Machine Learning II
A crash course on Optimization

Le Thi Khanh Hien
UMONS, thikhanhhien.le@umons.ac.be

Mons, February 2023

What will be covered in this crash course?

Lecture 1:

- Prerequisites of linear algebra and mathematical analysis
- A brief introduction to convex optimization

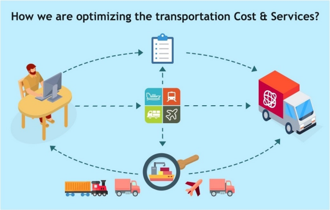
Lecture 2:

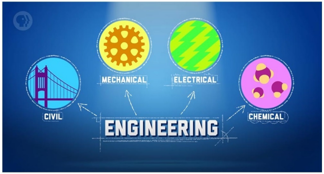
- Gradient descent method
- Newton method
- Proximal point algorithm

Lecture 3:

- Accelerated proximal point algorithm
- Stochastic gradient method

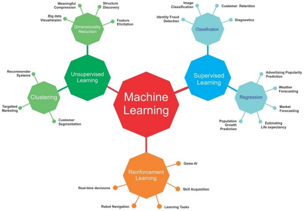
Where do we meet optimization?





OPTIMIZATION

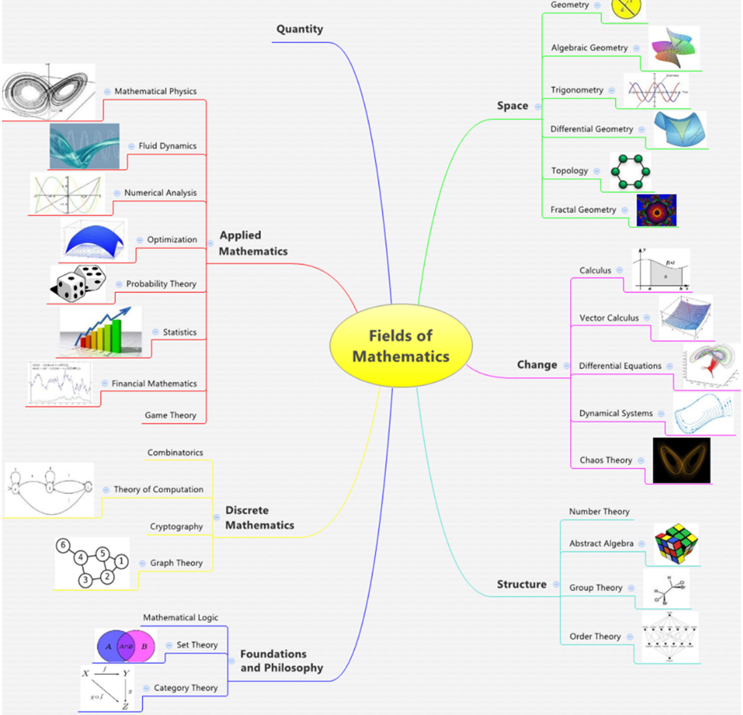
is everywhere



An example - diet problem

Design a diet plan that minimizes the expense per day

Food	Energy (kcal)	Protein (g)	Calcium (mg)	Price per serving (€)	Daily limit
Oatmeal	110	4	2	3	3
Chicken	205	32	12	24	3
Eggs	160	13	54	13	5
Milk	160	8	285	9	8
Pie	420	4	22	24	3
Pork	260	14	80	13	2


Fields of Mathematics
Quantity
Space
Geometry
Algebraic Geometry
Trigonometry
Differential Geometry
Topology
Fractal Geometry
Change
Calculus
Vector Calculus
Differential Equations
Dynamical Systems
Chaos Theory
Structure
Number Theory
Abstract Algebra
Group Theory
Order Theory
Applied Mathematics
Mathematical Physics
Fluid Dynamics
Numerical Analysis
Optimization
Probability Theory
Statistics
Financial Mathematics
Game Theory
Discrete Mathematics
Combinatorics
Theory of Computation
Cryptography
Graph Theory
Foundations and Philosophy
Mathematical Logic
Set Theory
Category Theory

source

Optimization - how?

- **Mathematical Modelling**: describing a real world problem in mathematical terms, and defining the corresponding optimization problem.
- **Computational Optimization**: using an appropriate optimization algorithm to find an approximate solution to the optimization problem.

Optimization - how?

- **Mathematical Modelling**: describing a real world problem in mathematical terms, and defining the corresponding optimization problem.
- **Computational Optimization**: using an appropriate optimization algorithm to find an approximate solution to the optimization problem.

Optimization for Machine Learning

- **Mathematical Modelling**: mathematically modelling the machine learning problem.
- **Computational Optimization**: learn the model parameters.
  - in practice, many libraries are available but practitioners consider optimization algorithms as "black box".
  - in this course, we study the algorithms and try to understand how they work.

General optimization problem

$$\begin{aligned} \min_{x} \quad & f(x) \\ s.t. \quad & x \in \mathcal{X}. \end{aligned}$$

Example:

- Regularized linear regression

$$\min_{x \in \mathbb{R}^d} \frac{1}{2}\|Ax - y\|_2^2 + \lambda R(x),$$

where $\{(a_i, y_i)\}$ for $i = 1, \ldots, n$, are n pairs of training data and A is a matrix whose i-th row is a_i^T.

General optimization problem

$$\min_{x} \quad f(x)$$
$$s.t. \quad x \in \mathcal{X}.$$

Example:

- Regularized linear regression

$$\min_{x\in\mathbb{R}^d} \frac{1}{2}\|Ax - y\|_2^2 + \lambda R(x),$$

where $\{(a_i, y_i)\}$ for $i = 1, \ldots, n$, are n pairs of training data and A is a matrix whose i-th row is a_i^T.

- Regularized logistic regression

$$\min_{w\in\mathbb{R}^m} \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i\left\langle x_i, w\right\rangle\right)\right) + \lambda\|w\|_1,$$

where $\{(x_i, y_i)\}$ for $i = 1, \ldots, n$, $x^i \in \mathbb{R}^m$ and $y_i \in \{-1, 1\}$ are n pairs of training data.

