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

Lecture 8: Sep 24, 2020

Lecturer: Aram Harrow *Scribe: Thao Dinh, Zeyang Li*

8.1 Shannon's Noisy Coding Theorem (cont)

Last class's proof we have two key features of Shannon's noisy coding theorem are random encoding and jointly typical decoding. The probability of error averaged over all the messages m, codebook C, and actions of the channel N^n is small.

$$\Pr_{m,C,N^n}[\text{error}] \leq \epsilon$$

The average is always greater than minimum, and therefore the LHS, i.e., the expectation value over C of the probability of error given the choice of C is greater than the minimum over C of the probability of error.

$$\mathop{\mathbb{E}}_{C}\left[\Pr_{m,N^n}[\text{error}|C]\right] \geq \min_{C} \Pr_{m,N^n}[\text{error}]$$

Fix the codebook C to be the one with minimum probability of error, but here we want it works for all message rather than some particular ones. In this case we can use the Markov's inequality, i.e. given a non negative random variable X, the probability to have $X \geq a$ is:

$$\Pr[X \geq a] \leq \frac{\mathbb{E}(X)}{a}$$

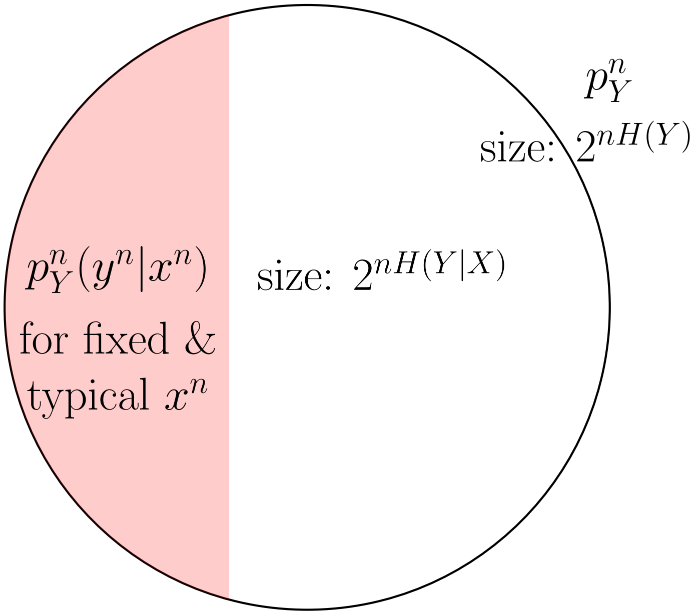
Applying the Markov's inequality for $a = 2\epsilon$, we have:

$$\Pr_{m}[\Pr_{N^n}[\text{error}|m] \geq 2\epsilon] \leq \Pr_{m}\left[\frac{\mathop{\mathbb{E}}_{N^n}([\text{error}|m])}{2\epsilon}\right] \leq \frac{1}{2}$$

Let's say the messages with $\Pr_{N^n}[\text{error}|m] \geq 2\epsilon$ are bad messages, then based on Markov's inequality, at most half of the messages are bad. But because the number of messages is exponential the number of channels, so it's no big deal to get rid of half of the bad messages ("expurgation").

The reduced codebook now will have at least half the size of the original codebook. For all the message m in $C_{reduced}$, we have $\Pr_{N^n}[\text{error}|m] \leq 2\epsilon$

8.1.1 Proof intuition of the theorem



Bob receives strings from $p_Y^{\otimes n}$, i.e. p_Y^n for classical case. The typical set of those has the size $2^{nH(Y)}$.

For any given message analysis, i.e. fixed typical X^n, then over a small subset we have the distribution $p_Y^n(y^n|x^n)$. The size of that subset is $2^{nH(Y|X)}$.

