Chapter 21
Mathematical Background

In this chapter, we gather some key mathematical concepts for better understanding this book. This is not intended to be an introductory tutorial, and it is assumed that the reader already has some background on probability theory, linear algebra, and optimization.

When writing this chapter, we have referred to [1–4] to a large extent. Note that we will not add explicit citations in the remaining part of this chapter. The readers are highly encouraged to read the aforementioned material.

21.1 Probability Theory

Probability theory plays a key role in machine learning and information retrieval, since the design of learning methods and ranking models often relies on the probability assumption on the data.

21.1.1 Probability Space and Random Variables

When we talk about probability, we often refer to the probability of an uncertain event. Therefore, in order to discuss probability theory formally, we must first clarify what the possible events are to which we would like to attach a probability.

Formally, a probability space is defined by the triple $(\Omega, \mathscr{F}, P)$, where Ω is the space of possible outcomes, $\mathscr{F} \subseteq 2^{\Omega}$ is the space of (measurable) events; and P is the probability measure (or probability distribution) that maps an event $E \in \mathscr{F}$ to a real value between 0 and 1.

Given the outcome space Ω, there are some restrictions on the event space $\mathscr{F}$:

- The trivial event Ω and the empty event $\emptyset$ are all in $\mathscr{F}$;
- The event space $\mathscr{F}$ is closed under (countable) union, i.e., if $E_1 \in \mathscr{F}$ and $E_2 \in \mathscr{F}$, then $E_1 \cup E_2 \in \mathscr{F}$;

DOI 10.1007/978-3-642-14267-3_21, © Springer-Verlag Berlin Heidelberg 2011

- The event space $\mathscr{F}$ is closed under complement, i.e., if $E \in \mathscr{F}$, then $\Omega \backslash E \in \mathscr{F}$.

Given an event space $\mathscr{F}$, the probability measure P must satisfy certain axioms.

- For all $E \in \mathscr{F}$, $P(E) \geq 0$.
- $P(\Omega) = 1$.
- For all $E_1, E_2 \in \mathscr{F}$, if $E_1 \cap E_2 = \emptyset$, $P(E_1 \cup E_2) = P(E_1) + P(E_2)$.

Random variables play an important role in probability theory. Random variables allow us to abstract away from the formal notion of event space, because we can define random variables that capture the appropriate events that we are interested in. More specifically, we can regard the random variable as a function, which maps the event in the outcome space to real values. For example, suppose the event is "it is sunny today". We can use a random variable X to map this event to value 1. And there is a probability associated with this mapping. That is, we can use $P(X = 1)$ to represent the probability that it is really sunny today.

21.1.2 Probability Distributions

Since random variables can take different values (or more specifically map the event to different real values) with different probabilities, we can use a probability distribution to describe it. Usually we also refer to it as the probability distribution of the random variable for simplicity.

21.1.2.1 Discrete Distribution

First, we take the discrete case as an example to introduce some key concepts related to probability distribution.

In the discrete case, the probability distribution specifies the probability for a random variable to take any possible values. It is clear that $\sum_a P(X = a) = 1$.

If we have multiple random variables, we will have the concept of joint distribution, marginal distribution, and conditional distribution. Joint distribution is something like $P(X = a, Y = b)$. Marginal distribution is the probability distribution of a random variable on its own. Its relationship with joint distribution is

$$P(X = a) = \sum_b P(X = a, Y = b).$$

Conditional distribution specifies the distribution of a random variable when the value of another random variable is known (or given). Formally conditional probability of $X = a$ given $Y = b$ is defined as

$$P(X = a | Y = b) = \frac{P(X = a, Y = b)}{P(Y = b)}.$$

With the concept of conditional probability, we can introduce another concept, named independence. Here independence means that the distribution of a random variable does not change when given the value of another random variable. In machine learning, we often make such assumptions. For example, we say that the data samples are independently and identically distributed.

