Classical Capacities of Classical and Quantum Channels

by

Silvester Borsboom

To obtain the degree of Bachelor of Science in Applied Mathematics and Applied Physics
at Delft University of Technology

Student number:	4707958
Supervisors:	Dr. B. Janssens (EEMCS)
	Dr. D. Elkouss (AS)
Other committee members:	Prof. Dr. D. Gijswijt (EEMCS)
	Dr. S. Groeblacher (AS)

Delft, August 2020

Abstract

This thesis investigates two types of classical capacities of both classical and quantum channels, giving rise to four different settings. The first type of classical capacity investigated is the ordinary capacity of a channel to transmit classical information with a probability of error which becomes arbitrarily small as the channel is used arbitrarily many times. The second type of classical capacity investigated is the capacity of a channel to transmit information with zero probability of error, called the zero-error classical capacity. The first setting which is studied is the ordinary capacity of a classical channel. The noisy channel coding theorem is proven in two different ways: one using the Markov inequality and the Law of Large Numbers and one using typical sets. The additivity of this capacity is also discussed. The second setting is the zero-error capacity of a classical channel. Lower and upper bounds on this capacity are proven, and its superadditivity is discussed. The third setting is the ordinary classical capacity of a quantum channel. The Holevo-Schumacher-Westmoreland theorem is proven using typical subspaces and the packing lemma, and the superadditivity of the Holevo information is discussed in terms of entanglement at the encoder. The fourth and last setting investigated is that of the zero-error classical capacity of a quantum channel. It is shown that this capacity can be achieved using only pure input states and that this capacity never exceeds the ordinary classical capacity. Moreover, a detailed investigation of superactivation of the zero-error classical capacity is presented. A topic for further research would be an exposition of the analogous concepts in the case of the quantum capacity of quantum channels. Another topic for further research would be an explicit construction of two quantum channels whose zero-error classical capacity is superactivated.

Contents

1
Introduction

Almost a century ago, the theory of quantum mechanics was invented by physicists such as Dirac, Heisenberg, Schrödinger and others, and was then given a rigorous mathematical foundation, chiefly by von Neumann. Quantum mechanics exhibits a plethora of counter-intuitive phenomena, one of which is called entanglement, which was described by Einstein as "spooky action at a distance".

Today, quantum theory has found direct applications in modern technology. An important example of such an application is quantum communication. The idea of quantum communication is to transmit information over so-called quantum channels, rather than over classical channels. By exploiting quantum phenomena such as entanglement, quantum communication has great advantages over classical communication, such as unbreakable encryption of messages.

Communication, in the above sense, is simply the transmission of information over some kind of channel. If every channel were perfectly reliable, then such transmission of information would pose few problems. In reality, however, channels are almost never completely devoid of noise. The point of communication theory, therefore, is to understand how one can reliably transmit information over a noisy channel.

An important property of a channel in communication theory is its capacity. The capacity is, generally speaking, the amount of information that can be reliably transmitted per use of the channel. The exact definition of the capacity, however, depends crucially on the type of information one wishes to transmit, on whether one uses a classical or quantum channel and on what one calls reliable communication. Indeed, there are many different types of capacities that can be defined for different channels. These can all serve as a measure of the performance of a channel.

Perhaps the most important distinction between different types of capacities is the one between classical and quantum capacity: one can transmit both classical and quantum information over a quantum channel. In this thesis we focus solely on the transmission of classical information and therefore on classical capacities. Now, a pivotal point to understand is that classical information over quantum channels can still make use of quantum phenomena. The classical information is encoded into quantum states which are subsequently transmitted over the channel. The quantum states are then decoded again into classical information. Overall, one thus transmits classical information, but by encoding into quantum states and subsequently decoding into classical information again, one can still exploit the features of the quantum world.

Although we only study classical capacities in this thesis, we do make two other distinctions. First of all, we study classical capacities of both classical channels and quantum channels. In the case of classical channels, one does not encode classical information into quantum states like with quantum channels. Instead, one encodes classical messages into classical symbols, which are subsequently transmitted over the channel. The classical output of the channel is then decoded again into classical information.