Table of content

Some preliminaries of linear algebra and mathematical analysis

Some notations

- Sets $\mathcal{X}$, (a, b), $[a, b)$, $[a, b]$, $(a, b]$, $\mathbb{R}$, $\mathbb{R}_+$, $x \in \mathcal{X}$.
- Real-valued functions: $f : \mathcal{X} \to \mathcal{Y} \subset \mathbb{R}$, $\mathcal{X}$ is called the domain, $\mathcal{Y}$ is called the range.

- Matrices
  - Matrix addition

  - Matrix product

  - Square matrix, trace of square matrix

  - Eigenvalues of a square matrix, spectral radius

  - Singular values of a matrix

  - Positive definite and positive semidefinite matrix

Real vector space

A vector space $\mathbb{E}$ over $\mathbb{R}$ is a set of elements (which are called "vectors") such that:

(A) For any $\mathbf{x}, \mathbf{y} \in \mathbb{E}$, there corresponds a "sum" $\mathbf{x} + \mathbf{y}$ that satisfies the following properties

- $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{E}$.
- $\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z}$ for any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{E}$.
- There exists a unique "zero vector" $\mathbf{0}$ in $\mathbb{E}$ such that $\mathbf{x} + \mathbf{0} = \mathbf{x}$ for any $\mathbf{x}$.
- For any $\mathbf{x} \in \mathbb{E}$, there exists $-\mathbf{x} \in \mathbb{E}$ such that $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$.

(B) For any scalar (real number) $a \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{E}$, there corresponds a "scalar multiplication" $a\mathbf{x}$ satisfying the following properties

- $a(b\mathbf{x}) = (ab)\mathbf{x}$ for any $a, b \in \mathbb{R}, \mathbf{x} \in \mathbb{E}$.
- $1\mathbf{x} = \mathbf{x}$ for any $\mathbf{x} \in \mathbb{E}$.

(C) The summation and scalar multiplication satisfy the following properties

- $a(\mathbf{x} + \mathbf{y}) = a\mathbf{x} + a\mathbf{y}$ for any $a \in \mathbb{R}$, $\mathbf{x}, \mathbf{y} \in \mathbb{E}$.
- $(a + b)\mathbf{x} = a\mathbf{x} + b\mathbf{x}$ for any $a, b \in \mathbb{R}$, $\mathbf{x} \in \mathbb{E}$.

Basis. A basis of a vector space $\mathbb{E}$ is a set of linearly independent vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ that spans $\mathbb{E}$: for any $\mathbf{x} \in \mathbb{E}$, there exists $\beta_1, \ldots, \beta_n \in \mathbb{R}$ such that

$$\mathbf{x} = \sum_{i=1}^{n} \beta_i \mathbf{v}_i.$$

Dimension. The dimension of a vector space $\mathbb{E}$ is the number of vectors in a basis of $\mathbb{E}$.

Norm

A norm $\|\cdot\|$ on a vector space $\mathbb{E}$ is a function $\|\cdot\| : \mathbb{E} \to \mathbb{R}_+$ satisfying the following properties:

- (nonnegativity) $\|\mathbf{x}\| \geq 0$ for any $\mathbf{x} \in \mathbb{E}$ and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = 0$.
- (positive homogeneity) $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$ for any $\mathbf{x} \in \mathbb{E}$ and $a \in \mathbb{R}$.
- (triangle inequality) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{E}$.

Norm

A norm $\|\cdot\|$ on a vector space $\mathbb{E}$ is a function $\|\cdot\| : \mathbb{E} \to \mathbb{R}_+$ satisfying the following properties:

- (nonnegativity) $\|\mathbf{x}\| \geq 0$ for any $\mathbf{x} \in \mathbb{E}$ and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = 0$.
- (positive homogeneity) $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$ for any $\mathbf{x} \in \mathbb{E}$ and $a \in \mathbb{R}$.
- (triangle inequality) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{E}$.

Example: l_p-norm of $\mathbb{R}^n$ (with $p \geq 1$):

Inner product

An inner product of $\mathbb{E}$ is a function that associates to each pair of $\mathbf{x}, \mathbf{y}$ a real number denoted by $\langle \mathbf{x}, \mathbf{y} \rangle$ and satisfying the following properties:

- (commutativity) $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{E}$.
- (linearity) $\langle a_1 \mathbf{x}_1 + a_2 \mathbf{x}_2, \mathbf{y} \rangle = a_1 \langle \mathbf{x}_1, \mathbf{y} \rangle + a_2 \langle \mathbf{x}_2, \mathbf{y} \rangle$ for any $a_1, a_2 \in \mathbb{R}$ and $\mathbf{x}_1, \mathbf{x}_2, \mathbf{y} \in \mathbb{E}$.
- (positive definite) $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ for any $\mathbf{x} \in \mathbb{E}$ and $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ if and only if $\mathbf{x} = 0$.

Euclidean Spaces. A finite dimensional real vector space equipped with an inner product $\langle \cdot, \cdot \rangle$ and the norm $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ is called a Euclidean space.

Example.

- $\mathbb{R}^n$ is an Euclidean space with
  - dot product
  - Q-inner product (where Q is a positive definite matrix)
- $\mathbb{R}^{m\times n}$ is an Euclidean space with inner product

Orthogonality $\mathbf{x} \perp \mathbf{y}$ if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$.

Cauchy-Schwarz inequality $\langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2$.

General version of Cauchy-Schwarz inequality: Let $\mathbb{E}$ be an inner product vector space endowed with a norm $\|\cdot\|$. Then we have

$$|\langle \mathbf{y}, \mathbf{x} \rangle| \leq \|\mathbf{y}\|_* \|\mathbf{x}\|, \text{for any } \mathbf{y} \in \mathbb{E}^*, \mathbf{x} \in \mathbb{E},$$

where $\mathbb{E}^*$ is the dual space of $\mathbb{E}$, and $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

General version of Cauchy-Schwarz inequality: Let $\mathbb{E}$ be an inner product vector space endowed with a norm $\|\cdot\|$. Then we have

$$|\langle \mathbf{y}, \mathbf{x} \rangle| \leq \|\mathbf{y}\|_* \|\mathbf{x}\|, \text{for any } \mathbf{y} \in \mathbb{E}^*, \mathbf{x} \in \mathbb{E},$$

where $\mathbb{E}^*$ is the dual space of $\mathbb{E}$, and $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.
(Optional reading) What is a dual space? What is a dual norm?

