Review of Proof Techniques and Probability

CSE 547 / STAT 548 at the University of Washington

Acknowledgment: This document has been adapted from a similar review session for CS 224W at Standard with substantial modifications by Yikun Zhang in Winter 2023 and Spring 2024 for CSE 547 / STAT 548 at UW. Some good references about writing proofs and basic probability theory include the followings:

- Greg Baker, "Introduction to Proofs": https://www.cs.sfu.ca/~ggbaker/zju/math/proof.html.
- CS103 Winter 2016 at Stanford, "Guide to Proofs": http://stanford.io/2dexnf9.
- Eugenia Cheng, "How to write proofs: a quick guide": https://deopurkar.github.io/teaching/algebra1/cheng.pdf.
- "Quick tour to Basic Probability Theory": http://snap.stanford.edu/class/cs224w-2015/recitation/prob_tutorial.pdf.

1 Proof Techniques

In this section, we will review several mathematical techniques for writing rigorous proofs. They are useful in the field of machine learning when we want to formally state and verify some theoretical properties of our proposed algorithms.

1.1 Terminologies

- **Definition:** an explanation of the mathematical concept in words.
- **Conjecture:** a statement that we think might be true and can be proven (but hasn't been proven yet).
- **Theorem:** a key statement/result that has been rigorously proved.
- **Proof:** a valid argument for showing why a statement/result is true.
- **Premise:** a condition for the theorem.
- **Lemma:** a small theorem (or preliminary result) used in proving the main theorems or other true statements.
- **Proposition:** a less important but interesting true statement with a short proof.
- **Corollary:** a true statement that is a simple deduction from a theorem or proposition.

- **Axiom:** a basic assumption about a mathematical situation, which is also a statement that we assume to be true.

1.2 Universal and Existence Statements

Universal statement: To *prove* a universal statement, like "the square of any odd number is odd", it is always easy to show that this is true for some specific cases – for example, $3^2 = 9$, which is an odd number, and $5^2 = 25$, which is another odd number. However, to rigorously prove the statement, we must show that it works for *all* odd numbers, which is difficult as we cannot enumerate all of them.

On the contrary, if we want to *disprove* a universal statement, we only need to find one counterexample. For instance, if we want to disprove the statement "the square of any odd number is even", it suffices to provide a specific example of an odd number whose square is not even. (For example, $3^2 = 9$, which is not an even number.)

As a summary, it leads to a *rule of thumb* for proving or disproving

- To *prove* a universal statement, we must show that it works for all cases.
- To *disprove* a universal statement, it suffices to find one counterexample.

Example 1. Prove that the square of any odd number is odd.

Proof. Let x be an arbitrary odd number. By definition, an odd number is an integer that can be written in the form $2k+1$ for some integer k. This means that we can write $x = 2k+1$, where k is some integer. Thus,

$$x^2 = (2k+1)^2 = 4k^2 + 4k + 1 = 2(2k^2 + 2k) + 1.$$

Since k is an integer, $2k^2+2k$ is also an integer, so we can write $x^2 = 2\ell+1$, where $\ell = 2k^2+2k$ is an integer. Therefore, x^2 is odd.

Since the above argument works for *any* odd number x, we have shown that the square of any odd number is odd. □

Remark 1. Using the statement that "the square of any odd number is odd" as an example, we showcase how to prove a universal statement. In particular, we pick an *arbitrary* odd number x and try to prove the statement for that number. In the proof, we cannot assume anything about x other than that it is an odd number. (In other words, we cannot simply set x to be a specific number, like 3, because then our proof might rely on special properties of the number 3 that do not generalize to all odd numbers).

Existence statement: The above rule of thumb is reversed when it comes to the "existence" statements. For example, if the statement to be proved is that "there exists at least one odd number whose square is odd, then proving the statement just requires finding one specific case, *e.g.*, $3^2 = 9$, while disproving the statement would require showing that none of the odd numbers have squares that are odd.

1.3 Special Proof Techniques

In addition to the technique of "picking an arbitrary element", here are several other techniques commonly seen in proofs.

Proof by contrapositive: Consider the statement that

(a) *"If it is raining today, then I do not go to class."*

This is logically equivalent to the statement that

(b) *"If I go to class, then it is not raining today."*

Therefore, if we want to prove statement (a), then it suffices (or equivalent) to prove statement (b). Statement (b) is called the **contrapositive** of statement (a).

It is worth mentioning that statement (a) is **not** logically equivalent to the statement:

(c) *"If I do not go to class, then it is raining today."*

which is a **converse** of statement (a). This non-equivalence is also known as the fallacy of the converse.

Example 2. Let x be an integer. Prove that x^2 is an odd number if and only if x is an odd number.

Proof. The "if and only if" in this statement requires us to prove both directions of the implication. First, we must prove that ($\Rightarrow$) if x is an odd number, then x^2 is an odd number. Then, we should also prove that ($\Leftarrow$) if x^2 is an odd number, then x is an odd number.

As we already proved the first statement ($\Rightarrow$) in Example 1, we only need to prove the second statement ($\Leftarrow$). The second statement is logically equivalent to its contrapositive, so it suffices to prove that "if x is an even number, then x^2 is even."

Suppose x is an even number. Then, we can write $x = 2k$ for some integer k. It implies that $x^2 = 4k^2 = 2(2k^2)$. Since k is an integer, $2k^2$ is also an integer, so we can write $x^2 = 2\ell$ for the integer $\ell = 2k^2$. By definition, this means that x^2 is an even number. □

Proof by contradiction: When proving a statement by contradiction, we would assume that our statement is not true and then derive a contradiction. This is a special case of proving by contrapositive (where our "if" is all of mathematics, and our "then" is the statement to be proved).

Example 3. Prove that $\sqrt{2}$ is irrational.