Why is that? For frequency typical, the number of x appear in x^n is approximately equal to $np_x(x)$. Suppose they are equal. For a string x^n, we have:

$$p_Y^n(y^n|x^n) = p(y_1|x_1)p(y_2|x_2)...p(y_n|x_n)$$

We expect y^n to have:

$$np_x(x_1) \text{ positions in the typical subspace } T^n_{p(\cdot|x_1),\delta}$$
$$np_x(x_2) \text{ positions in the typical subspace } T^n_{p(\cdot|x_2),\delta}$$

Then we can group all the p with same x together and have (but neglecting the δ stuffs just to provide intuitions):

$$\begin{aligned} p_Y^n(y^n|x^n) &= \prod_x \exp[-np_x(x)H(p(\cdot|x))] \\ &= \exp\left[-n\sum_x p_x(x)H(p(\cdot|x))\right] \\ &= \exp(-nH(Y|X)) \end{aligned}$$

If every string in subset has that probability, then the size of strings in the subset is roughly $2^{nH(Y|X)}$.

8.2 Converse of the Noisy Coding Theorem

8.2.1 Properties of entropy

- If X is deterministic (for quantum, ρ is pure), then $H(X) = 0$.
- If Y is completely determined by X, i.e $Y = f(X)$, then $H(Y|X) = 0$.
- If X and Y are independent, i.e $p(X, Y) = p_x(X)p_y(Y)$ (for quantum, $\rho_{xy} = \rho_x \otimes \rho_y$), then $I(X : Y) = 0$.
- Conditional mutual information (CMI): If $X - Z - Y$ is a Markov chain, i.e $p(x, y, z) = p_Z(z)p(x|z)p(y|z)$, then $I(X : Y|Z) = 0$.

8.2.2 Properties of CMI

- Chain rule: $I(X : YZ) = I(X : Y) + I(X : Z|Y)$

  Proof: If we denote $H(\alpha)$ as α where α could be single or joint distribution and could also be conditional. Then we expand:

  $$\begin{aligned} \text{LHS} &= X - X|YZ = X - XYZ + YZ \\ \text{RHS} &= (X - X|Y) + (X|Y + Z|Y - XZ|Y) \\ &= (X - XY + Y) + (XY - Y + YZ - Y - XYZ + Y) \\ &= X + YZ - XYZ \end{aligned}$$

  LHS and RHS are equal, thus the chain rule is proved.

- Generalized chain rule:

  $$I(X : Y_1...Y_n) = I(X : Y_1) + I(X : Y_2|Y_1) + ... + I(X : Y_n|Y_1...Y_{n-1})$$

- Data processing inequality: If $X - Z - Y$ is a Markov chain, then as we move along the Markov chain, that should only degrade the mutual information, i.e $I(X : Z) \geq I(X : Y)$

Proof: Applying the chain rule above, we have:

$$I(X:Z) = I(X:YZ) - I(X:Y|Z)$$
$$I(X:Y) = I(X:YZ) - I(X:Z|Y)$$

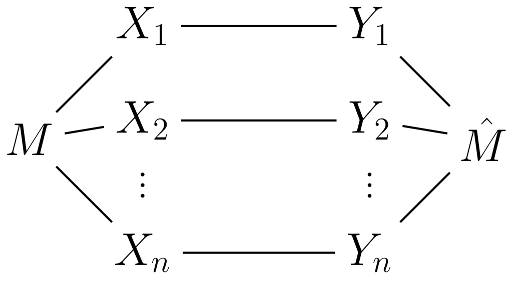
Take the difference on both sides of two equations:

$$\begin{aligned} I(X:Z) - I(X:Y) &= -I(X:Y|Z) + I(X:Z|Y) \\ &= I(X:Z|Y) \geq 0 \end{aligned}$$

In the second line, we used the property of Markov chain $I(X:Y|Z) = 0$. Thus $I(X:Z) \geq I(X:Y)$.

All of the above properties are true quantumly as well.

8.2.3 Converse of the Noisy Coding Theorem



If the noisy coding theorem says that we can send nR bits and R can get right up to the mutual information, the converse theorem says that we cannot do much better than that.

Consider a most general possible coding scheme: Alice sends message M, encodes it and inputs to the channels X^n. The input channels are mapped to output channels Y^n. Bob gets the outputs and decodes $\hat{M}$. i.e Markov chain $M - X^n - Y^n - \hat{M}$. We assume M is uniformly distributed in $\{0,1\}^{nR}$, then $H(M) = nR$. Note that here we choose the uniform distribution here just for simplicity, and the theorem should apply to all possible distributions.

