PHYS 7895: Quantum Information Theory Fall 2015

Lecture 23 — November 11, 2015

Prof. Mark M. Wilde *Scribe: Mark M. Wilde*

This document is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

1 Overview

In the last lecture, we proved the entanglement concentration theorem, which states that the entanglement concentration limit for a bipartite pure state is equal to its entropy of entanglement.

In this lecture, we discuss classical communication over quantum channels. We focus mainly on the achievability part of the classical capacity theorem, using a relatively recent method called sequential decoding.

2 Introduction

This lecture begins our exploration of "dynamic" information-processing tasks in quantum Shannon theory, where the term "dynamic" indicates that a quantum channel connects a sender to a receiver and their goal is to exploit this resource for communication. We specifically consider the scenario where a sender Alice would like to communicate classical information to a receiver Bob, and the capacity theorem that we prove here is one particular generalization of Shannon's noisy channel coding theorem from classical information theory In later chapters, we will see other generalizations of Shannon's theorem, depending on what resources are available to assist their communication or depending on whether they are trying to communicate classical or quantum information. For this reason and others, quantum Shannon theory is quite a bit richer than classical information theory.

The naive approach to communicate classical information over a quantum channel is for Alice and Bob simply to mimic the approach used in Shannon's noisy channel coding theorem. That is, they select a random classical code according to some distribution $p_X(x)$, and Bob performs individual measurements of the outputs of a noisy quantum channel according to some POVM. The POVM at the output induces some conditional probability distribution $p_{Y|X}(y|x)$, which we can in turn think of as an induced noisy classical channel. The classical mutual information $I(X;Y)$ of this channel is an achievable rate for communication, and the best strategy for Alice and Bob is to optimize the mutual information over all of Alice's inputs to the channel and over all measurements that Bob could perform at the output. The resulting quantity is equal to Bob's optimized accessible information.

If the aforementioned coding strategy were optimal, then there would not be anything much interesting to say for the information-processing task of classical communication. This is perhaps one first clue that the above strategy is not necessarily optimal. Furthermore, we know from that the Holevo information is an upper bound to the accessible information, and this bound might prompt

us to wonder if it is also an achievable rate for classical communication, given that the accessible information is achievable.

The main theorem of this lecture is the classical capacity theorem (also known as the Holevo–Schumacher–Westmoreland theorem), and it states that the Holevo information of a quantum channel is an achievable rate for classical communication. The Holevo information is easier to manipulate mathematically than is the accessible information. The proof of its achievability demonstrates that the aforementioned strategy is not optimal, and the proof also shows how performing collective measurements over all of the channel outputs allows the sender and receiver to achieve the Holevo information as a rate for classical communication. Thus, this strategy fundamentally makes use of quantum-mechanical effects at the decoder and suggests that such an approach is necessary to achieve the Holevo information. Although this strategy exploits collective measurements at the decoder, it does not make use of entangled states at the encoder. That is, the sender could input quantum states that are entangled across all of the channel inputs, and this encoder entanglement might potentially increase classical communication rates.

One major drawback of the classical capacity theorem (also the case for many other results in quantum Shannon theory) is that it only demonstrates that the Holevo information is an achievable rate for classical communication—the converse theorem is a "multi-letter" converse, meaning that it might be necessary in the general case to evaluate the Holevo information over a potentially infinite number of uses of the channel. The multi-letter nature of the capacity theorem implies that the optimization task for general channels is intractable and thus further implies that we know very little about the actual classical capacity of general quantum channels. Now, there are many natural quantum channels such as the depolarizing channel and the dephasing channel for which the classical capacity is known (the Holevo information becomes "single-letter" for these channels), and these results imply that we have a complete understanding of the classical information-transmission capabilities of these channels. All of these results have to do with the additivity of the Holevo information of a quantum channel, which is discussed in the book.

We mentioned that the Holevo–Schumacher–Westmoreland coding strategy does not make use of entangled inputs at the encoder. But a natural question is to wonder whether entanglement at the encoder could boost classical information-transmission rates, given that it is a resource for many quantum protocols. This question was known as the additivity conjecture and went unsolved for many years, but recently Hastings offered a proof that entangled inputs can increase communication rates for certain channels. Thus, for these channels, the single-letter Holevo information is not the proper characterization of classical capacity (however, this is not to say that there could be some alternate characterization of the classical capacity other than the Holevo information which would be single-letter). These recent results demonstrate that we still know little about classical communication in the general case and furthermore that quantum Shannon theory is an active area of research.

