Learner's Space 2025
Information Theory and Coding
Week 1, Handout 1

Suketu Patni

June 9, 2025

Abstract

For this Learner's Space course in information theory and coding, for the first two weeks we will be focussing on information theory. A strong understanding of probability is quintessential for any serious information theory student. This document is a compilation of some major definitions, theorems, and lemmas related to the same. However, this should not be treated as a first course in probability or even as an incomplete reference to the vast field. The text references provided on Moodle should serve that purpose. The purpose of this document is to fix all the probability theory that we'll be needing, so that it will be a refresher for the ones among you already aware of this and a guide of study for those not that well-acquainted. Beginners, do NOT try and learn probability theory from this document.

On a less serious note, some of you might find this handout a bit "dry". This is by purpose, since I do not wish to spend a lot of time on probability. The second handout of this week (and onwards), when we'll really get into information theory, will naturally be less "dry" and will help you build an intuition of it as well.

Lastly, all the exercises mentioned in this document will not be graded. They are only for your understanding.

1 Probability Spaces

Definition 1.1. *A **σ-algebra** over a non empty set S is a subset $\mathcal{A}$ of the power set of S such that.*

1. $S \in \mathcal{A}$
2. *If some set $T \in \mathcal{A}$ then $T^c = (S \setminus T) \in \mathcal{A}$ i.e. $\mathcal{A}$ is* closed under complementation.
3. *For a sequence of elements of $\mathcal{A}$ say $\{T_i\}_{i\in\mathbb{N}}$, their union $\bigcup\limits_{i\in\mathbb{N}} T_i$ is also in $\mathcal{A}$ i.e. $\mathcal{A}$ is closed under countable union.*

By countable here we mean countably infinite. It is not hard to show that being closed under countable union implies being closed under finite union also (by defining a sequence with all but finitely many members of it being non-empty. The empty set is anyways a member of all σ-algebras by the first two properties). Note that the union of a (possibly infinite) collection of sets is defined as the set of all elements which belong to atleast one of the members of that collection.

Exercise 1.1. *Show that $\sigma-$algebras are closed under countable intersection also.*

Definition 1.2. *A* ***probability space*** *is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where:*

- *Ω is a non-empty set called the* sample space.
- *$\mathcal{F}$ is a σ-algebra over Ω. The elements of $\mathcal{F}$ are called* events.
- *$P : \mathcal{F} \to [0, 1]$ is a function such that:*
  1. *$P(\Omega) = 1$,*
  2. *For any sequence of elements of $\mathcal{F}$, say $\{A_i\}_{i \in \mathbb{N}}$ where $A_i \cap A_j = \phi$ for all $i \neq j$, we have*
  $$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

2 Random Variables

Definition 2.1. *The* ***Borel σ-algebra*** *$\mathcal{B}(\mathbb{R})$ on $\mathbb{R}$ is the smallest σ-algebra over $\mathbb{R}$ that contains all the open intervals (a, b) for $a, b \in \mathbb{R}$, $a < b$. The members of $\mathcal{B}(\mathbb{R})$ we shall refer to as Borel sets.*

It is smallest in the sense that every σ-algebra on $\mathbb{R}$ that contains the open intervals of the above form contains the Borel $\sigma-$algebra. We will not prove its existence/uniqueness here.

Exercise 2.1. *Show that all closed intervals, half-open-half-closed intervals, finite subsets of $\mathbb{R}$, countable subsets of $\mathbb{R}$ are all Borel sets. Read up on the Vitali set, perhaps the most famous example of a set that is not Borel.*

Definition 2.2. *Let $(\Omega, \mathcal{F}, P)$ be a probability space. A* ***random variable*** *is a measurable function $X : \Omega \to \mathbb{R}$, meaning that for every Borel set $B \subseteq \mathbb{R}$, the preimage $X^{-1}(B) \in \mathcal{F}$.*

Definition 2.3. *A random variable X is said to be* ***discrete*** *if its image $\mathcal{X}$ is finite or countable. In this case, we define the* ***support*** *of the random variable X as the set $\{x \in \mathbb{R} : P(X = x) > 0\}$. Also, we commonly refer to the function $x \mapsto P(X = x)$ as a* ***probability mass function****.*

The notation $P(X = x)$ is shorthand for $P(\{w \in \Omega : X(w) = x\})$. So anytime we assign a probability to a random variable being equal to (or less than or greater than) some real number, we really are referring to the value of the probability function applied on the set of all elements of the sample space for which the real value that the random variable maps those elements to is equal to (or less than or greater than) that real number.

Definition 2.4. *A random variable X is said to be* ***continuous*** *if there exists a continuous function $F_X : \mathbb{R} \to [0,1]$ such that $F_X(x) = P(X \leq x)$. Such a function is called the* ***cumulative distribution function*** *(or cdf) of X. In case F_X is differentiable on $\mathbb{R}$, its derivative (denoted as f_X) is called the* ***probability density function*** *(or pdf).*

Note that cdfs are defined for all random variables. In case they are continuous, the random variable is said to be continuous. In any case cdfs are non-decreasing, which means that pdfs (if they exist) are always non-negative. Further note that for all Borel sets $A \subseteq \mathbb{R}$, for a continuous random variable X,

$$P(X \in A) = \int_A \mathrm{d}F_X(x)$$

where the integral is to be understood in the Riemann-Stieltjes sense (the reader is encouraged to read up on this and see the similarities it has with regular Riemann integration. We do not bother to define it here). In case X has a density, $dF_X(x)$ may be replaced by $f_X(x)dx$ for convenience.