Proof. Suppose that $\sqrt{2}$ was rational. By definition, this means that $\sqrt{2}$ can be written as m/n for some integers m and n. Since $\sqrt{2} = m/n$, it follows that $2 = m^2/n^2$, which in turn shows that $m^2 = 2n^2$. Now any square number x^2 must have an even number of prime factors, since any prime factor found in the first x must also appear in the second x. Therefore, m^2 must have an even number of prime factors. However, since n^2 must also have an even number of prime factors, and 2 is a prime number, $2n^2$ must have an odd number of

prime factors. This is a contradiction, since we claimed that $m^2 = 2n^2$, and no number can simultaneously have an even number of prime factors and an odd number of prime factors. Therefore, our initial assumption was wrong, and $\sqrt{2}$ must be irrational. □

Proof by cases: Sometimes, it might be difficult to prove the entire theorem at once. As a result, we consider splitting the proof into several cases and proving the theorem separately for each case.

Example 4. Let n be an integer. Show that if n is not divisible by 3, then $n^2 = 3k + 1$ for some integer k.

Proof. If n is not divisible by 3, then either $n = 3m + 1$ or $n = 3m + 2$ for some integer m.

Case 1: Suppose $n = 3m + 1$. Then $n^2 = (3m + 1)^2 = 9m^2 + 6m + 1 = 3(3m^2 + 2m) + 1$. Since $3m^2 + 2m$ is an integer, it follows that we can write $n^2 = 3k + 1$ for $k = 3m^2 + 2m$.

Case 2: Suppose $n = 3m+2$. Then $n^2 = (3m+2)^2 = 9m^2+12m+4 = 9m^2+12m+3+1 = 3(3m^2 + 4m + 1) + 1$. Hence, we can write $n^2 = 3k + 1$ for $k = 3m^2 + 4m + 1$.

Since Case 1 and Case 2 reflect all possible possibilities, the proof is completed. □

1.4 Proof by induction

We can prove a statement by induction when showing that the statement is valid for all positive integers n. Note that this is not the only situation in which we can use induction, and that induction is (usually) not the only way to prove a statement for all positive integers.

To use induction, we need to establish two results:

(i) **Base case:** The statement is true when $n = 1$.

(ii) **Inductive step:** If the statement is true for $n = k$, then the statement is also true for $n = k + 1$.

It allows for an infinite chain of implications:

- The statement is true for $n = 1$
- If the statement is true for $n = 1$, then it is also true for $n = 2$
- If the statement is true for $n = 2$, then it is also true for $n = 3$
- If the statement is true for $n = 3$, then it is also true for $n = 4$
- ...

Together, these implications prove the statement for all positive integer values of n. (It does not prove the statement for non-integer values of n, or values of n less than 1.)

Example 5. Prove that $1 + 2 + \cdots + n = n(n + 1)/2$ for all integers $n \geq 1$.

Proof. We proceed by induction.

Base case: If $n = 1$, then the statement becomes $1 = 1(1 + 1)/2$, which is true.

Inductive step: Suppose that the statement is true for $n = k$. This means $1+2+\cdots+k = k(k+1)/2$. We want to show the statement is true for $n = k + 1$, *i.e.*,

$$1 + 2 + \cdots + k + (k + 1) = (k + 1)(k + 2)/2.$$

By the induction hypothesis (*i.e.*, because the statement is true for $n = k$), we have $1 + 2 + \cdots + k + (k + 1) = k(k + 1)/2 + (k + 1)$. This equals $(k + 1)(k/2 + 1)$, which is equal to $(k + 1)(k + 2)/2$. This proves the inductive step.

Therefore, the statement is true for all integers $n \geq 1$. □

1.4.1 Strong induction

Strong induction (or complete induction) is a useful variant of induction. Here, the inductive step is changed to

(i) **Base case:** The statement is true when $n = 1$.

(ii) **Inductive step:** If the statement is true for all values of $1 \leq n < k$, then the statement is also true for $n = k$.

This also produces an infinite chain of implications:

- The statement is true for $n = 1$
- If the statement is true for $n = 1$, then it is true for $n = 2$
- If the statement is true for both $n = 1$ and $n = 2$, then it is true for $n = 3$
- If the statement is true for $n = 1$, $n = 2$, and $n = 3$, then it is true for $n = 4$
- ...

Strong induction works on the same principle as weak induction, but is generally easier to prove theorems under its stronger induction hypothesis.

Definition 1. A *prime number* (or a *prime*) is a natural number strictly greater than 1 that is not a product of two smaller natural numbers. A natural number greater than 1 that is not prime is called a *composite number*.

Example 6. Prove that every integer n greater than or equal to 2 can be factored into prime numbers.

Proof. We proceed by (strong) induction.

Base case: If $n = 2$, then n is a prime number, and its factorization is itself.

Inductive step: Suppose that k is some integer larger than 2, and assume that the statement is true for all numbers $n < k$. Then, there are two cases:

Case 1: k is prime. Then, its prime factorization is k itself.

Case 2: k is composite. This means that it can be decomposed into a product xy, where x and y are both greater than 1 and less than k. Since x and y are both less than k, both x and y can be factored into prime numbers (by the inductive hypothesis). That is, $x = p_1 \cdots p_s$ and $y = q_1 \cdots q_t$, where $p_1, \ldots, p_s$ and $q_1, \ldots, q_t$ are prime numbers.

Thus, k can be written as $(p_1 \ldots p_s) \cdot (q_1 \ldots q_t)$, which is a factorization into prime numbers. It also completed the proof of the statement. □

2 Useful Results in Calculus

The definition of the exponential function states that

$$e^x = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n.$$

In particular, it indicates that $\lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = e$ and $\lim_{n\to\infty} \left(1 - \frac{1}{n}\right)^n = \frac{1}{e}$.

Gamma Function and Stirling's Formula: For $x \in (0, \infty)$, the Gamma function is $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$. While the exact value of $\Gamma(x+1)$ is intractable for some $x \in (0, \infty)$, one can approximate $\Gamma(x+1)$ when x is large by *Stirling's formula*:

$$\lim_{x\to\infty} \frac{\Gamma(x+1)}{(x/e)^x \sqrt{2\pi x}} = 1.$$

This implies that when $x = n$ is a sufficiently large integer, we can approximate $\Gamma(n+1) = n!$ by $\sqrt{2\pi n} \left(\frac{n}{e}\right)^n$. More precisely, the following bound for $n!$ holds for all $n \geq 1$ rather than only asymptotically:

$$\sqrt{2\pi n} \left(\frac{n}{e}\right)^n e^{\frac{1}{12n+1}} < n! < \sqrt{2\pi n} \left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}.$$

3 Basic Probability Theory

Parts of this section are adapted from Chapter 1 of Casella and Berger (2002) and lecture notes of STAT 512 at UW by Professor Yen-Chi Chen's [1] and Professor Michael D. Perlman (Perlman, 2020).

