *stationary* if for all shift $l \in \mathbb{N}$,

$$\mathsf{P}_{X[1],X[2],\ldots,X[n]} = \mathsf{P}_{X[1+l],X[2+l],\ldots,X[n+l]},\ \forall\, n \in \mathbb{N}.$$

There are other important notions such as ergodicity. We talk about them later when needed.

Convergence of random variables (1/4)

With a sequence of random variables $X_1, X_2, ...$, it is natural to ask if the random sequence converges to another random variable X "in some sense":

$$X_n \to X \quad \text{as } n \to \infty.$$

Due to the inherent randomness, there are different notions of convergences. Let us recap them as follows.

1. Viewing a random variable as a mapping from Ω to $\mathbb{R}$, we can talk about *pointwise* convergence and *uniform* convergence:
   - $X_n \to X$ pointwise: $\forall\, \omega \in \Omega$, $\forall\, \epsilon > 0$, there exists a sufficiently large $N_\epsilon(\omega) \in \mathbb{N}$ such that $\forall\, n \geq N_\epsilon(\omega)$, $|X_n(\omega) - X(\omega)| < \epsilon$.
   - $X_n \to X$ uniformly: $\forall\, \epsilon > 0$, there exists a sufficiently large $N_\epsilon \in \mathbb{N}$ such that $\forall\, \omega \in \Omega$, $\forall\, n \geq N_\epsilon$, $|X_n(\omega) - X(\omega)| < \epsilon$.

   But under the probability measure P, it makes little sense to ask for convergence for every ω.

Convergence of random variables (2/4)

2. Almost sure convergence:

$$X_n \xrightarrow{\text{a.s.}} X \iff \Pr\left\{\lim_{n\to\infty} X_n = X\right\} = 1.$$

There could be some sample points ω such that $\{X_n(\omega)\}$ does not converge to $X(\omega)$. However, the total probability of those points is zero.

3. Convergence in L^p, $p \geq 1$: sometimes we want to use the expected L^p-norm to measure the distance, and making sure it converges to zero.

$$X_n \xrightarrow{L^p} X \iff \lim_{n\to\infty} \mathsf{E}\left[|X_n - X|^p\right] = 0.$$

Convergence of random variables (3/4)

4. Convergence in probability: sometimes we only care about the probability of slight deviation, and want to know if it is vanishing.

$$X_n \xrightarrow{\mathrm{p}} X \iff \forall\, \epsilon > 0,\ \lim_{n\to\infty} \Pr\{|X_n - X| > \epsilon\} = 0.$$

5. Convergence in distribution: sometime we only care about what the distribution of X_n will look like and other related probabilistic notions when n is large and do not care about its relation with X.

$$X_n \xrightarrow{\mathrm{d}} X \iff \lim_{n\to\infty} \mathsf{F}_{X_n}(x) = \mathsf{F}_X(x),\ \forall\, x \in \mathbb{R} \text{ s.t. } \mathsf{F}_X(x) \text{ is continuous.}$$

Convergence of random variables (4/4)

How do these notions relate to one another? Here we summarize some properties in the following.

$$\begin{array}{ccccccc} X_n \xrightarrow{\text{unif}} X & \Longrightarrow & X_n \xrightarrow{\text{pw}} X & \Longrightarrow & X_n \xrightarrow{\text{a.s.}} X & & \\ & & & & \Downarrow & & \\ X_n \xrightarrow{L^q} X & \underset{(q>p\geq 1)}{\Longrightarrow} & X_n \xrightarrow{L^p} X & \Longrightarrow & X_n \xrightarrow{\text{p}} X & \Longrightarrow & X_n \xrightarrow{\text{d}} X \end{array}$$

- Convergence to a constant in distribution implies convergence in probability:

$$X_n \xrightarrow{\text{d}} c \quad \Longrightarrow \quad X_n \xrightarrow{\text{p}} c.$$

- Almost sure convergence of a sequence dominated by another r.v. implies L^1 convergence:

$$X_n \xrightarrow{\text{a.s.}} X,\ |X_n| < Y,\ \mathsf{E}[Y] < \infty \quad \Longrightarrow \quad X_n \xrightarrow{L^1} X.$$

Limit theorems (1/2)

Law of large numbers (LLN): For a sequence of random variables

$$X_i \overset{\text{i.i.d.}}{\sim} \mathsf{P},\ i = 1, 2, ...,$$

which are independent and identically distributed (i.i.d.) copies[2] of an absolutely integrable r.v. $X \sim \mathsf{P}$, that is, $\mathsf{E}_{X\sim\mathsf{P}}[|X|] < \infty$, their sample means converge almost surely to the expected value $\mu = \mathsf{E}_{X\sim\mathsf{P}}[X]$:

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n} S_n \xrightarrow{\text{a.s.}} \mu, \quad \text{that is,} \quad (\bar{X}_n - \mu) \xrightarrow{\text{a.s.}} 0. \tag{1}$$

The i.i.d. condition can be relaxed. In some cases, it is possible to get convergence in probability but not almost sure convergence.

