**8.S372/18.S996 Quantum Information Science III** **Fall 2020**

# Lecture 6: Sep 17, 2020

*Lecturer: Aram Harrow* *Scribe: Andrew Tan, Changnan Peng*

## 6.1 Relative Entropy

In the previous lectures, we introduced information entropy. An alternative interpretation of entropy is as "average surprise". In our daily experience, a more likely event contains less information and brings us less surprise. For example, if the weather forecast said there would be 90% probability of raining and it rains, we would not be very surprised. If it said 10% and rains, we would be more surprised. It is similar for the events in Huffman coding. We would be more surprised when an event with probability $2^{-10}$ appears than when one with $2^{-1}$ does. Huffman coding offers us a quantification of surprise. Given $n$ bits, we can identify one of $2^n$ events each with probability $2^{-n}$. This suggests that an event with probability $p(x)$ need $\log \frac{1}{p(x)}$ bits.

We can define

$$\text{surprise}(x) \equiv \log \frac{1}{p(x)} \tag{6.1}$$

and therefore the "average surprise"

$$\mathbb{E}[\text{surprise}(x)] = \sum_x p(x) \log \frac{1}{p(x)} = H(p) \tag{6.2}$$

Also in Huffman coding, $x$ uses $\lceil \text{surprise}(x) \rceil$ bits. That's how entropy as "average surprise" measures information.

In the previous example of Huffman coding, we have assumed that we know the true distribution of the events and encode them accordingly. What if we use the wrong distribution (i.e. $x \sim p$, but we encode according to $q$)? For example, we compress a piece of text by encoding the letters according to their appearance probability in English, but actually the text is written in French. In such cases, we cannot have optimal compression. The compressed message is longer than the one encoded with the correct distribution. The message length $\sum_x p(x) \log \frac{1}{q(x)} \geq \sum_x p(x) \log \frac{1}{p(x)}$.

The excess, denoted $D(p||q)$ is known as the *relative entropy* or *Kullback-Leibler divergence*

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{6.3}$$

The relative entropy is always non-negative. One can see this roughly by noting that Shannon's coding theorem implies that we cannot compress a source beyond its entropy, and therefore the excess must be $\geq 0$. However, this conclusion is not obvious from viewing the formula. Unlike in the definition of entropy where every term is non-negative, here the terms have mixed signs, being non-negative on the support where $p(x) \geq q(x)$, and negative otherwise. The non-negativity of relative entropy comes from the positive terms outweighing the negative ones.

Showing this more rigorously, we make use of the fact that

$$1 + z \leq e^z,$$

which can be shown from the convexity of $f(z) = e^z - (z+1)$, and $f(0) = f'(0) = 0$.

Replacing $z$ by $\log y$, we get the equivalent forms

$$\begin{aligned} \log y &\leq y - 1 \\ \log \frac{1}{y} &\geq 1 - y. \end{aligned}$$

Applying this inequality,

$$\begin{aligned} D(p||q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &\geq \sum_x p(x) \left(1 - \frac{q(x)}{p(x)}\right) \\ &= \sum_x p(x) - q(x) = 0, \end{aligned}$$

in the last step we used $\sum_x p(x) = \sum_x q(x) = 1$.

From the definition of relative entropy, we can see that it is zero when $p = q$. Are there any other cases? Tracing through the derivation of non-negativity, the inequality is tight only at one point, $z = 0$, equivalently $y = 1$ or $p(x) = q(x)$. A little tricky point is that at the terms with $p(x) = 0$, the inequality may also be tight, as those terms are zero. However, $\sum_x p(x) = \sum_x q(x) = 1$ forces $q(x)$ to be 0 when $p(x) = 0$, given $p(x) = q(x)$ when $p(x) \neq 0$.

Therefore, $D(p||q) = 0$ if and only if $p = q$.

Another note is that although the relative entropy describes the difference between two distributions, it is not a true distance in a metric sense – it is neither symmetric, $D(p||q) \neq D(q||p)$, nor satisfying the triangular inequality.

### 6.1.1 Corollary: Subadditivity of entropy

Using the non-negativity of relative entropy, we can prove the subadditivity of information entropy. Consider a joint distribution $p_{XY}$, and the direct product of its marginals, $p_X \otimes p_Y$. We calculate the relative entropy between them, and we can group the terms in different ways.

$$\begin{aligned} D(p_{XY}||p_X \otimes p_Y) &= \sum_{x,y} p(x,y) \left(\log p(x,y) - \log p_X(x) - \log p_Y(y)\right) \\ &= -H(XY) + H(X) + H(Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &\equiv I(X:Y) \geq 0. \end{aligned}$$

The second line tells us the subadditivity of entropy, i.e. $H(X) + H(Y) \geq H(XY)$. The third and fourth lines tell us conditioning on other systems will decrease the entropy, i.e. $H(X) \geq H(X|Y)$ and $H(Y) \geq H(Y|X)$.