3.1 Probability Space

Sample space: The *sample space* Ω is the collection of all possible outcomes of a random experiment. For example, if we roll a die with 6 faces, then the sample space will be $\{1, 2, 3, 4, 5, 6\}$.

[1] See http://faculty.washington.edu/yenchic/20A_stat512.html.

Events: A subset of the sample space, $A \subset \Omega$, is called an *event*. For example, the event "the number from the above die is less than 4" can be represented by the subset $\{1, 2, 3\}$. The event "we roll a 6 from the above die" can be represented by the subset $\{6\}$.

σ-algebra[2]: While the collection of all possible events (*i.e.*, all subsets of Ω) is sometimes too large to define a valid probability space, we introduce the concept of σ-algebra $\mathcal{F}$ as a collections of subsets of Ω satisfying:

(A1) (*Nonemptiness*) $\Omega \in \mathcal{F}$ and $\emptyset \in \mathcal{F}$, where $\emptyset$ is the empty set.

(A2) (*Closure under complementation*) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.

(A3) (*Closure under countable unions*) If $A_1, A_2, ... \in \mathcal{F}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

The subsets of Ω in $\mathcal{F}$ are said to be *measurable*, and $(\Omega, \mathcal{F})$ is called *measurable space*.

Probability measure: A probability measure (or probability function) P is a mapping from σ-algebra $\mathcal{F}$ to real numbers in $[0, 1]$ satisfying the following three axioms:

- $P(A) \geq 0$ for all $A \in \mathcal{F}$.
- $P(\Omega) = 1$
- If $A_1, A_2, ... \in \mathcal{F}$ are mutually exclusive events, then $P\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{n} P(A_i)$.

The triplet $(\Omega, \mathcal{F}, P)$ is called a *probability space*.

Example 7. For a fair dice with 6 faces, we can define the probability function as:

$$P\left(\{\text{we roll the face } i\}\right) = \frac{1}{6} \quad \text{for i = 1,..., 6.}$$

Any event in the probability space can be represented as unions of these six disjoint events. For instance,

$$P\left(\text{we roll an odd number}\right) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.$$

Note that we can add probabilities here because the events {we roll the face 1}, {we roll the face 3}, and {we roll the face 5} are disjoint.

3.2 Properties of Probability Measure

Theorem 1 (Principle of inclusion-exclusion)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space. Given two subsets $A, B \in \mathcal{F}$ that are not necessarily disjoint, we have that*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Proof. We can derive this theorem from the probability axioms. Note that $A \cup B$ can be split into three disjoint events: $A \setminus B = A \cap B^c$, $A \cap B$, and $B \setminus A = B \cap A^c$. Furthermore, A can be split into $A \setminus B$ and $A \cap B$, and B can be split into $B \setminus A$ and $A \cap B$. Thus,

$$P(A \cup B) \;=\; P(A \setminus B) + P(A \cap B) + P(B \setminus A)$$

[2]This is an advanced concept and can be skipped.

$$\begin{aligned} &= P(A \setminus B) + P(A \cap B) + P(B \setminus A) + P(A \cap B) - P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

The result follows. □

Example 8. Suppose k is chosen uniformly at random from the integer set $\{1, 2, \ldots, 100\}$. (This means that the probability of getting each integer is 1/100.) Find the probability that k is divisible by 2 or 5.

Solution. By the principle of inclusion-exclusion (Theorem 1),

$$\begin{aligned} P(\{k \text{ is divisible by 2 or 5}\}) = P(\{k \text{ is divisible by 2}\}) + P(\{k \text{ is divisible by 5}\}) \\ - P(\{k \text{ is divisible by both 2 and 5}\}). \end{aligned}$$

There are 50 numbers divisible by 2, 20 numbers divisible by 5, and 10 numbers divisible by 10 (*i.e.*, divisible by both 2 and 5). Therefore, the probability is

$$P(\{k \text{ is divisible by 2 or 5}\}) = \frac{50}{100} + \frac{20}{100} - \frac{10}{100} = 0.6.$$

Theorem 2 (Union bound or Boole's inequality)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space. For any collection of n events $A_1, \ldots, A_n \in \mathcal{F}$, we have that*

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i).$$

Proof. We can prove this result by induction (for finite n).

Base case: When $n = 1$, the statement becomes $P(A_1) \leq P(A_1)$, which is true.

Inductive step: Suppose that the statement is true for $n = k$. We must prove that the statement is true for $n = k + 1$. Note that

$$\bigcup_{i=1}^{k+1} A_i = \left(\bigcup_{i=1}^{k} A_i\right) \cup A_{k+1}$$

and by the principle of inclusion-exclusion (Theorem 1),

$$P\left(\bigcup_{i=1}^{k+1} A_i\right) \leq P\left(\bigcup_{i=1}^{k} A_i\right) + P(A_{k+1}).$$

By the induction hypothesis, the first term is less than or equal to $\sum_{i=1}^{k} P(A_i)$. Hence,

$$P\left(\bigcup_{i=1}^{k+1} A_i\right) \leq \sum_{i=1}^{k+1} P(A_i).$$

The proof is completed. □

Example 9. Suppose that the chance of winning a Mega Million is 1 in 100000 every time a person buys a lottery ticket. If Tim buys one ticket every day of the year, how likely will he win the Mega Million at least once?

Answer. The union bound will not tell us the exact probability for Tim winning the Mega Million. However, it gives us an upper bound of this probability as 365/100000.

Other useful properties of probability measure: Let $(\Omega, \mathcal{F}, P)$ be a probability space.

