Lecture 4: Sep 10, 2020

Lecturer: Aram Harrow *Scribe: Annie Wei, Zane Rossi*

4.1 Information Theory

Classical information theory

- Shannon entropy, typical sets, and compression
- Mutual information and noisy channel coding
- Relative entropy and hypothesis testing

Quantum information theory

- von Neumann entropy, Schumacher-Jozsa compression
- Mutual information and HSW coding
- Relative entropy and hypothesis testing
- Quantum capacity and LSD theorem

4.1.1 Entropy

For random variable $X \sim p$:

$$H(X) = H(p) = -\sum_x p(x) \log p(x)$$

Quantifies uncertainty: for d the alphabet size of X,

$$0 \leq H(X) \leq \log d,$$

where lower bound corresponds to deterministic $p = (0, 0, 1, 0, 0)$, upper bounds corresponds to uniform $p = (1/d) \cdot (1, 1, 1, 1, 1)$. Note $0 \log 0 = 0$.

Note: ℓ_α norms work also, i.e.

$$\|p\|_{1+\epsilon} = 1 - \epsilon H(p) + O(\epsilon^2)$$

But $\|p\|_0$, $\|p\|_2$, $\|p\|_\infty$ also valid.

In the case of binary entropy, for $\Pi \in [0,1]$,

$$H_2(\Pi) = H\begin{pmatrix} \Pi \\ 1-\Pi \end{pmatrix}$$

4.1.1.1 Convexity Properties

Note that $H(p)$ is concave:

$$H(\Pi p + (1-\Pi)q) \geq \Pi H(p) + (1-\Pi)H(q)$$

This inequality is maximized by the uniform distribution. For example, assume that $(0.51, 0.49)$ maximizes entropy. Then $(0.49, 0.51)$ also does. But $H(\text{uniform}) \geq (1/2)H((0.51, 0.49)) + (1/2)H((0.49, 0.51))$.

We can also consider the convexity/concavity properties of fidelity and trace distance. In particular, fidelity is jointly concave:

$$F(\Pi\rho_1 + (1-\Pi)\rho_2, \Pi\sigma_1 + (1-\Pi)\sigma_2) \geq \Pi F(\rho_1, \sigma_1) + (1-\Pi)F(\rho_2, \sigma_2)$$

Trace distance is jointly convex:

$$T(\Pi\rho_1 + (1-\Pi)\rho_2, \Pi\sigma_1 + (1-\Pi)\sigma_2) \leq \Pi T(\rho_1, \sigma_1) + (1-\Pi)T(\rho_2, \sigma_2)$$

To see why this is true, define

$$\rho^{AB} = \Pi\left|1\right\rangle\left\langle 1\right| \otimes \rho_1 + (1-\Pi)\left|2\right\rangle\left\langle 2\right| \otimes \rho_2$$
$$\sigma^{AB} = \Pi\left|1\right\rangle\left\langle 1\right| \otimes \sigma_1 + (1-\Pi)\left|2\right\rangle\left\langle 2\right| \otimes \sigma_2$$

Then use the fact that

$$F(\rho, \sigma) = \Pi F(\rho_1, \sigma_1) + (1-\Pi)F(\rho_2, \sigma_2)$$
$$T(\rho, \sigma) = \Pi T(\rho_1, \sigma_1) + (1-\Pi)T(\rho_2, \sigma_2)$$

to get the right hand side of the inequalities. The left hand side comes from

$$\rho^B = \Pi\rho_1 + (1-\Pi)\rho_2$$
$$\sigma^B = \Pi\sigma_1 + (1-\Pi)\sigma_2$$

4.1.1.2 Joint and Conditional Entropies

For $X, Y \sim p(x, y)$, define joint entropy

$$H(XY) = H(p) = -\sum_{xy} p(x, y) \log p(x, y)$$

and conditional entropy

$$H(Y|X) = \sum_{x} p(X = x) H(Y|X = x)$$

For a classical distribution $p^{XY} = \Pi_1 |1\rangle \otimes p_1 + \Pi_2 |2\rangle \otimes p_2$,

$$\begin{aligned} H(Y|X = 1) &= H(p_1) \\ H(Y|X = 2) &= H(p_2) \\ \Rightarrow H(Y|X) &= \Pi_1 H(p_1) + \Pi_2 H(p_2) \end{aligned}$$

Note that we can rewrite the conditional entropy as

$$\begin{aligned} H(Y|X) &= -\sum_{x} p(x) \sum_{y} p(y|x) \log p(y|x) \\ &= -\sum_{xy} p(x) \cdot \frac{p(x, y)}{p(x)} \cdot \log \frac{p(x, y)}{p(x)} \\ &= -\sum_{xy} p(x, y) \log p(x, y) + \sum_{xy} p(x, y) \log p(x) \\ &= H(XY) + \sum_{x} p(x) \log p(x) \\ H(Y|X) &= H(XY) - H(X) \end{aligned}$$

Note also that

$$H(Y|X) \geq 0 \Leftrightarrow H(XY) \geq H(X)$$

although this is not always true quantumly. Also,

$$H(Y|X) \leq H(Y)$$

This statement, that conditioning reduces entropy, is also true quantumly. Note that it's also equivalent to concavity of entropy since

$$\begin{aligned} H(Y|X) &= \Pi_1 H(p_1) + \Pi_2 H(p_2) \\ H(Y) &= H(\Pi_1 p_1 + \Pi_2 p_2) \end{aligned}$$

4.1.2 Application: Compression

Say $X \sim p$, and $X^n = (x_1, x_2, ..., x_n) \sim p^{\otimes n}$ are iid samples from p. Can I compress X?

To do so with 0 error we need $\lceil \log |\text{supp}(p)| \rceil = \log \|p\|_0$ bits. To do so with ϵ error we need to throw away the smallest elements of p up to weight ϵ.

4.1.2.1 Shannon's Noiseless Coding Theorem

$X^n \sim p^{\otimes n}$, can compress to $n(H(X) + \delta)$ bits with error ϵ s.t. $\epsilon, \delta \to 0$ as $n \to \infty$.

The converse states that we can't do better. Compressing to $n(H(X) - \delta)$ bits means $\epsilon \to 1$.

Define a **typical set**:

$$T^n_{p,\delta} = \left\{ x^n = (x_1, ..., x_n), \left| -\frac{1}{n} \log p^{\otimes n}(x^n) - H(X) \right| \leq \delta \right\}$$

Define $p^{\otimes n}(x^n) = p(x_1)p(x_2)....p(x_n)$, then

$$\log p^{\otimes n}(x^n) = \sum_{i=1}^{n} \log p(x_i) \to -nH(p)$$

by the law of large numbers. This comes from the fact that

$$E[\log p(x_i)] = \sum_{x_i} p(x_i) \log p(x_i) = -H(p)$$

Thus by the law of large numbers, for all $\delta > 0$,

$$p^{\otimes n}(T^n_{p,\delta}) \to 1$$

as $n \to \infty$. Specifically, for $x^n \in T^n_{p,\delta}$,

$$\exp(-n(H(X) + \delta)) \leq p^{\otimes n}(x^n) \leq \exp(-n(H(X) - \delta))$$

and

$$p^{\otimes n}(T^n_{p,\delta}) \exp(n(H(X) - \delta)) \leq |T^n_{p,\delta}| \leq \exp(n(H(X) + \delta)$$

where the upper bound is used in the coding theorem, and the lower bound is used in the converse. Thus the number of bits needed is

$$\log |T^n_{p,\delta}| \leq n(H(X) + \delta)$$

Next time we'll look at Shannon's noisy coding theorem.