Exercise 2.2. *Construct a random variable that is neither discrete nor continuous.*

Exercise 2.3. *Show that a continuous random variable cannot assume any given real number with positive probability.*

Exercise 2.4. *There is a large number of commonly used cdfs like Bernoulli, Poisson, exponential, Gaussian, gamma etc. Get familiar with their cdfs, and if they have so, their densities.*

3 Independence

Definition 3.1. *Two random variables X and Y defined on the same probability space are said to be* ***independent*** *if their joint cdf $F_{XY}(x, y) = P(X \leq x, Y \leq y)$ is equal to the product of the individual cdfs $F_X(x)F_Y(y) = P(X \leq x)P(Y \leq y)$.*

Definition 3.2. *For two elements A and B of $\mathcal{F}$ with $P(B) \neq 0$, we define*

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

to be the probability of "A given B".

Exercise 3.1. *Read up on Bayes' theorem and develop a practical intuition for it.*

4 Expectation

Definition 4.1. *Let X be a random variable on $(\Omega, \mathcal{F}, P)$.*

- *If X is discrete with probability mass function $p(x)$, then*

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x p(x),$$

where $\mathcal{X}$ is the support of X, provided the sum converges. In case it converges we say that the expectation is finite.

- *If X is continuous with cdf $F_X(x)$, then*

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \, dF_X(x),$$

provided the integral exists.

Expectation is a linear operator on random variables i.e. it is both additive and homogenous.

Definition 4.2. *The variance of a random variable X is*

$$var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

and its standard deviation is $\sigma[X] = \sqrt{var[X]}$.

The variance can be shown to be nonnegative always. An equivalent definition of variance is $\mathbb{E}[(X - \mathbb{E}[X])^2]$

Exercise 4.1. *Prove that for two independent random variables X and Y, $var[X + Y] = var[X] + var[Y]$*

Exercise 4.2. *Show that for a non-negative continuous random variable X,*

$$\mathbb{E}[X] = \int\limits_0^\infty (1 - F_X(x)) \, dx$$

5 Basic Concentration Inequalities

Theorem 5.1 (Markov's Inequality)**.** *Let X be a non-negative random variable with finite expectation. Then, for any $a > 0$,*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Proof. Left as an exercise to the reader. For a hint, try splitting the region of integration $[0, \infty)$ as $[0, a] \cup (a, \infty)$. □

Theorem 5.2 (Chebyshev's Inequality)**.** *Let X be a random variable with finite expectation $\mu = \mathbb{E}[X]$ and finite variance σ^2. Then, for any $\varepsilon > 0$,*

$$\mathbb{P}(|X - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

Proof. Also left as an exercise to the reader. Markov's inequality proves useful here. □

6 Convergence of Random Variables

Let $\{X_n\}_{n \in \mathbb{N}}$ be a sequence of real-valued random variables defined on a common probability space $(\Omega, \mathcal{F}, P)$, and let X be another random variable on the same probability space.

6.1 Everywhere Convergence

We say $X_n \to X$ *everywhere* or *pointwise*, denoted as

$$X_n \xrightarrow{\text{e}} X,$$

if

$$\lim_{n \to \infty} X_n(w) = X(w)$$

for all $w \in \Omega$. This is the strongest mode of convergence.

6.2 Almost Sure Convergence

We say $X_n \to X$ *almost surely* (or almost everywhere), denoted as

$$X_n \xrightarrow{\text{a.s.}} X,$$

if

$$P\left(\left\{w \in \Omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)\right\}\right) = 1.$$

6.3 Convergence in Probability

We say $X_n \to X$ *in probability*, denoted as

$$X_n \xrightarrow{P} X,$$

if for all $\varepsilon > 0$,

$$\lim_{n \to \infty} P(|X_n - X| > \varepsilon) = 0.$$

6.4 Convergence in L^p

For $p \geq 1$, we say that $X_n \to X$ *in* L^p, denoted as

$$X_n \xrightarrow{L^p} X,$$

if

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^p] = 0.$$

On a sidenote, the notation L^p comes from the identical nomenclature of function spaces in measure theory.

6.5 Convergence in Distribution

We say that $X_n \to X$ *in distribution*, denoted as

$$X_n \xrightarrow{d} X,$$

if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

for all points x where F_X is continuous.

6.6 Relationships Between Modes of Convergence

- $X_n \xrightarrow{\text{e}} X \Rightarrow X_n \xrightarrow{\text{a.s.}} X \Rightarrow X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{d} X$
- $X_n \xrightarrow{L^p} X \Rightarrow X_n \xrightarrow{P} X$

None of the converse implications hold in general.

7 Laws of Large Numbers

We will look at these without proof. The term “i.i.d” is shorthand for ”independent and identically distributed” i.e. when we say that a sequence $\{X_i\}_{i\in\mathbb{N}}$ is i.i.d, we mean that the cdfs of all the X_i are the same, and that the joint cdf of any finite number of them is the product of their cdfs.

Theorem 7.1 (Weak Law of Large Numbers)**.** *Let $\{X_i\}_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables with finite expectation. Then,*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mathbb{E}[X_1],$$

Theorem 7.2 (Strong Law of Large Numbers)**.** *Let $\{X_i\}_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables with finite expectation. Then,*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mathbb{E}[X_1],$$

Exercise 7.1. *Develop some intuition for the laws of large numbers and the distinctions between the various modes of convergence.*