8.372 Quantum Information Science III **Fall 2024**

Lecture 8: October 1st 2024

Scribe: Adam Wills and Daniel Lee *Classical Messages over Quantum Channels*

8.0 Lecture Outline

The aim of today's lecture is to cover three things. First, we'd like to establish a quantum version of Shannon's noisy channel coding theorem that we discussed last time. In particular, we'd like to discuss sending classical messages over quantum channels, and establish the rate at which we can send information in such a situation. Having established the achievability of this, in a fairly analogous way to what we did last time for Shannon's noisy channel coding theorem, we will turn to the converse for both Shannon's noisy channel coding theorem and for this quantum situation. We won't have time to prove this completely, but as preparation for the upcoming proof, we will introduce the Conditional Mutual Information (CMI).

8.1 Classical-Quantum (CQ) Channels

We will talk about channels with classical input and quantum output, otherwise known as CQ channels; for example,

$$\mathcal{N} : x \mapsto \rho_x. \tag{8.1}$$

These can be imagined as special cases of usual CPTP quantum channels, for example,

$$\mathcal{N}(|x\rangle \langle y|) = \delta_{xy}\rho_x. \tag{8.2}$$

This also corresponds to a quantum channel which measures its input, before sending on some quantum states dependent on the outcome of this measurement. By specialising to classical-quantum channels, we avoid many of the difficulties experienced with general quantum-quantum channels, represented by a general CPTP map, for which entangled inputs are allowed, and many results become very hard to prove.

8.1.1 The HSW Theorem

The HSW Theorem tells us the capacity of such a channel. It is

$$C(\mathcal{N}) = \max_p I(X : Q)_\omega, \tag{8.3}$$

where the maximum is taken over all probability distributions p, and ω is the "classical-quantum state"

$$\omega^{XQ} = \sum_x p(x) |x\rangle \langle x|^X \otimes \rho_x^Q. \tag{8.4}$$

Comments:

1. As a special case, if ρ_x is diagonal, then this reduces to Shannon's noisy channel coding theorem.

2. We have $I(X:Q) = S(Q) - S(Q|X)$ as always. It's worth appreciating that asking for the entropy of Q conditioned on X makes more sense for this state than for a general quantum state; conditioning on the outcome of some classical random variable is much more meaningful than conditioning on some quantum state. Notice that

$$I(X:Q)_\omega = S(\rho) - \sum_x p(x) S(\rho_x) \tag{8.5}$$

where $\rho = \sum_x p(x)\rho_x$ is the average state. It is common to denote a quantity called Holevo's χ-quantity as

$$\chi(\mathcal{N}) = \max_p \left[S(\rho) - \sum_x p(x) S(\rho_x) \right], \tag{8.6}$$

so that $C(\mathcal{N}) = \chi(\mathcal{N})$.

3. As is typically the case, the theorem comes with an achievability part and a converse, so a full proof shows that information transmission at a rate of $C(\mathcal{N})$ is possible, and that attempting to transmit information at any faster rate will fail (i.e. the error rate in communication will become large).

4. The HSW theorem has an interesting relation to the scenario of *accessible information*. Suppose Alice wants to send classical information to Bob via a quantum channel. She wishes to encode some input x, corresponding to the value of some random variable $X \sim p$, into some quantum state ρ_x which then gets sent to Bob. Bob then tries to learn about x by performing some measurement $\{M_y\}_y$ and deducing the outcome y as a result, with the hope that $y = x$.

   This whole thing can be considered as a classical channel, and the corresponding mutual information (maximised over the best possible measurement by Bob) is known as the accessible information of the ensemble $\{p_x, \rho_x\}$:

$$I_{acc}(\{p_x, \rho_x\}) = \sup_{\{M_y\}_y} I(X:Y). \tag{8.7}$$

   It is true in general that

$$I_{acc}(\{p_x, \rho_x\}) \leq S(\rho) - \sum_x p(x) S(\rho_x). \tag{8.8}$$

Example 8.1.1. *Consider a C-Q channel that sends the classical input i to the quantum output ρ_i for $i = 1, 2, 3$. Suppose the ρ_i are pure qubit states lying in the equator of the Bloch sphere (see Figure 8.1), and the three of them are mutually equally spaced. In this case, we have that the average state ρ is the maximally mixed state, and so $S(\rho) = 1$. We also have that the entropy of each individual state is 0 (because the states are pure). As such, we have $C(\mathcal{N}) = 1$. We might consider this to be somehow surprising, because given only one transmission of ρ_1, ρ_2, ρ_3, it is impossible to reliably distinguish between them. However, the definition of $C(\mathcal{N})$ is an asymptotic statement. We are seeing that asymptotically, $\mathcal{N}$ is as good as a classical noiseless channel just transmitting one bit.*