In the last line we introduce a new quantity $I(X:Y)$, known as the *mutual information*. It describes the correlation of $X$ and $Y$ in a joint distribution $p_{XY}$. The mutual information $I(X:Y) = 0$ if and only if $X$ and $Y$ are independent, i.e. $p_{XY} = p_X \otimes p_Y$.

### 6.1.2 Corollary: Uniform distribution has largest entropy

We can also use the non-negativity of relative entropy to show that the uniform distribution has the largest entropy. Consider the special case of a distribution $p$ with the uniform distribution $u = \left(\frac{1}{d}, \frac{1}{d}, \ldots, \frac{1}{d}\right)$ on $d$ outcomes,

$$\begin{aligned} D(p||u) &= \sum_x p(x) \left(\log p(x) - \log \frac{1}{d}\right) \\ &= \log d - H(p) \geq 0. \end{aligned}$$

It tells us $H(p) \leq \log d$, and this maximum is reached if and only if $p = u$.

## 6.2 Hypothesis testing

The key application to understand information entropy is the message compression. We will see in this section that the key application for the relative entropy is the

hypothesis testing.

In a hypothesis testing, we are given several hypotheses of distributions and a sample of data. We would like to find out which distribution the sample comes from. When two hypotheses are given, it is called *binary hypothesis testing*. When there are more than two hypotheses, it is *multiple hypothesis testing*. Here we only talk about binary hypothesis testing.

Suppose we get $x$ sampled from $p$ or $q$ and want to guess which distribution $x$ comes from. There are two kind of errors we can make – $x$ sampled from $p$ but we guess $q$ (type 1), or $x$ sampled from $q$ but we guess $p$ (type 2). We define the probability of these two types of errors as

$$\begin{aligned} \alpha &= \Pr[\text{guess } q | x \sim p] \qquad \text{(type 1)} \\ \beta &= \Pr[\text{guess } p | x \sim q] \qquad \text{(type 2)}\,. \end{aligned}$$

We want to do the hypothesis testing that can minimize these errors. There are several ways to formulate the problem

1. Symmetric hypothesis testing: minimize $\alpha + \beta$. Answer is $\|p - q\|$.
2. Bayesian hypothesis testing: minimize $\pi\alpha + (1-\pi)\beta$. Answer on problem set 1.
3. Asymmetric hypothesis testing: minimize $\beta$ such that $\alpha \leq \epsilon$. Minimum is $\beta_\epsilon$.

As usual in the information theory, we consider the asymptotic case of $n$-copies with $n \to \infty$. Intuitively, with more samples in hand, we can distinguish the distributions better. When we cap $\alpha$ by a fixed value, $\beta$ should decrease exponentially with $n$. The question left is the coefficient in front of $n$ in the exponent, and the answer is the relative entropy.

Define $\beta_\epsilon^n$ to be the minimum of type-2 error for the binary hypothesis testing between $p^{\otimes n}$ and $q^{\otimes n}$. We expect $\lim_{n\to\infty} \beta_\epsilon^n \sim \exp\left(-nD(p||q)\right)$. Formally, we have the following theorem.

**Theorem 5 (Chernoff-Stein's Lemma)**

$$\lim_{n\to\infty} -\frac{1}{n} \log \beta_\epsilon^n = D(p||q), \quad \forall \epsilon \in (0,1). \tag{6.1}$$

We will not give a proof here. Instead, let's see some examples.

1. $p = q \iff D(p||q) = 0$. It is obvious that we cannot distinguish two distributions when they are identical. On the other direction, when two distributions are different, no matter how they are alike, we can do the hypothesis testing with exponentially small error given large enough number of samples.

2. $q$ is the uniform distribution $u$. $D(p||u) = \log d - H(p)$. We can do the following hypothesis testing. If the samples are in the typical set of $p$, i.e. $x^n \in T^n_{p,\delta}$, we guess $p$, otherwise we guess $q$. The appearance of the type-1 error is when $p$ generate samples outside the typical set. From the property of the typical set, we know the probability of type-1 error is $\alpha = 1 - p^{\otimes n}(T^n_{p,\delta}) \to 0$, for all $\delta \geq 0$. The appearance of the type-2 error is when the samples generated from the uniform distribution happen to be in the typical set. The probability is $\beta = \frac{|T^n_{p,\delta}|}{d^n} \leq \exp\left(n\left(H(p) + \delta\right) - n\log d\right) \leq \exp\left(-n\left(D(p||u) - \delta\right)\right)$. The minimal error must be small than this, i.e. $\beta^n_\epsilon \leq \beta \leq \exp\left(-n\left(D(p||u) - \delta\right)\right)$. We can smoothly reach the bound stated in the theorem by making $\delta$ slowly goes to zero.

