Marco Tomamichel

Lecture Notes for EE5139:
Information Theory for Communication

(Fall 2021)

Disclaimer: They are not yet complete, free of typos, or always presented in the most clear way. Any comments that help reduce these deficiencies are very much appreciated. Chapters 0 and 7 are based on notes by Vincent Y. F. Tan.

Contents

$\emptyset$	empty set $\{\}$
$[M]$	the set $\{1, 2, \ldots, M\}$
$\mathcal{P}(\mathcal{X})$	the power set of $\mathcal{X}$, i.e. $\{A : A \subseteq \mathcal{X}\}$
$\mathcal{X} \times \mathcal{Y}$	the set of tuples $\{(x, y) : x \in \mathcal{X}, y \in \mathcal{Y}\}$
$\mathcal{X}^n$	the set of n-tuples with each element taking values in $\mathcal{X}$, e.g., $\mathcal{X}^2 = \mathcal{X} \times \mathcal{X}$
$\{0, 1\}^n$	the set of n-bit strings
$\{0, 1\}^*$	the set of bit strings of arbitrary length
$\max \mathcal{X}$	largest $x^* \in \mathcal{X}$, might not always exist
$\sup \mathcal{X}$	smallest $x^* \in \mathbb{R}$ such that $x \leq x^*$ for all $x \in \mathcal{X}$; equals the maximum, $\max \mathcal{X}$, if it exists
$\min \mathcal{X}$	smallest $x^* \in \mathcal{X}$, might not always exist
$\inf \mathcal{X}$	largest $x^* \in \mathbb{R}$ such that $x \geq x^*$ for all $x \in \mathcal{X}$; equals the minimum, $\min \mathcal{X}$, if it exists
$\mathbf{1}\{x = y\}$	indicator function, evaluates to 1 if the condition is true and 0 otherwise, so that, for example, $\mathbf{1}\{x = y\} + \mathbf{1}\{x \neq y\} = 1$
δ_{xy}	shorthand for $\mathbf{1}\{x = y\}$
$P_X(x)$	probability mass function (pmf), $P_X(x) = P[X = x]$
$p_X(x)$	probability density function (pdf), i.e. $P[X \in (1, 2)] = \int_1^2 p_X(x) \mathrm{d}x$
$P[X \in \mathcal{A}]$	probability of a random variable X being in some set $\mathcal{A}$, i.e. $P[X \in \mathcal{A}] = \mathbb{P}(\{\omega : X(\omega) \in \mathcal{A}\}) = \sum_{x \in \mathcal{A}} P_X(x)$
$P[5 \leq X < 6]$	another way of writing $P[X \in [5, 6)]$
$\log$	logarithm; in these notes we take the logarithm to base 2, i.e. $\log = \log_2$

Table 1: Some basic notation used in this module.

pmf	probability mass function
pdf	probability density function
cdf	cumulative density function
rv	random variable
DMS	discrete memoryless source
DMC	discrete memoryless channel

Table 2: Some abbreviations used in this module.

Chapter 0

Review of mathematical notation and foundations

[Week 1]

Intended learning outcomes:

- You are familiar with common notation used throughout the lecture.
- You are comfortable with the main mathematical concepts needed in this module, namely basic probability theory including random variables, conditional probabilities and Markov chains.
- You can apply basic bounds on tail probabilities, and can prove the weak law of large numbers.
- You can compute vector norms and apply the Cauchy-Schwarz inequality.
- You know what convex and concave functions are and can apply Jensen's inequality.

0.1 Notation

We will use standard notation and abbreviations that you should be familiar with from other modules. Some of the less frequently encountered mathematical expressions are summarised in Tables 1 and 2.

0.2 Probability theory

We will not directly need the framework of probability theory in its most abstract formulation as presented in the following, but it is good to know that both discrete and continuous random variables can be seen as emanating from a shared mathematical framework.

