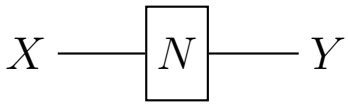
8.S372/18.S996 Quantum Information Science III **Fall 2020**

Lecture 7: Sep 22, 2020

Lecturer: Aram Harrow *Scribe: Zeyang Li, Andrew Tan*

7.1 Noisy Channel Coding

7.1.1 Classical noisy channels



We would like to study communication in a realistic setting where the medium over which messages are transmitted can (partially) corrupt the signal. We model a classical noisy channel N as a mapping between random variables X to Y, $N(y|x)$,

$$p_Y(y) = p_X(x)N(y|x)$$

This leads to a natural question: what is the largest amount of information that can be communicated per use of a given channel N? This quantity, known as the *channel capacity*, is defined as the highest rate of reliable communication that can be achieved, measured in bits per channel use, over all possible coding strategy; reliability in this case refers to the probability of error $\epsilon \to 0$ in the asymptotic data limit $n \to \infty$.

Mathmatically,

$$C(N) \equiv \lim_{\epsilon \to 0} \lim_{n \to \infty} \frac{1}{n} \log M^* \tag{7.1}$$

where

$$M^* = \max \left\{ M : \exists E : [M] \to X^n, \exists D : Y^n \to [M], \text{s.t.} \forall m, \Pr[m = D(N^{\otimes n}(E(m)))] \geq 1 - \epsilon \right\} \tag{7.2}$$

where the notation $[M] \equiv \{1, 2, \ldots, M\}$.

Shannon's noisy coding theorem provides the answer in simple expression:

$$C(N) = \max_{p_x} I(X : Y)_p \tag{7.3}$$

where $p(x, y) = p_X(x)N(y|x)$

Here the p is the joint input-output distribution. We can understand the mutual information in number of equivalent ways:

$$I(X;Y) = H(X) - H(X|Y)$$

The first interpretation of the mutual information is as the amount of information in the random variable X less the amount of uncertainty in X that still remains after observing the, potentially (partially) corrupted Y

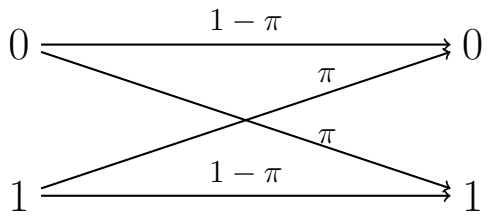
$$I(X;Y) = H(Y) - H(Y|X)$$

Another interpretation is as the amount of information carried in the observed Y less the randomness in Y that carries no information about X (i.e. the noise injected by the channel).

$$I(X;Y) = D(p_{XY}||p_X \otimes p_Y)$$

Yet another interpretation is as the relative entropy between the correlated joint distribution p_{XY} and the independent product of the marginals $p_X \otimes p_Y$

7.1.1.1 Example: Binary Symmetric Channel (BSC)



A commonly studied noisy channel model is the *binary symmetric channel* (BSC). The BSC has binary inputs and outputs with a probability π of the sent bit being flipped. That is the output

$$Y = X \oplus e, \; e = \begin{cases} 0, & \text{w/ prob. } 1-\pi \\ 1, & \text{w/ prob. } \pi \end{cases}$$

The Shannon limit defined in Equation 7.3, for the BSC is $C(N_{BSC}) = 1 - H_2(\pi)$; where $H_2(\pi)$ is the known as the binary entropy function

$$H_2(\pi) \equiv -\pi \log \pi - (1-\pi)\log(1-\pi)$$

we can see this by noting that the entropy is a concave function of p_X and is symmetric about the mid-point $\pi = \frac{1}{2}$ and is therefore maximized for $p_X = \left(\frac{1}{2}, \frac{1}{2}\right)$ (also note that

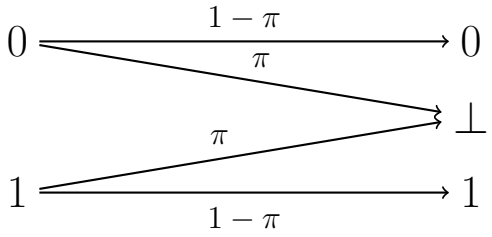
$p_Y = \left(\frac{1}{2}, \frac{1}{2}\right)$). From this, we can obtain the joint distribution

$$p = \frac{1}{2}\begin{pmatrix} 1-\pi & \pi \\ \pi & 1-\pi \end{pmatrix}$$

where we write the joint distribution p in the form of a 2×2 matrix.

Intuitively, this appears to be the best possible rate since, by correctly decoding the output Y, one obtains all of the information that has been input: one bit of information corresponding to the sum of the entropy of X as well as the entropy of the noise e.

