Lecture 3: September 12, 2024

Scribe: Aditi Venkatesh and Jin Ming Koh *Classical information theory*

3.1 Introduction

There are parallels between classical and quantum information theory.

Application	**Classical information theory**	**Quantum information theory**
Data compression	Shannon entropy	von Neumann entropy
Channel coding	Mutual information	Quantum mutual information (for *classical* information over noisy channel); coherent information (for *quantum* information over noisy channel).
Hypothesis testing	Relative entropy	Quantum relative entropy

3.2 Entropy

Entropy is a measure of uncertainty.

3.2.1 Shannon entropy

Definition 1 (Shannon entropy of single variable)**.** *For random variable $X \sim p$ such that $\mathbb{P}(x) = p(x)$, the Shannon entropy of X is*

$$H(X) = H(p) = -\sum_{x \in X} p(x) \log_2 p(x). \tag{3.1}$$

Some properties:

1. *Bounds.* The Shannon entropy satisfies $0 \leq H(X) \leq \log_2 d$ for d the size of the alphabet of X. The lower bound is attained for deterministic X. The upper bound is attained for uniformly random X.

2. *Concavity.* For all $0 \leq \lambda \leq 1$,

$$\lambda H(p_1) + (1-\lambda) H(p_2) \leq H[\lambda p_1 + (1-\lambda) p_2]. \tag{3.2}$$

3. *Norm power series expansion.*

$$\|p\|_{1+\epsilon} = \left(\sum_{x \in X} p(x)^{1+\epsilon} \right)^{\frac{1}{1+\epsilon}} = 1 + \epsilon H(p) + \mathcal{O}\big(\epsilon^2\big). \tag{3.3}$$

Definition 2 (Shannon entropy of two variables)**.** *For random variables $(X, Y) \sim p$ such that $\mathbb{P}(x, y) = p(x, y)$, the Shannon entropy of the joint distribution*

$$H(X, Y) = -\sum_{(x,y)\in(X,Y)} p(x, y) \log_2 p(x, y). \tag{3.4}$$

For a product distribution, $H(X, Y) = H(X) + H(Y)$.

3.2.2 Conditional entropy

Definition 3 (Conditional entropy)**.** *For random variables $(X, Y) \sim p$ such that $\mathbb{P}(x, y) = p(x, y)$, the conditional entropy of Y given X is*

$$\begin{aligned} H(Y|X) &= -\sum_{x\in X} \mathbb{P}(X = x) H(Y|X = x) && (3.5)\\ &= -\sum_{x\in X} p(x) \sum_{y\in Y} p(y|x) \log p(y|x) && (3.6)\\ &= H(X, Y) - H(X), && (3.7) \end{aligned}$$

where we have noted $p(y|x) = p(x, y)/p(x)$.

The physical intuition is that $H(X)$ is the uncertainty of X, whereas $H(Y|X)$ is the uncertainty of Y when we know X. Therefore $H(X, Y) = H(X) + H(Y|X)$.

Remark 1 (Chain rule)**.** *For random variables (X_1, X_2, X_3),*

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2). \tag{3.8}$$

Remark 2 (Non-negativity of conditional entropy)**.** *Classically, $H(Y|X) \geq 0$. But not so quantumly. An example is an EPR pair. Then $H(X, Y) = 0$ as the two subsystems jointly are in a pure state, but $H(X) = H(Y) = 1$ as the reduced density matrix of each subsystem is maximally mixed. That is, quantumly, the joint probability distribution can possess less entropy than its marginal distributions.*

3.2.3 Mutual information

Definition 4 (Mutual information)**.** *The mutual information between random variables (X, Y) is*

$$I(X; Y) = H(X) + H(Y) - H(X, Y) \tag{3.9}$$

The physical interpretation is that $I(X; Y)$ is how much information one learns about X when one looks at Y, or symmetrically, how much information one learns about Y when one looks at X.

Remark 3 (Non-negativity of mutual information)**.** *Both classically and quantumly,*

$$I(X; Y) \geq 0 \iff H(X) + H(Y) \geq H(X, Y) \iff H(Y|X) \leq H(Y). \tag{3.10}$$

3.2.4 Entropy of density matrices

Definition 5 (Shannon entropy of density matrix)**.** *The Shannon entropy of a density matrix ρ is*

$$H(\rho) = -\sum_k \lambda_k \log_2 \lambda_k \tag{3.11}$$

where $\{\lambda_k\}_k$ are the eigenvalues of the matrix.

The entropy $H(\rho)$ is the Shannon entropy of the probability distribution of measurement outcomes obtained when ρ is measured in its eigenbasis.