3. $D(p||q) = \infty$. From the definition of the relative entropy, this occurs when there is an element $x$ such that $p(x) \neq 0$ and $q(x) = 0$, i.e. when $\text{supp}(p) - \text{supp}(q) \neq \varnothing$. We can do the following hypothesis testing. If an element in $\text{supp}(p) - \text{supp}(q)$ is seen, we guess $p$, otherwise we guess $q$. The type-1 error appears when those elements happen not to be seen, which occurs with probability that decreases exponentially. Note that we can always guess $p$ with certainty. Therefore, the probability of the type-2 error $\beta = 0$.

## 6.3 Quantum relative entropy

Now let's consider the quantum case. Unlike in the classical case where we can divide two probability distributions, the quantum analog of the relative entropy is defined as follows:

$$D(\rho||\sigma) \equiv \text{tr}\left[\rho\log\rho - \rho\log\sigma\right] = \text{tr}\left[\rho\left(\log\rho - \log\sigma\right)\right] \tag{6.1}$$

There is $D(\rho||\sigma) \geq 0$. A consequence of this is that the quantum mutual information $I(X:Y) \equiv D(\rho_{XY}||\rho_X \otimes \rho_Y) \geq 0$ is still non-negative by the same arguments as in the classical case.

We have a similar theorem for the binary hypothesis testing in the quantum case.

**Theorem 6 (Quantum Stein's lemma)** *Given $\rho^{\otimes n}$, $\sigma^{\otimes n}$. For any possible two-outcome measurement $\{M, 1 - M\}$, define the minimal type-2 error given a capped*

*type-1 error*

$$\beta_\epsilon^n = \min\left\{\mathrm{tr}[M\sigma^{\otimes n}] \quad | \quad \forall M \,\mathrm{s.t.}\, \mathrm{tr}[M\rho^{\otimes n}] \geq 1-\epsilon\right\}.$$

*The following limit holds*

$$\lim_{n\to\infty} -\frac{1}{n}\log\beta_\epsilon^n = D(\rho||\sigma).$$

We will not give a proof here. Instead, we consider the special case $D(\rho||\sigma) = \infty$. This time there is no clear meaning of probability on an element as in the classical case. Instead, the support is defined as the span of all eigenvectors. In the quantum case, $D(\rho||\sigma) = \infty \iff \mathrm{supp}\,\rho \not\subseteq \mathrm{supp}\,\sigma$.

When $\rho$ and $\sigma$ are pure states, e.g. $\rho = |\psi\rangle\langle\psi|$ and $\sigma = |\phi\rangle\langle\phi|$. The support of $\rho$ is $\{|\psi\rangle\}$ and the support of $\sigma$ is $\{|\phi\rangle\}$. This implies that $D(\rho||\sigma) = 0$ for identical pure states, or $D(\rho||\sigma) = \infty$ otherwise. Therefore, the relative entropy is not a well description for the difference between two pure states.

The optimal measurement in this case, is to choose $M = |\psi^\perp\rangle\langle\psi^\perp|$ and $1-M$. Note that the optimal measurement is not parallel to the state, but instead perpendicular. With this measurement we can rule out one of the hypothesis definitely. The proof is similar as in the classical case.

### 6.3.1 Quantum versus classical entropies, Conditional mutual information

Here are some properties of the quantum entropy

1. $0 \leq S(X) \leq \log d$ with equality on the lower bound only for pure states and equality for the upper bound only for the maximally mixed state $I/d$.
2. $0 \not\leq S(X|Y) \leq S(X)$. the non-negativity of the conditional entropy only holds in the classical case.
3. $D(\rho||\sigma) \geq 0$
4. $I(X:Y) \geq 0$

There is another quantity we have not yet introduced in this family. The *conditional mutual information* is the amount of mutual information conditioned on another random variable. It combines the idea of conditional entropy and mutual information.

Classically,

$$I(X:Y|Z) = \sum_z p_z(z) I(X:Y)_{p(\cdot,\cdot|z)} \geq 0 \tag{6.2}$$

and $I(X:Y|Z) \geq 0$ follows directly from subadditivity. The following equivalent definitions hold in both the classical and quantum cases

$$\begin{aligned} I(X:Y|Z) &= H(X|Z) + H(Y|Z) - H(XY|Z) \\ &= H(XZ) - H(Z) + H(YZ) - H(Z) - H(XYZ) + H(Z) \\ &= H(XZ) + H(YZ) - H(XYZ) - H(Z) \\ &= I(X:YZ) - I(X:Z). \end{aligned}$$

In the quantum case, it is still true that $I(X:Y|Z) \geq 0$ but does not follow obviously from subadditivity. This property is known as the "*strong subadditivity* (SSA) of quantum entropy". The proof is far more complicated than in the classical case.

However, the relation between $I(X:Y|Z)$ and $I(X:Y)$ is not definite. It can be "$\geq$", "$=$", or "$\leq$". For example, if $Z$ describe the noise that is added on both $X$ and $Y$, conditioning on the noise can increase the mutual information between the signals. On the other hand, it is also possible that $Z$ exactly determines $X$ and $Y$. In this case, the conditional mutual information equals zero.