7.1.1.2 Example: Erasure Channel



Another commonly studied channel is the erasure channel where a bit is lost with probability π. By the same argument as above, the maximum of Equation 7.3 is again attained for $p_X = \left(\frac{1}{2}, \frac{1}{2}\right)$.

Here we have $H(Y|X) = H_2(\pi)$ and $H(Y) = 1-\pi+H_2(\pi)$ giving a channel capacity $C(N_{ERASURE}) = 1 - \pi$.

Once again, this appears to be the best possible rate. One can imagine a protocol where Alice communicates to Bob using the channel $N_{ERASURE}$ and Bob has a noiseless channel to Alice that can be used to confirm the reception of a bit or the loss of a bit. In the case that Bob loses the bit, he communicates this noiselessly to Alice asking for her to send it again. This occurs with rate $1 - \pi$. This appears to be the best-case scenario as it requires access to an unphysical noiseless channel; and it is somewhat surprising that the channel capacity saturates the upper-bound given by this idealized scenario.

7.1.1.3 Example: Gaussian Noise Channel

A common model for analog communication is that of the Gaussian noise channel.

For concreteness, consider $x^n \in \mathbb{R}^n$ and $e \sim \mathcal{N}(0, \sigma^2)$. Finding the channel capacity in this case is close to the problem of sphere packing – especially as the Gaussian balls

become less 'fuzzy' in high dimensions. Note here that Bob learns y^n and also gets x^n if decoding works, and then therefore reconstruct the noise e^n.

7.1.1.4 Is the channel capacity achievable?

From the examples above, it seems clear that the channel capacities in these cases are as high as one could expect; however, it is not immediately clear that we can find an encoding scheme that achieves the channel capacity.

Consider using the repetition code over the BSC: encode $0 \mapsto 0^k, 1 \mapsto 1^k$, and the error rate $\Pr[\text{error}] \sim e^{-\mathcal{O}(k)}$ – this can be shown more rigorously using Chernoff bounds.

Consider the case where Alice would like to send a message of length l, with each bit encoded by a k-fold repetition. The total length is $n = kl$. The error rate per encoded bit goes as e^{-k}. To reliably transmit the entire message, we require the error rate per block to be less than $1/l$ corresponding to a choice of $k \sim \log l$. This gives a rate of $R \sim \frac{1}{\log n}$ showing that as $n \to \infty$, the rate of this encoding scheme goes to zero. We will need to be more clever if we are to achieve the channel capacity.

7.1.2 Proof of Shannon's Noisy-Channel Coding Theorem

Now we will prove that the channel capacity defined in Equation 7.3 is achievable by providing constructing a suitable encoding scheme.

First we define the *jointly typical set* $J^n_{p,\delta}$ as follows:

$$J^n_{p,\delta} \equiv \left\{(x^n, y^n) : (x_1y_1, \ldots, x_ny_n) \in T^n_{p_{XY},\delta}, (x_1, \ldots, x_n) \in T^n_{p_X,\delta}, (y_1, \ldots, y_n) \in T^n_{p_Y,\delta}, \right\} \tag{7.4}$$

that is the $J^n_{p,\delta}$ is the set of length n pairs of (x^n, y^n) such that x^n is typical with respect to p_X, y^n is typical with respect to p_Y and (x^n, y^n) is typical with respect to p_{XY} simultaneously.

From the results on the typical set, we have the following properties of strings in the jointly typical set

$$\begin{aligned} p^n_{XY}(x^n, y^n) &\approx \exp\left(-nH(XY)\right) \\ p^n_X(x^n) &\approx \exp\left(-nH(X)\right) \\ p^n_Y(y^n) &\approx \exp\left(-nH(Y)\right) \end{aligned}$$

and $p^n_{XY}(J^n_{p,\delta}) \to 1$ as $n \to \infty$.

In encoding process, we have a random codebook, $C = \{E(1), \cdots, E(M)\}$, where $M = |C| = 2^{nR}$, and R is the rate. Each $E(m)$ is drawn independently from $p_x^{\otimes n}$. To decode, we perform *joint typicality decoding*: given output $y^n = N(x^n)$, $D(y^n) = \hat{m}$ s.t. $(E(\hat{m}), y^n) \in J_{p,\delta}^n$. This can fail if

1. $\hat{m}$ does not exist, or
2. $\exists \hat{m} \neq m$ satisfying $(E(m'), y^n) \in J$.

We now show that both of these are unlikely for $R \leq C(N)$. Consider uniform distributed $m \in [M]$:

$$\Pr[\underbrace{(E(\hat{m}), y^n)}_{\sim p_{xy}^{\otimes n}} \in J_{p,\delta}^n] = p_{xy}^{\otimes n}(J_{p,\delta}^n) \to 1 \text{ as } n \to \infty$$

demonstrating that the first failure mode is unlikely.