- Called the weak law (WLLN) if the convergence in (1) is “in probability”.
- Called the strong law (SLLN) if the convergence in (1) is “almost sure”.

[2] In fact, it suffices to have pairwise independence.

Limit theorems (2/2)

Central limit theorem (CLT): For a sequence of i.i.d. random variables

$$X_i \overset{\text{i.i.d.}}{\sim} \mathrm{P},\ i = 1, 2, ...,$$

with finite mean μ and variance σ^2, the deviation of the sample mean $\bar{X}_n$ from its true mean μ normalized appropriately converges to a zero-mean Gaussian:

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \sigma^2).$$

The above classical CLT can be extended in various ways:

- Lyapunov/Lindeberg CLT relax the identically distributed condition to *Lyapunov/Lindeberg* condition on moments respectively.
- Berry-Esseen theorem tells the speed of concentration for CLT.
- And much more.

Weak law of large numbers

Theorem 1 (Khinchin WLLN)

For a sequence of i.i.d. RVs $X_1, X_2, \ldots$ *with* $\mathsf{E}[|X_i|] < \infty$ ~~*and variance* $\sigma^2 < \infty$~~

$$\lim_{n\to\infty} \Pr\{|\bar{X}_n - \mu| \geq \epsilon\} = 0 \quad \forall\, \epsilon > 0,$$ *where* $\mu = \mathsf{E}[X_i]$ *denotes the mean.*

Remark (faster convergence):

In the proof of WLLN, it only guarantees that the chance that the estimation is bad will vanish to zero like $1/n$.

However, $1/n$ is pretty slow, meaning that you may need to get a lot of samples to ensure the estimation is good with high probability.

Large deviation theory tells that for i.i.d. RVs with "well-behaved" MGF, the chance that the estimation is bad vanishes to 0 *exponentially fast*!

Central limit theorem

Theorem 2 (Central limit theorem)

For i.i.d. RVs $X_1, X_2, \ldots$ with $\mathsf{E}[|X_i|] < \infty$, mean μ, and variance $\sigma^2 < \infty$,

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{\mathrm{d}} Z \sim \mathrm{N}(0,1) \quad \textit{as } n \to \infty.$$

CLT restated

For i.i.d. RVs $X_1, X_2, \ldots$ with $\mathsf{E}[|X_i|] < \infty$, mean μ, and variance $\sigma^2 < \infty$, the CDFs of $\frac{S_n - n\mu}{\sqrt{n}\sigma}$ converges to the standard normal CDF, that is,

$$\lim_{n\to\infty} \mathsf{F}_{\frac{S_n - n\mu}{\sqrt{n}\sigma}}(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}t^2} \,\mathrm{d}t \quad \forall\, z \in \mathbb{R}.$$

Rate of convergence

CLT tells us that the approximation error converges to 0. Now the question is, *how fast* does it converge to zero? How does it depend on the sample size n?

The following theorem says that the convergence is of the order of $O(1/\sqrt{n})$.

Theorem 3 (Berry-Esseen)

For i.i.d. RVs $X_1, X_2, \ldots$ with finite mean μ, variance σ^2, and absolutely third central moment $\rho_3 := \mathsf{E}[|X_i - \mu|^3]$, let $Z_n := \frac{S_n - n\mu}{\sqrt{n}\sigma}$ and $Z \sim \mathrm{N}(0,1)$. Then,

$$|\mathsf{F}_{Z_n}(z) - \mathsf{F}_Z(z)| \le c\frac{\rho_3}{\sigma^3} n^{-1/2} \quad \forall\, z \in \mathbb{R} \text{ and } n \in \mathbb{N},$$

for some universal constant $c \in (0.4, 0.5)$.

Note: extensions to non-identically-distributed cases (triangular array), random vectors, functions of independent RVs, etc., are available.

1 Probability and Random Processes

2 Optimization

Optimization problems

Finding a minimizer/maximizer of a function, subject to certain constraints.