3 Naive Approach: Product Measurements

We begin by discussing in more detail the most naive strategy that a sender and receiver can exploit for the transmission of classical information over many uses of a quantum channel. Figure 1 depicts this naive approach. This first approach mimics certain features of Shannon's classical approach without making any use of quantum-mechanical effects. Alice and Bob agree on a codebook be-

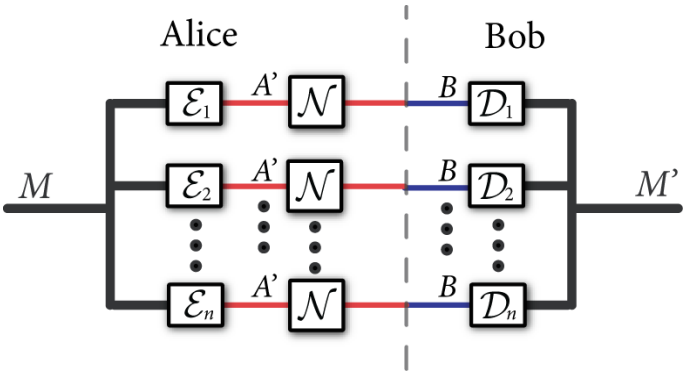

Figure 1: The most naive strategy for Alice and Bob to communicate classical information over many independent uses of a quantum channel. Alice wishes to send some message M and selects some tensor product state to input to the channel conditional on the message M. She transmits the codeword over the channel, and Bob then receives a noisy version of it. He performs individual measurements of his quantum systems and produces some estimate M' of the original message M. This scheme is effectively a classical scheme because it makes no use of quantum-mechanical features such as entanglement.

forehand, where each classical codeword $x^n(m)$ in the codebook corresponds to some message m that Alice wishes to transmit. Alice can exploit some alphabet $\{\rho_x\}$ of density operators to act as input to the quantum channel. That is, the quantum codewords are of the form

$$\rho_{x^n(m)} \equiv \rho_{x_1(m)} \otimes \rho_{x_2(m)} \otimes \cdots \otimes \rho_{x_n(m)}. \tag{1}$$

Bob then performs individual measurements of the outputs of the quantum channel by exploiting some POVM $\{\Lambda_y\}$. This scheme induces the following conditional probability distribution:

$$\begin{aligned}
&p_{Y_1\cdots Y_n|X_1\cdots X_n}(y_1\cdots y_n|x_1(m)\cdots x_n(m)) \\
&= \operatorname{Tr}\left\{\Lambda_{y_1}\otimes\cdots\otimes\Lambda_{y_n}\left(\mathcal{N}\otimes\cdots\otimes\mathcal{N}\right)\left(\rho_{x_1(m)}\otimes\cdots\otimes\rho_{x_n(m)}\right)\right\} &(2)\\
&= \operatorname{Tr}\left\{\left(\Lambda_{y_1}\otimes\cdots\otimes\Lambda_{y_n}\right)\left(\mathcal{N}\left(\rho_{x_1(m)}\right)\otimes\cdots\otimes\mathcal{N}\left(\rho_{x_n(m)}\right)\right)\right\} &(3)\\
&= \prod_{i=1}^{n}\operatorname{Tr}\left\{\Lambda_{y_i}\mathcal{N}\left(\rho_{x_i(m)}\right)\right\}, &(4)
\end{aligned}$$

which we immediately realize is many i.i.d. instances of the following classical channel:

$$p_{Y|X}(y|x) \equiv \operatorname{Tr}\left\{\Lambda_y\mathcal{N}(\rho_x)\right\}. \tag{5}$$

Thus, if they exploit this scheme, the optimal rate at which they can communicate is equal to the following expression:

$$I_{\text{acc}}(\mathcal{N}) \equiv \max_{\{p_X(x),\rho_x,\Lambda\}} I(X;Y), \tag{6}$$

where the maximization of the classical mutual information is over all input distributions, all input density operators, and all POVMs that Bob could perform at the output of the channel. This information quantity is known as the accessible information of the channel.

The above strategy is not necessarily an optimal strategy if the channel is truly a quantum channel—it does not make use of any quantum effects such as entanglement. A first simple modification of the

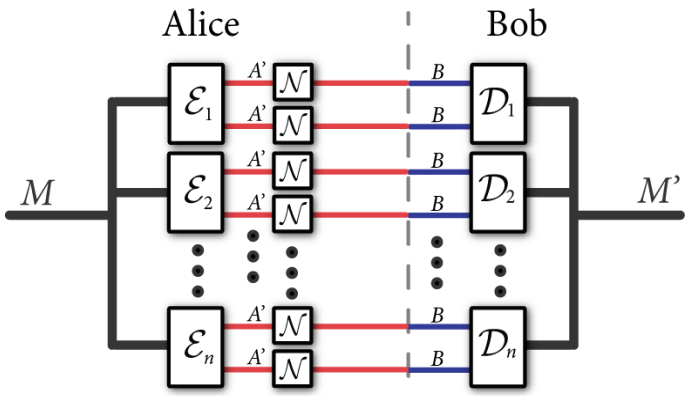

Figure 2: A coding strategy that can outperform the previous naive strategy, simply by making use of entanglement at the encoder and decoder.