0.2.1 Probability space

A probability space is represented by a triple $(\Omega, \Sigma, \mathbb{P})$. Here Ω is a set that is called the *sample space*. Moreover, Σ is a σ-algebra, i.e. a collection of subsets of Ω, called events, with the following properties:

- $\Omega \in \Sigma$
- If $A \in \Sigma$, then its *complement*, $A^c = \Omega \setminus A$ is also in Σ, i.e. $A^c \in \Sigma$.
- If $A_1, A_2, \ldots, A_n, \ldots \in \Sigma$, then $\bigcup_{i=1}^{\infty} A_i \in \Sigma$

Question 0.1. *Show that the above also implies that $\emptyset \in \Sigma$ and $\bigcap_{i=1}^{\infty} A_i \in \Sigma$.*

For example, let $\Omega = [0, 1]$, and we are interested in the probability of subsets of Ω that are intervals of the form $[a, b]$ where $0 \leq a < b \leq 1$, but not individual points in Ω. Then we should also be able to say something about the probability of the union, intersections, complement and so on of such intervals. This is captured by the definition of a σ-algebra. Think of Σ as the properties of Ω that can actually be observed.

Example 0.2. *If your random variable is the location an athlete lands after a long jump then it makes sense to take Ω to be positive real numbers, $\mathbb{R}_+$ indicating the distance jumped (say, in meters). However, even with arbitrarily good equipment we cannot actually measure a real number, we can only ever say that he landed in some interval, the size of which is given by our measurement precision. Thus, Σ, comprised of the events we can actually observe, is built up by including all (arbitrarily small) intervals in $\mathbb{R}_+$ and their unions and complements. Or another way of looking at this is that the probability of the jumper landing exactly at $9m$ is always zero — it is simply the wrong question to ask. But the probability of landing within $1cm$ or some arbitrarily small interval around $9m$ might very well be nonzero.*

Finally, the probability measure $\mathbb{P}$ is a function $\mathbb{P} : \Sigma \to [0, 1]$ defined on the measurable space (Ω, Σ), and represents your "belief" about the events in Σ. In order for $\mathbb{P}$ to be called a probability measure, it must satisfy the following two properties:

1. $\mathbb{P}(\Omega) = 1$
2. For $A_1, A_2, \ldots$ such that $A_i \cap A_j = \emptyset$ for all $i \neq j$, i.e. for mutually *disjoint* sets, we have

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) \tag{1}$$

Some basic and very useful properties that can be derived from the above definition. The union bound in particular is very often used when analysing problems in information theory.

Proposition 0.3. *Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space. The following holds true:*

1. *$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$*

2. *If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$*

3. *$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \leq \mathbb{P}(A) + \mathbb{P}(B)$, which is called the* union bound. *Clearly, by induction, the union bound works for finitely many sets $A_i, i = 1, \ldots, k$, namely*

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i). \tag{2}$$

Proof. Property 1 follows since $A^c \cap A = \emptyset$, and thus $\mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(\Omega) = 1$ by (1). For Property 2, note that $B \setminus A = B \cap A^c \in \Sigma$ and since $A \cap (B \setminus A) = \emptyset$ we again argue that $\mathbb{P}(A) + \mathbb{P}(B \setminus A) = \mathbb{P}(B)$, from which the desired inequality follows.

For Property 3 note that $A \cup B$ can be decomposed in three different ways into mutually disjoint sets:

$$A \cup B = A \cup (B \setminus A) = B \cup (A \setminus B) = (A \setminus B) \cup (B \setminus A) \cup (A \cap B)\,. \tag{3}$$

Again using (1) for each of these decompositions we have

$$\begin{aligned} 2\mathbb{P}(A \cup B) &= \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(B \setminus A) + \mathbb{P}(A \setminus B) & (4)\\ &= \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(A \cup B) - \mathbb{P}(A \cap B)\,, & (5) \end{aligned}$$

which implies the desired equality. □

Question 0.4. *Show that $0 \leq \mathbb{P}(A) \leq 1$ for every $A \in \Sigma$.*

Sometimes we have two conflicting beliefs, or models, about the underlying probability distribution, and so we will consider two compatible probability spaces $(\Omega, \Sigma, \mathbb{P})$ and $(\Omega, \Sigma, \mathbb{Q})$. They offer different predictions about the probability with which the events in Σ occur, and one fundamental task in statistics is to find out which model is the correct one from the frequency with which certain events occur. We will cover this later in the module.

0.2.2 Random variables

We will usually not deal directly with the probability space but with random variables. A *random variable* (rv) $X : \Omega \to \mathcal{X}$ is a function from the space (Ω, Σ) to a measurable space $(\mathcal{X}, \Sigma_X)$. In order for X to make any sense, the mapping has to ensure that $\{\omega \in \Omega : X(\omega) \in \mathcal{B}\} \in \Sigma$ for all $\mathcal{B} \in \Sigma_X$, because we are restricted to observing events in Σ and our random variable can thus not be more fine-grained than what Σ allows. Functions satisfying this property are called a *measurable function*. A random variable is then more formally defined as a measurable mapping from (Ω, Σ) to $(\mathcal{X}, \Sigma_X)$.