Since $E(m)$ and $E(m')$ are independent, $E(m')$ and y^n are independently distributed, therefore for a fixed m',

$$\Pr[m' \neq m \cap (\underbrace{E(m')}_{\sim p_x^n}, \underbrace{y^n}_{\sim p_y^n}) \in J_{p,\delta}^n] = (p_x^n \otimes p_y^n)(J_{p,\delta}^n).$$

Since $p^n(x^n, y^n) \geq \exp(-nH(XY) - n\delta)$, we have $|J_{p,\delta}^n| \leq \exp(nH(XY) + n\delta)$, and therefore the r.h.s is

$$\begin{aligned}(p_x^n \otimes p_y^n)(J_{p,\delta}^n) &\leq |J_{p,\delta}^n| \max p_x^n \max p_y^n \\ &= \exp(-nH(X) + n\delta) \exp(-nH(Y) + n\delta) \exp(nH(XY) + n\delta) \\ &= \exp(-nI(X:Y) + 3n\delta)\end{aligned}$$

for a fixed m'.

For all m',

$$\Pr[\exists_{m' \neq m} \text{s.t.} (E(m'), y^n) \in J_{p,\delta}^n] \leq M(p_x^n \otimes p_y^n)(J_{p,\delta}^n) \leq \exp(nR - nI(X;Y) + 3n\delta)$$

which $\to 0$ when $R < I(X:Y) - 3\delta$.

Note that this provides an another interpretation of the mutual information $I(X;Y)$.

Although we have proven the achievability of the Shannon limit, the use of random codebook and joint typicality decoding is quite messy. Next class we're going to get rid of the random codebook and random message.

7.2 Quantum analogues

7.2.1 CQ channel capacity

Consider classical input, quantum ouput or CQ channel N. This can be thought of a channel that takes as input a number $x \in [M]$ and outputs a quantum state ρ_x; it can also be thought of as a channel that takes a quantum state σ but immediately decoheres it:

$$N(\sigma) = \sum_x \langle x|\sigma|x\rangle \, \rho_x$$

What is the classical capacity of this? The answer to this is given by the Holevo-Schumacher-Westmoreland (HSW) theorem:

$$C(N) = \max_p I(X;Q)_\omega \tag{7.1}$$

where $\omega^{XQ} = \sum_x p(x)|x\rangle\langle x|^X \otimes \rho_X^Q$; and the CQ joint entropy is

$$\begin{aligned} S(XQ) &= -\mathrm{tr}\left[\omega^{XQ} \sum_x |x\rangle\langle x| \otimes (\log p(x) I + \log \rho_x)\right] \\ &= H(p) + \sum_x p_x S(\rho_x) = H(X) + H(Q|X) \end{aligned}$$

and the CQ mutual information is

$$\begin{aligned} I(X;Q) &= H(Q) - H(Q|X) \\ &= S\left(\sum_x p(x)\rho_x\right) - \sum_x p(x) S(\rho_x) = \chi \end{aligned}$$

7.2.1.1 Example: simple application of the HSW theorem

Consider qubit states $\rho_i = |v_i\rangle\langle v_i|$ for $i \in \{1,2,3\}$. Assume that they related by $\frac{2\pi}{3}$ rotation so that $\frac{1}{3}(\rho_1+\rho_2+\rho_3) = I_2/2$. In this case, the ρ_i are pure states and therefore $S(\rho_i) = 0$ giving $S = 1$ and $\chi = I(X;Q) = 1$. This means that we can reliably transmit 1 bit of information per use of the channel. This would be clear if the output states were orthogonal such as $|0\rangle\langle 0|$ and $|1\rangle\langle 1|$, but is not as obvious in this case where the output states are not orthogonal.

7.2.2 Quantum joint typicality

The quantum analogue for typical sets are projectors into jointly typical subspaces: $T_X \mapsto \Pi_X$, $T_Y \mapsto \Pi_Y$, and $T_{XY} \mapsto \Pi_{XY}$. The problem, however, is that these projectors

do not commute in general and therefore we cannot directly define joint typicality.

To get around this, we can first purify the state $\rho_{XY} \mapsto |\psi\rangle_{XYZ}$ and consider Π_Z which should be fine as the projectors have the same spectrum. Another way is to consider quantities of the form $\Pi_{XY}\Pi_X\rho^{\otimes n}\Pi_X\Pi_{XY}$ although one needs to be careful of the ordering of the operators. This will be expanded on in the next lectures as we prove the HSW theorem.