- If $A \subset B$, then $P(A) \leq P(B)$. More generally, $P(B \setminus A) = P(B) - P(A \cap B)$.
- For any $A \subset \mathcal{F}$ and some mutually exclusive $C_1, C_2, ...$ with $\cup_{i=1}^{\infty} C_i = \Omega$,

$$P(A) = \sum_{i=1}^{\infty} P(A \cap C_i).$$

- *Monotone continuity*: For a sequence of subsets $\{A_n\}_{n=1}^{\infty} \subset \mathcal{F}$ with $A_n \subset A_{n+1}$ for all n, we have that $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} P(A_n)$. Similarly, if $A_n \supset A_{n+1}$ for all n, we have that $P\left(\bigcap_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} P(A_n)$.

3.3 Conditional Probability, Independence, and Bayes' Rule

We motivate the concept of conditional probability through the following example.

Example 10. Suppose that you are administering the GRE, and you discover that 2.5% of students get a perfect score on the math section.[3] By itself, this is not a very useful statistic, because the scores on the math section vary substantially by major. You dive a little deeper and find that 7.5% of physical sciences students get a perfect score, 6.3% of engineering students get a perfect score, and most other majors do substantially worse.

In the language of conditional probability, we would say that the probability of getting a perfect score conditional on engineering majors is 6.3%, *i.e.*,

$$P(\text{perfect score} \mid \text{engineering major}) = 0.063.$$

If we want to compute this probability, we would take the number of engineering majors that receive a perfect score, and divide it by the total number of engineering majors. This is equivalent to computing the formula:

$$P(\text{perfect score} \mid \text{engineering major}) = \frac{P(\text{perfect score} \cap \text{engineering major})}{P(\text{engineering major})}.$$

In general, we can replace "perfect score" and "engineering major" with any two events and obtain the formal definition of conditional probability.

[3] See https://www.ets.org/s/gre/pdf/gre_guide_table4.pdf for a breakdown by specific majors. For some reason, computer science is counted as part of the physical sciences, and not as engineering.

Conditional Probability: For two events A and B with $P(B) > 0$, the conditional probability of A given B is defined as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Notice that when the event B is fixed, $P(\cdot|B)$ is another probability measure.

Example 11. Suppose that we toss a fair coin three times. What is the probability that all three tosses come up heads, given that the first toss came up heads?

Answer. This probability is

$$\frac{P\left(\{\text{all three tosses come up heads and the first toss came up heads}\}\right)}{P\left(\{\text{the first toss came up heads}\}\right)} = \frac{1/8}{1/2} = \frac{1}{4}.$$

Independence and conditional independence: Two events A and B are *independent* if

$$P(A|B) = P(A) \quad \text{or equivalently,} \quad P(A \cap B) = P(A) \cdot P(B).$$

In other words, the occurrence of event B does not affect the probability that event A happens.

For three events A, B, C, we say that A and B are *conditionally independent* given C if

$$P(A \cap B|C) = P(A|C) \cdot P(B|C).$$

There are no implications between independence and conditional independence!!

Bayes' rule: Given an event A and some mutually exclusive events $B_1, ..., B_k$ with $\cup_{i=1}^k B_i = \Omega$, the *Bayes' rule* states that

$$P(B_j|A) = \frac{P(B_j \cap A)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^k P(A|B_i)P(B_i)} \quad \text{for all } j = 1, ..., k.$$

Example 12. Suppose that 5% of students enrolled in CSE 547 at UW will get 4.0, and a student with 4.0 in CSE 547 has 80% chance of getting recruited by Google. A student without getting 4.0 in CSE 547 still has 40% chance of getting recruited by Google. What is the probability of a student getting 4.0 in CSE 547, given that he/she has been recruited by Google?

Answer. By Bayes' Rule,

$$\begin{aligned}
&P(\text{Get 4.0} \mid \text{Recruited by Google}) \\
&= \frac{P(\text{Recruited by Google} \mid \text{Get 4.0}) \cdot P(\text{Get 4.0})}{P(\text{Recruited by Google})} \\
&= \frac{P(\text{Recruited by Google} \mid \text{Get 4.0}) \cdot P(\text{Get 4.0})}{P(\text{Recruited by Google}|\text{Get 4.0})P(\text{Get 4.0}) + P(\text{Recruited by Google}|\text{Not get 4.0})P(\text{Not get 4.0})} \\
&= \frac{0.8 \times 0.05}{0.8 \times 0.05 + 0.4 \times 0.95} \\
&\approx 9.52\%.
\end{aligned}$$

3.4 Random variables

Random variable: Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathbb{R} = (-\infty, \infty)$ be the set of all real numbers. A *random variable* $X : \Omega \to \mathbb{R}$ is a (measurable) function satisfying

$$X^{-1}((-\infty, c]) := \{\omega \in \Omega : X(\omega) \leq c\} \in \mathcal{F} \quad \text{for all } c \in \mathbb{R}.$$

The probability that X takes on a value in a Borel set[4] $B \subset \mathbb{R}$ is written as:

$$P(X \in B) = P(\{\omega \in \Omega : X(\omega) \in B\}).$$

Example 13. Suppose that we are tossing three fair coins. Let X be the number of coins that come up heads. Then, $P(X = 0) = 1/8$.

Cumulative distribution function (CDF): The CDF $F : \mathbb{R} \to [0, 1]$ of a random variable X is a right continuous and nondecreasing function with left limits satisfying

$$F(x) := P(X \leq x) = P(\{\omega \in \Omega : X(\omega) \leq x\}).$$

In particular, $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

Probability mass function (PMF) and probability density function (PDF):

- If the range $\mathcal{X} \subset \mathbb{R}$ of a random variable X is countable, it is called a *discrete* random variable, whose distribution can be characterized by the PMF as:

$$P(X = x) = F(x) - \lim_{\epsilon \to 0^+} F(x - \epsilon) \quad \text{for all } x \in \mathcal{X}.$$

- If the range $\mathcal{X} \subseteq \mathbb{R}$ of a random variable X has an absolutely continuous CDF F, then we can describe its distribution through the PDF as:

$$p(x) = F'(x) = \frac{d}{dx} F(x).$$

  In this case, $F(x) = P(X \leq x) = \int_{-\infty}^{x} p(u)\, du$, and $P(X = x) = 0$ for a single number $x \in \mathbb{R}$.