The only two examples of interest for us in the following are discrete and continuous random variables:

discrete rv: $\mathcal{X}$ is a discrete set and Σ_X is the power set $\mathcal{P}(\mathcal{X})$ of $\mathcal{X}$, i.e. the set of all subsets of $\mathcal{X}$.

continuous rv: $\mathcal{X} = \mathbb{R}$ and $\Sigma_X = \mathcal{B}$, the Borel σ-algebra. This is the smallest σ-algebra containing all open intervals in $\mathbb{R}$.

The probability measure $\mathbb{P}$ induces a probability measure P_X on $(\mathcal{X}, \Sigma_X)$, given by

$$P_X(B) = P[X \in B] = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\}) \tag{6}$$

for all $B \in \Sigma'$. P_X is called the *distribution* of the random variable X.

If $\mathcal{X} = \{a_1, \ldots, a_d\}$ is discrete (and Σ_X the power set of $\mathcal{X}$), then we say that X is a *discrete random variable.* The distribution of X is then also known as the *probability mass function (pmf)* of X and is fully characterised by all the events consisting of a single value, i.e. the values $P_X(a_1), P_X(a_2), \ldots, P_X(a_d)$.

Question 0.5. *Can you give a formal argument why the values at these points are sufficient?*

Some random variables are not random at all. If there is an a_i with $P_X(a_i) = 1$ (and thus $P_X(a_j) = 0$ for all $j \neq i$), then we call this random variable *deterministic.* On the other extreme we have *uniformly distributed* random variables, where $P_X(a_i) = \frac{1}{d}$ for all $i \in [d]$.

Example 0.6. *The simplest example is the Bernoulli random variable. It is defined on a binary alphabet $\mathcal{X} = \{0,1\}$ and we write $X \sim \mathrm{Bern}(\epsilon)$ to denote the rv with $P[X = 1] = \epsilon$ and $P[X = 0] = 1 - \epsilon$.*

Let us now consider a real-valued random variable X. If there exists a function $p_X : \mathbb{R} \to [0, \infty)$ such that for all $A \in \Sigma_X$, we have

$$P[X \in A] = \int_A p_X(x)\,\mathrm{d}x \tag{7}$$

then we say that X is a *continuous random variable.* The function p_X is called the *probability density function (pdf)* of X. We also define the *cumulative distribution function* (cdf) by integrating $p_X(x)$, that is, the cdf is given by $F_X(a) = \mathbb{P}[X \leq a] = \int_{-\infty}^{a} f_X(x)\,\mathrm{d}x$.

Question 0.7. *Show that $\int_{\mathcal{X}} p_X(x) = 1$. Moreover, if p_X is continuous at some point x, it must satisfy $p_X(x) \geq 0$. Can $p_X(x)$ ever be larger than 1?*

In this class, we deal mainly with discrete rvs, although we will also encounter Gaussian random variables, which are continuous, later on.

Example 0.8. *We denote the pdf of a Gaussian as*

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}. \tag{8}$$

A standard Gaussian distribution *is one in which the mean $\mu = 0$ and the standard deviation $\sigma = 1$. The corresponding cdf is denoted as*

$$\Phi(y) = \int_{-\infty}^{y} \mathcal{N}(x; 0, 1)\,\mathrm{d}x. \tag{9}$$

Some additional notations and definitions for discrete random variables are given below. The counterparts for continuous random variables can be obtained by simply replacing pmfs with pdfs. Thus assume now that X and Y are discrete random variables taking on values in $\mathcal{X}$ and $\mathcal{Y}$ respectively. The joint pmf of X and Y is defined as

$$P_{X,Y}(x,y) = P[X = x \wedge Y = y] = \mathbb{P}\big(\{\omega \in \Omega : X(\omega) = x \wedge Y(\omega) = y\}\big)\,. \tag{10}$$

Question 0.9. *Verify that* $P_Y(y) = \sum_{x' \in \mathcal{X}} P_{X,Y}(x', y)$*.*

With this in hand we can define conditional pmf's and a notion of independence of random variables.