protocol to allow for such effects would be to consider coding for the tensor product channel $\mathcal{N} \otimes \mathcal{N}$ rather than the original channel. The input states would be entangled across two channel uses, and the output measurements would be over two channel outputs at a time. In this way, they would be exploiting entangled states at the encoder and collective measurements at the decoder. Figure 2 illustrates the modified protocol, and the rate of classical communication that they can achieve with such a strategy is $\frac{1}{2} I_{\text{acc}}(\mathcal{N} \otimes \mathcal{N})$. This quantity is always at least as large as $I_{\text{acc}}(\mathcal{N})$ because a special case of the strategy for the tensor product channel $\mathcal{N} \otimes \mathcal{N}$ is to choose the distribution $p_X(x)$, the states ρ_x, and the POVM Λ to be tensor products of the ones that maximize $I_{\text{acc}}(\mathcal{N})$. We can then extend this construction inductively by forming codes for the tensor product channel $\mathcal{N}^{\otimes k}$ (where k is a positive integer), and this extended strategy achieves the classical communication rate of $\frac{1}{k} I_{\text{acc}}\left(\mathcal{N}^{\otimes k}\right)$ for any finite k. These results then suggest that the ultimate classical capacity of the channel is the regularization of the accessible information of the channel:

$$I_{\text{reg}}(\mathcal{N}) \equiv \lim_{k \to \infty} \frac{1}{k} I_{\text{acc}}\left(\mathcal{N}^{\otimes k}\right). \tag{7}$$

The regularization of the accessible information is intractable for general quantum channels, but the optimization task could simplify immensely if the accessible information is additive. In this case, the regularized accessible information $I_{\text{reg}}(\mathcal{N})$ would be equivalent to the accessible information $I_{\text{acc}}(\mathcal{N})$. However, even if the quantity is additive, the optimization could still be difficult to perform in practice. A simple upper bound on the accessible information is the Holevo information $\chi(\mathcal{N})$ of the channel, defined as

$$\chi(\mathcal{N}) \equiv \max_{\rho} I(X;B), \tag{8}$$

where the maximization is over classical–quantum states ρ_{XB} of the following form:

$$\rho_{XB} \equiv \sum_x p_X(x) |x\rangle\langle x|_X \otimes \mathcal{N}_{A' \to B}(\psi^x_{A'}). \tag{9}$$

The Holevo information is a more desirable quantity to characterize classical communication over a quantum channel because it is always an upper bound on the accessible information and because Theorem **??** states that it is sufficient to consider pure states $\psi^x_{A'}$ at the channel input for maximizing the Holevo information.

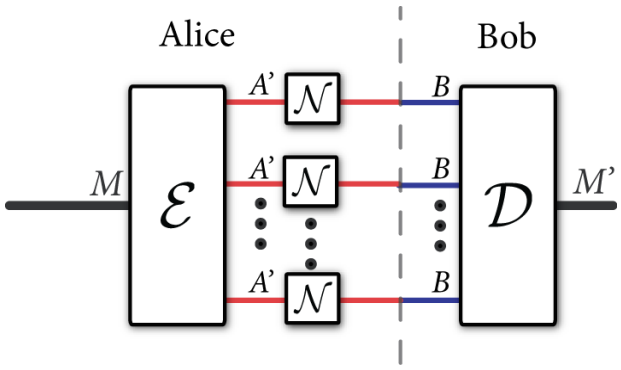

Figure 3: The most general protocol for classical communication over a quantum channel. Alice selects some message M and encodes it as a quantum codeword for input to many independent uses of the noisy quantum channel. Bob performs some POVM over all of the channel outputs to determine the message that Alice transmits.

Thus, a natural question to ask is whether Alice and Bob can achieve the Holevo information rate, and the main theorem of this chapter states that it is possible to do so. The resulting coding scheme bears some similarities with the techniques in Shannon's noisy channel coding theorem, but the main difference is that the decoding POVM is a collective measurement over all of the channel outputs.

4 The Information-Processing Task

4.1 Classical Communication

We now discuss the most general form of the information-processing task and give the criterion for a classical communication rate C to be achievable—i.e., we define an $(n, C - \delta, \varepsilon)$ code for classical communication over a quantum channel. Alice begins by selecting some classical message m that she would like to transmit to Bob—she selects from a set of messages $\{1, \ldots, |\mathcal{M}|\}$. Let M denote the random variable corresponding to Alice's choice of message, and let $|\mathcal{M}|$ denote its cardinality. She then prepares some state $\rho^m_{A'^n}$ as input to the many independent uses of the channel—the input systems are n copies of the channel input system A'. She transmits this state over n independent uses of the channel $\mathcal{N}$, and the state at Bob's receiving end is

$$\mathcal{N}^{\otimes n}(\rho^m_{A'^n}). \tag{10}$$

Bob has some decoding POVM $\{\Lambda_m\}$ that he can exploit to determine which message Alice transmits. Figure 3 depicts such a general protocol for classical communication over a quantum channel.

Let M' denote the random variable for Bob's estimate of the message. The probability that he determines the correct message m is as follows:

$$\Pr\left\{M = m | M' = m\right\} = \operatorname{Tr}\left\{\Lambda_m \mathcal{N}^{\otimes n}(\rho^m_{A'^n})\right\}, \tag{11}$$

and thus the probability of error for a particular message m is

$$p_e(m) \equiv 1 - \Pr\left\{M = m | M' = m\right\} \tag{12}$$

$$= \mathrm{Tr}\left\{(I - \Lambda_m)\,\mathcal{N}^{\otimes n}(\rho^m_{A'^n})\right\}. \tag{13}$$

The maximal probability of error for any coding scheme is then

$$p^*_e \equiv \max_{m \in \mathcal{M}} p_e(m). \tag{14}$$

The rate C of communication is

$$C \equiv \frac{1}{n}\log_2 |\mathcal{M}| + \delta, \tag{15}$$

where δ is some arbitrarily small positive number, and the code has ε error if $p^*_e \leq \varepsilon$. A rate C of classical communication is *achievable* if there exists an $(n, C - \delta, \varepsilon)$ code for all $\delta, \varepsilon > 0$ and sufficiently large n.