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathcal{X}} \quad & f_0(\boldsymbol{x}) \\ \text{subject to} \quad & f_i(\boldsymbol{x}) \le 0,\ i = 1, ..., m \\ & h_j(\boldsymbol{x}) = 0,\ j = 1, ..., p \end{aligned}$$

1. Optimization variable[3] $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$.
2. Objective function $f_0(\cdot) : \mathcal{X} \to \mathbb{R}$. (domain of the objective function $\mathcal{X} \equiv \mathrm{dom} f_0$)
3. Constraints:
   - Inequality constraints $f_i(\cdot) : \mathcal{X} \to \mathbb{R}$.
   - Equality constraints $h_j(\cdot) : \mathcal{X} \to \mathbb{R}$.

   Feasible set: $\mathcal{F} := \{\boldsymbol{x} \in \mathcal{X} \mid f_i(\boldsymbol{x}) \le 0, h_j(\boldsymbol{x}) = 0, \forall\, i, j\}$.

[3] In general it can be a scaler, a vector, a matrix, etc.. Here we consider real vectors only.

Local optimality conditions

Local optimum:

$\boldsymbol{x}^*$ is a local minimum iff there exists a small *closed* ball centered at $\boldsymbol{x}^*$ such that for all $\boldsymbol{y}$ in the ball, $f(\boldsymbol{y}) \geq f(\boldsymbol{x}^*)$.

Local optimum cannot be on the *boundary* of a closed feasible set.

For differentiable functions, there are conditions to check local optimality. Below are some conditions for $\boldsymbol{x}^*$ being a local minimum of $f(\cdot)$.

- First-order condition: $f(\cdot)$ is differentiable.

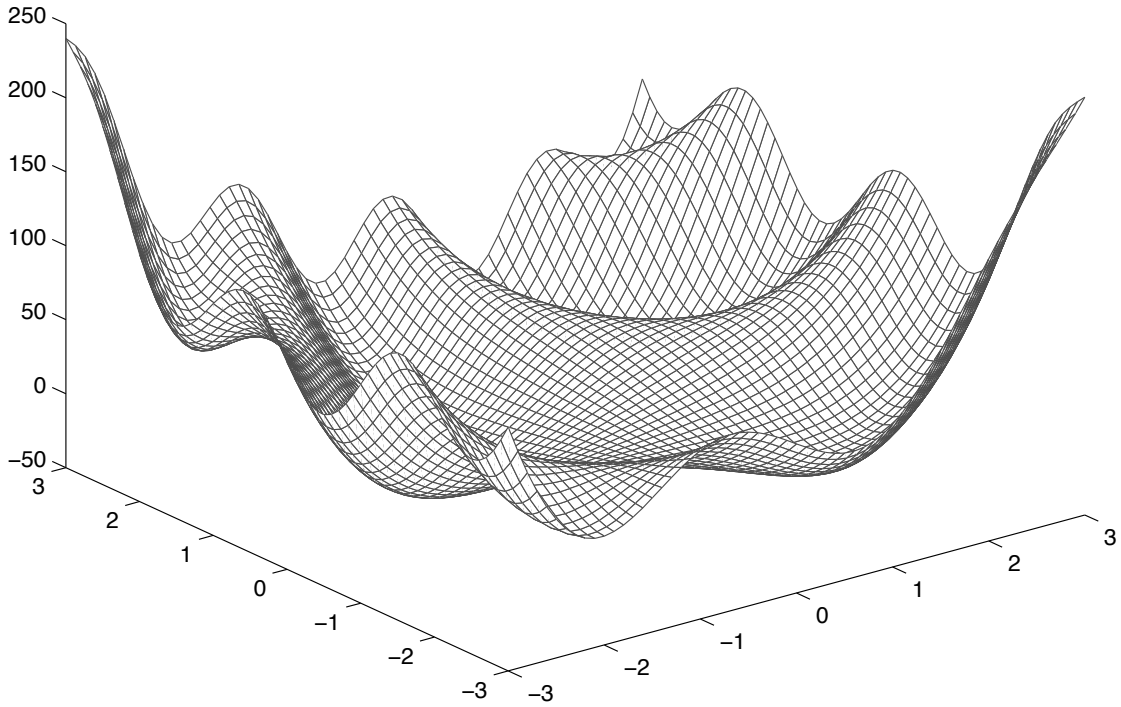
  A necessary condition: $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.

- Second-order condition: $f(\cdot)$ is twice differentiable.

  A necessary condition: $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ and $\nabla^2 f(\boldsymbol{x}^*) \succeq \mathbf{0}$.

- Second-order condition: $f(\cdot)$ is twice differentiable.