- The conditional pmf is given by

  $$P_{X|Y}(x|y) = \frac{P_{X,Y}(x,y)}{P_Y(y)} = \frac{P_{Y|X}(y|x)P_X(x)}{P_Y(y)}, \quad \text{for} \quad P_Y(y) > 0\,, \tag{11}$$

  where the second expression is often referred to as Bayes' rule. If $P_Y(y) = 0$ then the conditional pmf is simply not defined.

- X and Y are *independent random variables*, if and only if, for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

  $$P_{X,Y}(x,y) = P_X(x)P_Y(y) \tag{12}$$

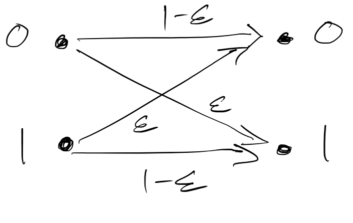
  or equivalently $P_{X|Y}(x|y) = P_X(x)$. The latter condition simply states that the conditional distribution $P_{X|Y}(x|y)$ does not depend on y.

Example 0.10 (Binary symmetric channel)**.** *$X \sim$ Bern(p) is a bit that is sent over channel and is corrupted by additive noise $Z \sim$ Bern(ϵ), where X and Z are independent. The output of the channel is $Y = X \oplus Z$. The channel is fully defined by the conditional distribution $P_{Y|X}$, which we can compute as follows:*

$$P_{Y|X}(y|x) = P[X \oplus Z = y \mid X = x] = P[Z = y \oplus x \mid X = x] = P[Z = y \oplus x] = P_Z(y \oplus x) \tag{13}$$

Hence, the channel can be given as a matrix or pictorially as follows:

x	y	$P_{Y\|X}$
0	*0*	$1-\epsilon$
1	*0*	ϵ
0	*1*	ϵ
1	*1*	$1-\epsilon$



0.2.3 Expectation and variance

The expectation of a random variable X is defined to be

$$\mathbb{E}[X] = \int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}(\omega). \tag{14}$$

This definition has a very precise mathematical meaning in measure theory, but here we are only interested in two special cases. If X is a discrete random variable this reduces to the familiar formula

$$\mathbb{E}[X] = \sum_{x\in\mathcal{X}} xP_X(x)\,. \tag{15}$$

If X is a continuous random variable with pdf $f_X(x)$, we have

$$\mathbb{E}[X] = \int_\mathbb{R} xp_X(x)\,\mathrm{d}x\,. \tag{16}$$

Note that the expectation is a statistical summary of the distribution of X, rather than depending on the realised value of X. If there are two different models $\mathbb{P}$ and $\mathbb{Q}$ we need to specify which probability measure we are using. We only do this when necessary (because the the model is not obvious from context) by adding a subscript $\mathbb{E}_P$ or $\mathbb{E}_Q$.

If g is a function, the expectation of $g(X)$ is given by

$$\mathbb{E}[g(X)] = \int_\mathbb{R} g(x)p_X(x)\,\mathrm{d}x. \tag{17}$$

Question 0.11. *Show that the expectation is linear, i.e.* $\mathbb{E}[aX+b] = a\mathbb{E}[X)+b$.

The variance of X is the expectation of $g(X) = (X-\mathbb{E}[X])^2$. Thus,

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mathbb{E}[X])^2] = \int_\mathbb{R}(x-\mathbb{E}[X])^2 p_X(x)\,\mathrm{d}x\,. \tag{18}$$

Question 0.12. *Check from the above definition that the variance can also be expressed as*

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2. \tag{19}$$

Question 0.13. *Verify that* $\mathcal{N}(x;\mu,\sigma^2)$ *indeed has expectation* μ *and variance* σ^2*.*

0.2.4 Markov chains

Markov chains describe a notion of conditional independence. Let's start with the three random variables X, Y and Z. They are said to form a *Markov chain in the order*

$$X - Y - Z$$

if their joint distribution P_{XYZ} satisfies

$$P_{XYZ}(x,y,z) = P_X(x)P_{Y|X}(y|x)P_{Z|Y}(z|y) \qquad \text{for all} \qquad (x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}. \tag{20}$$

This the same as saying that X and Z are *conditionally independent given* Y.