A linear transformation from $\mathbb{E}$ to $\mathbb{R}$ is called a linear functional. The dual space of $\mathbb{E}$, denoted by $\mathbb{E}^*$ is the set of all linear functionals on $\mathbb{E}$. For inner product spaces, given a linear functional $f \in \mathbb{E}^*$, there always exists $\mathbf{y} \in \mathbb{E}$ such that $f(\mathbf{x}) = \langle \mathbf{y}, \mathbf{x} \rangle$.
Suppose $\mathbb{E}$ is endowed with a norm $\|\cdot\|$, then the dual norm of the dual space is given by

$$\|\mathbf{y}\|_* := \max_{\mathbf{x}:\|\mathbf{x}\| \leq 1} \langle \mathbf{y}, \mathbf{x} \rangle = \max_{\mathbf{x}:\|\mathbf{x}\|=1} \langle \mathbf{y}, \mathbf{x} \rangle .$$

General version of Cauchy-Schwarz inequality: Let $\mathbb{E}$ be an inner product vector space endowed with a norm $\|\cdot\|$. Then we have

$$|\langle \mathbf{y}, \mathbf{x} \rangle| \leq \|\mathbf{y}\|_* \|\mathbf{x}\|, \text{for any } \mathbf{y} \in \mathbb{E}^*, \mathbf{x} \in \mathbb{E},$$

where $\mathbb{E}^*$ is the dual space of $\mathbb{E}$, and $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.
(Optional reading) What is a dual space? What is a dual norm?

Example.

- l_p-norm of $\mathbb{R}^n$ (with $p \geq 1$): $\|x\|_p = \sqrt[p]{\sum_{i=1}^n |x_i|^p}$. The dual norm of l_p-norm with $p > 1$ is l_q-norm, where q satisfies $\frac{1}{p} + \frac{1}{q} = 1$.
- The dual norm of $\|\cdot\|_Q$ is $\|\cdot\|_{Q^{-1}}$.

Matrix norm Let $A \in \mathbb{R}^{m \times n}$.

- Frobenius norm $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2} = \sqrt{\sum_i \sigma_i^2}$.
- Nuclear norm $\|A\|_* = \sum_i \sigma_i$.

Matrix norm Let $A \in \mathbb{R}^{m \times n}$.

- Frobenius norm $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2} = \sqrt{\sum_i \sigma_i^2}$.
- Nuclear norm $\|A\|_* = \sum_i \sigma_i$.

(Optional reading)

(a, b)-norm

$$\|A\|_{a,b} = \max_x \left\{ \|Ax\|_b : \|x\|_a \leq 1 \right\}.$$

- Spectral norm: $\|A\|_2 = \|A\|_{2,2} = \sqrt{\lambda_{\max}(A^\top A)} = \sigma_{\max}(A)$
- 1-norm: $\|A\|_1 = \|A\|_{1,1} = \max_{j=1,\ldots,n} \sum_{i=1}^m |A_{i,j}|$.
- ∞-norm: $\|A\|_\infty = \|A\|_{\infty,\infty} = \max_{i=1,\ldots,m} \sum_{j=1}^n |A_{i,j}|$.

Matrix norm Let $A \in \mathbb{R}^{m \times n}$.

- Frobenius norm $\|A\|_F = \sqrt{\sum_{i,j} A_{ij}^2} = \sqrt{\sum_i \sigma_i^2}$.
- Nuclear norm $\|A\|_* = \sum_i \sigma_i$.

(Optional reading)

(a, b)-norm

$$\|A\|_{a,b} = \max_x \left\{ \|Ax\|_b : \|x\|_a \le 1 \right\}.$$

- Spectral norm: $\|A\|_2 = \|A\|_{2,2} = \sqrt{\lambda_{\max}(A^\top A)} = \sigma_{\max}(A)$
- 1-norm: $\|A\|_1 = \|A\|_{1,1} = \max_{j=1,\dots,n} \sum_{i=1}^m |A_{i,j}|$.
- ∞-norm: $\|A\|_\infty = \|A\|_{\infty,\infty} = \max_{i=1,\dots,m} \sum_{j=1}^n |A_{i,j}|$.

Useful inequalities:

- $\|Ax\|_2 \le \|A\|_2 \|x\|_2$.
- $\rho(A) \le \|A\|_2 \le \|A\|_F \le \|A\|_*$.
- $\langle A, B \rangle \le \|A\|_F \|B\|_F$.

Big O notations

When $x \to \infty$:

- $f(x) = O(g(x)) \Leftrightarrow \exists \alpha, x_0 > 0$ s.t $|f(x)| \leq \alpha |g(x)|, \forall x > x_0$.
- $f(x) = o(g(x)) \Leftrightarrow \forall \alpha > 0, \exists x_0 > 0$ s.t $|f(x)| \leq \alpha |g(x)|, \forall x > x_0$.

Big O notations

When $x \to \infty$:

- $f(x) = O(g(x)) \Leftrightarrow \exists \alpha, x_0 > 0$ s.t $|f(x)| \leq \alpha |g(x)|, \forall x > x_0$.
- $f(x) = o(g(x)) \Leftrightarrow \forall \alpha > 0, \exists x_0 > 0$ s.t $|f(x)| \leq \alpha |g(x)|, \forall x > x_0$.

When $x \to a$:

- $f(x) = O(g(x)) \Leftrightarrow \exists \alpha, d > 0$ s.t $|f(x)| \leq \alpha |g(x)|, \forall x : \|x - a\| < d$.
- $f(x) = o(g(x)) \Leftrightarrow \forall \alpha > 0, \exists d > 0$ s.t
  $|f(x)| \leq \alpha |g(x)|, \forall x : \|x - a\| < d$.