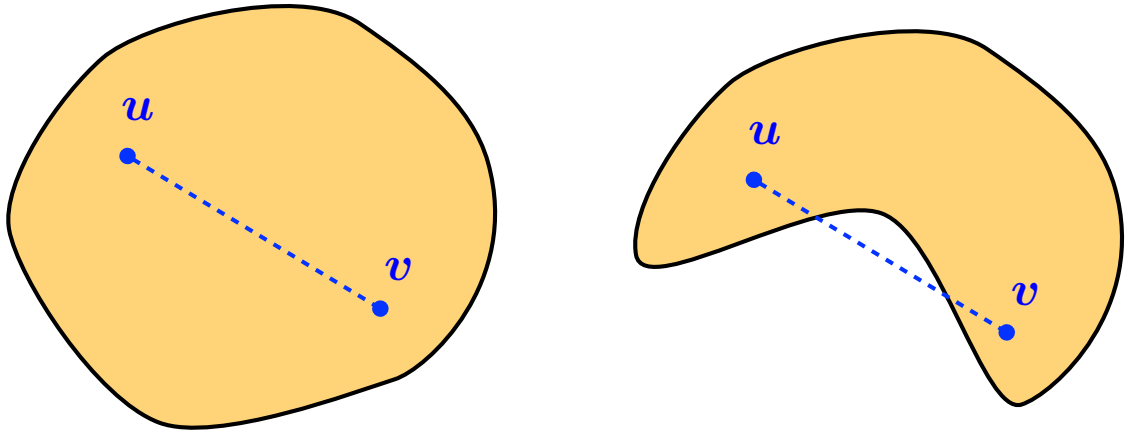
  A sufficient condition for $\boldsymbol{x}^*$ being a *strict* local minimum:
  $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ and $\nabla^2 f(\boldsymbol{x}^*) \succ \mathbf{0}$.



From local to global: we need additional problem structures

$\Longrightarrow$ **Convexity**

Convex sets



convex non-convex

Definition 2

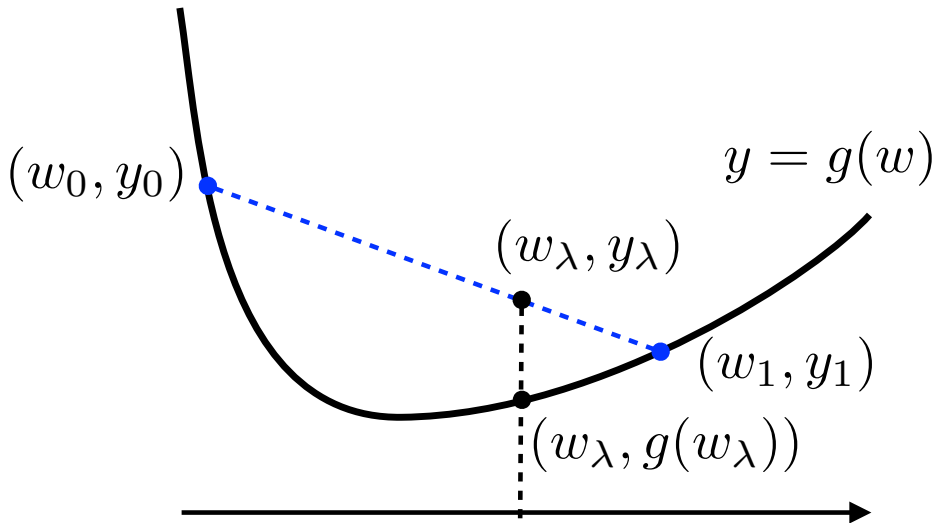
A set $\mathcal{C} \subseteq \mathbb{R}^d$ is convex if $\forall\, \boldsymbol{u}, \boldsymbol{v} \in \mathcal{C}$ and $\forall\, \lambda \in [0, 1]$,

$$\lambda \boldsymbol{u} + (1 - \lambda)\boldsymbol{v} \in \mathcal{C}.$$

Examples of convex sets

- $\mathbb{R}^d$.
- Hyperplane: $\{\boldsymbol{x} \mid \langle \boldsymbol{a}, \boldsymbol{x} \rangle = b\}$; Halfspace: $\{\boldsymbol{x} \mid \langle \boldsymbol{a}, \boldsymbol{x} \rangle \leq b\}$.
- Polyhedra: intersection of hyperplanes and halfspaces.
- Any linear subspace (recall your undergrad linear algebra).
- Any affine subspace (can be viewed as the translation of linear subspace).
- Probability simplex: $\mathcal{P}_d := \{\boldsymbol{x} \in \mathbb{R}^d \mid \sum_{i=1}^d x_i = 1,\ x_i \geq 0\ \forall\, i\}$.
- Norm ball $\{\boldsymbol{x} \mid \|\boldsymbol{x}\| \leq 1\}$; norm cone $\{(\boldsymbol{x}, t) \mid \|\boldsymbol{x}\| \leq t\}$.
- Intersection of convex sets.
- Positive semidefinite matrices (PSD cone): $\{\mathbf{a} \mid \langle \boldsymbol{x}^\intercal \mathbf{a} \boldsymbol{x} \geq 0,\ \forall\, \boldsymbol{x} \in \mathbb{R}^n\}$.