Secondly, we distinguish classical capacities by their measures of reliability. The first type of reliability is one of an asymptotically vanishing probability of error. This means that we define the classical capacity of a channel as the greatest amount of information one can transmit per channel use with an arbitrarily small probability of error when one is allowed to use the channel arbitrarily many times. In other words, the communication does not have to be perfect, but it must be such that the probability of an error occurring approaches zero when the channel is used arbitrarily many times. We call the capacity defined this way the ordinary classical capacity.

The second sense of reliability which we consider is more straightforward to define: we call communication reliable if the probability of an error occurring is zero. That is, we do not even allow a very small probability of error, but we require zero probability of error. This type of capacity is called the zero-error classical capacity.

In short, we investigate four types of classical capacity in this thesis. These four types are characterised by two parameters: whether we use classical or quantum channels and whether we allow a very small error or require no error whatsoever. Each of these four capacities has its own properties, of which we focus on two main ones. The first of these is simply the expression for the capacity of a channel in terms of the properties of that channel. These expressions sometimes provide a precise characterisation of the capacity, as is the case for the ordinary capacity of a classical channel. In other cases, however, the best we can do is bound a particular capacity.

The second property of a capacity which we investigate is what happens when two channels are used together. One might expect the capacity of two channels to simply be the sum of the individual capacities, but it turns out that this is the case only for the ordinary capacity of a classical channel. The other capacities exhibit a phenomenon called superadditivity, which is when two channels together have a greater capacity than the sum of their individual capacities. In the case of quantum channels this phenomenon can be attributed to entanglement, but in the case of zero-error capacities this is not ncessarily true, as can be seen from the fact that superadditivity even occurs for zero-error communication over classical channels.

Since both the quantum and zero-error aspects of communication seem to trigger non-additivity phenomena, it is natural to wonder if these two aspects combined give rise to an even more extreme phenomenon. As was shown in [CCH11], this is the case. Indeed, the zero-error capacity of a quantum channel exhibits a property called superactivation, which is when two channels individually have no zero-error capacity, but their combined channel has strictly positive zero-error capacity.

Before we outline the structure of this thesis we remark that quantum information theory is a field of physics that relies heavily on a wide range of mathematical concepts. The linear algebra of vectors in and operators on Hilbert spaces always plays a paramount role in quantum mechanics, but in this thesis we encounter a lot of probability theory and graph theory as well. We also mention or sketch proofs which rely heavily on abstract algebra and even some algebraic geometry. Moreover, quantum information theoretic concepts can naturally be formulated in the theory of operator algebras. We develop most notions along the way, but we do assume familiarity with linear algebra, probability theory, basic graph theory and basic abstract algebra.

This thesis is divided into two parts: classical communication and quantum communication, which respectively investigate classical channels and quantum channels. Each part is then again subdivided into a chapter on the ordinary capacity and a chapter on the zero-error capacity. Part II also contains a chapter which introduces the quantum mechanical notions of quantum information theory.

Chapter 2 introduces the basic concepts of communication theory and presents a pivotal theorem in classical communication theory: the noisy channel coding theorem. This theorem gives the ordinary capacity of a classical channel. We proof this theorem in two ways: using the Markov inequality and the Law of Large numbers, and using the notion of a typical set. The first of these proofs is simplest, but the second generalises to quantum channels.

In chapter 3 we then treat the zero-error capacity of classical channels. We show that this capacity can be expressed in graph-theoretical language. This formulation in terms of graphs allows us to prove several results on the zero-error capacity. We also consider the zero-error capacity of a product of two channels and show that the capacity of the product channel can be greater than the sum of the capacities of the individual channels. This is the first example of superadditivity.

Chapter 4 then presents the basic notions of quantum mechanics and of quantum information theory in order to prepare us to investigate the classical capacities of quantum channels. It starts with the postulates of quantum mechanics, then treats the theory of composite quantum systems and entanglement, then introduces the quantum entropy and ends by introducing the trace distance between two quantum states, which is a measure of how close two quantum states are.