Big O notations

When $x \to \infty$:

- $f(x) = O(g(x)) \Leftrightarrow \exists \alpha, x_0 > 0$ s.t $|f(x)| \leq \alpha |g(x)|, \forall x > x_0$.
- $f(x) = o(g(x)) \Leftrightarrow \forall \alpha > 0, \exists x_0 > 0$ s.t $|f(x)| \leq \alpha |g(x)|, \forall x > x_0$.

When $x \to a$:

- $f(x) = O(g(x)) \Leftrightarrow \exists \alpha, d > 0$ s.t $|f(x)| \leq \alpha |g(x)|, \forall x : \|x - a\| < d$.
- $f(x) = o(g(x)) \Leftrightarrow \forall \alpha > 0, \exists d > 0$ s.t $|f(x)| \leq \alpha |g(x)|, \forall x : \|x - a\| < d$.

(Optional reading)

When $x \to \infty$:

- $f(x) = \Omega(g(x)) \Leftrightarrow \exists \alpha, x_0 > 0$ s.t $|f(x)| \geq \alpha |g(x)|, \forall x > x_0$.
- $f(x) = \omega(g(x)) \Leftrightarrow \forall \alpha > 0, \exists x_0 > 0$ s.t $|f(x)| \leq \alpha |g(x)|, \forall x > x_0$.

When $x \to a$:

- $f(x) = \Omega(g(x)) \Leftrightarrow \exists \alpha, d > 0$ s.t $|f(x)| \geq \alpha |g(x)|, \forall x : \|x - a\| < d$.
- $f(x) = \omega(g(x)) \Leftrightarrow \forall \alpha > 0, \exists d > 0$ s.t $|f(x)| \leq \alpha |g(x)|, \forall x : \|x - a\| < d$.

Example.

- $sin(x) = O(1)$ as $x \to \infty$.
- $x^4 + 100x^2 = O(\quad)$ as $x \to 0$.
- $\frac{3}{n} + \frac{5}{n^2} = O(\quad)$ as $n \to +\infty$.

Example.

- $sin(x) = O(1)$ as $x \to \infty$.
- $x^4 + 100x^2 = O(\quad)$ as $x \to 0$.
- $\frac{3}{n} + \frac{5}{n^2} = O(\quad)$ as $n \to +\infty$.

Table: Classes of functions commonly encountered when analyzing algorithms

Notation	Name
$O(1)$	constant
$O(\log(x))$	logarithmic
$O(x)$	linear
$O(x^q), 1 < q < 2$	super-linear
$O(x^2)$	quadratic
$O(x^c)$ (for some constant c)	polynomial
$O(c^x)$ (for some constant c)	exponential

A few basic differentiation rules

Derivatives

- Scalar function of scalar variable $f : \mathbb{R} \to \mathbb{R}$

$$f'(x) := \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{\Delta f}{\Delta x}.$$

A few basic differentiation rules

Derivatives

- Scalar function of scalar variable $f : \mathbb{R} \to \mathbb{R}$

$$f'(x) := \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{\Delta f}{\Delta x}.$$

- Multivariate scalar function $f : \mathbb{R}^n \to \mathbb{R}$. Suppose f is a continuously twice differentiable function.
  - Partial derivative

$$\frac{\partial f}{\partial x_i} :=$$

  - Gradient

$$\nabla f(x) := (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n})^\top.$$

  - Hessian matrix

$$\nabla^2 f(x) := [\frac{\partial^2 f(x)}{\partial x_i \partial x_j}] =$$

Example.

Find gradient and Hessian of $f : \mathbb{R}^3 \to \mathbb{R}, f(x) = x_1^2 + 3x_1x_2 + 2x_2^2 + x_1$.

(**Optional reading**) Frechet derivative on norm vector spaces

`https://link.springer.com/content/pdf/10.1007/3-7643-7357-1_4.pdf`

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously twice differentiable function.

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously twice differentiable function.
Taylor Expansion:

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + o(\|y - x\|),$$

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2}(y - x)^\top \nabla^2 f(x)(y - x) + o(\|y - x\|^2).$$

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously twice differentiable function.
Taylor Expansion:

$$f(y) = f(x) + \nabla f(x)^\top (y - x) + o(\|y - x\|),$$
$$f(y) = f(x) + \nabla f(x)^\top (y - x) + \frac{1}{2}(y - x)^\top \nabla^2 f(x)(y - x) + o(\|y - x\|^2).$$

Jacobian matrix of a multivalued function $f : \mathbb{R}^n \to \mathbb{R}^m$,
$f(x) = [f_1(x), \dots, f_m(x)]$ is

$$\nabla f(x) = [\nabla f_1(x), \nabla f_2(x), \dots, \nabla f_m(x)] \in \mathbb{R}^{n \times m}.$$

Example. Find Jacobian matrix of $f(x_1, x_2) = [x_1^2, x_1 x_2, 2]^\top$.

Chain rule. Let $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^p$ be two differentiable mappings. Define $h(x) := g(f(x))$ (or $h = g \circ f$, h is a composition of g and f). Then we have

$$\nabla h(x) = \nabla f(x) \nabla g(f(x)).$$

(Optional reading)

Chain rule in general case

https://link.springer.com/content/pdf/10.1007/3-7643-7357-1_4.pdf

Chain rule. Let $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^p$ be two differentiable mappings. Define $h(x) := g(f(x))$ (or $h = g \circ f$, h is a composition of g and f). Then we have

$$\nabla h(x) = \nabla f(x) \nabla g(f(x)).$$

Example

- $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \log(1 + \exp(-x))$,
  $\nabla f(x) = ?$,
  $\nabla^2 f(x) = ?$.

Chain rule. Let $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^p$ be two differentiable mappings. Define $h(x) := g(f(x))$ (or $h = g \circ f$, h is a composition of g and f). Then we have

$$\nabla h(x) = \nabla f(x) \nabla g(f(x)).$$

Example

- $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \log(1 + \exp(-x))$,
  $\nabla f(x) =?$,
  $\nabla^2 f(x) =?$.