Question 0.14. *Assume $X - Y - Z$. Show that it is also true that $Z - Y - X$.*

Notice that if we do not assume anything about the joint distribution P_{XYZ}, then it factorizes (by repeated applications of Bayes rule) as

$$P_{XYZ}(x, y, z) = P_X(x)P_{Y|X}(y|x)P_{Z|XY}(z|x, y) \qquad \text{for all} \qquad (x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \tag{21}$$

so what Markovianty in the order $X - Y - Z$ buys us is that $P_{Z|XY}(z|x, y) = P_{Z|Y}(z|y)$ (i.e., we can drop the conditioning on X). In essence all the information that we can learn about Z is already contained in Y. No other information about Z can be gleaned from knowing X if we already know Y. Another way of saying this is that the conditional distribution of X and Z given $Y = y$ can be factorised as

$$P_{XZ|Y}(x, z|y) = P_{X|Y}(x|y)P_{Z|Y}(z|y) \qquad \text{for all} \qquad (x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}. \tag{22}$$

Notice that this is in direct analogy to the situation where X and Z are (marginally) independent. Simply set Y to be a deterministic random variable (with only one possible outcome) to recover the definition of independence.

Question 0.15. *If Z is a deterministic function of Y, show that $X - Y - Z$ is true.*

Question 0.16. *If X and Z are conditionally independent given Y, this does not imply that X and Z are marginally independent (in general). Construct a counterexample.*

0.3 Tail bounds

In this section, we summarise some bounds on probabilities that we use extensively in the sequel. More precisely, we are interested in showing that the probability of a random variable deviating too far from its expectation value is small.

0.3.1 Basic bounds

We start with the familiar Markov and Chebyshev inequalities.

Proposition 0.17 (Markov's inequality)**.** *Let X be a real-valued non-negative random variable with pdf p_X. Then for any $a > 0$, we have*

$$P[X > a] \leq \frac{\mathbb{E}[X]}{a}. \tag{23}$$

Proof. By the definition of the expectation, we have

$$\mathbb{E}[X] = \int_0^\infty x p_X(x) \, \mathrm{d}x \geq \int_a^\infty x p_X(x) \, \mathrm{d}x \geq a \int_a^\infty p_X(x) \, \mathrm{d}x = aP[X > a]. \tag{24}$$

and we are done. □

Note that this bound only becomes nontrivial if a exceeds the expectation value $\mathbb{E}[X]$.

Question 0.18. *In which step is non-negativity of X used?*

Question 0.19. *Can you do the proof also for discrete random variables?*

If we let X above be the non-negative random variable $(X - \mathbb{E}[X])^2$, we obtain Chebyshev's inequality.

Proposition 0.20 (Chebyshev's inequality)**.** *Let X be a real-valued random variable with mean μ and variance σ^2. Then for any $a > 0$, we have*

$$P\big[|X - \mu| > a\sigma\big] \leq \frac{1}{a^2}. \tag{25}$$

Proof. Let X in Markov's inequality be the random variable $g(X) = (X - \mathbb{E}[X])^2$. This is clearly non-negative and the expectation of $g(X)$ is $\mathrm{Var}(X) = \sigma^2$. Thus, by Markov's inequality, we have

$$P[g(X) > a^2\sigma^2] \leq \frac{\sigma^2}{a^2\sigma^2} = \frac{1}{a^2}. \tag{26}$$

Now, $g(X) > a^2\sigma^2$ if and only if $|X - \mu| > a\sigma$ so the claim is proved. □

We now consider a collection of real-valued random variables that are independent and identically distributed (i.i.d.). In particular, let $X^n = (X_1, \ldots, X_n)$ be a collection of independent random variables where each X_i has distribution P with zero mean and finite variance σ^2.

Proposition 0.21 (Weak Law of Large Numbers)**.** *For every $\epsilon > 0$, we have*

$$\lim_{n\to\infty} P\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| > \epsilon\right] = 0. \tag{27}$$

Consequently, the average $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges to 0 in probability.

Note that for a sequence of random variables $\{S_n\}_{n=1}^{\infty}$, we say that this sequence *converges to a number $b \in \mathbb{R}$ in probability* if for all $\epsilon > 0$,

$$\lim_{n\to\infty} P\big[|S_n - b| > \epsilon\big] = 0. \tag{28}$$

We also write this as $S_n \xrightarrow{\mathrm{p}} b$. Contrast this to convergence of numbers: We say that a sequence of numbers $\{s_n\}_{n=1}^{\infty}$ *converges to a number $b \in \mathbb{R}$* if we have $\lim_{n\to\infty} |s_n - b| = 0$.