3.5 Expectation and Variance

Expectation: The *expected value* (or mean) of a random variable X with range $\mathcal{X} \subset \mathbb{R}$ can be interpreted as a weighted average.

- For a discrete random variable, $E(X) = \sum_{x \in \mathcal{X}} x \cdot P(X = x)$.
- For a continuous random variable with PDF p_X, $E(X) = \int_{-\infty}^{\infty} x \cdot p_X(x)\, dx$

[4] A Borel set in $\mathbb{R}$ is a set that can be formed from open sets through the operations of countable unions/intersections and complements.

Example 14. Suppose that Tim's happiness scores 10 when it is sunny outside and 2 when it is raining outside. It is sunny 60% of the time at Seattle and raining 40%. What is the expected value of Tim's happiness at Seattle?

Answer. $10 \times 0.6 + 2 \times 0.4 = 6.8$.

Linearity of expectation: If X, Y are two random variables and a is a constant in $\mathbb{R}$, then

$$E(X+Y) = E(X) + E(Y) \quad \text{and} \quad E(aX) = a \cdot E(X).$$

This is true even if X and Y are not independent.

Variance and covariance: The *variance* of a random variable measures how far away it is, on average, from the mean. It is defined as

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

The covariance between random variables X, Y is defined as:

$$\mathrm{Cov}(X,Y) = E\left[(X - E(X))\,(Y - E(Y))\right].$$

For a random variable X and a constant $a \in \mathbb{R}$, we have $\mathrm{Var}(X+a) = \mathrm{Var}(X)$ and $\mathrm{Var}(aX) = a^2 \cdot \mathrm{Var}(X)$. We **do not** have $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ unless X and Y are uncorrelated (which means they have covariance 0). In particular, independent random variables are always uncorrelated, although the reverse doesn't hold in general.

Pearson's correlation coefficient: For two random variables X and Y, their (Pearson's) correlation coefficient is defined as:

$$\rho_{XY} = \mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}},$$

where $\rho_{XY} \in [-1, 1]$ by the Cauchy-Schwarz inequality; see Section 3.7. It measures the *linear* relation between two random variables.

3.6 Known Probability Distributions

3.6.1 Discrete random variables

Bernoulli: If X is a Bernoulli random variable denoted by $X \sim \texttt{Bernoulli}(p)$, then

$$P(X = 1) = p \quad \text{and} \quad P(X = 0) = 1 - p.$$

A Bernoulli random variable with parameter p can be interpreted as a coin flip that comes up heads with probability p and tails with probability $1 - p$. We know that

$$E(X) = p \quad \text{and} \quad \mathrm{Var}(X) = p(1-p).$$

Binomial: If X is a binomial random variable denoted by $X \sim \texttt{Binomial}(n, p)$, then

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, ..., n.$$

A binomial random variable with parameters n and p models the number of successes in n trials, each of which has a successful probability p. When $n = 1$, it reduces to a Bernoulli random variable. We know that

$$E(X) = np \quad \text{and} \quad \text{Var}(X) = np(1-p).$$

Geometric: If X is a geometric random variable denoted by $X \sim \texttt{Geometric}(p)$, then

$$P(X = k) = (1-p)^{k-1}p \quad \text{for } k = 0, 1, ...$$

A geometric random variable with parameter p models the number of trials until the first success occurs, where each trial has a successful probability p. We know that

$$E(X) = \frac{1}{p} \quad \text{and} \quad \text{Var}(X) = \frac{1-p}{p^2}.$$

Poisson: If X is a Poisson random variable denoted by $X \sim \texttt{Poisson}(\lambda)$, then

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{for } k = 0, 1, ...$$

A Poisson random variable often appears in counting processes. For instance, the number of laser photons hitting a detector in a particular time interval can be modeled as a Poisson random variable. We know that

$$E(X) = \lambda \quad \text{and} \quad \text{Var}(X) = \lambda.$$

Indicator random variable: For an event A, an indicator random variable takes value 1 when A occurs and 0 otherwise, *i.e.*,

$$I_A = \begin{cases} 1 & \text{if event } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

The expectation of an indicator random variable is just the probability of the event occurring, *i.e.*,

$$\begin{aligned} E[I_A] &= 1 \cdot P(I_A = 1) + 0 \cdot P(I_A = 0) \\ &= P(I_A = 1) \\ &= P(A), \end{aligned}$$

and its variance is $\text{Var}(I_A) = P(A)\,[1 - P(A)]$.

Multinomial: Suppose that Z is a categorical random variable with range $\{1, ..., k\}$ and $P(Z = j) = p_j$ for $j = 1, ..., k$. We generate independently and identically distributed data $Z_1, ..., Z_n$ with the above distribution and take

$$X_j = \sum_{i=1}^{n} I_{\{Z_i = j\}} = \text{Number of observations in Category } j.$$

Then, the random vector $\boldsymbol{X} = (X_1, ..., X_k)$ follows a multinomial distribution denoted by $\boldsymbol{X} \sim \texttt{Multinomial}(n; p_1, ..., p_k)$ with $\sum_{j=1}^{k} p_j = 1$, whose PMF is given by

$$P(X_1 = x_1, ..., X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} \cdot p_1^{x_1} \cdots p_k^{x_k}.$$

Here, $\boldsymbol{X}$ takes integer values within a simplex $\left\{(x_1, ..., x_k) \in \{0, 1, ..., k\}^n : \sum_{j=1}^{n} x_j = n\right\}$.