- Exercise Given a matrix $A \in \mathbb{R}^{m \times n}$ and a differentiable function $f : \mathbb{R}^m \to \mathbb{R}$. Let $h : \mathbb{R}^n \to \mathbb{R}$ be defined by $h(x) = f(Ax)$.
  $\nabla h(x) =?$ $\nabla^2 h(x) =?$

General optimization problem

$$
\begin{aligned}
\min_{x} \quad & f(x) \\
s.t. \quad & x \in \mathcal{X}.
\end{aligned}
$$

- Existence of a solution?
- Global or local minimum?
- Optimality conditions?
- Convex/nonconvex problem?
- Constrained/unconstrained problem?
- Continuous/discrete problem?

II. A brief introduction to convex optimization

References:

- Sebastien Bubeck, " Convex Optimization: Algorithms and Complexity", Foundations and Trends in Machine Learning, 8(3-4): 231-357, 2015.
- Stephen Boyd and Lieven Vandenberghe, "Convex Optimization". Web-page of the book: `https://stanford.edu/~boyd/cvxbook/`
- Boris Mordukhovich and Nguyen Mau Nam, "An Easy Path to Convex Analysis and Applications" (2013).

Convexity, subgradient, Fermat's optimality condition

Definition 1

A set $C \subset \mathbb{E}$ is convex if for any $x, y \in C$ and $\lambda \in [0, 1]$, the point $\lambda x + (1 - \lambda)y \in C$. Equivalently, C is convex if

$$[x, y] := \Big\{\lambda x + (1 - \lambda)y \,|\, \lambda \in [0, 1]\Big\} \subset C.$$

Example

- Let $a \in \mathbb{E}^* \setminus \{0\}$ and $\beta \in \mathbb{R}$. The following sets are convex:
  - (i) the hyperplane $H = \{x \in \mathbb{E} |\ \langle a, x\rangle = \beta\}$,
  - (ii) the half-space $H^- = \{x \in \mathbb{E} |\ \langle a, x\rangle \leq \beta\}$,
- Let $c \in \mathbb{E}$ and $\varepsilon > 0$. Let $\|\cdot\|$ be an arbitrary norm on $\mathbb{E}$. The closed ball

  $$\mathbb{B}_r(c) := \{x \in \mathbb{E} |\ \|x - c\| \leq \varepsilon\}$$

  is a convex set.

Algebraic Operations with convex sets

Proposition 2.1 (Intersections of convex sets)

Let $\{C_i\}_{i\in I}$ be a collection of convex sets in $\mathbb{E}$. Then $\bigcap_{i\in I} C_i$ is also a convex set.

Corollary. Let A be an $m \times n$ matrix and $b \in \mathbb{R}^m$. The polyhedral set $\{x \in \mathbb{R}^n | \; Ax \leq b\}$ is convex.

Optional reading

Proposition 2.2

Suppose that $\mathbb{E}$ and $\mathbb{F}$ be two Euclidean spaces.

(i) *Let A and B be two convex sets in $\mathbb{E}$. Then*

- $\lambda A := \{\lambda a |\ a \in A\}$ *is convex.*
- $A + B := \{a + b |\ a \in A, b \in B\}$ *is convex.*

(ii) *Let $A \subset \mathbb{E}$ and $B \subset \mathbb{F}$ be convex sets. Then $A \times B$ is convex on $\mathbb{E} \times \mathbb{F}$.*

(iii) *Let $A \subset \mathbb{E}$ be a convex set and $\Gamma : \mathbb{E} \to \mathbb{F}$ be an affine mapping. Then the image*

$$\Gamma(A) := \{\Gamma a |\ a \in A\}$$

is convex.

(iv) *Let $B \subset \mathbb{F}$ be a convex set and $\Gamma : \mathbb{E} \to \mathbb{F}$ be an affine mapping. Then the pre-image*

$$\Gamma^{-1}(A) := \{x \in \mathbb{E} |\ \Gamma x \in A\}$$

is convex.

Extended real-valued functions

That function can take value in $\mathbb{R} \cup \{-\infty, \infty\}$. Conventions: for $a \in \mathbb{R}$ we have

- $a + \infty = \infty + a = \infty$,
- $a - \infty = -\infty + a = \infty$,
- $a \cdot \infty = \infty \cdot a = \infty$ for $0 < a$,
- $a \cdot (-\infty) = (-\infty) \cdot a = -\infty$ for $0 < a$,
- $a \cdot \infty = \infty \cdot a = -\infty$ for $a < 0$,
- $a \cdot (-\infty) = (-\infty) \cdot a = \infty$ for $a < 0$,
- $0 \cdot \infty = \infty \cdot 0 = 0 \cdot (-\infty) = (-\infty) \cdot 0 = 0$.

For an extended real-valued function f, we define:

- $\mathrm{dom}(f) := \{\mathbf{x} \in \mathbb{E} : f(\mathbf{x}) < \infty\}$.
- $\mathrm{epi}(f) := \{(\mathbf{x}, t) : f(\mathbf{x}) \leq t, \mathbf{x} \in \mathbb{E}, t \in \mathbb{R}\}$.
- $\mathrm{Lev}(f, \alpha) := \{\mathbf{x} \in \mathbb{E} : f(\mathbf{x}) \leq \alpha\}$ for any $\alpha \in \mathbb{R}$.

Optional reading

Proper functions. f is called proper if f does not take the value $-\infty$ and $\mathrm{dom}(f)$ is nonempty.
Closed functions. A function $f : \mathbb{E} \to [-\infty, \infty]$ is closed if its epigraph is closed.
Lower semicontinuity. A function $f : \mathbb{E} \to [-\infty, \infty]$ is called lower semicontinuous at $\bar{\mathbf{x}} \in \mathbb{R}$ if for any sequence $\{\mathbf{x}_n\} \to \bar{\mathbf{x}}$ we have

$$f(\bar{\mathbf{x}}) \le \liminf_{n\to\infty} f(\mathbf{x}_n).$$

Or equivalently, for every $\alpha \in \mathbb{R}$ with $f(\bar{\mathbf{x}}) > \alpha$ there exists $\delta > 0$ such that

$$f(\mathbf{x}) > \alpha \quad \text{for all} \quad \mathbf{x} \in \mathbb{B}_\delta(\bar{\mathbf{x}}).$$

A function $f : \mathbb{E} \to [-\infty, \infty]$ is called lower semicontinuous if it is lower semicontinuous at each point in $\mathbb{E}$.

