

## Handout 1: Introduction

*Instructor: Shashank Vatedka*

**Disclaimer:** *These notes have not been subjected to the usual scrutiny reserved for formal publications. Please email the course instructor in case of any errors.*

## 1.1 Introduction to information theory

Welcome to this course on information theory! This course will introduce you to fundamental concepts that deal with the processing of information. The principles introduced in this course are useful in a variety of applications including

- Data compression
- Digital communication
- Cryptography
- Gambling/investment
- Machine learning
- Deriving bounds in computer science
- Physics
- Neuroscience, bioinformatics, ...

In this course, we will see a number of quantities used to measure information, and their connections to data compression and digital communication.

## 1.2 Uncertainty and information

What is uncertainty? What is information? Can we measure information?

- Heuristically, more randomness implies more uncertainty
- Example: predicting the next letter in an English word

Q-----

probably 'U'

A-----

more unsure

- We gain information if there is some reduction in uncertainty
  - Saying that Q is followed by U gives less information than saying that A is followed by N.



### 1.2.1 Randomness and data compression

- More random a file is, higher the compressed file size, i.e., more the uncertainty, the harder it is to compress
- Intuitively, more uniform a source is, the more “random” it is.
  - E.g.: A fair coin vs a biased coin with  $\Pr[H] = 0.9$ .

## 1.3 Communication over a noisy channel

Communication over wireless/wireline media is hampered by: (a) attenuation (b) noise<sup>1</sup>. The simplest model used in practice is the following

$$y(t) = ax(t) + n(t)$$

where  $y(t)$  is the received signal,  $x(t)$  is the transmitted signal, and  $n(t)$  is a noise signal. The variance of  $n(t)$  is bounded, and known (through estimation) in practice. The attenuation factor  $a$  typically depends on the distance between the transmitter and receiver (in many situations, inversely proportional to the square of the distance).

In such a scenario, how do we ensure reliable long-range communication?

- Increasing noise power improves reliability, but not practical.
- Solution used prior to the 80’s: repeaters. These act as relays, which amplify the received signal and retransmit. But this amplifies both signal and noise.
- If the signal power and noise power are fixed, can we obtain an arbitrarily small probability of error?
- (Shannon, 1948): YES! Using channel coding.

<sup>1</sup>There are other factors as well, such as self and inter-user interference, but we will ignore that for now.

## 1.4 Refresher in probability

Some references to refresh your memory:

- Papoulis, *Probability, random variables and stochastic processes*
- Ross, *Introduction to probability models*

Concepts that you will need:

- What is a random variable
- Discrete and continuous random variables
- Probability mass function, probability density function, cumulative distribution function
- Expectation, conditional expectation, properties of expectation
- Variance, standard deviation
- Moments, moment generating function

The following bound is frequently used:

**Lemma 1.1** (Union bound). *If  $\mathcal{E}_1$  and  $\mathcal{E}_2$  are two events, then*

$$\Pr[\mathcal{E}_1 \cup \mathcal{E}_2] \leq \Pr[\mathcal{E}_1] + \Pr[\mathcal{E}_2].$$

## 1.5 Concentration inequalities

A concentration inequality typically tries to answer the following question: What is the probability that a random variable deviates significantly from its mean?

The following results are extremely useful in several situations:

**Lemma 1.2** (Markov's inequality). *Suppose that  $X$  is a nonnegative random variable, and  $\mathbb{E}X = \mu > 0$ . Then, for all  $a > 0$ , we have*

$$\Pr[X \geq a] \leq \frac{\mu}{a}.$$

*Proof.* Suppose that  $X$  has pdf  $f$ . Then,

$$\begin{aligned} \mu = \mathbb{E}X &= \int_{t=0}^{\infty} x f(x) dx \\ &= \int_{t=0}^a x f(x) dx + \int_{t=a}^{\infty} x f(x) dx \\ &\geq \int_{t=a}^{\infty} x f(x) dx && \text{(why?)} \\ &\geq a \int_{t=a}^{\infty} f(x) dx && \text{(why?)} \\ &= a \Pr[X \geq a]. \end{aligned}$$

Rearranging the above gives the desired inequality. □

The above lemma is used as a starting point in proving many other inequalities.

