Lecture 7: September 26, 2024

Scribe: Jonathan Lu *Noisy channel coding*

7.1 Aside: concavity of quantum entropy

Suppose we have two density matrices ρ_0 and ρ_1. We can mix them together with some probability weight π to obtain $\rho := \pi\rho_0 + (1-\pi)\rho_1$. The concavity property of the quantum entropy S tells us that $S(\rho)$ is at least as large as the mixed entropy $\pi S(\rho_0) + S(\rho_1)$. Because it's so important, let's prove the concavity of S.

Theorem 7.1.1. $S(\rho) = S(\pi\rho_0 + (1-\pi)\rho_1) \geq \pi S(\rho_0) + (1-\pi)S(\rho_1)$.

Proof. Let $\sigma^{AB} = \pi\rho_0^A \otimes |0\rangle\langle 0|^B + (1-\pi)\rho_1^A \otimes |1\rangle\langle 1|^B$. This is the *labeled* mixture, so that if we have access to the B system we know which density matrix we have. Note now that

$$S(A) = S(\rho),\ S(B) = H_2(\pi) := -\pi\log\pi - (1-\pi)\log(1-\pi). \tag{7.1}$$

Also, by definition, $S(A|B) = S(AB) - S(B)$ and $S(AB) = -\operatorname{tr}[\sigma\log\sigma]$. The structure of σ makes it block diagonal, since

$$\sigma = \left(\begin{array}{c|c} \pi\rho_0 & 0 \\ \hline 0 & (1-\pi)\rho_1 \end{array}\right), \quad \log\sigma = \left(\begin{array}{c|c} \log\rho_0 + (\log\pi)I & 0 \\ \hline 0 & \log\rho_1 + \log(1-\pi)I \end{array}\right). \tag{7.2}$$

This block diagonal structure makes the calculation of the joint entropy simple:

$$\begin{aligned} S(AB) = -\operatorname{tr}[\sigma\log\sigma] &= -\pi\operatorname{tr}[\rho_0\log\rho_0] - \pi\log\pi - (1-\pi)\operatorname{tr}[\rho_1\log\rho_1] - (1-\pi)\log(1-\pi) && (7.3)\\ &= H_2(\pi) + \pi S(\rho_0) + (1-\pi)S(\rho_1) && (7.4)\\ &= S(B) + S(A|B). && (7.5)\end{aligned}$$

We observe that the conditional entropy takes a simple form because the system being conditioned upon is just a classical probability distribution:

$$S(A|B) = \pi S(\rho_0) + (1-\pi)S(\rho_1). \tag{7.6}$$

Therefore, $S(\rho) - [\pi S(\rho_0) + (1-\pi)S(\rho_1)] = S(A) - S(A|B) = I(A;B) \geq 0$. The last inequality follows from the fact that $I(A;B) = D(\rho_{AB}||\rho_A \otimes \rho_B) \geq 0$, as we saw in the classical case. The proof that quantum relative entropy is non-negative is delegated to Problem Set 3. □

7.2 Classical noisy channel coding

In lecture 3, we stated Shannon's noisy coding theorem. Today we will prove it. Recall that a *channel* is a conditional probability distribution $N(y|x)$, so that if your input source is the distribution $\pi(x)$, the joint distribution of input-output pairs is $p(x,y) = \pi(x)N(y|x)$. The *capacity* of a channel is defined to encode the most amount of information you can send through the channel with asymptotically small noise.

Definition 7.2.1. *For a channel N, the capacity is given by*

$$C(N) = \lim_{\epsilon \to 0} \lim_{n \to \infty} \frac{1}{n} \log |M(\epsilon, N)|, \tag{7.7}$$

where $M(\epsilon, N)$ is the set of messages that can be sent through N^n with error probability $\leq \epsilon$.

Figure 7.1 shows the model we will adopt, in which we encode a set of messages M into bits before it is sent through a noisy channel, after which the noisy message is decoded into something that is hopefully the original message.

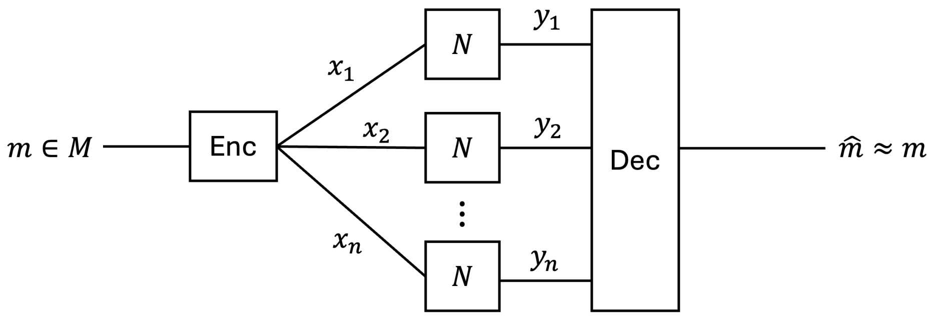

Figure 7.1: Encoder-decoder model with a channel in bewteen them.

Theorem 7.2.1 (Shannon's noisy coding theorem)**.** $C(N) = \max_\pi I(X;Y)$.

Before we prove the theorem, let's assume it's true and look at some illustrative examples of channels. For just these examples, let π be the probability that $X = 0$; we will only do examples with a single bit.

1. Binary symmetric channel with error probability η. Let x, y be single bits. Then $y = x \oplus \rho$, where $\Pr[\rho = 1] = \eta$. So, the bit gets flipped with probability η. Then

   $$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H_2(\eta). \tag{7.8}$$

   To maximize, note that $H_2(\eta)$ does not depend on π, and $H(Y) \leq 1$. But if $\pi = 1/2$, then $H(Y) = 1$. Hence for any δ, we can asymptotically send $n(1 - H_2(\eta) - \delta)$ bits of information to the output using n bits of input.