Proof. Let $\frac{1}{n}\sum_{i=1}^n X_i$ take the role of X in Chebyshev's inequality. Clearly, the mean is zero. The variance of X is

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{\sigma^2}{n}. \tag{29}$$

Thus, we have

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| > \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2} \longrightarrow 0 \tag{30}$$

as $n \to \infty$, which proves the claim. □

Some further useful bounds are derived in the homework.

0.3.2 Central limit theorem

We can actually say quite a bit more than the weak law of large numbers dictates. If the scaling in front of the sum in the statement of the law of large numbers Proposition 0.21 is $1/\sqrt{n}$ instead of $1/n$, the resultant random variable $\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$ converges in distribution to a Gaussian random variable. As in Proposition 0.21, let X^n be a collection of i.i.d. random variables where each X_i is zero mean with finite variance σ^2.

Proposition 0.22 (Central limit theorem)**.** *For any $a \in \mathbb{R}$, we have*

$$\lim_{n\to\infty} P\left(\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^n X_i < a\right) = \Phi(a). \tag{31}$$

In other words,

$$\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^n X_i \xrightarrow{\mathrm{d}} Z \tag{32}$$

where $\xrightarrow{\mathrm{d}}$ means convergence in distribution and Z is a standard Gaussian random variable.

For a sequence of random variables $\{S_n\}_{n=1}^\infty$, we say that this sequence of random variables *converges in distribution* to another random variable $\bar{S}$ if

$$\lim_{n\to\infty} P(S_n < a) = P(\bar{S} < a)$$

for all $a \in \mathbb{R}$. The proof of this statement requires tools that are outside the scope of these notes, but can be found in any textbook on probability theory.

0.4 Vector norms and Cauchy-Schwarz inequality

We can naturally interpret pmf's on an alphabet with d symbols as row vectors in a d-dimensional inner-product space. Without loss of generality we take the alphabet to be $\mathcal{X} = \{1, 2, \ldots, d\} = [d]$ and define the vector $p \in \mathbb{R}^d$ by its elements $p_x = P_X(x)$ for $x \in [d]$. The inner product is denoted by $\langle \cdot, \cdot \rangle$. For two general vectors $u, v \in \mathbb{R}^d$, it evaluates to

$$\langle u, v \rangle = uv^T = \sum_{i=1}^{d} u_i v_i \,, \tag{33}$$

where v^T denotes the transpose of the vector v, and is a column vector. The Cauchy-Schwarz inequality then states that for any two vectors we have

$$|\langle u, v \rangle|^2 \leq \langle u, u \rangle \langle v, v \rangle \,. \tag{34}$$

On these vector spaces we can also define the p-norms for $p \geq 1$ as

$$\|u\|_p = \left(\sum_{x=1}^{d} |u_x|^p \right)^{\frac{1}{p}} \tag{35}$$

We will mostly encounter the 1-norm and the 2-norm, the latter being the usual Euclidian norm of the vector. The following special case of the Cauchy-Schwarz inequality will be encountered later.

Lemma 0.23. *Let $u, v \in \mathbb{R}^d$. Then,*

$$\|u \cdot v\|_1 \leq \|u\|_2\| \, \|v\|_2 \,, \tag{36}$$

where $\cdot$ denotes the element-wise product of the vectors, i.e. $(u \cdot v)_i = u_i v_i$.

Proof. Define $k \in$ using $k_i = \mathrm{sgn}^*(u_i v_i)$, where sgn^* is the modified sign function, i.e.

$$\mathrm{sgn}^*(x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases} \,. \tag{37}$$

Then since $\mathrm{sgn}^*(x)^2 = 1$ for all $x \in \mathbb{R}$, the Cauchy-Schwarz inequality yields

$$\big|\langle k \cdot u, v \rangle\big| \leq \langle u, u \rangle \langle v, v \rangle = \|k \cdot u\|_2 \|v\|_2 = \|u\|_2 \|v\|_2 \,. \tag{38}$$

Moreover, we have

$$\langle k \cdot u, v \rangle = \sum_{x=1}^{d} k_i u_i v_i = \sum_{x=1}^{d} |u_i v_i| = \|u \cdot v\|_1 \,. \tag{39}$$

□

Question 0.24. *Using the above, can you show that $\|u\|_1 \leq \sqrt{d} \|u\|_2$?*