According to the above definition, we can clearly see that if two random variables X and Y are independent, then

$$P(X, Y) = P(X)P(Y).$$

Sometimes, we will also use conditional independence, which means that if we know the value of a random variable, then some other random variables will become independent of each other. That is,

$$P(X, Y|Z) = P(X|Z)P(Y|Z).$$

There are two basic rules in probability theory, which are widely used in various settings. The first is the Chain Rule. It can be represented as follows:

$$P(X_1, X_2, \dots, X_n) = P(X_1)P(X_2|X_1)\cdots P(X_n|X_1, X_2, \dots, X_{n-1}).$$

The chain rule provides a way of calculating the joint probability of some random variables, which is especially useful when there is independence across some of the variables.

The second rule is the Bayes Rule, which allows us to compute the conditional probability $P(X|Y)$ from another conditional probability $P(Y|X)$. Intuitively, it actually inverses the cause and result. The Bayes Rule takes the following form:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} = \frac{P(Y|X)P(X)}{\sum_a P(Y|X=a)P(X=a)}.$$

21.1.2.2 Continuous Distribution

Now we generalize the above discussions to the continuous case. This time, we need to introduce the concept of the probability density function (PDF). A probability density function, p, is a non-negative, integrable function such that

$$\int p(x)\,dx = 1.$$

The probability of a random variable distributed according to a PDF f is computed as follows:

$$P(a \le x \le b) = \int_a^b p(x)\,dx.$$

As for the probability density function, we have similar results to those for probabilities. For example, we have

$$p(y|x) = \frac{p(x, y)}{p(x)}.$$

21.1.2.3 Popular Distributions

There are many well-studied probability distributions. In this subsection, we list a few of them for example.

- Bernoulii distribution: $P(X = x) = p^x(1 - p)^{1-x}, x = 0$ or 1.
- Poisson distribution: $P(X = k) = \frac{\exp(-\lambda)\lambda^k}{k!}$.
- Gaussion distribution: $p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$.
- Exponential distribution: $p(x) = \lambda \exp(-\lambda x), x \geq 0$.

21.1.3 Expectations and Variances

Expectations and variances are widely used concepts in probability theory. Expectation is also referred to as mean, expected value, or first moment. The expectation of a random variable, denoted as $E[X]$, is defined by

$$E[X] = \sum_a aP(X = a), \quad \text{for the discrete case;}$$

$$E[X] = \int_{-\infty}^{\infty} xp(x)\,dx, \quad \text{for the continuous case.}$$

When the random variable X is an indicator variable (i.e., it takes a value from $\{0, 1\}$), we have the following result:

$$E[X] = P(X = 1).$$

The expectation of a random variable has the following three properties.

- $E[X_1 + X_2 + \cdots + X_n] = E[X_1] + E[X_2] + \cdots + E[X_n]$.
- $E[XY] = E[X]E[Y]$, if X and Y are independent.
- If f is a convex function, $f(E[X]) \leq E[f(x)]$.

The variance of a distribution is a measure of the spread of it. It is also referred to as the second moment. The variance is defined by

$$\mathrm{Var}(X) = E\big[\big(X - E[X]\big)^2\big].$$

The variance of a random variable is often denoted as σ^2. And σ is called the standard deviation.

The variance of a random variable has the following properties.

- $\mathrm{Var}(X) = E[X^2] - (E[X])^2$.
- $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X)\mathrm{Var}(Y)$, if X and Y are independent.

21.2 Linear Algebra and Matrix Computation

Linear algebra is an important branch of mathematics concerned with the study of vectors, vector spaces (or linear spaces), and functions with vector input and output. Matrix computation is the study of algorithms for performing linear algebra computations, including the study of matrix properties and matrix decompositions. Linear algebra and matrix computation are useful tools in machine learning and information retrieval.