5 Sequential Decoding

In this section (taken from arXiv:1202.0518), we describe the operation of a sequential decoder that can reliably recover classical information encoded into a pure state ensemble. We follow with a full error analysis, demonstrating that the scheme achieves the Holevo rate for pure-state channels. Suppose that a classical-quantum channel of the form

$$x \to |\phi_x\rangle$$

connects a sender Alice to a receiver Bob. For our purposes here, it does not matter whether the classical input x is discrete or continuous.

Theorem 1. *Let $x \to |\phi_x\rangle$ be a classical-quantum channel and let $\rho \equiv \sum_x p_X(x)|\phi_x\rangle\langle\phi_x|$ for some distribution $p_X(x)$. Then the rate $H(\rho)$ bits per channel use is achievable for communication over this channel by having the receiver employ a sequential decoding strategy.*

Proof. We break the proof into several steps.

Codebook Construction. Before communication begins, Alice and Bob agree upon a codebook. We allow them to select a codebook randomly according to the distribution $p_X(x)$. So, for every message $m \in \mathcal{M} \equiv \left\{1, \ldots, 2^{nR}\right\}$, generate a codeword $x^n(m) \equiv x_1(m) \cdots x_n(m)$ randomly and independently according to

$$p_{X^n}(x^n) \equiv \prod_{i=1}^{n} p_X(x_i).$$

Sequential Decoding. Transmitting the codeword $x^n(m)$ through n uses of the channel $x \to |\phi_x\rangle$ leads to the following quantum state at Bob's output:

$$|\phi_{x^n(m)}\rangle \equiv |\phi_{x_1(m)}\rangle \otimes \cdots \otimes |\phi_{x_n(m)}\rangle.$$

Upon receiving the quantum codeword $|\phi_{x^n(m)}\rangle$, Bob performs a sequence of binary-outcome quantum measurements to determine the classical codeword $x^n(m)$ that Alice transmitted. He

first "asks," "Is it the first codeword?" by performing the measurement $\{|\phi_{x^n(1)}\rangle\langle\phi_{x^n(1)}|, I^{\otimes n} - |\phi_{x^n(1)}\rangle\langle\phi_{x^n(1)}|\}$. If he receives the outcome "yes," then he performs no further measurements and concludes that Alice transmitted the codeword $x^n(1)$. If he receives the outcome "no," then he performs the measurement $\{|\phi_{x^n(2)}\rangle\langle\phi_{x^n(2)}|, I^{\otimes n} - |\phi_{x^n(2)}\rangle\langle\phi_{x^n(2)}|\}$ to check if Alice sent the second codeword. Similarly, he stops if he receives "yes," and otherwise, he proceeds along similar lines.

Error Analysis. We now provide an error analysis demonstrating that this scheme works well, i.e., the word error goes to zero as $n \to \infty$, as long as $R < H(\rho)$. In general, if Alice transmits the m^{th} codeword, then the probability for Bob to decode correctly with this sequential decoding strategy is as follows:

$$\text{Tr}\left\{\phi_{x^n(m)}\hat{\Pi}_{m-1}\cdots\hat{\Pi}_1\phi_{x^n(m)}\hat{\Pi}_1\cdots\hat{\Pi}_{m-1}\phi_{x^n(m)}\right\},$$

where we make the abbreviations

$$\begin{aligned}\phi_{x^n(m)} &\equiv |\phi_{x^n(m)}\rangle\langle\phi_{x^n(m)}|,\\ \hat{\Pi}_i &\equiv I^{\otimes n} - |\phi_{x^n(i)}\rangle\langle\phi_{x^n(i)}|.\end{aligned}$$

So the probability that Bob makes an error when decoding the m^{th} codeword is just

$$1 - \text{Tr}\left\{\phi_{x^n(m)}\hat{\Pi}_{m-1}\cdots\hat{\Pi}_1\phi_{x^n(m)}\hat{\Pi}_1\cdots\hat{\Pi}_{m-1}\phi_{x^n(m)}\right\}.$$

To further simplify the error analysis, we consider the expectation of the above error probability, under the assumption that Alice selects a message uniformly at random according to a random variable M and that the codeword x^n is selected at random according to the distribution $p_{X^n}(x^n)$ (as described above):

$$1 - \mathop{\mathbb{E}}_{X^n,M}\text{Tr}\left\{\phi_{X^n(M)}\hat{\Pi}_{M-1}\cdots\hat{\Pi}_1\phi_{X^n(M)}\hat{\Pi}_1\cdots\hat{\Pi}_{M-1}\right\}. \tag{16}$$