Theorem 2.1

Let f be an extended real-valued function. Then the following properties are equivalent:

(i) *f is lower semicontinuous.*

(ii) *f is closed.*

(iii) *For any $\alpha \in \mathbb{R}$, the α-level set of f is closed.*

Convex functions.

Definition 2

Let $f : \Omega \to \bar{\mathbb{R}}$ be an extended real-valued function defined on a convex set $\Omega \subset \mathbb{E}$. We say f is convex on Ω (or convex relative to Ω) if

$$f(tx + (1-t)y) \le tf(x) + (1-t)f(y) \quad \text{for all} \quad x, y \in \Omega, t \in [0,1].$$

If the inequality is strict for $x \neq y$ then f is strictly convex on Ω.

Proposition 2.3 (Convexity of epigraph for convex functions)

The extended real-valued function $f : \mathbb{E} \to \bar{\mathbb{R}}$ is convex if and only if its epigraph $\operatorname{epi}(f)$ is convex.

Convex functions.

Definition 2

Let $f : \Omega \to \bar{\mathbb{R}}$ be an extended real-valued function defined on a convex set $\Omega \subset \mathbb{E}$. We say f is convex on Ω (or convex relative to Ω) if

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \quad \text{for all} \quad x, y \in \Omega, t \in [0,1].$$

If the inequality is strict for $x \neq y$ then f is strictly convex on Ω.

Proposition 2.3 (Convexity of epigraph for convex functions)

The extended real-valued function $f : \mathbb{E} \to \bar{\mathbb{R}}$ is convex if and only if its epigraph $\mathrm{epi}\,(f)$ is convex.

- Affine function $f(x) = a^\top x + b$
- Every norm on $\mathbb{R}^n$ is convex.

Characterizations of differentiable convex functions.

Theorem 2.2 (Derivative tests)

Suppose f is a differentiable function on an open convex set $\Omega \subset \mathbb{R}^n$. f is convex on Ω if and only if one of the following conditions holds

(i) $\langle y - x, \nabla f(y) - \nabla f(x) \rangle \geq 0$ *for all* $x, y \in \Omega$.

(ii) $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$ *for any* $x, y \in \Omega$.

(iii) $\nabla^2 f(x)$ *is positive-semidefinite for all* x *in* Ω.

Example

- $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^p$.
- $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = \frac{1}{2}\alpha x^2 + \frac{1}{2}\beta y^2 - y$.

Example.

- $f : [\varepsilon, \infty) \to \mathbb{R}$, $f(x) = -log(x)$.

- $f : \mathbb{R} \to \mathbb{R}$, $f(x) = log(1 + \exp(-x))$.

- $f : [\varepsilon, \infty) \to \mathbb{R}$, $f(x) = 1/x$.

- Quadratic function $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = x^\top A x + x^\top b + c$.

- Consider $f : \mathbb{R} \to \bar{\mathbb{R}}$ defined by $f(x) = \begin{cases} +\infty, & \text{if } x < 0, \\ 1, & \text{if } x = 0, \\ x^2, & \text{if } x > 0. \end{cases}$

  Find $\operatorname{dom}(f)$. Is f convex?

- Exercise. Find all $a \in \mathbb{R}$ such that $f : \mathbb{R}^2 \to \bar{\mathbb{R}}$ defined by $f(x, y) = |xy| + a(x^2 + y^2)$ is convex.

Operations preserving convexity.

Proposition 2.4

(i) *(Linear operation) Let $f, g : \mathbb{E} \to \bar{\mathbb{R}}$ be convex function. Then αf and $f + g$ are convex for any $\alpha \geq 0$.*

(ii) *(Supremum operation) Let $f_i : \mathbb{E} \to \bar{\mathbb{R}}$, $i \in I$, be convex functions, where I is an arbitrary index set. Then $\sup_{i \in I} f_i(x)$ is convex.*

(iii) *(Linear change of variable) Let $f : \mathbb{F} \to \bar{\mathbb{R}}$ be a convex function, let $A : \mathbb{E} \to \mathbb{F}$ be a linear operator between two Euclidean spaces, and $b \in \mathbb{F}$. Then the function $g(x) := f(Ax + b)$ is convex on $\mathbb{E}$.*

(iv) *Let $f : \mathbb{E} \to \mathbb{R}$ be convex and let $g : \mathbb{R} \to \bar{\mathbb{R}}$ be non-decreasing and convex on a convex containing the range of the function f. Then the composition $g \circ f$ is convex.*

Example. Is $f : \mathbb{R}^2 \to \mathbb{R}, f(x, y) = (|x| + |y|)^2$ a convex function?

Optimization problem. Consider the following optimization problem

$$\min_{x\in\Omega} \quad f(x), \tag{1}$$

where $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is an extended real-valued function (cost function, objective function), and Ω is a nonempty, convex set.

Example.

- $\min_{x\in\mathbb{R}^n} \|Ax - b\|_2^2$ s.t. $x \geq 0$.

Definition 3 (Local/global minimizers)

Let $\bar{x} \in \mathbb{E}$, we say

- $\bar{x}$ is a local minimizer/optimal solution to Problem (1) if $f(\bar{x}) < \infty$ and there exists $\varepsilon > 0$ such that

$$f(x) \geq f(\bar{x}) \quad \text{for all} \quad x \in \mathbb{B}_\varepsilon(\bar{x}) \cap \Omega.$$

  In this case, $f(\bar{x})$ is called the local optimal value of f.

- $\bar{x}$ is a global/absolute minimizer/optimal solution to Problem (1) if $f(x) \geq f(\bar{x})$ for all $x \in \Omega$. In this case, $f(\bar{x})$ is called the optimal value of f.