This is already giving us the indication that to get the most out of this communication scenario, the receiver, Bob, must perform entangled measurements on the outputs of the n uses of the channel

$$\rho_{x_1} \otimes \rho_{x_2} \otimes \ldots \otimes \rho_{x_n}. \tag{8.9}$$

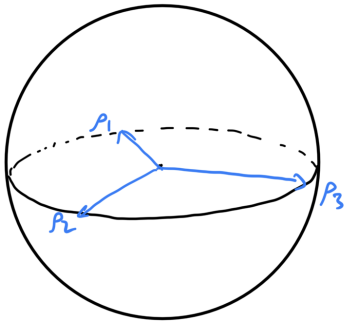

Figure 8.1: Three pure states on the equator of the Bloch sphere, which average to the maximally mixed state

8.2 Quantum-Quantum Channels

We will quite shortly go on to prove the achievability portion of the HSW theorem, and then direct ourselves towards the converse. First, however, let us make some comments on general quantum-quantum channels, which are in general harder to deal with.

A general quantum-quantum channel is represented by a CPTP map $\mathcal{N}$, and in this case we define the χ-quantity as

$$\chi(\mathcal{N}) = \max_{\{p_x, \sigma_x\}} I(X:Q)_\omega, \tag{8.10}$$

where

$$\omega = \sum_x p_x \left|x\right\rangle \left\langle x\right| \otimes \mathcal{N}(\sigma_x). \tag{8.11}$$

The capacity of the channel is then in fact

$$C(\mathcal{N}) = \lim_{n\to\infty} \frac{1}{n} \chi(\mathcal{N}^n). \tag{8.12}$$

It is an interesting and very much non-trivial fact, known as the superadditivity of quantum channel capacity, that in general one can send more information through a channel $\mathcal{N}$ by sending entangled inputs through multiple uses of it, then just using it separately multiple times, i.e., $C(\mathcal{N}) > \chi(\mathcal{N})$ in general.

Comments:

1. The above formula for quantum channel capacity is very hard to work with in practice. To see this, let us start by considering the formula given by the HSW theorem. This is a relatively easy expression to work with, because it is a maximisation of a concave function ($I(X:Q)_\omega$) over a convex set (the set of probability distributions). This makes it easy to prove things about in theory, and also our computers can handle it easily with standard optimisation techniques.

   Conversely, the general quantum channel capacity is not a maximisation of a concave function. It is in fact NP-hard in general to computationally find the maximising solution of $\chi(\mathcal{N})$, so we do not expect to find efficient techniques for performing this optimisation.

2. By applying this formula to a noiseless qubit channel, one can see that it is impossible to reliably send more than n bits of information in n qubits, despite the fact that a full description of an n-qubit state requires 2^n complex numbers. Qubits are therefore no better carriers of information than classical bits, even if they are better at certain other tasks, like secret sharing.

8.3 Proof of Achievability for the HSW Theorem

To prove the achievability part of the HSW theorem, we want another tool.

8.3.1 Non-Commutative Union Bound

Classically, suppose we have some bad events that are unlikely to happen. Via the standard union bound, it is easy to say that the probability that at least one bad event happens is the sum of the probabilities of all the individual bad events.

Quantumly, the analogous question might be to consider some density matrix ρ and some two-outcome measurements, each corresponding to operators $\{P_i, I - P_i\}$ for $i = 1, \ldots, l$. Supposing that each P_i is fairly likely to be measured, what is the probability of measuring all $P_1, \ldots, P_l$ on ρ sequentially? We cannot just use the classical case because the state gets disturbed with each measurement. The statement of the lemma is this.