For the rest of the proof, it is implicit that the expectation $\mathbb{E}$ is with respect to random variables X^n and M. Our first observation is that, for the purposes of our error analysis, we can "smooth" the channel $x^n \to \phi_{x^n}$, by imagining instead that we are coding for a projected version of the channel $\Pi\ \phi_{x^n}\ \Pi$, where Π is the typical projector for the average state $\rho \equiv \sum_x p_X(x)\phi_x$. Doing so simplifies the error analysis by cutting off large eigenvalues that reside outside of the high-probability typical subspace. Furthermore, we expect that doing so should not affect the error analysis very much because most of the probability tends to concentrate in this subspace anyway. That we can do so follows from the fact that

$$\begin{aligned}1 &= \mathbb{E}\text{Tr}\left\{\phi_{X^n(M)}\right\}\\ &= \mathbb{E}\text{Tr}\left\{\Pi\phi_{X^n(M)}\right\} + \mathbb{E}\text{Tr}\left\{\hat{\Pi}\phi_{X^n(M)}\right\}\\ &= \mathbb{E}\text{Tr}\left\{\Pi\phi_{X^n(M)}\Pi\right\} + \text{Tr}\left\{\hat{\Pi}\mathbb{E}\phi_{X^n(M)}\right\}\\ &= \mathbb{E}\text{Tr}\left\{\Pi\phi_{X^n(M)}\Pi\right\} + \text{Tr}\left\{\hat{\Pi}\rho^{\otimes n}\right\},\end{aligned}$$

where $\hat{\Pi} \equiv I - \Pi$. Furthermore, we know that

$$\begin{aligned}&\mathbb{E}\text{Tr}\left\{\phi_{X^n(M)}\hat{\Pi}_{M-1}\cdots\hat{\Pi}_1\phi_{X^n(M)}\hat{\Pi}_1\cdots\hat{\Pi}_{M-1}\right\}\\ &= \mathbb{E}\text{Tr}\left\{\hat{\Pi}_1\cdots\hat{\Pi}_{M-1}\phi_{X^n(M)}\hat{\Pi}_{M-1}\cdots\hat{\Pi}_1\phi_{X^n(M)}\right\}\\ &\geq \mathbb{E}\text{Tr}\left\{\hat{\Pi}_1\cdots\hat{\Pi}_{M-1}\phi_{X^n(M)}\hat{\Pi}_{M-1}\cdots\hat{\Pi}_1\Pi\phi_{X^n(M)}\Pi\right\} - \mathbb{E}\left\|\phi_{X^n(M)} - \Pi\phi_{X^n(M)}\Pi\right\|_1,\end{aligned}$$

where the inequality follows from the following lemma:

Lemma 2. *Let ρ and σ be such that $0 \leq \rho, \sigma$ and $\mathrm{Tr}\{\rho\}, \mathrm{Tr}\{\sigma\} \leq 1$. Let Λ be such that $0 \leq \Lambda \leq I$. Then*

$$\mathrm{Tr}\left[\Lambda\rho\right] \leq \mathrm{Tr}\left[\Lambda\sigma\right] + \left\|\rho - \sigma\right\|_1. \tag{17}$$

Proof. This follows from a variational characterization of trace distance as the distinguishability of the states under an optimal measurement M: $\left\|\rho - \sigma\right\|_1 = 2\max_{0\leq M\leq I} \mathrm{Tr}\left[M(\rho - \sigma)\right]$. □

We need another lemma, known as the Gentle Operator Lemma:

Lemma 3 (Gentle Operator Lemma for Ensembles)**.** *Given an ensemble $\{p_X(x), \rho_x\}$ with expected density operator $\rho \equiv \sum_x p_X(x)\rho_x$, suppose that an operator Λ such that $I \geq \Lambda \geq 0$ succeeds with high probability on the state ρ:*

$$\mathrm{Tr}\left\{\Lambda\rho\right\} \geq 1 - \varepsilon.$$

Then the subnormalized state $\sqrt{\Lambda}\rho_x\sqrt{\Lambda}$ is close in expected trace distance to the original state ρ_x:

$$\mathbb{E}_X\left\{\left\|\sqrt{\Lambda}\rho_X\sqrt{\Lambda} - \rho_X\right\|_1\right\} \leq 2\sqrt{\varepsilon}.$$

Using the above observations and the facts that

$$\mathbb{E}\left\|\phi_{X^n(M)} - \Pi\phi_{X^n(M)}\Pi\right\|_1 \leq 2\sqrt{\varepsilon}, \tag{18}$$

$$\mathrm{Tr}\left\{\hat{\Pi}\rho^{\otimes n}\right\} \leq \varepsilon, \tag{19}$$

for all $\varepsilon > 0$ whenever n is sufficiently large (these are from the properties of typicality and the Gentle Operator Lemma, we obtain the following upper bound on (16):

$$\mathbb{E}\mathrm{Tr}\left\{\Pi\phi_{X^n(M)}\Pi\right\} - \mathbb{E}\mathrm{Tr}\left\{\phi_{X^n(M)}\hat{\Pi}_{M-1}\cdots\hat{\Pi}_1\Pi\phi_{X^n(M)}\Pi\hat{\Pi}_1\cdots\hat{\Pi}_{M-1}\phi_{X^n(M)}\right\} + \varepsilon + 2\sqrt{\varepsilon}.$$

