8.372 Quantum Information Science III **Fall 2024**

Lecture 5: September 19, 2024

Scribe: John Blue and Jonathan Lu *Relative entropy*

5.1 Huffman codes as an interpretation of entropy

Here's an interesting interpretation of entropy given by Shannon. Suppose $X \sim p$. How surprised would you be to see that $X = x$? For example, in the English language, you wouldn't be very surprised if $X = \texttt{e}$, but you would be pretty surprised if $X = \texttt{q}$. Define

$$\text{surprise}(x) = \log \frac{1}{p(x)}. \tag{5.1}$$

The idea of this definition is that $1/p(x)$ gets larger as $p(x)$ gets smaller, so that as things are less probable we are more surprised. But why the log? This comes from an explicit construction of an information compression scheme known as *Huffman* coding. The idea is to map an outcome x into a bitstring

$$x \longrightarrow \text{Enc}(x) \text{ s.t. } |\text{Enc}(x)| = \left\lceil \log \frac{1}{p(x)} \right\rceil = \lceil \text{surprise}(x) \rceil. \tag{5.2}$$

If you pretend for a moment that all the probabilities are dyadic (i.e. 2^{-k} for some k depending on x), then the log gives an immediate interpretation of representing a number as a bitstring. In making the encoding, you have to be careful to ensure that it can be decodable. One straightforward way to do this is by ensuring that the code is *prefix-free*, i.e. that no codeword is the prefix of another codeword. If this weren't the case, we would get lost trying to decode locally. As an example, consider the Table 5.1 below. If we instead encoded `a` with `1`, and we have a stream of bits coming in that look like `111`, we couldn't distinguish `aaa` from `d`. (We could add separation characters between each encoded bitstring, but that would increase the encoding size!)

x	$p(x)$	Enc(x)
`a`	1/2	0
`b`	1/4	10
`c`	1/8	110
`d`	1/8	111

Table 5.1: Huffman coding for a dyadic distribution over 4 characters.

Note that if p is a dyadic distribution, $H(p) = \mathbb{E}[|\text{Enc}(X|]$, giving a constructive interpretation of entropy as the average Huffman encoding length. In general, the ceiling gives a few off-by-one errors that makes the Huffman code slightly more annoying to deal with. We won't get into that here, but it doesn't make any practical impact on the fundamental concepts we have discussed.

5.2 Relative entropy

Let's now imagine that we are trying to follow the Huffman procedure to encode our data into bits. The data comes from a distribution p, but we don't know p. Instead, we guess a distribution q and encode according to q instead. How good is our Huffman code now? Define the Huffman encoding map using q as Enc_q. The new average length of the encoding is given by

$$\mathbb{E}_{X\sim p}[\text{Enc}_q(X)] = \sum_x p(x) \log \frac{1}{q(x)}. \tag{5.3}$$

To study this quantity, we would like to write it in terms of the actual entropy $H(p)$ and some kind of measure of how much q deviates from p.

Definition 1. *The relative entropy of q relative to p is given by $D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$.*

In a very loose sense, the relative entropy $D(p||q)$ is meant to give a "distance" between distributions p and q. However, note that $D(p||q)$ is *not* symmetric. Also, p is supported on a character in which q is not, $D(p||q)$ is infinite! So what can we say about it?

Theorem 1. $D(p||q) \geq 0$.

Proof. One way to prove this is by applying Shannon's noiseless coding theorem. But we'll do this by a direct algebraic proof because of how important this bound is. First, note that $1 + z \leq e^z\ \forall z \in \mathbb{R}$. In particular, $z \leq e^z - 1$. Now let $z = \ln y$, so that $\ln y \leq y - 1$ and thus $\ln \frac{1}{y} \geq 1 - y$. Hence,