Theorem 2.3 (Weierstrass existence theorem)

Let $f : \Omega \to \mathbb{R}$ be a continuous function, where Ω is a nonempty, compact subset of $\mathbb{R}^n$. Then the following optimization problem has global optimal solution

$$\min_{x\in\Omega} f(x) \quad \text{and} \quad \max_{x\in\Omega} f(x).$$

Example.

- Given $X \in \mathbb{R}^{m\times n}$, find

$$\min_{W\in\mathbb{R}^{m\times r}, H\in\mathbb{R}^{r\times n}} \frac{1}{2}\|X - WH\|^2 \quad \text{s.t} \quad W_{ij}, H_{ij} \in [0,1].$$

- Given training data (x_i, y_i), $y_i \in \{-1, 1\}$, $i = 1, \ldots, n$, find

$$\min_{w} \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-y_i\langle x_i, w\rangle)) + \frac{\lambda}{2}\|w\|_1.$$

Optional reading

Theorem 2.4

Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be a lower semicontinuous function (l.s.c). The following hold:

- *Then the optimization problem $\min_{x \in \Omega} f(x)$, where Ω is a nonempty compact subset of $\mathbb{R}^n$ that intersects $\operatorname{dom} f$, attains its absolute minimum.*
- *Assume that $\inf\{f(x), x \in \mathbb{R}^n\} < \infty$ and there exists $\alpha \in \mathbb{R}$ for which $\inf\{f(x), x \in \mathbb{R}^n\} < \alpha$ and the level set $\{x \in \mathbb{R}^n \big| f(x) < \alpha\}$ is bounded. Then the optimization problem $\min_{x \in \mathbb{R}^n} f(x)$ attains its absolute minimum at some point $\bar{x} \in \operatorname{dom} f$.*

Theorem 2.5

Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be convex and assume that $\inf\{f(x), x \in \mathbb{R}^n\} < \infty$. The following properties are equivalent.

- *There exists $\alpha \in \mathbb{R}$ such that $\inf\{f(x), x \in \mathbb{R}^n\} < \alpha$ and the level set $\{x \in \mathbb{R}^n \big| f(x) < \alpha\}$ is bounded.*
- *All the level sets $\{f(x), x \in \mathbb{R}^n\} < \alpha$ of f are bounded.*
- $\lim_{\|x\| \to \infty} f(x) = \infty$.
- $\liminf_{\|x\| \to \infty} \frac{f(x)}{\|x\|} > 0$.

Theorem 2.6

If f is a convex function, a local minimizer of f is also a global minimizer.

Proof.
Suppose $\bar{x}$ is a local minimizer of f. We have $f(\bar{x}) \leq f(x)$ for all $x \in B_\varepsilon(\bar{x})$ for some ϵ. For all y, let us define $y^k = \frac{1}{k}y + (1 - \frac{1}{k})\bar{x}$. Then we have

$$y^k - \bar{x} = \frac{1}{k}(y - \bar{x}).$$

Hence, when k is big enough, we have $y^k \in B_\varepsilon(\bar{x})$. By convexity of f, we have

$$f(\bar{x}) \leq f(y^k) \leq \frac{1}{k}f(y) + (1 - \frac{1}{k})f(\bar{x}).$$

This implies $f(\bar{x}) \leq f(y)$ for all y.

Subdifferential of a convex function.

Definition 4

Let $f : \mathbb{E} \to \overline{\mathbb{R}}$ be a proper convex function (not necessarily differentiable) and $\bar{\mathbf{x}} \in \mathrm{dom} f$. A vector $v \in \mathbb{E}^*$ is called a subgradient of f at $\bar{\mathbf{x}}$ if

$$f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \langle v, \mathbf{x} - \bar{\mathbf{x}} \rangle \quad \text{for all} \quad \mathbf{x} \in \mathbb{E}.$$

The subdifferential of f at $\bar{\mathbf{x}}$ is defined by

$$\partial f(\bar{\mathbf{x}}) := \{ v \in \mathbb{E}^* | \ f(\mathbf{x}) \geq f(\bar{\mathbf{x}}) + \langle v, \mathbf{x} - \bar{\mathbf{x}} \rangle \, \forall \mathbf{x} \in \mathbb{E} \}.$$

Proposition 2.5

Let $f : \mathbb{E} \to \bar{\mathbb{R}}$ be a proper, l.s.c. convex function.

- *If f is differentiable at $\bar{x}$ then $\partial f(\bar{x}) = \nabla f(\bar{x})$.*
- **(Optional reading)** *Suppose that $\bar{x} \in \mathrm{int}\,(\mathrm{dom}\, f)$, i.e., f is continuous at $\bar{x}$. Then $\partial f(\bar{x})$ is nonempty and is a compact convex set.*

Example.

- $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$. Find $\partial f(x)$.

Example.

- $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$. Find $\partial f(x)$.

- Exercise. Let $f(x) = \max\{f_1(x), f_2(x)\}$, where $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ are differentiable convex functions. Prove that
  - If $f_1(x) > f_2(x)$, f has unique subgradient $v = \nabla f_1(x)$
  - If $f_2(x) > f_1(x)$, f has unique subgradient $v = \nabla f_2(x)$.
  - If $f_1(x) = f_2(x)$, then any point on the line segment between $\nabla f_1(x)$ and $\nabla f_2(x)$ is a subgradient of f at x.

Example.

- $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$. Find $\partial f(x)$.

- Exercise. Let $f(x) = \max\{f_1(x), f_2(x)\}$, where $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ are differentiable convex functions. Prove that
    - If $f_1(x) > f_2(x)$, f has unique subgradient $v = \nabla f_1(x)$
    - If $f_2(x) > f_1(x)$, f has unique subgradient $v = \nabla f_2(x)$.
    - If $f_1(x) = f_2(x)$, then any point on the line segment between $\nabla f_1(x)$ and $\nabla f_2(x)$ is a subgradient of f at x.

- Let $f_i(w) = \max(0, -y_i w^\top x_i)$. Find $\partial f_i(w)$.

Theorem 2.7 (Fermat's optimality condition)

Let $f \in \mathbb{E} \to \overline{\mathbb{R}}$ *be a proper convex function. Then* $\mathbf{x}^*$ *is a minimizer to f if and only if* $0 \in \partial f(\mathbf{x}^*)$.