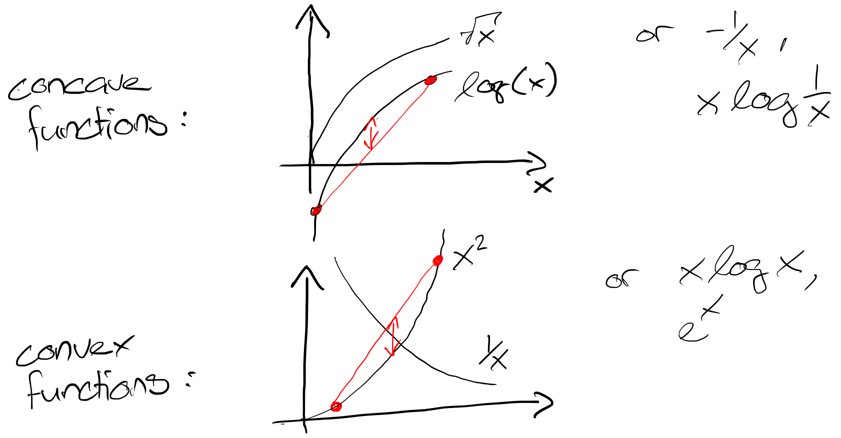


Figure 0.1: Examples of concave (convex) functions. The straight line between two points of the curve is below (above) the plot of the function, which is exactly what the definition requires.

0.5 Convexity and Jensen's inequality

A function $f(x)$ is said to be *convex* on [a, b] if for all $x, y \in [a, b]$ and $\lambda \in [0, 1]$, we have

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y). \tag{40}$$

If we do not mention any interval then we mean that the function is convex on its full domain, i.e. the statement $\log(x)$ is concave should be understood as $\log(x)$ is concave on $(0, \infty)$.

The function f is *strictly convex* if equality in (40) holds only if $\lambda = 0$ or 1, or $x = y$. The function f is *concave* if $-f$ is convex, and *strictly concave* if $-f$ is strictly convex.

In the homework you will show the following lemma:

Lemma 0.25. *If f is convex on $[a, b]$, then for any $a \leq x_1 < x_2 \leq x_3 < x_4 \leq b$, we have*

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(x_4) - f(x_3)}{x_4 - x_3} \tag{41}$$

Proposition 0.26 (Jensen's inequality)**.** *If $f(x)$ is convex and X is a random variable on $\mathbb{R}$, then*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)] \tag{42}$$

We only give a proof for discrete distributions here.

Proof. We give a proof by induction. Due to convexity, we have

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2), \tag{43}$$

which proves the statement if $|\mathcal{X}| = 2$.

Suppose the statement $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$ is true when $|\mathcal{X}| = k-1$. Then consider a pmf with k mass points $\{p_1, p_2, \ldots, p_k\}$. Define another pmf on $k-1$ points given by the probabilities

$$p_i' = \frac{p_i}{1-p_k}, \quad i = 1, \ldots, k-1\,. \tag{44}$$

We then have

$$\begin{aligned}
\sum_{i=1}^{k} p_i f(x_i) &= p_k f(x_k) + (1-p_k) \sum_{i=1}^{k-1} p_i' f(x_i) && (45)\\
&\geq p_k f(x_k) + (1-p_k) f\left(\sum_{i=1}^{k-1} p_i' x_i\right) && (46)\\
&\geq f\left(p_k x_k + (1-p_k) \sum_{i=1}^{k-1} p_i' x_i\right) && (47)\\
&= f\left(\sum_{i=1}^{k} p_i x_i\right) && (48)
\end{aligned}$$

where the first inequality is from the induction hypothesis and the second by convexity (of two points). By the definition of expectation we have $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$. □

Often it is hard to check convexity directly. But for twice differentiable functions, this is easy.

Proposition 0.27. *Let $f : [a,b] \to \mathbb{R}$ be twice differentiable. The function f is convex if and only if $f''(x) \geq 0$ for all $x \in (a,b)$, and strictly convex if $f''(x) > 0$ for all $x \in (a,b)$.*

Proof. Assume $f''(x) > 0$ for all $x \in [a,b]$. By Taylor expansion of f around $x_0 \in (a,b)$, we have

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x^*)}{2}(x - x_0)^2 \tag{49}$$

where $x^* \in [x_0, x]$. By assumption $f''(x^*) > 0$ so the quadratic term is strictly positive unless $x = x_0$, in which case it is still non-negative. Now let $x_0 = \lambda x_1 + (1-\lambda) x_2$. Further let $x = x_1$. Then we have

$$f(x_1) \geq f(x_0) + f'(x_0)((1-\lambda)(x_1 - x_2)). \tag{50}$$

Now let $x = x_2$. Then we have

$$f(x_2) \geq f(x_0) + f'(x_0)(\lambda(x_2 - x_1)). \tag{51}$$

Both of these inequalities are strict unless $\lambda \in [0,1]$ or $x_1 = x_2$. Multiplying the first inequality by λ and the second by $1-\lambda$ and adding them up, we recover the definition of strict convexity. If we instead had assumed only $f''(x) \geq 0$ the same argument would ensure convexity (but no longer strict convexity).