8.2.3.1 Fano's inequality

Obviously, the conditional entropy for M based on $\hat{M}$ is small because for most cases they're equal, and similarly the mutual information between them is high. Quantitatively, we have the Fano's inequality says:

$$\begin{aligned} H(M|\hat{M}) &\le \epsilon nR + 1 \\ \rightarrow I(M:\hat{M}) = H(M) - H(M|\hat{M}) &\ge (1-\epsilon)nR - 1 \end{aligned}$$

Proof:

If the alphabet has size d, which in our real applications it's nR. And suppose that the probability of one element $p(m) \ge 1-\epsilon$, which corresponds to $M = \hat{M}$ in the above scenario. Then the entropy $H(p) \le 1 + \epsilon \log d$.

We name the $p(m) = 1-\delta, \delta \le \epsilon$. Rewrite the distribution $p = (1-\delta)1_m + \delta q$, where q is another distribution which satisfies $q(m) = 0$. Then, the entropy can be rewritten as a sum of entropy of mixing being m or not being m, and the entropy of the rest components:

$$\begin{aligned} H(p) &= -(1-\delta)\log(1-\delta) - \sum_x \delta q(x) \log \delta q(x) \\ &= -(1-\delta)\log(1-\delta) - \delta \log \delta - \delta \sum_x q(x) \log q(x) \\ &= H_2(\delta) + \delta H(q) \\ &\le 1 + \delta \log d \end{aligned}$$

Fannes' inequality (generalized version of Fano's inequality): If p, q are distributions on alphabet of size d, then

$$\begin{gathered} |H(p) - H(q)| \le H_2(\epsilon) + \epsilon \log d \\ \epsilon = \frac{1}{2}||p-q||_1 \end{gathered}$$

Where in the quantum version of we just replace H by S and p, q by the density matrix.

8.2.3.2 Proof of converse theorem

Here we want to relate the above inequality to channels to work with Shannon's theorem:

$$(1-\epsilon)nR - 1 \underbrace{\le}_{\text{Fano's inequality}} I(M:\hat{M}) \overbrace{\le}^{\substack{\text{data processing worsen information} \\ \text{in Markov chain } M-X-Y-\hat{M}}} I(X^n:Y^n) \underbrace{\le}_{\star} \sum_{j=1}^{n} I(X_j:Y_j) \le nC \tag{8.1}$$

$$\rightarrow R \le \frac{C}{1-\epsilon}$$

The second inequality results from data processing: in the Markov's chain, the mutual information of two ends is less than or equal to the mutual information of the middles.

The last inequality: the mutual information of each input-output channels pair is at most C (the mutual information obtained by maximizing over all the inputs).

The third inequality is a little bit unique because basically all other properties we mentioned in this section can be naturally generalized to quantum cases but this one not[1]. The major difference happens when the input has quantum entanglement. The reason we will mention the quantum capacity theorem in CQ channels in Sec. 7.2.1 is specifically to avoid such things to happen.

Now we try to prove ($\bigstar$) in classical regime. The mutual information is $I(X^n : Y^n) = H(Y^n) - H(Y^n|X^n)$. Because there is no correlation between different pairs of input-output channels, using chain rule, we have:

$$\begin{aligned} H(Y^n|X^n) &= \sum_{j=1}^{n} H(Y_j|X^n Y_1 ... Y_{j-1}) \\ &= \sum_{j=1}^{n} H(Y_j|X_j), \end{aligned}$$

where the last equality is based on the fact that the Markov chain only connects directly related pairs, so once condition on X_j, the Y_j becomes conditionally independent on everything else. One can imagine that this fails quantumly when different X_j are entangled and therefore can all contribute to Y_j.

The entropy of the sum is less than sum of the entropies of the part, i.e., the sub-additivity of entropy, so we have:

$$H(Y^n) \leq \sum_{j=1}^{n} H(Y_j)$$

[1] The conditional entropy, however, is also different in quantum since it can go to negative and therefore being equal to 0 does not have unique properties. The CMI and the corresponding Markov chain state can be generalized to quantum Markov states which we will revisit later, but in short the chain rule holds in quantum cases.