Convex functions



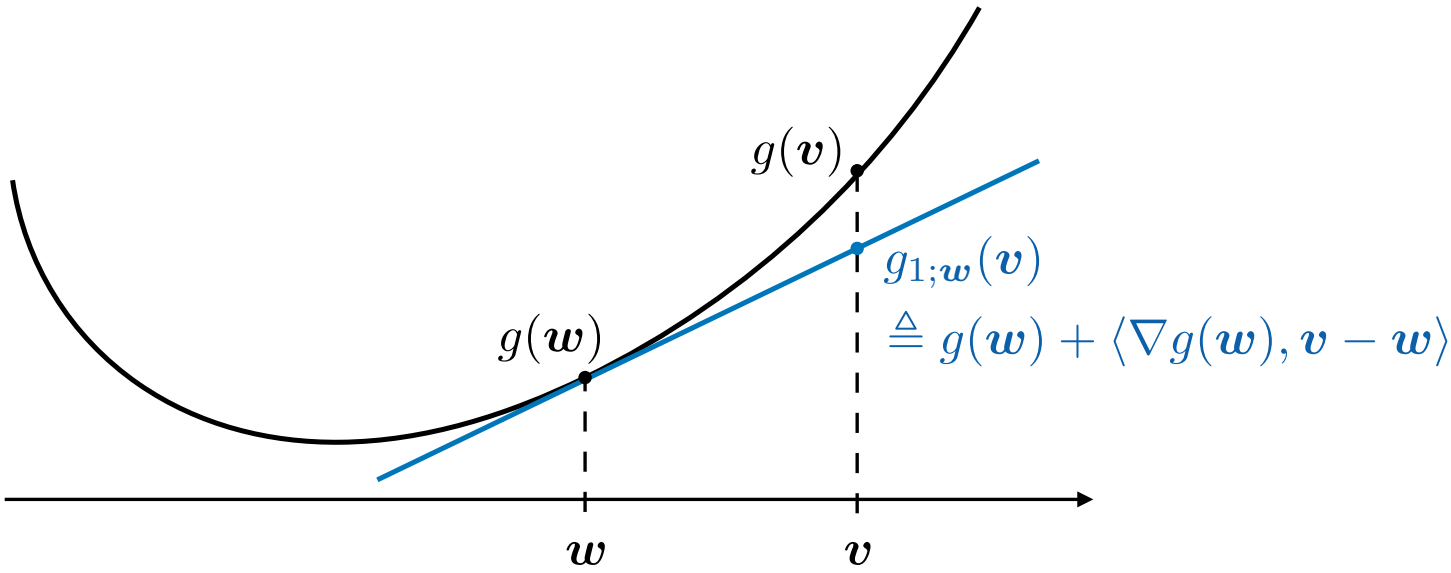
Definition 3

A function $g : \mathcal{W} \to \mathbb{R}$ is convex if its domain $\mathcal{W}$ is convex and $\forall\, \boldsymbol{w}_0, \boldsymbol{w}_1 \in \mathcal{W}$ and $\forall\, \lambda \in [0, 1]$,

$$y_\lambda - g(\boldsymbol{w}_\lambda) \geq 0,$$

where $\boldsymbol{w}_\lambda := (1-\lambda)\boldsymbol{w}_0 + \lambda \boldsymbol{w}_1$, $y_i := g(\boldsymbol{w}_i)$, and $y_\lambda := (1-\lambda)y_0 + \lambda y_1$.

