Information and Coding Theory **Autumn 2022**

Lecture 1: September 27, 2022

Lecturer: Madhur Tulsiani

1 Administrivia

This course will cover some basic concepts in information and coding theory, and their applications to statistics, machine learning and theoretical computer science.

- The course will have 4-5 homeworks (60 % of the grade) and a final (40 %). The homeworks will be posted on the course homepage and announced in class, and will be due about 7-10 days after they are posted.
- The pre-requisites for the course are familiarity with discrete and continuous probability and random variables, and algorithmic notions. Some knowledge of finite fields will help with the coding theory part though we will briefly review the relevant concepts from algebra.
- We will not follow any single textbook, though the book *Elements of Information Theory* by T. M. Cover and J. A. Thomas is a good reference for most of the material we will cover. The "Resources" section on the course page also contains links to some other similar courses.

2 A quick reminder about random variables and convexity

2.1 Random variables

Let Ω be a finite set. Let $\mu : \Omega \to [0, 1]$ be a function such that

$$\sum_{\omega \in \Omega} \mu(\omega) = 1.$$

We often refer to Ω as a sample space and the function μ as a probability distribution on this space. When Ω is not finite, μ may need to be replaced by an object called a probability measure (we will discuss this later). Ω and μ are together said to define a probability space (for infinite Ω, pobability spaces need an additional component called a σ-algebra).

A real-valued random variable over Ω is any function $X : \Omega \to \mathbb{R}$. We define

$$\mathbb{E}[X] \;=\; \sum_{\omega \in \Omega} \mu(\omega) \cdot X(\omega)\,.$$

We will also think of a random variable X as given by its distribution. If $\mathcal{X}$ is the (finite) set of values taken by X, we can think of the probability distribution on $\mathcal{X}$ given by

$$p(x) \;=\; \mathbb{P}[X = x] \;=\; \sum_{\omega : X(\omega) = x} \mu(\omega)\,,$$

for all values $x \in \mathcal{X}$.

A word on notation

Note that a random variable X is defined simply as a function on Ω, and the distribution of X is induced by the distribution (or measure) μ on Ω. We use the notation $P(X)$ to denote the distribution P for the random variable X. Note that changing the underlying probability space can result in a different distribution (say) $Q(X)$ for the same function X.

In information theory notation, it is common to only talk of distributions P, Q, if they are for the same X which is clear from context. Similarly, it is common to define quantities (such as entropy) which depend on the *distribution*, simply in terms of random variables X, Y when the underlying probability space is fixed. When we need to talk of multiple random variables, and also of multiple distributions for the same variable X, we will use the more explicit notation $P(X)$.

We will use uppercase letters X, Y, Z for random variables, lowercase letters x, y, z to denote values for these random variables, and caligraphic letters $\mathcal{X}, \mathcal{Y}, cZ$ to denote the sets of *possible* values for random variables (also known as the support of a random variable). We will also use uppercase letters P, Q to denote the names of distributions, and lowercase letters p, q to denote probabilities. Thus, a random variable X with distribution $P(X)$ and support $\mathcal{X}$ may satisfy that for a specific value $x \in \mathcal{X}$, we have $p(x) := \mathbb{P}[X = x] = 1/2$.

2.2 Convexity and Jensen's inequality

A set $S \subset \mathbb{R}^n$ is said to be convex subset of $\mathbb{R}^n$ if the line segment joining any two points in S lies entirely in S i.e., for all $x, y \in S$ and for all $\alpha \in [0,1]$, $\alpha \cdot x + (1-\alpha) \cdot y \in S$. For a convex set $S \subseteq \mathbb{R}^n$, a function $f : S \to \mathbb{R}$ is said to be a convex function on S, if for all $x, y \in S$ and for all $\alpha \in [0,1]$, we have

$$f(\alpha \cdot x + (1-\alpha) \cdot y) \;\leq\; \alpha \cdot f(x) + (1-\alpha) \cdot f(y)\,.$$