Thus,

$$\begin{aligned} I(X^n : Y^n) &= H(Y^n) - H(Y^n|X^n) \\ &\leq \sum_{j=1}^{n} H(Y_j) - \sum_{j=1}^{n} H(Y_j|X_j) \\ &\leq \sum_{j=1}^{n} I(X_j : Y_j) \end{aligned}$$

8.3 Quantum Capacity Theorem

Idea: Find achievability via Packing Lemma

Example: Suppose that Alice has a menu of pure states as output $|0\rangle, |1\rangle, |+\rangle, |-\rangle$ to send

- Can send 1 classical bit ($0 \rightarrow |0\rangle$ and $1 \rightarrow |1\rangle$) or ($0 \rightarrow |+\rangle$ and $1 \rightarrow |-\rangle$)
- Can send 2 classical bits ($00 \rightarrow |0\rangle$, $01 \rightarrow |+\rangle$, $10 \rightarrow |-\rangle$ and $11 \rightarrow |1\rangle$)

Can Bob extract two classical bits from one quantum bit? No.

If Q is the quantum system, then we have:

$$I(M : \hat{M}) \leq I(M : Q) \leq \log(\text{dim}Q) = 1$$

Therefore, Bob can extract at most one classical bit. So Alice should choose a distinguishable subset instead.

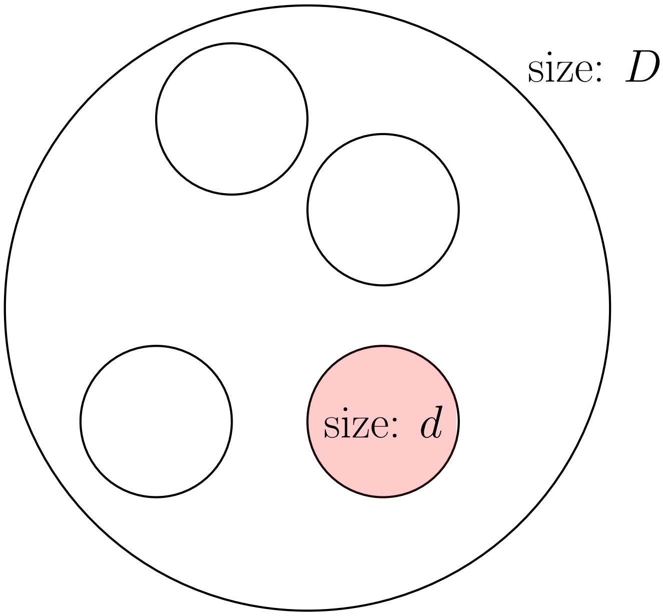
8.3.1 Packing Lemma

Given $\{\rho(x), \sigma(x)\}_{x \in X}$ with probability distribution $\rho(x)$ and signal state $\sigma(x)$. And $\sigma = \sum_x \rho(x)\sigma(x)$ is the average state.

Suppose there exists a projector Π and family of projectors $\{\Pi_x\}_{x \in X}$ such that if d and D are dimensions of subspace and space then:

- $\text{Tr}[\Pi \sigma_x] \geq 1 - \epsilon$ for all x
- $\text{Tr}[\Pi_x \sigma_x] \geq 1 - \epsilon$ for all x
- $\text{Tr}[\Pi_x] \leq d$

- $\Pi\sigma\Pi \leq \frac{\Pi}{D}$



Choose the codebook $C = \{C_1, ..., C_M\} \sim p^n(n$ the number of letters in each code word), then there exists a positive operator-valued measure (POVM) $\{\Lambda_m\}$ such that averaging over the codebooks, averaging over messages in codebook, the probability to get the right outcome when measuring the state σ_{C_m} is close to 1:

$$\mathop{\mathbb{E}}_{C} \mathop{\mathbb{E}}_{m\in[M]} \mathrm{Tr}(\Lambda_m \sigma_{C_m}) \geq 1 - 2\epsilon - 4\sqrt{\epsilon} - 4M\frac{d}{D}$$

Here whenever M the total amount of message in codebook is less than the order of D/d can give a small enough error probability. Corresponds to fill size $d \geq \mathrm{Tr}(\Pi_x)$ sphere inside of size $D(1-\epsilon) \geq \mathrm{Tr}(\Pi)$, like the Gaussian Noise channel in Sec. 7.1.1.3.