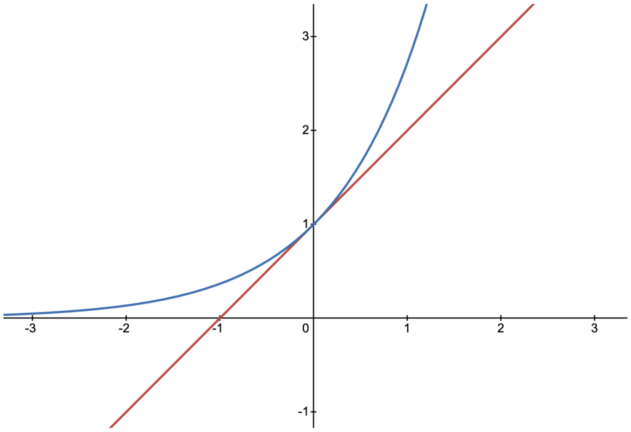

Figure 5.1: e^z (blue) is lower bounded by $1 + z$ (red).

$$\begin{aligned} D(p||q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} = \frac{1}{\ln 2} \sum_x p(x) \ln \frac{p(x)}{q(x)} \qquad (5.4) \\ &\geq \sum_x p(x) \left(1 - \frac{q(x)}{p(x)}\right) = \sum_x p(x) - q(x) = 1 - 1 = 0. \qquad (5.5) \end{aligned}$$

□

There are a number of important corollaries that follow almost immediately from this result.

Corollary 1. $H(p) \leq \log d$ *where d is the size of the character set from which p draws.*

Proof. First, define $u \in \mathbb{R}^d$ to be the uniform distribution, that is $u = (1/d, \ldots, 1/d)$. Then

$$\begin{aligned} 0 \leq D(p||u) &= \sum_x p(x)(\log p(x) + \log d) \qquad (5.6) \\ &= \log d - H(p) \qquad (5.7) \end{aligned}$$

□

This quick proof emphasizes the the asymmetry of the relative entropy is telling you something - the mixed thing should always go second!

We give the following inequality without proof

Theorem 2 (Pinsker's Inequality)**.**

$$D(p||q) \geq \frac{1}{2\ln 2}||p-q||_1^2 \tag{5.8}$$

Using this, we can make rigorous an intuition that $H(p)$ being close to $\log d$ means that p is close to uniform. Suppose $H(p) \geq \log d - \delta$. Then $D(p||u) \leq \delta$, which by Pinsker's inequality implies that $||p-u||_1 \leq \sqrt{2\ln(2)\delta}$.

We'll next use relative entropy to prove things about mutual information.

Corollary 2.

$$\begin{aligned} 0 \leq I(X;Y) &= H(X) + H(Y) - H(XY) && (5.9)\\ &= H(X) - H(X|Y) && (5.10)\\ &= H(Y) - H(Y|X) && (5.11) \end{aligned}$$

We can interpret this as saying that a) mutual information is a correlation (non-negative) b) conditioning reduces entropy.

Proof. Consider the relative entropy between a joint distribution p_{XY}, and the product of it's marginals:

$$D(p_{XY}||p_X \otimes p_Y) = \sum_{x,y} p_{XY}(x,y)\left(\log(p_{XY}(x,y)) - \log p_X(x) - \log p_Y(y)\right) \tag{5.12}$$

The first term is simply $-H(XY)$. Looking at the second term, $\sum_{x,y} p_{XY}(x,y)\log p_X(x)$, we see that by summing over y, we recover the marginal $p_X(x)$, and so this is just equal to $H(X)$. Similarly, the third term is just $H(Y)$. Putting it all together yields

$$\begin{aligned} D(p_{XY}||p_X \otimes p_Y) &= -H(XY) + H(X) + H(Y) && (5.13)\\ &\geq 0 && (5.14) \end{aligned}$$

where the inequality follows because relative entropy is non-negative. □

As a final application, we will prove that entropy is concave.

Corollary 3.