Lemma 1. *If $\rho \geq 0$, $Tr(\rho) \leq 1$, and $P_1, \ldots, P_l$ are projectors, then*

$$Tr(\rho) - Tr(P_1, \ldots P_l \rho P_l \ldots P_1] \leq 2\sqrt{\sum_{i=1}^{l} Tr((I - P_i)\rho)} \tag{8.13}$$

8.3.2 Remainder of the Proof

Let p achieve the maximum in

$$C(\mathcal{N}) = \max_p (I(X : Q))_\omega. \tag{8.14}$$

Just as in the classical case, let us take a random codebook, so Alice chooses some codewords

$$X^n(1), \ldots, X^n(M) \sim p^n \tag{8.15}$$

identically and independently. For each $m \in [M]$, the state Bob receives is

$$\sigma_m = \rho_{X_1(m)} \otimes \ldots \otimes \rho_{X_n(m)}, \tag{8.16}$$

where $X_i(m)$ is the i-th symbol in the codeword corresponding to $m \in [M]$. Notice that if you average over the choice of codeword $X^n(m)$, you get

$$\mathbb{E}_{X^n(m)} \sigma_m = \left(\sum_x p(x) \rho_x \right)^{\otimes n} = \rho^{\otimes n}. \tag{8.17}$$

Let us define the conditionally typical projector Π_m, for σ_m. Letting t be the type of $X^n(m)$[1], Π_m is defined by

$$\Pi_m = \Pi \bigotimes_{i=1}^{d} \Pi_{\rho_i, \delta}^{n t_i} \Pi^{-1}, \tag{8.18}$$

[1]It is worth appreciating that the type of $X^n(m)$ will be very close to p (for large n).

where Π is the permutation mapping the states ρ_i as they appear in ascending order corresponding to the type, to the order in which they appear in σ_m, i.e. it maps

$$\bigotimes_{i=1}^{d} \rho_i^{nt_i} \mapsto \bigotimes_{i=1}^{n} \rho_{X_i(m)} = \sigma_m. \tag{8.19}$$

Since Π_m is a typical projector for σ_m, we have that

$$Tr(\Pi_m \sigma_m) \geq 1 - \epsilon \tag{8.20}$$

for each m.

In complete analogy to Bob's decoding procedure for Shannon's noisy channel coding theorem, Bob will do nothing other than to sequentially measure $\Pi_1, \Pi_2, \ldots$, and accept the first $m \in [M]$ for which the measurement of Π_m succeeds. We know that the chance of making the correct measurement is high, since

$$Tr(\Pi_m \sigma_m) \geq 1 - \epsilon \tag{8.21}$$

for each m. Let us consider the chance of failure. The non-commutative union bound justifies taking an upper bound for the chance of making the wrong measurement on a message m as simply the sum of making the wrong measurement $\Pi_{\hat{m}}$ on σ_m for each $\hat{m} \neq m$ (because we can ignore the factor of two and the square-root, because they asymptotically make no difference to the rate). The expectation (taken over all codebooks) of making the wrong measurement on a message m is

$$\sum_{\hat{m}:\hat{m}\neq m} \mathbb{E}_{X^n(m),X^n(\hat{m})} Tr(\Pi_{\hat{m}} \sigma_m) = \sum_{\hat{m}:\hat{m}\neq m} \mathbb{E}_{X^n(\hat{m})} Tr(\Pi_{\hat{m}} \rho^{\otimes n}) \tag{8.22}$$

$$= \underbrace{(M-1)}_{2^{nR}-1 \approx 2^{nR}} \exp\left(\underbrace{-\sum_{i=1}^{d} nt_i D(\rho_i||\rho)}_{-nI(X:Q)} \right) \tag{8.23}$$

$$\approx 2^{nR} 2^{-nI(X:Q)}, \tag{8.24}$$

so that indeed if $R < I(X:Q)$, then the probability of making the wrong measurement on m goes to zero. We need to, however, justify the claim that

$$-\sum_{i=1}^{d} nt_i D(\rho_i||\rho) \approx -nI(X:Q). \tag{8.25}$$

This is done quite straightforwardly, however, from the relation

$$I(X:Q)_\omega = \sum_i p(i) D(\rho_i||\rho) \tag{8.26}$$

and then the argument

$$|\sum_i nt_i D(\rho_i||\rho) - \sum_i np(i) D(\rho_i||\rho)| \leq n \underbrace{||t-p||}_{\text{Small with high probability}} \underbrace{\max_i D(\rho_i||\rho)}_{\leq \log d}. \tag{8.27}$$

This concludes the proof, although we comment that making this fully rigorous would mean taking care of various details that have been omitted. For example, taking the average over all codebooks

in the probability of making the wrong measurement only means that some codebook would work - we must fix such a codebook. Also, the last statement that $||t-p||$ is small with high probability is true, although we would need to get rid of all the m's for which this is large. By standard arguments that are very similar to that made last time (for the classical case), this does not mean getting rid of too many messages m.

This concludes our discussion of the proof of achievability.

8.4 Towards a Converse

We will not have full time to prove a converse in this lecture but will start to develop the tools to support the subsequent proof.