First-order conditions for convexity



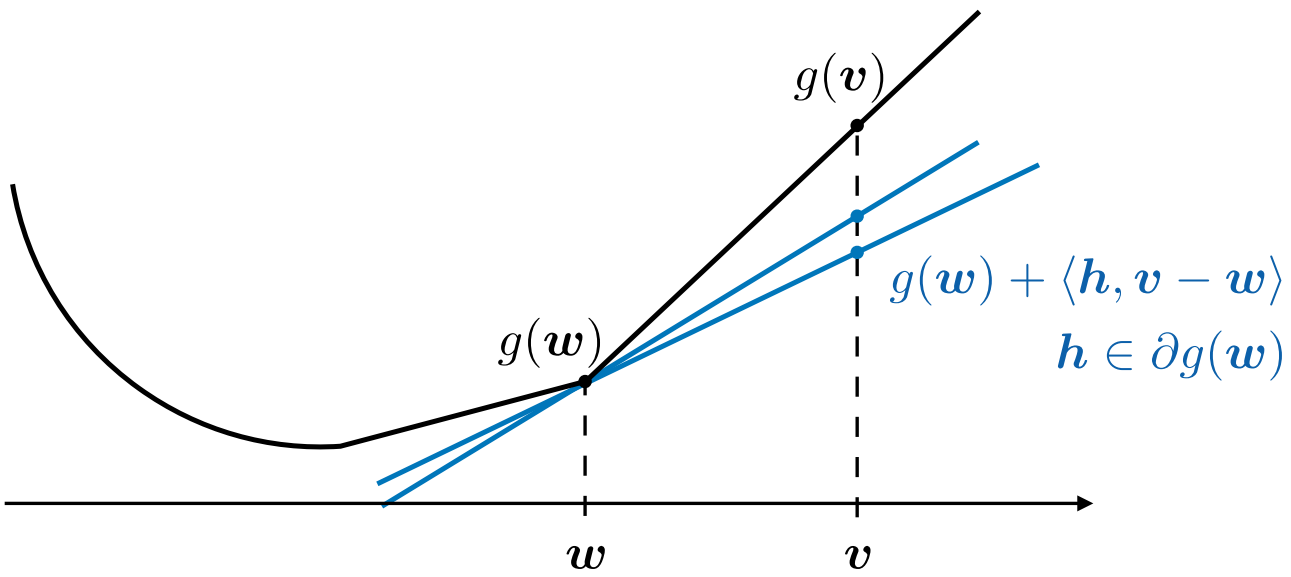
Fact 1

A differentiable function $g : \mathcal{W} \to \mathbb{R}$ is convex if and only if

$$g(\boldsymbol{v}) - g_{1;\boldsymbol{w}}(\boldsymbol{v}) \geq 0 \quad \forall\, \boldsymbol{w}, \boldsymbol{v} \in \mathcal{W},$$

where $g_{1;\boldsymbol{w}}(\boldsymbol{v}) := g(\boldsymbol{w}) + \langle \nabla g(\boldsymbol{w}), \boldsymbol{v} - \boldsymbol{w} \rangle$ denotes the first-order approximation of g at $\boldsymbol{v}$ with respect to $\boldsymbol{w}$.

First-order conditions for convexity (cont'd)



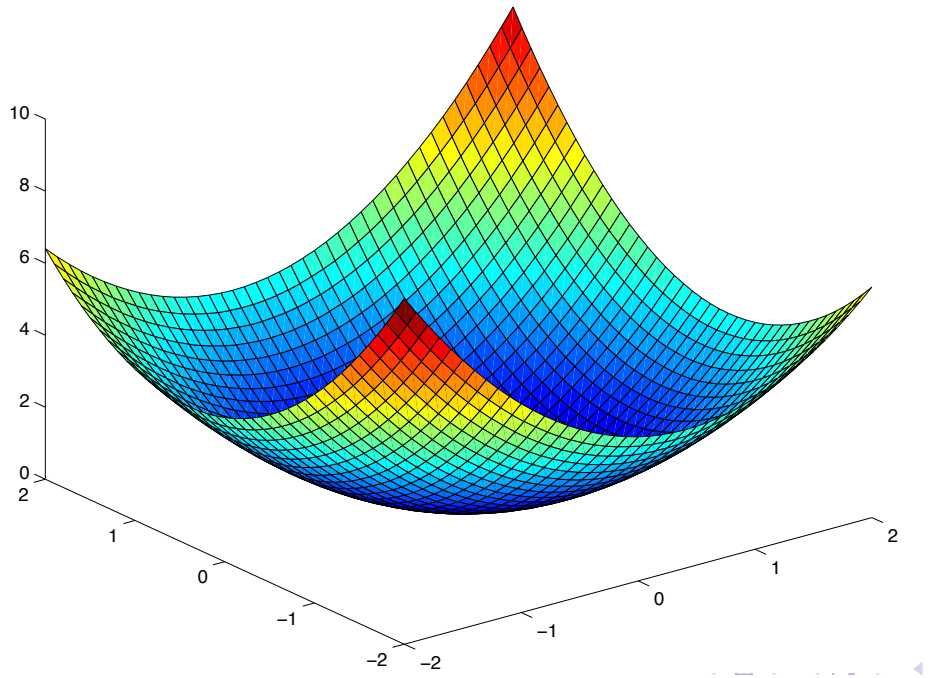
Fact 2

For a general function $g : \mathcal{W} \to \mathbb{R}$, define its subdifferential at point $\boldsymbol{w}$ as