21.2.1 Notations

We start with introducing the following notations:

- $\mathscr{R}^n$: the n-dimensional space of real numbers.
- $A \in \mathscr{R}^{m\times n}$: A is a matrix with m rows and n columns, with real number elements.
- $x \in \mathscr{R}^n$: x is a vector with n elements of real numbers. By default, x denotes a column vector, which is a matrix with n rows and one column. To denote a row vector (or a matrix with one row and n columns), we use the transpose of x, which is written as x^T.
- $x_i \in \mathscr{R}$: the ith element of a vector x.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^T.$$

- $a_{ij} \in \mathscr{R}$: the element in the ith row and jth column of a matrix A, which is also written by A_{ij}, $A_{i,j}$, etc.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

- a_j or $A_{:,j}$: the jth column of A.

$$A = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}.$$

- $\bar{a}_i^T$ or $A_{i,:}$: the ith row of A.

$$A = \begin{bmatrix} \bar{a}_1^T \\ \bar{a}_2^T \\ \vdots \\ \bar{a}_m^T \end{bmatrix}.$$

21.2.2 Basic Matrix Operations and Properties

21.2.2.1 Matrix Multiplication

Given two matrices $A \in \mathscr{R}^{m \times n}$ and $B \in \mathscr{R}^{n \times l}$, the matrix multiplication (product) of them is defined by

$$C = AB \in \mathscr{R}^{m \times l},$$

in which

$$C_{ik} = \sum_{j=1}^{n} A_{ij} B_{jk}.$$

Note that the number of columns of A and the number of rows of B should be the same in order to ensure the implementation of matrix multiplication.

The properties of matrix multiplication are listed as below:

- Associative: $(AB)C = A(BC)$.
- Distributive: $A(B + C) = AB + AC$.
- Generally not commutative: $AB \neq BA$.

Given two vectors $x, y \in \mathscr{R}^n$, the *inner product* (or *dot product*) of them is a real number defined as

$$x^T y = \sum_{i=1}^{n} x_i y_i = y^T x \in \mathscr{R}.$$

Given two vectors $x \in \mathscr{R}^m$, $y \in \mathscr{R}^n$, the *outer product* of them is a matrix defined as

$$xy^T = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m y_1 & x_m y_2 & \cdots & x_m y_n \end{bmatrix} \in \mathscr{R}^{m \times n}.$$

Given a matrix $A \in \mathscr{R}^{m\times n}$ and a vector $x \in \mathscr{R}^n$, their product is a vector given by $y = Ax \in \mathscr{R}^m$. The product can be expressed in multiple forms as below.

$$y = \begin{bmatrix} \bar{a}_1^T \\ \bar{a}_2^T \\ \vdots \\ \bar{a}_m^T \end{bmatrix} x = \begin{bmatrix} \bar{a}_1^T x \\ \bar{a}_2^T x \\ \vdots \\ \bar{a}_m^T x \end{bmatrix} = Ax = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$= x_1 a_1 + x_2 a_2 + \cdots + x_n a_n.$$

21.2.2.2 Identity Matrix

The *identity matrix* $I \in \mathscr{R}^{n\times n}$ is a square matrix with its diagonal elements equal to 1 and the others equal to 0.

$$I_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

Given a matrix $A \in \mathscr{R}^{m\times n}$, we have the following property.

$$AI = A = IA.$$

21.2.2.3 Diagonal Matrix

A *diagonal matrix* $D \in \mathscr{R}^{n\times n}$ is a square matrix with all its non-diagonal elements equal to 0, i.e.,

$$D = \mathrm{diag}(d_1, d_2, \ldots, d_n) \quad \text{or} \quad D_{ij} = \begin{cases} d_i, & i = j, \\ 0, & i \neq j. \end{cases}$$

21.2.2.4 Transpose

Given a matrix $A \in \mathscr{R}^{m\times n}$, the *transpose* of A (denoted by A^T) is obtained by swapping the rows and columns, i.e.,

$$A^T \in \mathscr{R}^{n\times m} \quad \text{where } \left(A^T\right)_{ij} = A_{ji}.$$