(In the next steps, we omit the terms $\varepsilon + 2\sqrt{\varepsilon}$ as they are negligible.) The most important step of this error analysis is to apply Pranab Sen's non-commutative union bound (Lemma 3 of arXiv:1109.0802), which holds for any subnormalized state σ ($\sigma \geq 0$ and $\mathrm{Tr}\{\sigma\} \leq 1$) and sequence of projectors $\Pi_1, \ldots, \Pi_N$:

$$\mathrm{Tr}\left\{\sigma\right\} - \mathrm{Tr}\left\{\Pi_N\cdots\Pi_1\sigma\Pi_1\cdots\Pi_N\right\} \leq 2\sqrt{\sum_{i=1}^{N}\mathrm{Tr}\left\{\left(I - \Pi_i\right)\sigma\right\}}$$

For our case, we take $\Pi\phi_{X^n(M)}\Pi$ as σ and $\phi_{X^n(M)}$, $\hat{\Pi}_{M-1}$, …, $\hat{\Pi}_1$ as the sequence of projectors. Applying Sen's bound and concavity of the square root function leads to the following upper bound on (**??**):

$$2\sqrt{\mathbb{E}\mathrm{Tr}\left\{\hat{\Pi}_M\Pi\phi_{X^n(M)}\Pi\right\} + \mathbb{E}\sum_{i=1}^{M-1}\mathrm{Tr}\left\{\phi_{X^n(i)}\Pi\phi_{X^n(M)}\Pi\right\}}$$

where $\hat{\Pi}_M = I^{\otimes n} - \phi_{X^n(M)}$ and $\phi_{X^n(i)} = I^{\otimes n} - \hat{\Pi}_i$. We now bound each of the above two terms individually. For the first term, consider that

$$\begin{aligned}
&\mathbb{E}\mathrm{Tr}\left\{\hat{\Pi}_M \Pi \phi_{X^n(M)} \Pi\right\} \\
&\leq \mathbb{E}\mathrm{Tr}\left\{\hat{\Pi}_M \phi_{X^n(M)}\right\} + \mathbb{E}\left\|\phi_{X^n(M)} - \Pi\phi_{X^n(M)}\Pi\right\|_1 \\
&\leq 2\sqrt{\varepsilon}.
\end{aligned}$$

where the last inequality follows from applying (18) and because

$$\begin{aligned}
\mathrm{Tr}\left\{\hat{\Pi}_M \phi_{X^n(M)}\right\} &= \mathrm{Tr}\left\{\left(I^{\otimes n} - \phi_{X^n(M)}\right)\phi_{X^n(M)}\right\} \\
&= 0.
\end{aligned}$$

For the second term, consider that

$$\begin{aligned}
&\mathbb{E}\sum_{i=1}^{M-1} \mathrm{Tr}\left\{\phi_{X^n(i)}\Pi\phi_{X^n(M)}\Pi\right\} \\
&\leq \mathbb{E}_M \sum_{i\neq M} \mathbb{E}_{X^n}\mathrm{Tr}\left\{\phi_{X^n(i)}\Pi\phi_{X^n(M)}\Pi\right\} \\
&= \mathbb{E}_M \sum_{i\neq M} \mathrm{Tr}\left\{\mathbb{E}_{X^n}\left\{\phi_{X^n(i)}\right\}\Pi\mathbb{E}_{X^n}\left\{\phi_{X^n(M)}\right\}\Pi\right\} \\
&= \sum_{i\neq M} \mathrm{Tr}\left\{\rho^{\otimes n}\Pi\rho^{\otimes n}\Pi\right\} \\
&\leq 2^{-n[H(\rho)-\delta]} \sum_{i\neq M} \mathrm{Tr}\left\{\rho^{\otimes n}\Pi\right\} \\
&\leq 2^{-n[H(\rho)-\delta]}\ |\mathcal{M}|
\end{aligned}$$

The first inequality follows by just adding in all of the future terms $i > M$ to the sum. The first equality follows because the random variables $X^n(i)$ and $X^n(M)$ are independent, due to the way that we selected the code (each codeword is selected independently of a different one). The second equality follows from averaging the state ϕ_{X^n} with respect to the distribution p_{X^n}, and we drop the expectation $\mathbb{E}_M$ because the quantities inside the trace no longer have a dependence on the message M. The second inequality follows from the entropy bound for the eigenvalues of $\rho^{\otimes n}$ in the typical subspace. The final inequality follows because $\mathrm{Tr}\{\rho^{\otimes n}\Pi\} \leq 1$. Thus, the overall upper bound on the error probability with this sequential decoding strategy is

$$\varepsilon' \equiv \varepsilon + 2\sqrt{\varepsilon} + 2\sqrt{2\sqrt{\varepsilon} + 2^{-n[H(\rho)-\delta]}\ |\mathcal{M}|},$$

which we can make arbitrarily small by choosing $|\mathcal{M}| = 2^{n[H(\rho)-2\delta]}$ and n sufficiently large. The next arguments are standard. We proved a bound on the expectation of the average probability, which implies there exists a particular code that has arbitrarily small average error probability under the same choice of $|\mathcal{M}|$ and n. For this code, we can then eliminate the worst half of the codewords, ensuring that the error probability of the resulting code is no larger than $2\varepsilon'$. Furthermore, it should be clear that it is only necessary for the sequential decoder to process the remaining codewords when decoding messages. □