**Lemma 1.3** (Chebyshev inequality). *Suppose that  $X$  is a random variable with mean  $\mu$  and variance  $\sigma^2$ . Then, for any  $a > 0$ , we have*

$$\Pr[|X - \mu| > a] \leq \frac{\sigma^2}{a^2}$$

*Proof.* We can write

$$\Pr[|X - \mu| > a] = \Pr[(X - \mu)^2 > a^2]$$

We can now use Markov's inequality to bound the right hand side.

$$\Pr[|X - \mu| > a] \leq \frac{\mathbb{E}[(X - \mu)^2]}{a^2} \tag{1.1}$$

$$= \frac{\sigma^2}{a^2}. \tag{1.2}$$

□

**Lemma 1.4** (Chernoff bound). *If  $X$  is a random variable with mean  $\mu$ , then for every  $a > 0$ , we have*

$$\Pr[X \geq \mu + a] \leq \min_{t>0} \frac{\mathbb{E}e^{t(X-\mu)}}{e^{ta}}$$

$$\Pr[X \leq \mu - a] \leq \min_{t>0} \frac{\mathbb{E}e^{-t(X-\mu)}}{e^{-ta}}$$

*Proof.* Left as exercise. □

Use the above to prove the following result for Bernoulli random variables:

**Lemma 1.5.** *If  $X^n$  is an iid random sequence with Bernoulli( $p$ ) components, then*

$$\Pr\left[\frac{1}{n} \sum_{i=1}^n X_i \geq p(1 + \delta)\right] \leq \left(\frac{e^\delta}{(1 + \delta)^{1+\delta}}\right)^{np} \leq e^{-\frac{\delta^2 np}{3}}$$

$$\Pr\left[\frac{1}{n} \sum_{i=1}^n X_i \leq p(1 - \delta)\right] \leq \left(\frac{e^{-\delta}}{(1 - \delta)^{1-\delta}}\right)^{np} \leq e^{-\frac{\delta^2 np}{3}}$$

for any  $0 \leq \delta \leq 1$ .

To prove the above, you can use the inequality<sup>2</sup>  $\frac{2\delta}{2+\delta} \leq \log(1 + \delta)$ .

### 1.5.1 Sequences of random variables

Random variables  $X_1$  and  $X_2$  are independent if

$$\Pr[X_1 \in \mathcal{A}, X_2 \in \mathcal{B}] = \Pr[X_1 \in \mathcal{A}] \times \Pr[X_2 \in \mathcal{B}]$$

for all  $\mathcal{A}, \mathcal{B}$ .

<sup>2</sup>See [https://en.wikipedia.org/wiki/List\\_of\\_logarithmic\\_identities](https://en.wikipedia.org/wiki/List_of_logarithmic_identities) for more such identities.

We say that a sequence of random variables  $X^n \stackrel{\text{def}}{=} X_1, X_2, \dots, X_n$  (where each  $X_i$  is drawn from a finite alphabet  $\mathcal{X}$ ) is independent and identically distributed according to  $p_X$  (abbreviated as iid  $\sim p_X$ ) if

$$\Pr[X_1 \in \mathcal{A}_1, X_2 \in \mathcal{A}_2, \dots, X_n \in \mathcal{A}_n] = p_X(\mathcal{A}_1) \times p_X(\mathcal{A}_2) \times \dots \times p_X(\mathcal{A}_n) = \prod_{i=1}^n p_X(\mathcal{A}_i),$$

for all  $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_n$ .

We say that a sequence of random variables is a first-order time-homogeneous Markov chain with transition kernel  $p_{X_2|X_1}$  and initial distribution  $\pi$  if for all  $x^n \in \mathcal{X}^n$ , we have

$$\Pr[X^n = x^n] = \pi(x_1) \prod_{i=2}^n p_{X_2|X_1}(x_i|x_{i-1}).$$

We say that a sequence of random variables  $X^n$  forms a time-homogeneous  $k$ -th order Markov chain with transition kernel  $p_{X_k|X_1 \dots X_{k-1}}$  and initial distribution  $\pi_k$  if

$$\Pr[X^n = x^n] = \pi_k(x^k) \prod_{i=k+1}^n p_{X_k|X_1 \dots X_{k-1}}(x_i|x_{i-k} \dots x_{i-1})$$

for every  $x^n \in \mathcal{X}^n$ .

### 1.5.1.1 Weak law of large numbers

**Lemma 1.6** (Weak law of large numbers). *If  $X^n$  is a sequences of independent and identically distributed random variables with finite mean  $\mu$  and finite variance, then*

$$\lim_{n \rightarrow \infty} \Pr \left[ \left| \frac{1}{n} \sum_{i=1}^n X_i - \mu \right| > \epsilon \right] = 0$$

for all  $\epsilon > 0$ .