The properties of transpose are listed as below:

- $(A^T)^T = A$
- $(A + B)^T = A^T + B^T$
- $(AB)^T = B^T A^T$

21.2.2.5 Symmetric Matrix

If a square matrix $A \in \mathscr{R}^{n\times n}$ holds $A^T = A$, it is called a *symmetric matrix*; if it holds $A^T = -A$, it is called an *anti-symmetric matrix*. Given $A \in \mathscr{R}^{n\times n}$, it is not difficult to verify that $A + A^T$ is symmetric and $A - A^T$ is anti-symmetric. Therefore, any square matrix can be written as the sum of a symmetric matrix and an anti-symmetric matrix, i.e.,

$$A = \frac{1}{2}\left(A + A^T\right) + \frac{1}{2}\left(A - A^T\right).$$

21.2.2.6 Trace

The sum of diagonal elements in a square matrix $A \in \mathscr{R}^{n\times n}$ is called the trace of the matrix, denoted by $\operatorname{tr} A$:

$$\operatorname{tr} A = \sum_{i=1}^{n} A_{ii}.$$

The properties of trace are listed as below:

- Given $A \in \mathscr{R}^{n\times n}$, we have $\operatorname{tr} A = \operatorname{tr} A^T$.
- Given $A \in \mathscr{R}^{n\times n}$ and $\alpha \in \mathscr{R}$, we have $\operatorname{tr}(\alpha A) = \alpha \operatorname{tr} A$.
- Given $A, B \in \mathscr{R}^{n\times n}$, we have $\operatorname{tr}(A + B) = \operatorname{tr} A + \operatorname{tr} B$.
- Given A and B such that AB is a square matrix, we have $\operatorname{tr}(AB) = \operatorname{tr}(BA)$.
- Given matrices $A_1, A_2, \ldots, A_k$ such that $A_1 A_2 \cdots A_k$ is a square matrix, we have $\operatorname{tr} A_1 A_2 \cdots A_k = \operatorname{tr} A_2 A_3 \cdots A_k A_1 = \operatorname{tr} A_3 A_4 \cdots A_k A_1 A_2 = \cdots = \operatorname{tr} A_k A_1 \cdots A_{k-1}$.

21.2.2.7 Norm

A *norm* of a vector is a function $f : \mathscr{R}^n \to \mathscr{R}$ that respects the following four conditions:

- non-negativity: $f(x) \geq 0, \forall x \in \mathscr{R}^n$
- definiteness: $f(x) = 0$ if and only if $x = 0$
- homogeneity: $f(\alpha x) = |\alpha| f(x), \forall x \in \mathscr{R}^n$ and $\alpha \in \mathscr{R}$
- triangle inequality: $f(x) + f(y) \geq f(x + y), \forall x, y \in \mathscr{R}^n$

Some commonly-used norms for vector $x \in \mathscr{R}^n$ are listed as below:

- L_1 norm: $\|x\|_1 = \sum_{i=1}^n |x_i|$
- L_2 norm (or Euclidean norm): $\|x\|_2 = (\sum_{i=1}^n x_i^2)^{\frac{1}{2}}$
- L_∞ norm: $\|x\|_\infty = \max_i |x_i|$

The above three norms are members of the L_p norm family, i.e., for $p \in \mathscr{R}$ and $p \geq 1$,

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}.$$

One can also define norms for matrices. For example, the following norm is the commonly-used Frobenius norm. For a matrix $A \in \mathscr{R}^{m \times n}$,

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |A_{ij}|} = \sqrt{\operatorname{tr}\left(A^T A\right)}.$$

21.2.2.8 Inverse

A square matrix $A \in \mathscr{R}^{n \times n}$ is called *invertible* or *nonsingular* if there exists a square matrix $B \in \mathscr{R}^{n \times n}$ such that

$$AB = I = BA.$$

In this case, B is uniquely determined by A and is referred to as the *inverse* of A, denoted by A^{-1}. If such a kind of B (or A^{-1}) does not exist, we call A a *non-invertible* or *singular* matrix.