3.6.2 Continuous random variables

Uniform: If X is a uniform random variable over the interval $[a, b]$ denote by $X \sim \texttt{Uniform}[a, b]$, then its PDF is given by

$$p(x) = \frac{1}{b-a} \cdot I_{[a,b]}(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b, \\ 0 & \text{otherwise.} \end{cases}$$

We know that

$$E(X) = \frac{a+b}{2} \quad \text{and} \quad \text{Var}(X) = \frac{(b-a)^2}{12}.$$

Normal: If X is a normal random variable with parameters μ and σ^2 denoted by $X \sim N(\mu, \sigma^2)$, then its PDF is given by

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

where $x \in (-\infty, \infty)$. We know that

$$E(X) = \mu \quad \text{and} \quad \text{Var}(X) = \sigma^2.$$

Cauchy: If X is a Cauchy random variable with parameters μ, σ^2 denoted by $X \sim \texttt{Cauchy}(\mu, \sigma^2)$, then its PDF is given by

$$p(x) = \frac{1}{\pi\sigma} \left[\frac{\sigma^2}{\sigma^2 + (x-\mu)^2}\right],$$

where $x \in (-\infty, \infty)$. Note that both the expectation and variance of a Cauchy distribution *do not exist*. The parameter μ represents its median.

Student's t: If X is a Student's t random variable with parameter $\nu > 0$ denoted by $X \sim \texttt{t}(\nu)$, then its PDF is given by

$$p(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

where $x \in (-\infty, \infty)$. The parameter $\nu > 0$ is also called degrees of freedom. Note that when $\nu = 1$, the Student's t distribution reduces to a Cauchy distribution. The expectation of $X \sim \texttt{t}(\nu)$ is 0 for $\nu > 1$ and undefined for $0 < \nu \le 1$. The variance of $X \sim \texttt{t}(\nu)$ is

$$\text{Var}(X) = \begin{cases} \frac{\nu}{\nu-2} & \text{for } v > 2, \\ \infty & \text{for } 1 < \nu \le 2, \\ \text{undefined} & \text{for } 0 < \nu \le 1. \end{cases}$$

Exponential: If X is an exponential random variable with parameter λ denoted by $X \sim \texttt{Exp}(\lambda)$, then its PDF is given by

$$p(x) = \lambda e^{-\lambda x} \cdot I_{[0,\infty)}(x).$$

We know that

$$E(X) = \frac{1}{\lambda} \quad \text{and} \quad \text{Var}(X) = \frac{1}{\lambda^2}.$$

A *double exponential* random variable Y satisfies that $|Y| \sim \texttt{Exp}(\lambda)$, so its PDF is given by

$$p(y) = \frac{\lambda}{2} e^{-\lambda|y|} \quad \text{with} \quad y \in (-\infty, \infty).$$

In particular, $E(Y) = 0$ and $\text{Var}(Y) = \frac{2}{\lambda^2}$. Sometimes, Y is also called a *Laplace* random variable[5].

Gamma: A Gamma random variable X is characterized by two parameters $\alpha, \lambda > 0$ and has a PDF

$$p(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \cdot I_{[0,\infty)}(x),$$

where $\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du$ is the Gamma function; see Section 2. We denote $X \sim \texttt{Gamma}(\alpha, \lambda)$ and have that

$$E(X) = \frac{\alpha}{\lambda} \quad \text{and} \quad \text{Var}(X) = \frac{\alpha}{\lambda^2}.$$

Beta: A Beta random variable X with parameters $\alpha, \beta > 0$ has its PDF as:

$$p(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \cdot I_{[0,1]}(x),$$

where $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. Given that the Beta random variable $X \sim \texttt{Beta}(\alpha, \beta)$ has a continuous distribution on $[0, 1]$, it is often used to model a ratio or probability. We know that

$$E(X) = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Logistic: A logistic random variable X with parameters $\alpha \in \mathbb{R}, \beta > 0$ has its CDF with the form of a logistic function as:

$$F(x) = P(X \le x) = \frac{1}{1 + e^{-\alpha-\beta x}}.$$

Thus, its PDF is given by

$$p(x) = \frac{d}{dx} F(x) = \frac{\beta e^{-\alpha-\beta x}}{(1 + e^{-\alpha-\beta x})^2} = \frac{\beta e^{\alpha+\beta x}}{(1 + e^{\alpha+\beta x})^2}.$$

[5] See https://en.wikipedia.org/wiki/Laplace_distribution.

Dirichlet: A Dirichlet random vector $\boldsymbol{Z} = (Z_1, ..., Z_k)$ generalizes the Beta distribution to its multivariate version (or extend the multinomial distribution to its continuous version). It has a PDF defined on the simplex $\left\{(z_1, ..., z_k) \in [0,1]^k : \sum_{i=1}^k z_i = 1\right\}$ as:

$$p(z_1, ..., z_k; \alpha_1, ..., \alpha_k) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^k z_i^{\alpha_i - 1},$$

where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_k)$ with $\alpha_i > 0, i = 1, ..., k$ is a k-dimensional parameter vector. The Dirichlet distribution is particularly useful in modeling the prior probabilities of a multinomial distribution that generates the latent topics of a document (Blei et al., 2003). When $\boldsymbol{Z} \sim \texttt{Dirichlet}(\alpha_1, ..., \alpha_k)$, its mean vector is $E(\boldsymbol{Z}) = \left(\frac{\alpha_1}{\sum_{i=1}^k \alpha_i}, \ldots, \frac{\alpha_k}{\sum_{i=1}^k \alpha_i}\right)$.

3.7 Inequalities

Markov's inequality: Let X be a nonnegative random variable. Then, for any $\epsilon > 0$,

$$P(X \geq \epsilon) \leq \frac{E(X)}{\epsilon}.$$

Proof. For any $\epsilon > 0$, we consider splitting the expectation $E(X)$ into two parts as:

$$\begin{aligned} E(X) &= E\left(X \cdot I_{\{X \geq \epsilon\}}\right) + E\left(X \cdot I_{\{X < \epsilon\}}\right) \\ &\geq E\left(X \cdot I_{\{X \geq \epsilon\}}\right) \\ &\geq E\left(\epsilon \cdot I_{\{X \geq \epsilon\}}\right) \\ &= \epsilon \cdot P(X \geq \epsilon). \end{aligned}$$

The result follows by dividing $\epsilon > 0$ on both sides of the above inequality. □

Chebyshev's inequality: Let X be a random variable with $\text{Var}(X) < \infty$. Then, for any $\epsilon > 0$,

$$P\left(|X - E(X)| \geq \epsilon\right) \leq \frac{\text{Var}(X)}{\epsilon}.$$

The Chebyshev's inequality can be proved by applying Markov's inequality to the nonnegative random variable $[X - E(X)]^2$. It is a simple instance of general concentration inequalities that give a probabilistic bound on the deviation of X away from its mean.