Let $\{p_x\}$ be a set of probability distributions, and π_x be a a probability distribution. Then

$$\sum_x \pi_x H(p_x) \leq H(\sum_x \pi_x p_x) \tag{5.15}$$

Proof. Define $p(x,y) = \pi_x p_x(y)$. Then

$$\begin{aligned} H(Y|X) &= H(X,Y) - H(X) && (5.16)\\ &= -\sum_{x,y} \pi_x p_x(y)\log(\pi_x p_x(y)) - \sum_x \pi_x \log(1/\pi_x) && (5.17)\\ &= \sum_x \pi_x H(p_x) && (5.18) \end{aligned}$$

and

$$H(Y) = H\left(\sum_x \pi_x p_x\right). \tag{5.19}$$

Concavity then follows from the fact that $H(Y|X) \le H(Y)$. □

5.2.1 Hypothesis Testing

Hypothesis testing is concerned with the following question: suppose we have two distributions p and q, and we get a sample x that we are told came either from p or q. How can we decide which?

There are two possible mistakes you could make, which are very descriptively called "type 1 error" and "type 2 error".

- Type 1 error: You guess $x \sim q$ when actually $x \sim p$. We will use α to denote the probability of a type 1 error.
- Type 2 error: You guess $x \sim p$ when actually $x \sim q$. We will use β to denote the probability of a type 2 error.

There are a few different kinds of hypothesis testing:

- Symmetric hypothesis testing: Come up with a test that minimizes $(\alpha+\beta)/2$, which has a minimum of $\frac{1}{2}||p-q||_1$.
- Bayesian hypothesis testing: Come up with a test that minimizes $\pi\alpha + (1-\pi)\beta$, which has a minimum of $||\pi p - (1-\pi)q||_1 + f(\pi)$ for some function f (π is prior probability that it's p.)
- Asymmetric hypothesis testing: Minimize β, subject to the contraint that $alpha < \epsilon$.

  We are going to study asymmetric hypothesis testing. Let $\beta_\epsilon = \min\{\beta|\alpha \le \epsilon\}$, and $\beta_\epsilon^n = \beta_{epsilon}$ for distinguishing p^n vs q^n (i.e. you get n samples to distinguish p and q).

  As a first observation, note that as we increase n, we should get more confident, and β_ϵ should go decrease. One might hope that it will scale as e^{-nR} for some R, and this indeed turns out to be the case, with R being the relative entropy.

  Theorem 3 (Chernoff-Stein). *For all $\epsilon \in (0,1)$,*

  $$\lim_{n\to\infty} \frac{-1}{n} \log \beta_\epsilon^n = D(p||q) \tag{5.20}$$

  Instead of proving this theorem, we will look at a few examples to see the sorts of tests that yield the desired β_ϵ.

  Examples

  1. Suppose $p = q$. Then $\Leftrightarrow D(p||q) = 0$ and β_ϵ^n stays constant, as the distributions are identical and there's nothing that can distinguish them.
  2. Suppose q is the uniform distribution. Then $D(p||q) = \log d - H(p)$. Here is a test: take a sample and check if $x^n \in T_{p,\delta}^n$. If it is, guess p, otherwise, guess q. We see that

the probability of a type one error is the probability that the sample is not in $T^n_{p,\delta}$, i.e. $\alpha = p^n(\overline{T^n_{p,\delta}})$ which goes to zero as n goes to infinity. We can also see that

$$\begin{aligned}
\beta &= u^n(T^n_{p\delta}) && (5.21)\\
&= \frac{|T^n_{p,\delta}|}{d^n} && (5.22)\\
&\leq \exp\left(nH(p) + n\delta - n\log d\right) && (5.23)\\
&= \exp\left(-n(D(p||u) - \delta)\right) && (5.24)
\end{aligned}$$

Since we can make δ arbitrarily small, we see that our test obtains the scaling from theorem 3.

3. Suppose $D(p||q) = \infty$. Then, $A = \text{supp}(p) \setminus \text{supp}(q) \neq \emptyset$, i.e. the support of p is not contained in the support of q. We can then use a simple test: if there exists an x_i in the sample such that $x_i \in A$, then guess p. Otherwise, guess q. Since x_i being in p guarantees that the sample came from p, we get $\beta = 0$, and $\alpha = p(\text{supp }(q))^n \to 0$ as n gets big.