The properties of inverse are listed as below:

- given $A \in \mathscr{R}^{n \times n}$, we have $(A^{-1})^{-1} = A$
- given $A \in \mathscr{R}^{n \times n}$, we have $(A^{-1})^T = (A^T)^{-1} = A^{-T}$
- given $A, B \in \mathscr{R}^{n \times n}$, we have $(AB)^{-1} = B^{-1}A^{-1}$

21.2.2.9 Orthogonal Matrix

Given two vectors $x, y \in \mathscr{R}^n$, they are *orthogonal* if $x^T y = 0$.

A square matrix $A \in \mathscr{R}^{n \times n}$ is an *orthogonal matrix* if its columns are orthogonal unit vectors.

$$A^T A = I = AA^T.$$

The properties of orthogonal matrix are listed as below:

- the inverse of an orthogonal matrix equals to its transpose, i.e., $A^{-1} = A^T$
- $\|Ax\|_2 = \|x\|_2$ for any $x \in \mathscr{R}^n$ and orthogonal $A \in \mathscr{R}^{n \times n}$

21.2.2.10 Determinant

The *determinant* of a square matrix $A \in \mathscr{R}^{n \times n}$ is a function $\det : \mathscr{R}^{n \times n} \to \mathscr{R}^n$, denoted by $\det A$ or $|A|$, i.e.,

$$\det A = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} |A_{\backslash i, \backslash j}| \quad (\forall j = 1, 2, \ldots, n)$$

or

$$\det A = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} |A_{\backslash i, \backslash j}| \quad (\forall i = 1, 2, \ldots, n).$$

Here $A_{\backslash i, \backslash j} \in \mathscr{R}^{(n-1) \times (n-1)}$ is the matrix obtained by deleting the ith row and the jth column from A. Note that the above definition is recursive and thus we need to define $|A| = a_{11}$ for $A \in \mathscr{R}^{1 \times 1}$.

Given A, $B \in \mathscr{R}^{n \times n}$, we have the following properties for the determinant:

- $|A| = |A^T|$
- $|AB| = |A||B|$
- $|A| = 0$ if and only if A is singular
- $|A|^{-1} = \frac{1}{|A|}$ if A is nonsingular

21.2.2.11 Quadratic Form

Given a vector $x \in \mathscr{R}^n$ and a square matrix $A \in \mathscr{R}^{n \times n}$, we call the following scalar a *quadratic form*:

$$x^T A x = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j.$$

As we have

$$x^T A x = \left(x^T A x\right)^T = x^T A^T x = x^T \left(\frac{1}{2} A + \frac{1}{2} A^T\right) x,$$

we may as well regard the matrix in a quadratic form to be symmetric.

Given a vector $x \in \mathscr{R}^n$ and a symmetric matrix $A \in \mathscr{R}^{n \times n}$, we have the following definitions:

- If $x^T A x > 0$, $\forall x$, A is *positive definite*, denoted by $A \succ 0$.
- If $x^T A x \geq 0$, $\forall x$, A is *positive semidefinite*, denoted by $A \succeq 0$.
- If $x^T A x < 0$, $\forall x$, A is *negative definite*, denoted by $A \prec 0$.
- If $x^T A x \leq 0$, $\forall x$, A is *negative semidefinite*, denoted by $A \preceq 0$.
- Otherwise, A is *indefinite*.

21.2.3 Eigenvalues and Eigenvectors

Given a square matrix $A \in \mathscr{R}^{n\times n}$, a non-zero vector $x \in \mathscr{R}^n$ is defined as an *eigenvector* of the matrix if it satisfies the eigenvalue equation

$$Ax = \lambda x \quad (x \neq 0)$$

for some scalar $\lambda \in \mathscr{R}$. In this situation, the scalar λ is called an *eigenvalue* of A corresponding to the eigenvector x. It is easy to verify that αx is also an eigenvector of A for any $\alpha \in \mathscr{R}$.