6 Sequential Decoding for Optical Communication

We now provide a physical realization of the sequential decoding strategy in the context of optical communications. In this setting, we suppose that a lossy bosonic channel, specified by the following Heisenberg relations, connects Alice to Bob:

$$\hat{b} = \sqrt{\eta}\hat{a} + \sqrt{1-\eta}\hat{e}, \tag{20}$$

where $\hat{a}$, $\hat{b}$, and $\hat{e}$ are the respective field operators for Alice's input mode, Bob's output mode, and an environmental input mode (assumed to be in its vacuum state). The transmissivity $\eta \in [0,1]$ is the fraction of Alice's input photons that make it to Bob on average. We assume that Alice is constrained to using mean photon number N_S per channel use. The strategy for achieving the classical capacity of this channel is for Alice to induce a classical-quantum channel, by selecting $\alpha \in \mathbb{C}$ and preparing a coherent state $|\alpha\rangle$ at the input of the channel in (20). A coherent state in quantum optics is defined as the following coherent superposition of photon number states:

$$|\alpha\rangle \equiv \exp\left\{\frac{-|\alpha|^2}{2}\right\} \sum_{n=0}^{\infty} \frac{\alpha^n}{\sqrt{n!}} |n\rangle.$$

It is often described as being the ideal state of a single mode of the light field output from a laser. The most useful property of coherent states for classical communication over a pure-loss bosonic channel is that it retains its purity. That is, if Alice inputs the state $|\alpha\rangle$ to the pure-loss bosonic channel with transmissivity η, then the state output for Bob and Eve is

$$|\sqrt{\eta}\alpha\rangle \otimes |\sqrt{1-\eta}\alpha\rangle,$$

so that we recover a pure coherent state for Bob when tracing over the second mode. The resulting induced classical-quantum channel to Bob is of the following form:

$$\alpha \to |\sqrt{\eta}\alpha\rangle.$$

By choosing the distribution $p_X(x)$ in Theorem 1 to be an isotropic, complex Gaussian with variance N_S:

$$p_{N_S}(\alpha) \equiv (1/\pi N_S) \exp\left\{-|\alpha|^2/N_S\right\},$$

we have that $g(\eta N_S)$ is an achievable rate for classical communication, where

$$g(x) \equiv (x+1)\log(x+1) - x\log x.$$

The quantity $g(\eta N_S)$ is the entropy of the average state of the ensemble $\{p_{N_S}(\alpha),\ |\sqrt{\eta}\alpha\rangle\}$:

$$\int d^2\alpha\ p_{N_S}(\alpha)\, |\sqrt{\eta}\alpha\rangle\langle\sqrt{\eta}\alpha|,$$

which is a thermal state with mean photon number ηN_S. Each quantum codeword selected from the ensemble $\{p_{N_S}(\alpha),\ |\alpha\rangle\}$ has the following form:

$$|\alpha^n(m)\rangle \equiv |\alpha_1(m)\rangle \otimes \cdots \otimes |\alpha_n(m)\rangle.$$

We assume $\eta = 1$ above and for the rest of this section without loss of generality. Thus, the sequential decoder consists of measurements of the following form for all $m \in \mathcal{M}$:

$$\left\{|\alpha^n(m)\rangle\langle\alpha^n(m)|,\ I^{\otimes n} - |\alpha^n(m)\rangle\langle\alpha^n(m)|\right\}. \tag{21}$$

Observing that

$$|\alpha^n(m)\rangle = D(\alpha_1(m)) \otimes \cdots \otimes D(\alpha_n(m)) |0\rangle^{\otimes n},$$

where $D(\alpha) \equiv \exp\left\{\alpha \hat{a}^\dagger - \alpha^* \hat{a}\right\}$ is the unitary "displacement" operator from quantum optics and $|0\rangle^{\otimes n}$ is the n-fold tensor product vacuum state, we see that that the decoder can implement the measurement in (21) in three steps:

1. Displace the n-mode codeword state by

$$D(-\alpha_1(m)) \otimes \cdots \otimes D(-\alpha_n(m)),$$

   by employing highly asymmetric beam-splitters with a strong local oscillator [**?**].

2. Perform a "vacuum-or-not" measurement of the form

$$\left\{|0\rangle\langle 0|^{\otimes n}, \quad I^{\otimes n} - |0\rangle\langle 0|^{\otimes n}\right\}.$$

   If the vacuum outcome occurs, decode as the m^{th} codeword. Otherwise, proceed.