$$\partial g(\boldsymbol{w}) := \{\boldsymbol{h} : g(\boldsymbol{v}) \geq g(\boldsymbol{w}) + \langle \boldsymbol{h}, \boldsymbol{v} - \boldsymbol{w} \rangle, \ \forall\, \boldsymbol{v} \in \mathcal{W}\}.$$

g is convex if and only if $\partial g(\boldsymbol{w}) \neq \emptyset \ \forall\, \boldsymbol{w} \in \mathcal{W}$.

Second-order condition for convexity



Fact 3

For a twice-differentiable function $g : \mathcal{W} \to \mathbb{R}$, it is convex iff its Hessian is positive semi-definite:

$$\nabla^2 g(\boldsymbol{w}) \succeq \mathbf{0}, \ \forall\, \boldsymbol{w} \in \mathcal{W}.$$

Examples of convex functions

- Affine functions: $f(\boldsymbol{x}) = \langle \boldsymbol{b}, \boldsymbol{x} \rangle + c$.
- Quadratic functions: $f(\boldsymbol{x}) = \frac{1}{2}\langle \boldsymbol{x}, \mathbf{a}\boldsymbol{x} \rangle + \boldsymbol{b}^\mathsf{T}\boldsymbol{x} + c$.
- Norms: $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$.
- Composition of a convex function with an affine function:

$$f : \text{convex} \implies f(\mathbf{a}\boldsymbol{x} + \boldsymbol{b}) : \text{convex}.$$

- Hinge loss: $\ell\big(\boldsymbol{w}, (\boldsymbol{x}_i, y_i)\big) \equiv f_{(\boldsymbol{x}_i, y_i)}(\boldsymbol{w}) = \max(0, 1 - y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{w})$.
- Logistic loss: $\ell\big(\boldsymbol{w}, (\boldsymbol{x}_i, y_i)\big) \equiv f_{(\boldsymbol{x}_i, y_i)}(\boldsymbol{w}) = \log(1 + \mathrm{e}^{-y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{w}})$.

Verify convexity

1. Verify definition.
2. Check the second-order condition $\nabla^2 f(\boldsymbol{x}) \succeq \mathbf{0}$ if it is twice differentiable.
3. Show that f is obtained from simple convex functions by operations that preserve convexity:
   - nonnegative weighted sum
   - composition with affine function
   - pointwise maximum and supremum
   - minimization
   - etc.

Convex optimization problems (convex programs)

Definition 4

An optimization problem is convex if its objective function f_0 is convex, and the feasible set $\mathcal{F}$ is convex, that is,

- its inequality constraints f_i's are convex, and
- its equality constraints h_j's are affine.

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathcal{X}} \quad & f_0(\boldsymbol{x}) && \text{(convex)} \\ \text{subject to} \quad & f_i(\boldsymbol{x}) \le 0,\ i = 1, ..., m && \text{(convex)} \\ & h_j(\boldsymbol{x}) = 0,\ j = 1, ..., p && \text{(affine)} \end{aligned}$$

The objective function should be convex for minimization (concave for maximization), and the feasible set should be a convex set.

Global optimality of convex programs

Fact 4 (Local optima $\equiv$ Global optima)

For a convex optimization problem, a local minimizer is a global minimizer.

Fact 5 (First-order optimality condition)

The minimizer of a convex program

$$\min_{\boldsymbol{x}} \{f_0(\boldsymbol{x}) | \boldsymbol{x} \in \mathcal{F}\}$$

can be characterized by the following first-order optimality condition

$$\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}} \{f_0(\boldsymbol{x}) | \boldsymbol{x} \in \mathcal{F}\} \iff \langle \nabla f_0(\boldsymbol{x}^*), \boldsymbol{x} - \boldsymbol{x}^* \rangle \geq 0, \ \forall \boldsymbol{x} \in \mathcal{F}.$$

Lagrange duality

Goal:

1. Get certificate of optimality
2. Remove constraints
3. Reformulate problem

Lagrangian and dual

Given an optimization problem

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathcal{X}} \quad & f_0(\boldsymbol{x}) \\ \text{subject to} \quad & f_i(\boldsymbol{x}) \le 0,\ i=1,...,m \\ & h_j(\boldsymbol{x}) = 0,\ j=1,...,p \end{aligned} \quad , \tag{2}$$

its Lagrangian is formed by introducing *Lagrange multipliers* $\lambda_i \ge 0$ and $\nu_j \in \mathbb{R}$:

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\nu}) := f_0(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i f_i(\boldsymbol{x}) + \sum_{j=1}^{p} \nu_j h_j(\boldsymbol{x}).$$

Its dual function is

$$g(\boldsymbol{\lambda},\boldsymbol{\nu}) := \inf_{\boldsymbol{x}\in\mathcal{X}} \mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\nu}).$$

Dual problem

The dual problem is defined as

$$\max_{\boldsymbol{\lambda}\geq 0,\boldsymbol{\nu}} g(\boldsymbol{\lambda},\boldsymbol{\nu}).$$

Fact 6

Dual problem is always a convex optimization problem.

This is because $g(\boldsymbol{\lambda},\boldsymbol{\nu})$ is a concave function.

The original (primal) problem is just $\min_{\boldsymbol{x}} \left\{\sup_{\boldsymbol{\lambda}\geq 0,\boldsymbol{\nu}} \mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\nu})\right\}$.

The dual problem is just switching the order of min and max above:

$$\max_{\boldsymbol{\lambda}\geq 0,\boldsymbol{\nu}} \left\{\inf_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda},\boldsymbol{\nu})\right\}.$$

Weak duality

Fact 7

Let p^* and d^* denote the optimal value of the primal problem and the dual problem respectively. Then, for any kinds of primal problem,

$$d^* \leq p^*.$$

This is because $\forall\, \boldsymbol{\lambda} \geq 0$, $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^* = f_0(\boldsymbol{x}^*)$.

Main use – certificate of optimality (also known as duality gap): Suppose we have a feasible x and want to know how far away $f_0(\boldsymbol{x})$ is from the optimum.

- Get the dual (which is simpler to solve) $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$.
- Then, we know that $f_0(\boldsymbol{x}) - f_0(\boldsymbol{x}^*) \leq f_0(\boldsymbol{x}) - g(\boldsymbol{\lambda}, \boldsymbol{\nu})$.

Also used in developing more advanced algorithms (primal-dual algorithms).

Intuitions about Lagrange duality

Let

$$\mathbb{1}_-(a) := \begin{cases} \infty, & a > 0 \\ 0, & \text{otherwise} \end{cases}, \qquad \mathbb{1}_0(a) := \begin{cases} \infty, & a \neq 0 \\ 0, & \text{otherwise} \end{cases}.$$

The primal problem can be rewritten as

$$\min_{\boldsymbol{x}} \left\{ f_0(\boldsymbol{x}) + \sum_{i=1}^{m} \mathbb{1}_-(f_i(\boldsymbol{x})) + \sum_{j=1}^{p} \mathbb{1}_0(h_j(\boldsymbol{x})) \right\}.$$

The dual function is formed by replacing $\mathbb{1}_-(f_i(\boldsymbol{x}))$ by its lower bound $\lambda_i f_i(\boldsymbol{x})$, and replacing $\mathbb{1}_0(h_j(\boldsymbol{x}))$ by its lower bound $\nu_j h_j(\boldsymbol{x})$.

Strong duality

Fact 8

For "reasonable" convex problems, $d^* = p^*$.

- Reasonable conditions $\equiv$ Constraint qualifications
- Slater's constraint qualification: $\exists$ strictly feasible $\boldsymbol{x} \in$ interior of $\mathcal{X}$.

Note: There are nonconvex problems where strong duality holds.

Karush-Kuhn-Tucker (KKT) conditions

For a convex optimization problem (2) with differentiable objective function f_0 and inequality constraints $f_1, ..., f_m$, suppose that there exists $\boldsymbol{x}$ in the interior of the domain $\mathcal{X}$ that is *strictly feasible*:

$$f_i(\boldsymbol{x}) < 0,\ \forall\, i = 1, ..., m \quad \text{and} \quad h_j(\boldsymbol{x}) = 0,\ \forall\, j = 1, ..., p.$$

Then, a *feasible* $\boldsymbol{x}^*$ attains the optimality iff $\exists\, (\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ such that

$$\begin{aligned} &\lambda_i^* \geq 0\ \forall\, i = 1, 2, \ldots, m \\ &\lambda_i^* f_i(\boldsymbol{x}^*) = 0,\ \forall\, i = 1, 2, \ldots, m \\ &\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})\big|_{(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = (\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)} = \mathbf{0} \end{aligned} \tag{3}$$

(3) together with the feasibility of $\boldsymbol{x}^*$ are called the *Karush-Kuhn-Tucker (KKT) conditions*.