To compute the eigenvalues of matrix A, we have to solve the following equation,

$$(\lambda I - A)x = 0 \quad (x \neq 0),$$

which has a non-zero solution if and only if $(\lambda I - A)$ is singular, i.e.,

$$|\lambda I - A| = 0.$$

Theoretically, by solving the above *eigenpolynomial*, we can obtain all the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. To further compute the corresponding eigenvectors, we may solve the following linear equation,

$$(\lambda_i I - A)x = 0 \quad (x \neq 0).$$

Given a square matrix $A \in \mathscr{R}^{n\times n}$, its eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ and the corresponding eigenvectors $x_1, x_2, \ldots, x_n$, we have the following properties:

- $\operatorname{tr} A = \sum_{i=1}^{n} \lambda_i$
- $\det A = \prod_{i=1}^{n} \lambda_i$
- The eigenvalues of a diagonal matrix $D = \operatorname{diag}(d_1, d_2, \ldots, d_n)$ are exactly the diagonal elements $d_1, d_2, \ldots, d_n$
- If A is nonsingular, then the eigenvalues of A^{-1} are $1/\lambda_i$ $(i = 1, 2, \ldots, n)$ with their corresponding eigenvectors x_i $(i = 1, 2, \ldots, n)$.

If we denote

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}, \qquad \Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n),$$

we will have the matrix form of the eigenvalue equation

$$AX = X\Lambda.$$

If the eigenvectors $x_1, x_2, \ldots, x_n$ are linearly independent, then X is nonsingular and we have

$$A = X\Lambda X^{-1}.$$

In this situation, A is called *diagonalizable*.

21.3 Convex Optimization

When we solve a machine learning problem, it is almost unavoidable that we will use some optimization techniques. In this section, we will introduce some basic concepts and representative algorithms for convex optimization.

21.3.1 Convex Set and Convex Function

A set $C \subseteq \mathscr{R}^n$ is a *convex set* if $\forall x, y \in C$ and $\forall \theta \in [0, 1]$, there holds

$$\theta x + (1 - \theta) y \in C,$$

i.e., the line segments between any pairs of x and y lies in C.

A function $f : \mathscr{R}^n \to \mathscr{R}$ is a *convex function* if the domain of f (denoted by $\mathrm{dom}\, f$) is a convex set and there holds

$$f\big(\theta x + (1 - \theta) y\big) \le \theta f(x) + (1 - \theta) f(y)$$

$\forall x, y \in \mathrm{dom}\, f$ and $\theta \in [0, 1]$.

Note that (i) function f is *concave* if $-f$ is convex; (ii) function f is *strictly convex* if $\mathrm{dom}\, f$ is convex and there holds $f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta) f(y)$, $\forall x, y \in \mathrm{dom}\, f$, $x \neq y$, and $\theta \in [0, 1]$; (iii) from the definition of a convex function one can conclude that *any local minimum is a global minimum for convex functions*.

Here are some examples of convex functions:

- affine: $ax + b$ on $x \in \mathscr{R}$, $\forall a, b \in \mathscr{R}$
- exponential: e^{ax} on $x \in \mathscr{R}$, $\forall a \in \mathscr{R}$
- powers: x^α on $x \in (0, +\infty)$, for $\alpha \in (-\infty, 0] \cup [1, +\infty]$
- powers of absolute value: $|x|^p$ on $x \in \mathscr{R}$, $\forall p \in [1, +\infty]$
- negative logarithm: $-\log x$ on $x \in (0, +\infty)$
- negative entropy: $x \log x$ on $x \in (0, +\infty)$
- affine function: $a^T x + b$ on $x \in \mathscr{R}^n$, $\forall a, b \in \mathscr{R}^n$
- norms: $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$ on $x \in \mathscr{R}^n$ for $p \in [1, +\infty]$; $\|x\|_\infty = \max_i |x_i|$ on $x \in \mathscr{R}^n$
- quadratic: $x^2 + bx + c$ on $x \in \mathscr{R}$, $\forall b, c \in \mathscr{R}$