Equivalently, we say that the function f is convex if the set $S_f = \{(x,z) \mid z \geq f(x)\}$ is a convex subset of $\mathbb{R}^{n+1}$. f is said to be strictly convex when the inequality above is strict for all x, y, α. A function which satisfies the opposite inequality i.e., for all $x, y \in S$ and $\alpha \in [0,1]$

$$f\left(\alpha \cdot x + (1-\alpha) \cdot y\right) \;\geq\; \alpha \cdot f(x) + (1-\alpha) \cdot f(y)\,,$$

is said to be a concave function (and strictly concave if the inequalities are strict). Note that if f is a convex function then $-f$ is a concave function (and vice-versa). For a single variable function $f : \mathbb{R} \to \mathbb{R}$ which is twice differentiable, we can also use the easier criterion that f is convex on $S \subseteq \mathbb{R}$ if and only if $f''(x) \geq 0$ for all $x \in S$. We will frequently use the following inequality about convex functions.

Lemma 2.1 (Jensen's inequality)**.** *Let $S \subseteq \mathbb{R}^n$ be a convex set and let X be a random variable taking values only inside S. Then, for a convex function $f : S \to \mathbb{R}$, we have that*

$$\mathbb{E}\left[f(X)\right] \;\geq\; f\left(\mathbb{E}\left[X\right]\right)\,.$$

Equivalently, for a concave function $f : S \to \mathbb{R}$, we have

$$\mathbb{E}\left[f(X)\right] \;\leq\; f\left(\mathbb{E}\left[X\right]\right)\,.$$

Note that the definition of convexity is the same as the statement of Jensen's inequality for a random variable taking only two values: x with probability α and y with probability $1-\alpha$. You can try the following exercises to familiarize yourself with this inequality.

Exercise 2.2. *Prove Jensen's inequality when the random variable X has a finite support.*

Exercise 2.3. *Check that the $f(x) = x^2$ is a convex function on $\mathbb{R}$. Also show that the functions $\log(x)$ and $x\log(x)$ are respectively, concave and convex functions on $(0,\infty)$.*

Exercise 2.4. *Prove the Cauchy-Schwarz inequality using Jensen's inequality.*

3 Entropy

The concepts from information theory are applicable in many areas as it gives a precise mathematical way of stating and answering the following question: How much information is revealed by the outcome of a random event? Let us begin with a few simple examples. Let X be a random variable which takes the value a with probability $1/2$ and b with probability $1/2$. We can then describe the value of X using one bit (say 0 for a and 1 for b). Suppose it takes one of the values $\{a_1, \ldots, a_n\}$, each with probability, then we can describe the outcome using $\lceil \log_2(n) \rceil$ bits. The n possible outcomes for this random variable each occur with probability $1/n$, and require $\approx \log_2(n)$ bits to describe.

The concept of entropy is basically an extrapolation of this idea when the different outcomes do not occur with equal probability. We think of the "information content" of an event that occurs with probability p as being $\log_2(1/p)$. If a random variable X is distributed over a universe $\mathcal{X} = \{a_1, \ldots, a_n\}$ such that it takes value $x \in \mathcal{X}$ with probability $p(x)$. Then, we define the *entropy* of the random variable X as

$$H(X) \quad = \quad \sum_{x \in \mathcal{X}} p(x) \cdot \log\left(\frac{1}{p(x)}\right) .$$

The following basic property of entropy is extremely useful in applications to counting problems.

Proposition 3.1. *Let X be a random variable supported on a finite set $\mathcal{X}$ as above. Then*

$$0 \;\le\; H(X) \;\le\; \log(|\mathcal{X}|) .$$

Proof: Since $p(x) \le 1$ we have $\log(1/p(x)) \ge 0$ for all $x \in \mathcal{X}$ and hence $H(X) \ge 0$. For the upper bound, consider a random variable Y which takes value $1/p(x)$ with probability $p(x)$. Since $\log(\cdot)$ is a concave function, we use Jensen's inequality to say that

$$\begin{aligned} \sum_{x \in \mathcal{X}} p(x) \cdot \log\left(\frac{1}{p(x)}\right) &= \mathbb{E}\left[\log(Y)\right] \\ &\le \log\left(\mathbb{E}\left[Y\right]\right) \\ &= \log\left(\sum_{x \in \mathcal{X}} p(x) \cdot \frac{1}{p(x)}\right) \;=\; \log(|\mathcal{X}|) . \end{aligned}$$

■