Chernoff bound: Suppose that there is a constant $b > 0$ such that the *moment generating function* $\varphi(\lambda) = E\left[e^{\lambda(X-\mu)}\right]$ of a random variable X exists when $\lambda \leq |b|$, where $\mu = E(X)$. Given that

$$P\left[(X - \mu) > t\right] = P\left[e^{\lambda(X-\mu)} \geq e^{\lambda t}\right] \leq \frac{E\left[e^{\lambda(X-\mu)}\right]}{e^{\lambda t}} \quad \text{for any } \lambda \in [0, b],$$

we can optimize our choice of λ to obtain the *Chernoff bound* as:

$$P\left[(X - \mu) > t\right] \leq \inf_{\lambda \in [0,b]} \frac{E\left[e^{\lambda(X-\mu)}\right]}{e^{\lambda t}}.$$

Cauchy-Schwarz inequality: Given two random variables X and Y,

$$|E(XY)|^2 \leq E(X^2) \cdot E(Y^2),$$

where equality holds if and only if either $P(X = 0) = 0$, or $P(Y = 0) = 0$, or $P(X = cY) = 1$ for some nonzero constant $c \in \mathbb{R}$. A useful corollary of the Cauchy-Schwarz inequality is that

$$|\mathrm{Cov}(X, Y)|^2 \leq \mathrm{Var}(X)\mathrm{Var}(Y).$$

Hölder inequality: Given two random variables X and Y,

$$E|XY| \leq \left(E|X|^p\right)^{\frac{1}{p}} \left(E|Y|^q\right)^{\frac{1}{q}} \equiv ||X||_p \, ||Y||_q$$

with $p, q \in [1, \infty]$ and $\frac{1}{p} + \frac{1}{q} = 1$, where equality holds if and only if $P\left(|X|^p = c|Y|^q\right) = 1$ for some nonzero constant c. Specifically, when $p = \infty$, $||X||_\infty = \inf\{M \geq 0 : P(|X| > M) = 0\}$.

Minkowski Inequality: Given two random variables X and Y,

$$\left[E|X + Y|^p\right]^{\frac{1}{p}} \leq \left[E|X|^p\right]^{\frac{1}{p}} + \left[E|Y|^p\right]^{\frac{1}{p}}$$

for $p \in [1, \infty)$, where equality holds if and only if $P(X = cY) = 1$ for some nonzero constant c, or $P(Y = 0) = 1$, or $P(X = 0) = 1$.

Jensen's Inequality: Given a convex function φ and a random variable X,

$$\varphi\left(E(X)\right) \leq E\left[\varphi(X)\right],$$

where equality holds if and only if either $P(X = c) = 1$ for some constant c or for every line $a + bx$ that is tangent to φ at $E(X)$, $P\left(\varphi(x) = a + bx\right) = 1$.

4 Big O and O_P Symbols

In machine learning, big O and little o symbols are used to characterize the time or space complexity of our algorithm with respect to the sample size or data dimension. In general, these symbols describe the growth rate of functions as follows.

Let $f(x)$ be the function to be estimated on $\mathbb{R}$ and $g(x)$ be the comparison function that is strictly positive when x is large on $\mathbb{R}$.

Big O symbol: We write $f(x) = O\left(g(x)\right)$ if there exist constants $M > 0$ and $x_0 > 0$ such that[6]

$$|f(x)| \leq M \cdot g(x) \quad \text{for all } x \geq x_0.$$

Using the limit superior notation, we know that

$$f(x) = O\left(g(x)\right) \quad \Longleftrightarrow \quad \limsup_{x \to \infty} \frac{|f(x)|}{g(x)} < \infty.$$

[6]See https://en.wikipedia.org/wiki/Big_O_notation.

Notice that $\limsup_{x\to\infty} h(x) = \lim_{x_0\to\infty}\left[\sup_{x\geq x_0} f(x)\right]$.

Little o symbol: Similarly, we write $f(x) = o\left(g(x)\right)$ if for any $\epsilon > 0$, there exists a constant $x_0 > 0$ such that

$$|f(x)| \leq \epsilon \cdot g(x) \quad \text{for all } x \geq x_0.$$

Under the limit notation, we have that

$$f(x) = o\left(g(x)\right) \iff \lim_{x\to\infty} \frac{|f(x)|}{g(x)} = 0.$$

Big Ω symbol: Sometimes, depending on the context, we may encounter the big Ω symbol in machine learning literature. In most cases, the definition of $f(x) = \Omega\left(g(x)\right)$ follows from Knuth (1976), so we write $f(x) = \Omega\left(g(x)\right)$ if there exist constants $m > 0$ and x_0 such that

$$|f(x)| \geq m \cdot g(x) \quad \text{for all } x \geq x_0,$$

or equivalently,

$$f(x) = \Omega\left(g(x)\right) \iff \liminf_{n\to\infty} \frac{f(x)}{g(x)} > 0.$$

Taking into account the randomness of input data, it may not be possible to bound a quantity or random variable in our algorithm through the above big O and little o symbols. We introduce the O_P and o_P symbols[7] to handle the stochastic rate of convergence for a sequence of random variables $\{X_n\}_{n=1}^{\infty}$; see also Section 2.2 in van der Vaart (1998).