2. Erasure channel. Regardless of the input bit, there is a probability η that the channel maps it to $\perp$ (the "erased" message). Then $H(Y|X) = H_2(\eta)$ as with the binary symmetric channel. To calculate $H(Y)$, we note that $Y \in \{0, 1, \perp\}$ with probabilities $\pi(1-\eta), (1-\pi)(1-\eta), \eta$. Therefore,

   $$\begin{aligned} H(Y) &= -\pi(1-\eta)\log[\pi(1-\eta)] - (1-\pi)(1-\eta)\log[(1-\pi)(1-\eta)] - \eta \log \eta \qquad (7.9) \\ &= H_2(\eta) + (1-\eta)H_2(\pi). \qquad (7.10) \end{aligned}$$

   So we want to maximize $I(X;Y) = H(Y) - H(Y|X) = (1-\eta)H_2(\pi)$, which occurs when again $\pi = 1/2$, giving a capacity $C = 1 - \eta$.

The erasure channel capacity result is particularly remarkable. Consider a situation in which you and your friend are talking over the phone. Sometimes, the phone glitches with probability η and erases whatever your friend said at that time. To remedy this, you might say "what?", asking your friend to repeat herself. This protocol has an obvious capacity of $1-\eta$, but it also involves *feedback*, allowing the receiver to send information back to the sender. Shannon's theorem implies by the above that *even without feedback*, you can achieve the same capacity!

Now we want to actually prove Theorem 7.2.1. Let Alice send input bits and Bob receive output bits. Alice will send bits $x_1, \dots, x_n$ and Bob receives $y_1, \dots, y_n$. The intuition for this proof is that we will only worry about the typical set of Y^n, of which there are about $2^{nH(Y)}$, since those are the only strings that will be sent asymptotically. On the other hand, for a given input string x^n, there are about $2^{nH(Y|X)}$ output strings y^n that could have reasonably come from x^n. To ensure decodability, we want these possible string sets for each distinct (typical) x^n not to overlap. That implies we can have at most $2^{nH(Y)}/2^{nH(Y|X)} = 2^{nI(I;Y)}$ codewords.

Today we prove the achievability portion of the theorem, and leave the converse to next time.

Lemma 7.2.1. *$C(N) \geq \max_\pi I(I;Y)$. That is, for any rate $R < \max_\pi I(I;Y)$, there exists an encoding procedure that decodes with asymptotically vanishing error probability.*

Proof. For the proof, we'll switch back to $\pi = \pi(x)$ being a distribution over x. We formalize the above intuition by using relative entropy. Define $q_x(y) = N(y|x)$ just for notation and let $q(y) = \sum_x \pi(x) q_x(y)$ be the marginal distribution on Y. Then

$$D(q_x||q) = \sum_y q_x(y) \log \frac{q_x(y)}{q(y)} = -H(q_x) - \sum_y q_x(y) \log q(y). \tag{7.11}$$

The relation between relative entropy of these distributions and mutual information becomes clear when we sum over x:

$$\sum_x \pi(x) D(q_x||q) = -\sum_x \pi(x) H(q_x) - \sum_{x,y} \pi(x) q_x(y) \log q(y) \tag{7.12}$$

$$= -H(Y|X) + H(Y) = I(X;Y). \tag{7.13}$$

Define $x^n(m) := \mathrm{Enc}(m)$ and consider only $x^n(m) \in T_\pi^n$ the typical space, i.e. where i appears $n\pi_i$ times. Then $N^n(x^n(m)) = q_{x_1} \otimes \cdots \otimes q_{x_n}$, which up to permutation is $q_1^{\otimes n\pi_1} \otimes \cdots \otimes q_d^{\otimes n\pi_d}$. Note that $N^n(x^n(m))$ is itself a probability function and can be evaluated on strings and sets of strings, so to avoid confusion we will write $N^n(x^n(m))[S]$ as the conditional probability of getting strings in S as output given $x^n(m)$ as input. Since relative entropies add for independent distributions,

$$D(N^n(x^n(m)), q^{\otimes n}) = nI(X;Y). \tag{7.14}$$

By Stein's lemma from last lecture, there exists for any choices of ϵ, δ, a test set $A(m) \subseteq [d]^n$ such that $N^n(x^n(m))[A(m)] \geq 1-\epsilon$ but $q^n(A(m)) \leq 2^{-n(I(X;Y)-\delta)}$.

With these guarantees, we are ready to write down our encoding and decoding procedures. For the encoding, for each $m \in M$, Alice chooses $x^n(m) \in T_\pi$ (or, nearly equivalently, randomly from π^n). Note that the marginal probability holds as expected:

$$\mathbb{E}_{x^n(m)\sim\pi^n} N^n(x^n(m)) \approx q^{\otimes n}. \tag{7.15}$$

To decode, Bob brute-force iterates through $A(m)$, $m \in M$ and outputs m when the test passes. The probability the test fails is

$$\Pr[\text{error}] = \Pr[\text{wrong test accepts}] + \Pr[\text{right test rejects}] \tag{7.16}$$

$$\leq |M| 2^{-n(I(X;Y)-\delta)} + \epsilon. \tag{7.17}$$

By construction, $|M| = 2^{nR}$. Thus, if $R < I(X;Y) - \delta$, the first term asymptotically vanishes and so the error probability will asymptotically be ϵ, as desired. □