8.4.1 Fano's Inequality and Fannes' Inequality

If M and $\hat{M}$ are two random variables (over the same alphabet) that are very likely to be equal, then their conditional entropy is small. In particular, suppose that M and $\hat{M}$ are random variables over the set $\{0,1\}^{nR}$. Then

$$\mathbb{P}[M \neq \hat{N}] \leq \epsilon \implies H(M|\hat{N}) \leq \epsilon nR + 1. \tag{8.28}$$

Proof. Let p be the distribution for M given $\hat{M}$. Then, letting η be the probability that M does not equal $\hat{M}$ ($\eta \leq \epsilon$), we have

$$p = (1-\eta)1_{\hat{M}} + \eta q, \tag{8.29}$$

where $1_{\hat{M}}$ is the distribution concentrated on the value of $\hat{M}$, and q is some distribution with $q(\hat{M}) = 0$. From this, we compute the entropy of p as

$$-(1-\eta)\log(1-\eta) - \sum_x \eta q(x)(\log\eta + \log q(x)) = \underbrace{H_2(\eta)}_{\leq 1} + \underbrace{\eta}_{\leq \epsilon}\underbrace{H(q)}_{\leq nR}, \tag{8.30}$$

which concludes the proof. □

A further useful result is that of Fannes' Inequality, which we will not prove. This says that

$$|S(\rho) - S(\sigma)| \leq H_2(\epsilon) + \epsilon \log d, \tag{8.31}$$

where

$$\epsilon = \frac{1}{2}||\rho - \sigma||_1 (= T(\rho,\sigma)). \tag{8.32}$$

Note that this can be interpreted as a continuity statement for the von Neumann entropy S.

8.4.2 Conditional Mutual Information and Markov Chains

Another very useful tool for the proof of the converse will be that of the conditional mutual information (CMI), which relates closely with the theory of Markov Chains. The CMI of two random variables X and Y given the random variable Z is the mutual information between the random variables X and Y averaged over the output of Z being fixed, i.e.,

$$I(X:Y|Z) \coloneqq \sum_z p_Z(z) I(X:Y|Z=z) \tag{8.33}$$

$$= H(X|Z) + H(Y|Z) - H(XY|Z) \tag{8.34}$$

$$= H(XZ) + H(YZ) - H(XYZ) - H(Z). \tag{8.35}$$

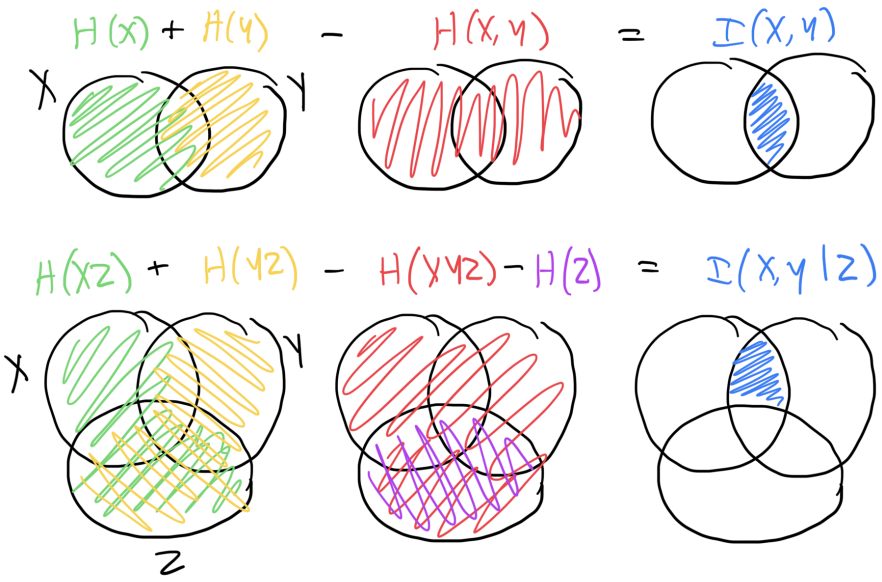

Figure 8.2: Mutual information and conditional mutual information as venn diagrams

These relations can be understood via Venn diagrams (Figure 8.2)

Whereas $I(X:Y)$ may be interpreted as a measure of the correlation between X and Y, $I(X:Y|Z)$ may be interpreted as the correlation between X and Y that remains once you have conditioned on Z. To illustrate this interpretation, we note that $I(X:Y|Z)=0$ if and only if the sequence

$$X \to Z \to Y \tag{8.36}$$

forms a Markov Chain.