3. Displace by $D(\alpha_1(m)) \otimes \cdots \otimes D(\alpha_n(m))$ with the same method as in Step 1.

The receiver just iterates this strategy for every codeword in the codebook, and Theorem 1 states this strategy is capacity-achieving.

7 Non-Commutative Union Bound Proof

Theorem 4 (Non-Commutative Union Bound)**.** *Let σ be such that $\sigma \geq 0$ and $\mathrm{Tr}\{\sigma\} \leq 1$. Let Π_1, $\ldots$, Π_L be Hermitian projectors. Then*

$$\mathrm{Tr}\{\sigma\} - \mathrm{Tr}\{\Pi_L \cdots \Pi_1 \sigma \Pi_1 \cdots \Pi_L\} \leq 2\sqrt{\sum_{i=1}^{L} \mathrm{Tr}\{(I - \Pi_i)\sigma\}}. \tag{22}$$

Proof. It suffices to prove the following bound for a vector $|\psi\rangle$ such that $\||\psi\rangle\|_2^2 \leq 1$:

$$\||\psi\rangle\|_2^2 - \|\Pi_L \cdots \Pi_1 |\psi\rangle\|_2^2 \leq 2\sqrt{\sum_{i=1}^{L} \|(I - \Pi_i)|\psi\rangle\|_2^2}. \tag{23}$$

This is because

$$\||\psi\rangle\|_2^2 = \mathrm{Tr}\{|\psi\rangle\langle\psi|\}, \tag{24}$$

$$\|\Pi_L \cdots \Pi_1 |\psi\rangle\|_2^2 = \mathrm{Tr}\{\Pi_L \cdots \Pi_1 |\psi\rangle\langle\psi| \Pi_1 \cdots \Pi_L\}, \tag{25}$$

$$\|(I - \Pi_i)|\psi\rangle\|_2^2 = \mathrm{Tr}\{(I - \Pi_i)|\psi\rangle\langle\psi|\}, \tag{26}$$

and any σ satisfying the conditions given can be written as a convex combination $\sigma = \sum_z p(z)|\psi_z\rangle\langle\psi_z|$ where $p(z)$ is a probability distribution and each $|\psi_z\rangle$ satisfies $\||\psi_z\rangle\|_2^2 \leq 1$. Then (22) follows from (23) by concavity of the square root function. So we now focus on proving (23).

We begin by showing that

$$\| |\psi\rangle - \Pi_L \cdots \Pi_1 |\psi\rangle \|_2^2 \leq \sum_{i=1}^{L} \| (I - \Pi_i) |\psi\rangle \|_2^2 . \tag{27}$$

To see this, consider that

$$\begin{aligned} \| |\psi\rangle - \Pi_L \cdots \Pi_1 |\psi\rangle \|_2^2 &= \| (I - \Pi_L) |\psi\rangle \|_2^2 + \| \Pi_L \left(|\psi\rangle - \Pi_{L-1} \cdots \Pi_1 |\psi\rangle \right) \|_2^2 && (28) \\ &\leq \| (I - \Pi_L) |\psi\rangle \|_2^2 + \| |\psi\rangle - \Pi_{L-1} \cdots \Pi_1 |\psi\rangle \|_2^2 && (29) \\ &\leq \sum_{i=1}^{L} \| (I - \Pi_i) |\psi\rangle \|_2^2 . && (30) \end{aligned}$$

The first equality follows from Pythagorean's theorem. The first inequality follows because a projection cannot increase the norm of a vector. The last inequality is by induction. Now we take the square root of (27):

$$\| |\psi\rangle - \Pi_L \cdots \Pi_1 |\psi\rangle \|_2 \leq \sqrt{\sum_{i=1}^{L} \| (I - \Pi_i) |\psi\rangle \|_2^2}, \tag{31}$$

from which we can conclude the following by the triangle inequality:

$$\| |\psi\rangle \|_2 - \| \Pi_L \cdots \Pi_1 |\psi\rangle \|_2 \leq \sqrt{\sum_{i=1}^{L} \| (I - \Pi_i) |\psi\rangle \|_2^2}. \tag{32}$$

Then rearrange this as follows:

$$\| |\psi\rangle \|_2 - \sqrt{\sum_{i=1}^{L} \| (I - \Pi_i) |\psi\rangle \|_2^2} \leq \| \Pi_L \cdots \Pi_1 |\psi\rangle \|_2 \tag{33}$$

and square both sides to get

$$\begin{aligned} &\left(\| |\psi\rangle \|_2 - \sqrt{\sum_{i=1}^{L} \| (I - \Pi_i) |\psi\rangle \|_2^2} \right)^2 \\ &= \| |\psi\rangle \|_2^2 - 2 \sqrt{\sum_{i=1}^{L} \| (I - \Pi_i) |\psi\rangle \|_2^2} + \sum_{i=1}^{L} \| (I - \Pi_i) |\psi\rangle \|_2^2 && (34) \\ &\leq \| \Pi_L \cdots \Pi_1 |\psi\rangle \|_2^2 . && (35) \end{aligned}$$

This then implies (23) by dropping the non-negative term $\sum_{i=1}^{L} \| (I - \Pi_i) |\psi\rangle \|_2^2$. □