Here are some examples of *concave* functions:

- affine: $ax + b$ on $x \in \mathscr{R}$, $\forall a, b \in \mathscr{R}$
- powers: x^α on $x \in (0, +\infty)$, for $\alpha \in [0, 1]$
- logarithm: $\log x$ on $x \in (0, +\infty)$

21.3.2 Conditions for Convexity

21.3.2.1 First-Order Condition

Function f is *differentiable* if dom f is open and the gradient $\nabla f(x) = (\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_n})$ exists $\forall x \in$ dom f. Differentiable f is convex if and only if dom f is convex and there holds

$$f(y) \geq f(x) + \nabla f(x)^T(y - x)$$

$\forall x, y \in$ dom f. This is the *first-order condition for convexity.*

21.3.2.2 Second-Order Condition

Function f is *twice differentiable* if dom f is open and the Hessian $\nabla^2 f(x)$ (which is an n-by-n symmetric matrix with $\nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, i, j = 1, 2, \ldots, n$) exists $\forall x \in$ dom f. Twice differentiable f is convex if and only if dom f is convex and Hessian $\nabla^2 f(x)$ is positive semidefinite (i.e., $\nabla^2 f(x) \succeq 0$), $\forall x \in$ dom f. This is the *second-order condition for convexity.*

21.3.3 Convex Optimization Problem

An optimization problem is convex if the objective function is convex and the feasible set is also convex. That is, a *convex optimization problem* is written as

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

where f and $g_i, i = 1, \ldots, m$ are convex.

Note that any local optimum of a convex optimization problem is globally optimal. Here are some examples of convex optimization problem.

Linear Programming (LP) A linear programming is a convex optimization problem with affine objective and constraint functions. Its feasible set is a polyhedron.

$$\begin{aligned} &\min_{x \in \mathscr{R}^n} \alpha^T x + \beta, \\ \text{s.t.} \quad & g_i(x) \leq h_i, \quad i = 1, \ldots, m, \\ & a_i^T x = b_i, \quad i = 1, \ldots, l, \end{aligned}$$

Quadratic Programming (QP) A quadratic programming is a convex optimization problem with a convex quadratic function as its objective function and with affine constraint functions.

$$\begin{aligned} &\min_{x\in\mathscr{R}^n} \frac{1}{2}x^T Qx + p^T x + r, \\ \text{s.t.} \quad & g_i(x) \le h_i, \quad i = 1, \dots, m, \\ & a_i^T x = b_i, \quad i = 1, \dots, l, \end{aligned}$$

where Q is a symmetric and positive semidefinite matrix.

Semidefinite Programming (SDP) A semidefinite programming is a convex optimization problem with a linear objective function optimized over the intersection of the cone of a group of positive semidefinite matrices with an affine space.

$$\begin{aligned} &\min_{x\in\mathscr{R}^n} \alpha^T x, \\ \text{s.t.} \quad & x_1 K_1 + \cdots + x_n K_n + G \preceq 0, \\ & a_i^T x = b_i, \quad i = 1, \dots, l \end{aligned}$$

where G and $K_i, i = 1, \dots, n$ are the cone of positive semidefinite matrices.

21.3.4 Lagrangian Duality

An optimization problem can be converted to its dual form, called the dual problem of the original optimization problem. Sometimes, it is much easier to solve the dual problem.