8.3.2 Application to channel coding

The HSW theorem (7.1) says the capacity of a noisy quantum channel is the maximal mutual information between input X and output Q, maximizing over input distribution p

$$C(N) = \max_p I(X : Q)$$

Choosing the codebook C_i from $p_T^n = p^n | T_{p,\delta}^n$ where $T_{p,\delta}^n$ is the frequncy-typical set.

Then, the corresponding string x^n and state ρ_{x^n} are:

$$\begin{aligned} x^n(i) &= C_i \\ \rho_{x^n} &= \rho_{x_1} \otimes \rho_{x_2} \otimes ... \otimes \rho_{x_n} \end{aligned}$$

We need to show this choice satisfies all the conditions of Packing Lemma. The average state:

$$\sigma = \mathbb{E}(\rho_{x^n}) = \sum_{x^n} p_T^n(x^n)\rho_{x^n} \approx \sum_{x^n} p^n(x^n)\rho_{x^n} = \bar{\rho}^{\otimes n} \text{ for } \bar{\rho} = \sum_x p(x)\rho_x$$

And let's take the total projector to be projecting into this average state: $\Pi = \Pi^n_{\bar{\rho},\delta}$. Does this projector satisfy the four conditions in packing Lemma?

- Condition 1:

$$\mathop{\mathbb{E}}_{x^n} \mathrm{Tr}[\Pi\rho_{x^n}] \geq \mathrm{Tr}[\Pi\bar{\rho}^{\otimes n}] - \epsilon \geq 1 - 2\epsilon$$

  Therefore, the best 1/2 of $x^n \in X^n$ have $\mathrm{Tr}[\Pi\rho_{x^n}] \geq 1 - 4\epsilon$

  Denote Π_{X^n} as the conditionally typical projector, it is calculated as follow:

$$\Pi_{X^n} = \bigotimes_{x \in X} \Pi^{\#x}_{\rho_x,\delta}$$

  This product is permuted according to x^n and $\#x$ is the count of occurrences of x in x^n.

- Condition 2:

$$\mathrm{Tr}[\Pi_{X^n}\rho_{X^n}] = \prod_{x\in X} \mathrm{Tr}[\rho_x^{\otimes \#x}\Pi^{\#x}_{\rho_x,\delta}] \geq 1 - |X|\epsilon$$

- Condition 3:

$$\begin{aligned} \mathrm{Tr}[\Pi_{X^n}] &= \prod_{x\in X} \mathrm{Tr}[\Pi^{\#x}_{\rho_x,\delta}] \\ &\leq \prod_{x\in X} \mathrm{Tr}[\Pi^{n(p(x)+\delta)}_{\rho_x,\delta}] \text{ because of the freq typicality} \\ &\leq \prod_{x\in X} \exp(n(p(x)+\delta)(S(\rho_x)+\delta)) \\ &\leq \exp(n(S(Q|X)+\delta')) \end{aligned}$$

- Condition 4: For $p_T^n \leq (1-\epsilon)^{-1}p^n$, we have the average of ρ_{X^n} is

$$E(\rho_{X^n}) \leq (1-\epsilon)^{-1} \sum_{X^n} p^n(X^n)\rho_{X^n} = (1-\epsilon)^{-1}\bar{\rho}^{\otimes n}$$

Now we need to calculate

$$\begin{aligned} \Pi\bar{\rho}^{\otimes n}\Pi &\leq 2^{-n(S(\rho)-\delta)} \\ \rightarrow \Pi E(\rho_{X^n})\Pi &\leq (1-\epsilon)^{-1}2^{-n(S(\rho)-\delta)} \end{aligned}$$

Here $D \approx 2^{nS(Q)}$, $d \approx 2^{nS(Q|X)}$, thus we can take $M \approx 2^{nI(X:Q)}$.