Remark 4 (Strong sub-additivity)**.** *For quantum systems A, B and C,*

$$H(A) + H(ABC) \leq H(AB) + H(AC). \tag{3.12}$$

3.3 Noiseless coding theorem

Theorem 1 (Noiseless coding theorem)**.** *It is possible to compress n length iid message $x_1, x_2, \ldots, x_n$, from x X, to $nH(X) + o(1)$ bits, with perfect recovery.*

Proof. Lets consider the probability distribution $X := \{x, p(x)\}$ where each letter x_i has probability $p(x_i)$. For an n-letter message,

$$p(x_1 x_2 ... x_n) = \prod_{i=1}^{n} p(x_i)$$

due to iid. Unless X is uniformly random it is possible to compress this distribution to an smaller string. Using the law of large numbers we know that for a string of n letters, x_i typically occurs $np(x_i)$ times. Therefore using Stirling's approximation we can say that the number of typical strings is

$$\frac{n!}{\prod_x (np(x))!} \approx 2^{nH(X)}$$

where,

$$H(X) = -\sum p(x) \log_2 p(x)$$

If we use a block code that relates integers to typical sequences of the n-letter message, then the information in the n-letter string can be conveyed in on average $nH(X)$ bits. We need the $+o(1)$ in order to prove achievability. □

3.4 Noisy coding theorem

Consider now that the channel over which we transmit information is noisy. We encode our input message, pass the encoded message over the channel, and decode at the destination.

Definition 6 (Rate)**.** *Using a message of length n sent over the channel to encode a message of length k, the rate of the transmission is $R = k/n$.*

Definition 7 (Channel capacity)**.** *Consider a noisy channel which receives random variable X as input and outputs random variable Y. Then the channel capacity is*

$$C = \max_X I(X; Y), \tag{3.13}$$

where the maximization is performed over the input random variable X.

Theorem 2 (Noisy coding theorem)**.** *Consider sending a message of length n over a noisy channel at rate R. It is possible to do so with vanishing probability of error as $n \to \infty$ as long as $R < C$ where C is the capacity of the channel. Otherwise, the probability of error approaches unity as $n \to \infty$.*

Proof. Consider a discrete memoryless channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$. The channel is characterized by a conditional probability distribution $P(y|x)$, which gives the probability of receiving symbol $y \in \mathcal{Y}$ when $x \in \mathcal{X}$ is sent from Alice to Bob. R and C of the code are defined as

$$R = \frac{\log_2 M}{n},$$
$$C = \max_{P(x)} I(X;Y),$$

where M is the number of codewords, and n is the length of each codeword. Construct a random codebook by selecting $M = 2^{nr}$ codewords $x_1, x_2, \ldots, x_M$ independently and uniformly from $\mathcal{X}^n$ according to the distribution $P(x)$. Bob observes the output $y^n \in \mathcal{Y}^n$ and decodes the received message to one of the M possible codewords. The goal is to show that for $r < C$, the probability of error can be made arbitrarily small as $n \to \infty$. Define the jointly typical set $T_\epsilon^{(n)}(X,Y)$ as the set of pairs (x^n, y^n) such that:

$$\left| -\frac{1}{n} \log P(x^n) - H(X) \right| < \epsilon,$$
$$\left| -\frac{1}{n} \log P(y^n) - H(Y) \right| < \epsilon,$$
$$\left| -\frac{1}{n} \log P(x^n, y^n) - I(X;Y) \right| < \epsilon.$$

Decoding is performed by finding the unique codeword x_i^n such that the pair (x_i^n, y^n) is jointly typical.

The probability of error can be decomposed into two types: 1. No codeword x_i^n is jointly typical with y^n. 2. There exists a codeword x_j^n (with $j \neq i$) that is jointly typical with y^n.

The probability of the first type of error vanishes as $n \to \infty$, by the law of large numbers and the properties of typical sets. For the second type of error, using the union bound and the independence of codewords, we get:

$$P(\text{error}) \leq P(\text{incorrect decoding}) \leq (M-1)P(\text{codeword typical with } y^n).$$

Since the number of codewords $M = 2^{nr}$ and the probability that a randomly chosen codeword is jointly typical with y^n is approximately $2^{-nI(X;Y)}$, the probability of error is bounded by:

$$P(\text{error}) \leq (M-1)2^{-nI(X;Y)} \approx 2^{n(R-I(X;Y))}.$$

Thus, for $R < C$, the probability of error tends to zero as $n \to \infty$. Conversely, if $r > C$, the probability of error approaches 1 as $n \to \infty$.

Therefore, reliable communication over a noisy channel is possible at any rate $R < C$, and the probability of error can be made arbitrarily small. Conversely, for rates $R > C$, the probability of error approaches 1.

□