Little o_P symbol: We write $X_n = o_P(a_n)$ for a sequence of constants $\{a_n\}_{n=1}^{\infty}$ if $\frac{X_n}{a_n}$ converges to 0 in probability as $n \to \infty$. That is, for any $\epsilon > 0$,

$$X_n = o_P(a_n) \iff \lim_{n\to\infty} P\left(\left|\frac{X_n}{a_n}\right| \geq \epsilon\right) = 0.$$

Big O_P symbol: We write $X_n = O_P(a_n)$ for a sequence of constants $\{a_n\}_{n=1}^{\infty}$ if $\frac{X_n}{a_n}$ is bounded in probability when n is large. That is, for any $\epsilon > 0$, there exist a constant $M > 0$ and an integer $N > 0$ such that

$$X_n = O_P(a_n) \iff P\left(\left|\frac{X_n}{a_n}\right| > M\right) < \epsilon \text{ when } n > N.$$

5 Basic Optimization and Lagrange Multiplier

For given function $f, g_i, h_j : \mathbb{R}^n \to \mathbb{R}$, a (mathematical) optimization problem has the following form

$$\begin{aligned} &\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \\ \text{subject to } \; &g_i(\boldsymbol{x}) \leq 0, \qquad i = 1, ..., m, \\ &h_j(\boldsymbol{x}) = 0, \quad j = 1, ..., p. \end{aligned} \tag{1}$$

[7] See https://en.wikipedia.org/wiki/Big_O_in_probability_notation.

Here, we call $\boldsymbol{x} \in \mathbb{R}^n$ the *optimization variable*, f the *objective function*, $g_i(\boldsymbol{x}) \leq 0, i = 1, ..., m$ the *inequality constraints*, and $h_j(\boldsymbol{x}) = 0, j = 1, ..., p$ the *equality constraints*.

Unconstrained Optimization: If there are no constraints (*i.e.*, $m = p = 0$), then we say that the problem (1) is *unconstrained*. Solving the unconstrained optimization problem is relatively easy. When the objective function f is differentiable, we can compute its gradient $\nabla f(\boldsymbol{x})$ and set it to 0 for a equation $\nabla f(\boldsymbol{x}) = 0$ whose roots consist of a candidate set of solutions. Practically, we will discuss how to use the (stochastic) gradient descent algorithm to solve for a solution/minimizer $\boldsymbol{x}^*$ in the lecture (Lecture 14: Large-Scale Machine Learning II).

Constrained Optimization: The general optimization problem (1) with constraints is more difficult to handle. A general approach is to convert (1) into an unconstrained optimization problem using the *method of Lagrange multipliers*. We define the *Lagrangian* $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ associated with the problem (1) as:

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{x}) + \sum_{j=1}^{p} \nu_j h_j(\boldsymbol{x}). \tag{2}$$

We refer to $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_m)^T \in \mathbb{R}^m$ and $\boldsymbol{\nu} = (\nu_1, ..., \nu_p)^T \in \mathbb{R}^p$ as the Lagrange multipliers.

Case 1: If there are no inequality constraints, then the Lagrangian (2) becomes

$$L(\boldsymbol{x}, \boldsymbol{\nu}) = f(\boldsymbol{x}) + \sum_{j=1}^{p} \nu_j h_j(\boldsymbol{x}),$$

and we can find the minimizer of f by identifying the stationary points of L. This means that we set all the partial derivatives of L to 0 and solve the following system of equations:

$$\frac{\partial}{\partial \boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\nu}) = \nabla f(\boldsymbol{x}) + \sum_{j=1}^{p} \nu_j \nabla h_j(\boldsymbol{x}) = \mathbf{0} \quad \text{and} \quad \frac{\partial}{\partial \nu_j} L(\boldsymbol{x}, \boldsymbol{\nu}) = h_j(\boldsymbol{x}) = 0 \quad \text{for} \quad j = 1, ..., p.$$

However, not all the stationary points yield a solution of the original problem, as the method of Lagrange multipliers only gives a necessary condition for optimality in constrained problems. Thus, we need to verify whether a yielded solution $\tilde{\boldsymbol{x}}$ is a minimizer or not by checking other sufficient conditions (if exist) or comparing $f(\tilde{\boldsymbol{x}})$ with the values of f (in a neighborhood of $\tilde{\boldsymbol{x}}$).

Case 2: If there are some inequality constraints, then we define the Lagrange dual function $D : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ as the minimum value of the Lagrangian over $\boldsymbol{x}$ as:

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\boldsymbol{x} \in \mathbb{R}^n} L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\boldsymbol{x} \in \mathbb{R}^n} \left[f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{x}) + \sum_{j=1}^{p} \nu_j h_j(\boldsymbol{x}) \right]. \tag{3}$$

When the Lagrangian is unbounded below in $\boldsymbol{x}$, the dual function takes on the value $-\infty$. The dual function will provide lower bounds on the optimal value p^* of the original problem (1), *i.e.*, for any $\boldsymbol{\lambda} \succeq \mathbf{0}$ and any $\boldsymbol{\nu} \in \mathbb{R}^p$, we have that

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^*.$$

Here, $\boldsymbol{\lambda} \succeq \mathbf{0}$ means that each entry of $\lambda_i, i = 1, ..., m$ is bigger or equal to 0. Under some conditions (such as Slater's condition[8]), the above equality holds, and by Karush–Kuhn–Tucker (KKT) conditions[9], we can relate the solution to the primal problem (1) with the solution to its dual problem

$$\begin{aligned} &\max_{(\boldsymbol{\lambda},\boldsymbol{\nu}) \in \mathbb{R}^m \times \mathbb{R}^p} g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\ &\text{subject to } \boldsymbol{\lambda} \succeq \mathbf{0}. \end{aligned} \tag{4}$$

See Chapter 5 of Boyd and Vandenberghe (2004) for more details.

Remark 2. Since the dual function is the pointwise infimum of a family of affine functions of $(\boldsymbol{\lambda}, \boldsymbol{\nu})$, it is concave, even when the problem (1) is not convex.

References

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004. URL https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf.

G. Casella and R. Berger. *Statistical Inference*. Duxbury advanced series. Thomson Learning, 2nd ed. edition, 2002.

D. E. Knuth. Big omicron and big omega and big theta. *ACM Sigact News*, 8(2):18–24, 1976.

M. Perlman. Probability and Mathematical Statistics I (STAT 512 Lecture Notes), 2020. URL https://sites.stat.washington.edu/people/mdperlma/STAT%20512%20MDP%20Notes.pdf.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[8] See https://en.wikipedia.org/wiki/Slater%27s_condition.

[9] See ttps://en.wikipedia.org/wiki/Karush\OT1\textendashKuhn\OT1\textendashTucker_conditions.