Suppose the original optimization problem is written as

$$\begin{aligned} &\min_{x} f(x), \\ \text{s.t.} \quad & g_i(x) \le 0, \quad i = 1, \dots, m, \\ & h_i(x) = 0, \quad i = 1, \dots, l. \end{aligned}$$

The main idea of *Lagrangian Duality* is to take the constraints in the above optimization problem into account by revising the objective function with a weighted sum of the constraint functions. The Lagrangian associated with the above optimization problem is defined as

$$L(x, \theta, \phi) = f(x) + \sum_{i=1}^{m} \theta_i g_i(x) + \sum_{i=1}^{l} \phi_i h_i(x),$$

where θ_i $(i = 1, \ldots, m)$ and ϕ_i $(i = 1, \ldots, l)$ are Lagrangian multipliers for the constraint inequalities and constraint equations.

The *Lagrangian Dual Function* f_d is defined as the minimum value of the Lagrangian over x,

$$f_d(\theta, \phi) = \min_x L(x, \theta, \phi).$$

Suppose the optimal value of the original optimization problem is f^*, obtained at x^*, i.e., $f^* = f(x^*)$. Then $\forall\theta \geq 0$, considering $g_i(x) \leq 0$ and $h_i(x) = 0$, we have,

$$f_d(\theta, \phi) = \min_x \left\{ f(x) + \sum_{i=1}^{m} \theta_i g_i(x) + \sum_{i=1}^{l} \phi_i h_i(x) \right\} \leq \min_x f(x) = f^*.$$

Therefore, $f_d(\theta, \phi)$ is a lower bound for the optimal value f^* of the original problem. To find the best (tightest) bound, we can solve the following *Lagrangian Dual Problem* to approach f^*,

$$\begin{aligned} &\max_{\theta,\phi} f_d(\theta, \phi), \\ \text{s.t.} \quad &\theta \geq 0. \end{aligned}$$

It is not difficult to see that the dual problem is equivalent to the original problem.

Suppose the optimal value of the dual problem is f_d^*, obtained at (θ^*, ϕ^*), i.e., $f_d^* = f_d(\theta^*, \phi^*)$. Generally, $f_d^* \leq f^*$ since $f_d(\theta, \phi) \leq f^*$. This is called as *weak duality*. If $f_d^* = f^*$ holds, then we claim that *strong duality* holds, indicating that the best bound obtained from the Lagrange dual function is tight.

It can be proven that strong duality holds if the optimization problem respects *Slater's Condition*. Slater's Condition requires that there exists a point $x \in \operatorname{dom} f$ such that x is strictly feasible, i.e., $g_i(x) < 0$, $\forall i = 1, \ldots, m$ and $h_i(x) = 0$, $\forall i = 1, \ldots, l$.

21.3.5 KKT Conditions

If a convex optimization problem respects Slater's Condition, we can obtain an optimal tuple (x^*, θ^*, ϕ^*) using the Karush–Kuhn–Tucker (KKT) conditions below. Then x^* is an optimal point for the original problem and (θ^*, ϕ^*) is an optimal point for the dual problem.

$$\begin{cases} g_i(x^*) \leq 0, \quad i = 1, \ldots, m, \\ h_i(x^*) = 0, \quad i = 1, \ldots, l, \\ \theta_i^* \geq 0, \quad i = 1, \ldots, m, \\ \theta_i^* g_i(x^*) = 0, \quad i = 1, \ldots, m, \\ \nabla f(x^*) + \sum_{i=1}^{m} \theta_i^* \nabla g_i(x^*) + \sum_{i=1}^{l} \phi_i^* \nabla h_i(x^*) = 0. \end{cases}$$

References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Berlin (2006)
2. Boyd, S., Vendenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2003)
3. Golub, G.H., Loan, C.F.V.: Matrix Computations, 3rd edn. Johns Hopkins University Press, Baltimore (1996)
4. Ng, A.: Lecture notes for machine learning. Stanford University cs229 (2010). http://see.stanford.edu/see/courseinfo.aspx?coll=348ca38a-3a6d-4052-937d-cb017338d7b1s