For the other direction, choose $a < x_1 < x_2 < x_3 < x_4 < b$. By the property shown in Lemma 0.25,

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq \frac{f(x_4) - f(x_3)}{x_4 - x_3} \tag{52}$$

Now let $x_2 \searrow x_1$ and $x_3 \nearrow x_4$. We see that $f'(x_1) \leq f'(x_4)$, and since these were arbitrary points, f' is increasing on (a,b). So $f''(x) \geq 0$ for all $x \in (a,b)$. □

0.6 Finite field arithmetic

This is a rather informal discussion, but it is sufficient for our purposes.

A finite field is a field (on which addition and multiplication is defined) with a finite number of elements. Such fields are denoted by F_q where q is the number of elements in the field, or its *dimension*. Such fields exist only for particular numbers of elements, for example when $q = p^\ell$ for some prime p and $\ell \in \mathbb{N}$. For our purposes, we mostly need that addition and multiplication are defined and follow the usual rules we are used to from $\mathbb{R}$ or $\mathbb{C}$ (like associativity and distributivity) and that there exist multiplicative and additive identities and inverses. For F_q where q is prime we can always simply denote the elements of F_q by the integers $\{0, 1, \ldots, q-1\}$ and use integer addition and multiplication modulo q as our operations.

Question 0.28. *Can you find the inverse of* 2 *if* q *is prime? Use that* $q+1$ *is even...*

When q is not a prime but a prime power we can construct the arithmetic by constructing a polynomial ring. To see how this works with an example, let us look at $q = 2^2 = 4$. Let us first try something that does not work. As for primes, we may denote the elements by $\{0,1,2,3\}$ and use addition modulo 4, which works fine, but the problem is that using multiplication modulo 4 we would, for example, get

$$2 \times 0 = 0 \tag{53}$$
$$2 \times 1 = 2 \tag{54}$$
$$2 \times 2 = 4 \quad \mod 4 = 0 \tag{55}$$
$$2 \times 3 = 6 \quad \mod 4 = 2\,, \tag{56}$$

and hence 2 does not have a multiplicative inverse.

The arithmetic is thus defined differently, by interpreting $\{0,1,2,3\}$ instead as the polynomials, 0, 1, x and $x+1$, respectively. We can add these polynomials modulo 2 for each coefficient individually, so in particular for the binary case the negation of each number is just the number itself. For multiplication, we simply do this modulo an irreducible polynomial, for example $x^2 + x + 1$ in this case. So for the above labelings $\{0, 1, 2, 3\}$ of elements (which, admittedly, is more confusing than helpful), we get

$$\begin{aligned}
&2 \times 0 \to x \times 0 = 0 \to 0 && \Longrightarrow\ 2 \times 0 = 0 \qquad (57)\\
&2 \times 1 \to x \times 1 = x \to 2 && \Longrightarrow\ 2 \times 1 = 2 \qquad (58)\\
&2 \times 2 \to x \times x = x^2 \mod x^2 + x + 1 = -x - 1 \to 3 && \Longrightarrow\ 2 \times 2 = 3 \qquad (59)\\
&2 \times 3 \to x \times (x + 1) = x^2 + x \mod x^2 + x + 1 = -1 \to 1 && \Longrightarrow\ 2 \times 3 = 1\,. \qquad (60)
\end{aligned}$$

Hence, 2 and 3 are multiplicative inverses of each other. The full addition and multiplication tables can then be written down as follows:

+	0	1	2	3
0	0	1	2	3
1	1	0	3	2
2	2	3	0	1
3	3	2	1	0

×	0	1	2	3
0	0	0	0	0
1	0	1	2	3
2	0	2	3	1
3	0	3	1	2

Similar constructions can be done for every prime power, and, quite importantly for practical applications, all of this arithmetic can be implemented highly efficiently.