#### Note

A sequence of random variables  $X^n$  is said to converge in probability to another random variable  $X$  if for every  $\epsilon > 0$ ,

$$\lim_{n \rightarrow \infty} \Pr[|X_n - X| > \epsilon] = 0.$$

Lemma 1.6 says that  $\frac{1}{n} \sum_{i=1}^n X_i$  converges to  $\mu$  in probability if  $X^n$  is iid.

### 1.5.2 Markov chains

$X^n$  is a first-order time-homogeneous Markov chain with transition probabilities  $p_{X'|X}$  and initial distribution  $p_X$  if

$$\Pr[X^n = x^n] = p_X(x_1) \prod_{i=2}^n p_{X'|X}(x_i|x_{i-1})$$

It is also called a *first-order Markov source*.

Some properties:

- Conditioned on the present, the future is independent of the past, i.e., given  $X_i$ , the random variable  $X_{i+k}$  is independent of  $(X_1, \dots, X_{i-1})$  for all  $i, k$ .
- If  $P$  denotes the transition probability matrix, then the stationary distribution is a pmf  $\pi$  such that

$$\pi P = \pi$$

$X^n$  is a  $k$ -th order Markov source with initial distribution  $p_{X^k}$  and transition probabilities  $p_{X_{k+1}|X^k}$  if

$$p_{X^n}(x^n) = p_{X^k}(x_1, \dots, x_k) \prod_{i=k+1}^n p_{X_{k+1}|X^k}(x_i | x_{i-1}, \dots, x_{i-k})$$

Markov sources are the simplest sources which capture “memory” in the source. These sources have no long-range dependencies. In fact, English text can be roughly approximated by a Markov source.

### 1.5.3 Big-O notation

Frequently in computer science and information theory, we use the big-O notation, or Bachmann-Landau notation<sup>3</sup> to study asymptotics.

Given two sequences in  $n$ , say  $f(n)$  and  $g(n)$ ,

- We say that  $f(n) = O(g(n))$  if there exist positive constants  $\alpha, n_0$  independent of  $n$  such that  $f(n) \leq \alpha g(n)$  for all  $n \geq n_0$ .
- We say that  $f(n) = \Omega(g(n))$  if there exist positive constants  $\alpha, n_0$  independent of  $n$  such that  $f(n) \geq \alpha g(n)$  for all  $n \geq n_0$ .
- We say that  $f(n) = \Theta(g(n))$  if there exist positive constants  $\alpha_l, \alpha_u, n_0$  independent of  $n$  such that  $\alpha_l g(n) \leq f(n) \leq \alpha_u g(n)$  for all  $n \geq n_0$ .
- We say that  $f(n) = o(g(n))$  if  $\lim_{n \rightarrow \infty} f(n)/g(n) = 0$ .
- We say that  $f(n) = \omega(g(n))$  if  $\lim_{n \rightarrow \infty} f(n)/g(n) = \infty$ .

Questions:

- If  $f(n) = O(1)$ , then what can you say?
- If  $f(n) = o(1)$ , then what can you say?
- If  $f(n) = \omega(1)$ , then what can you say?
- If  $f(n) = o(1/n)$ , then what can you say?
- If  $f(n) = 2^{3n(1+o(1))}$ , then what can you say?

<sup>3</sup>See [https://en.wikipedia.org/wiki/Big\\_O\\_notation](https://en.wikipedia.org/wiki/Big_O_notation) for more.

## 1.6 Commonly used notation

Vectors of length  $n$  (typically column vectors) will be denoted with a superscript  $n$ , e.g.,  $x^n, y^n, a^n$ . The  $i$ th component of this vector will be denoted using a subscript, e.g.,  $x_i$  is the  $i$ th component of  $x^n$ .

Random variables will be denoted in uppercase, e.g.,  $X, Y, Z$ . Random vectors in uppercase, with superscript denoting the dimension. E.g.,  $X^n, Y^n, Z^n$ .

Sets are usually denoted in calligraphic uppercase, e.g.,  $\mathcal{A}, \mathcal{B}$ , etc. Special sets:  $\mathbb{R}$  denotes the set of real numbers,  $\mathbb{Z}$  is the set of integers.

The probability mass function of a discrete rv  $X$  will be  $p_X$ , and that of  $X^n$  will be  $p_{X^n}$ . The probability density function of a continuous rv  $X$  will be  $f_X$ .

In this course, we will assume that the continuous random variables in consideration have a density. However, this is not true in general (there exist random variables having a CDF but not a PDF). Most definitions and results studied in this course also carry forward even when the rvs do not have PDF, but this requires some careful redefinition of various quantities that we will not get into.

We will use  $nC_k$  and  $\binom{n}{k}$  interchangeably to denote the binomial coefficient.