Proof.

Theorem 2.7 (Fermat's optimality condition)

Let $f \in \mathbb{E} \to \overline{\mathbb{R}}$ be a proper convex function. Then $\mathbf{x}^*$ is a minimizer to f if and only if $0 \in \partial f(\mathbf{x}^*)$.

Proof.

Example. Given $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, find

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax - b\|_2^2.$$

Basic subgradient calculus rules.

Theorem 2.8 (Sum rule I)

Let $f, g : \mathbb{E} \to \bar{\mathbb{R}}$ be proper convex functions. Suppose that f is differentiable at $\bar{x} \in \operatorname{dom} g$. Then we have

$$\partial(f + g)(\bar{x}) = \nabla f(\bar{x}) + \partial g(\bar{x}).$$

Basic subgradient calculus rules.

Theorem 2.8 (Sum rule I)

Let $f, g : \mathbb{E} \to \bar{\mathbb{R}}$ *be proper convex functions. Suppose that* f *is differentiable at* $\bar{x} \in \operatorname{dom} g$. *Then we have*

$$\partial(f+g)(\bar{x}) = \nabla f(\bar{x}) + \partial g(\bar{x}).$$

(**Optional reading**)

Theorem 2.9 (Sum rule II)

Let $f, g : \mathbb{E} \to \bar{\mathbb{R}}$ *be proper convex functions. Suppose that* $\operatorname{dom} f \cap \operatorname{int}(\operatorname{dom} g) \neq \emptyset$. *Then we have the sume rule*

$$\partial(f+g)(x) = \partial f(x) + \partial g(x) \quad \textit{for any} \quad x \in \operatorname{dom} f \cap \operatorname{dom} g.$$

Lagrangian Duality

Primal optimization problem

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ s.t. \quad & g_i(x) \leq 0, i = 1, \ldots, m \\ & h_i(x) = 0, i = 1, \ldots, q. \end{aligned}$$

The Lagrangian function is the function $L : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$

$$L(x, \lambda, \gamma) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{i=1}^{q} \gamma_i h_i(x).$$

Primal optimization problem

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ s.t. \quad & g_i(x) \leq 0, i = 1, \ldots, m \\ & h_i(x) = 0, i = 1, \ldots, q. \end{aligned}$$

The Lagrangian function is the function $L : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$

$$L(x, \lambda, \gamma) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{i=1}^{q} \gamma_i h_i(x).$$

Note that

$$\max_{\lambda \geq 0, \gamma} L(x, \lambda, \gamma) = \begin{cases} f(x) & \text{if } x \text{ is a feasible solution of the primal problem} \\ +\infty & \text{otherwise.} \end{cases}$$

Primal optimization problem

$$
\begin{aligned}
\min_{x \in \mathbb{R}^n} \quad & f(x) \\
s.t. \quad & g_i(x) \leq 0, i = 1, \ldots, m \\
& h_i(x) = 0, i = 1, \ldots, q.
\end{aligned}
$$

The Lagrangian function is the function $L : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$

$$
L(x, \lambda, \gamma) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{i=1}^{q} \gamma_i h_i(x).
$$

Note that

$$
\max_{\lambda \geq 0, \gamma} L(x, \lambda, \gamma) = \begin{cases} f(x) & \text{if } x \text{ is a feasible solution of the primal problem} \\ +\infty & \text{otherwise.} \end{cases}
$$

Hence, solving the primal problem is equivalent to solving

$$
\min_{x} \max_{\lambda \geq 0, \gamma} L(x, \lambda, \gamma)
$$

Dual problem

$$\max_{\lambda \geq 0, \gamma} \min_{x} L(x, \lambda, \gamma) = \max_{\lambda \geq 0, \gamma} \rho(\lambda, \gamma),$$

where $\rho(\lambda, \gamma) = \min_x L(x, \lambda, \gamma)$.

Dual problem

$$\max_{\lambda \geq 0, \gamma} \min_x L(x, \lambda, \gamma) = \max_{\lambda \geq 0, \gamma} \rho(\lambda, \gamma),$$

where $\rho(\lambda, \gamma) = \min_x L(x, \lambda, \gamma)$.

Weak duality Let p^* be the optimal value of the primal problem and q^* be the optimal value of the dual problem. It always holds that $p^* \geq d^*$, that is

$$\min_x \max_{\lambda \geq 0, \gamma} L(x, \lambda, \gamma) \geq \max_{\lambda \geq 0, \gamma} \min_x L(x, \lambda, \gamma).$$

Dual problem

$$\max_{\lambda \geq 0, \gamma} \min_{x} L(x, \lambda, \gamma) = \max_{\lambda \geq 0, \gamma} \rho(\lambda, \gamma),$$

where $\rho(\lambda, \gamma) = \min_x L(x, \lambda, \gamma)$.

Weak duality Let p^* be the optimal value of the primal problem and q^* be the optimal value of the dual problem. It always holds that $p^* \geq d^*$, that is

$$\min_{x} \max_{\lambda \geq 0, \gamma} L(x, \lambda, \gamma) \geq \max_{\lambda \geq 0, \gamma} \min_{x} L(x, \lambda, \gamma).$$

- For some certain problems, we have strong duality $p^* = d^*$.
- Conditions that guarantee strong duality in convex problems are called constraint qualifications

(**Optional reading**)

Standard convex optimization problem

$$\begin{aligned} \text{minimize}_{x \in \mathbb{R}^n} \quad & f(x) \\ s.t. \quad & g_i(x) \leq 0, i = 1, \ldots, m \\ & Ax = b, \end{aligned}$$

where f and g_i are convex functions. Strong duality holds for the standard convex optimization problem if there exists a point that is strictly feasible (this is called Slater's condition):

$$\exists x \in \operatorname{rel\,int} \cap_{i=1}^{m} \operatorname{dom} g_i, \text{ such that } g_i(x) < 0, i = 1, \ldots, m, Ax = b.$$

Soft-margin Support Vector Machine

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y_i(w^\top x_i)\} + R(w).$$