EE5847: Information Theory **2022**

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

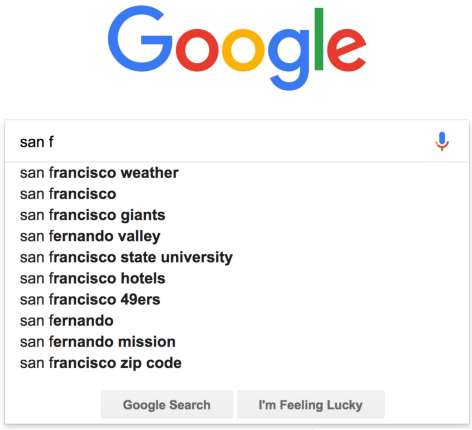
- Data compression: Given a certain probabilistic model for input/raw files, how small can the compressed file be?
- Digital communication: Given a model for a noisy communication medium, what is the maximum rate at which we can communicate under a desired probability of error constraint?
- Cryptography: When do we say that a protocol is perfectly secure?
- Gambling/investment: Given a model for the risk/reward, what is the maximum reward we can obtain?
- Machine learning: What are the fundamental limits on the loss/classification error for a given machine learning problem?
- Algorithms: What are the fundamental limits on the running time/performance for a given computational problem?
- Physics: How much randomness is present in a thermodynamic/quantum system?
- Neuroscience, bioinformatics: How much useful information is transferred from our eye to the brain?
- ...

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[1]. 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?

[1] There are other factors as well, such as self and inter-user interference, but we will ignore that for now.

- 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.

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
- Common probability distributions: Bernoulli, Binomial, Poisson, Uniform, Gaussian, Chi-square ,...

The following bound is frequently used:

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

$$\Pr[\mathcal{E}_1 \bigcup \mathcal{E}_2] \leqslant \Pr[\mathcal{E}_1] + \Pr[\mathcal{E}_2].$$

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*

$$\Pr[X \geqslant a] \leqslant \frac{\mu}{a}.$$

Proof. Suppose that X has pdf f. Then,

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

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 μ and variance σ^2. Then, for any $a > 0$, we have*

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

Proof. We can write

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

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

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

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

□

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

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

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

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*

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

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

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

You can use the general Chernoff bound, and then minimize for t. This gives a complicated looking expression which can then be upper bounded as above.

Here is a better approach: To prove $\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i \geqslant p(1+\delta)\right] \leqslant \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{np}$ and $\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i \leqslant p(1-\delta)\right] \leqslant \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{np}$, compute the moment generating function and then use the following inequality:

$$1 + p(e^t - 1) \leqslant e^{p(e^t-1)}.$$

Using this, we get

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i \geqslant p(1+\delta)\right] \leqslant \frac{e^{(e^t-1)np}}{e^{t(1+\delta)np}}$$

and

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

You can then use the following values for t: $t = \log_e(1+\delta)$ (for the first inequality) and $t = \log_e(1-\delta)$ (second inequality).

To show that $\left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{np} \leqslant e^{-\frac{\delta^2 np}{3}}$ and $\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{np} \leqslant e^{-\frac{\delta^2 np}{3}}$, you only need to show that

$$f(\delta) = \delta - (1+\delta)\log_e(1+\delta) + \frac{\delta^2}{3} \leqslant 0$$

$$g(\delta) = -\delta - (1-\delta)\log_e(1-\delta) + \frac{\delta^2}{3} \leqslant 0$$

Why?

Note that $f(0) = g(0) = 0$. If we can show that f and g are decreasing functions of δ, then we are done. To prove this, compute the slope and use the inequalities[2]

$$\frac{2\delta}{3} \leqslant \frac{2\delta}{2+\delta} \leqslant \log(1+\delta)$$

$$\frac{2\delta}{3} \leqslant -\log(1-\delta)$$

.

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

We say that a sequence of random variables $X^n \stackrel{\text{def}}{=} X_1, X_2, \ldots, X_n$ (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

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

[2]See `https://en.wikipedia.org/wiki/List_of_logarithmic_identities` for more such identities.

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

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 π 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 $p_{X_k|X_1 \ldots X_{k-1}}$ and initial distribution π_k if

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

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 μ and finite variance, then*

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

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 $\epsilon > 0$,

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

Lemma 1.6 says that $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges to μ 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 $(X_1, \ldots X_{i-1})$ for all i, k.

- If P denotes the transition probability matrix, then the stationary distribution is a pmf π 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

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

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 Stationary random sequences

A sequence of random variables $X_1, X_2, \ldots$ is said to be stationary (sometimes called strict-sense stationary) if for every positive integer k and positive integers $i_1 < i_2 < \cdots < i_k$, the joint distribution of $(X_{i_1}, X_{i_2}, \ldots, X_{i_k})$ is the same as $(X_1, X_{i_2-i_1+1}, \ldots, X_{i_k-i_1+1})$.

- Every iid sequence is stationary.
- Not every Markov chain is stationary. However, for specific choices of the initial distribution p_{X^k} (which is called the stationary distribution), we can make a Markov chain stationary.

1.5.4 Big-O notation

Frequently in computer science and information theory, we use the big-O notation, or Bachmann-Landau notation[3] 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 α, n_0 independent of n such that $f(n) \leqslant \alpha g(n)$ for all $n \geqslant n_0$.
- We say that $f(n) = \Omega(g(n))$ if there exist positive constants α, n_0 independent of n such that $f(n) \geqslant \alpha g(n)$ for all $n \geqslant n_0$.
- We say that $f(n) = \Theta(g(n))$ if there exist positive constants α_l, α_u, n_0 independent of n such that $\alpha_l g(n) \leqslant f(n) \leqslant \alpha_u g(n)$ for all $n \geqslant n_0$.
- We say that $f(n) = o(g(n))$ if $\lim_{n\to\infty} f(n)/g(n) = 0$.
- We say that $f(n) = \omega(g(n))$ if $\lim_{n\to\infty} f(n)/g(n) = \infty$.

Questions:

- If $f(n) = O(1)$, then what can you say?
- If $f(n) = o(1)$, then what can you say?

[3] See `https://en.wikipedia.org/wiki/Big_O_notation` for more.

- 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?
- If $f(n) = \log(2+n) - \sqrt{5 \log n}$, then is $f(n) = O(\log n)$? Is $f(n) = \omega(\log n)$? Is $f(n) = o(\log n)$? Is $f(n) = o(n)$?

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 it component of this vector will be denoted using a subscript, e.g., x_i is the ith 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.