In chapter 5 we first generalise the concept of typicality from chapter 2 to the quantum setting, and then state and prove the packing lemma. Quantum typicality and the packing lemma are then subsequently used to prove the generalisation of the noisy channel coding theorem to the case of quantum mechanics, which is called the Holevo-Schumacher-Westmoreland theorem and is one of the main results of this thesis. It expresses the ordinary classical capacity of a quantum channel in terms of a quantity called the Holevo information. We then end the chapter by discussing the superadditivity of the Holevo information.

Chapter 6 generalises the ideas developed in chapter 3 to quantum channels, i.e. it investigates the zero-error classical capacity of quantum channels. Like with the zero-error capacity of classical channels, we express this capacity in graph-theoretical language, which allows us to prove that the zero-error classical capacity of a quantum channel can be achieved using only pure input quantum states. Lastly, we present the main idea of the proof which shows that two quantum channels can individually have no zero-error classical capacity, but together can have positive zero-error classical capacity. This phenomenon is called superactivation.

This thesis was written as part of the double bachelor's degree in Applied Mathematics and Applied Physics at Delft University of Technology.

I

Classical Communication

2

Classical Shannon Theory

We begin our study of capacities of communication channels by introducing several ubiquitous concepts of classical information theory which were invented by Claude Shannon. For this reason, we use the term Shannon Theory. We feel obliged to remark, however, that Shannon also laid the foundations for other topics in information theory. He is, however, most famous for the results presented in this chapter, and his landmark paper [Sha48] remains his most cited paper by far.

The main result of this chapter is the noisy channel coding theorem. We first present a simple proof of this theorem based on the Markov inequality and law of large numbers. In the section after, however, we will introduce different tools for proving the direct part of the noisy channel coding theorem, which can be more easily generalised to the case of quantum channels.

2.1. Noisy Channel Coding Theorem

In this section we introduce the fundamental notions communication theory: channels, codes, rates, errors and capacities. We do this in the setting of the ordinary capacity of classical channels. We then present and prove the main theorem on the ordinary capacity of classical channels: the noisy channel coding theorem.

2.1.1. Codes, Rates and Errors

We first define a coding scheme over a noisy channel. To this end, we follow [Wil19]. Our situation is as follows: Alice wishes to send some messages to Bob, and she has a noisy channel $\mathcal{N}$ (we will shortly characterise this channel) which she can use multiple times and independently for this.

There is a finite set $\mathcal{M}$ of possible messages that Alice could send to Bob. Bob also knows the contents of this message set, i.e. he knows what messages Alice might send him. Moreover, there is a finite alphabet set $\mathcal{X}$, which consists of the symbols that Alice could put into the channel. Bob knows what this alphabet looks like. The finite alphabet of possible outcomes of the channel that Bob might receive is $\mathcal{Y}$. The channel is represented by a conditional probability distribution $p_{Y|X}$, so for every letter $x \in \mathcal{X}$ that Alice might put into the channel, the probability for Bob to receive a letter $y \in \mathcal{Y}$ is $P_{Y|X}(y|x)$. This type of channel is called a discrete memoryless channel, because its input is a discrete alphabet, and its output depends only on the current input. In the following discussion we will always be considering a fixed channel $p_{Y|X}(y|x)$. Alice sends a message to Bob by encoding this message into symbols which can be put into the channel:

Definition 2.1. An encoder is a map E^n that translates a message into a codeword:

$$E^n : \mathcal{M} \to \mathcal{X}^n. \tag{2.1}$$

She uses the encoder to encode some message $m \in \mathcal{M}$ and then exploits n uses of the channel $\mathcal{N}$ in order to send the codeword $x^n(m) = E^n(m)$ belonging to the message. Bob will receive some sequence $y^n \in \mathcal{Y}^n$ according to the conditional probability:

$$p_{Y^n|X^n}(y^n|x^n) = p_{Y|X}(y_1|x_1)...p_{Y|X}(y_n|x_n), \tag{2.2}$$

where $y^n = (y_1,...,y_n)$ and $x^n = (x_1,...,x_n)$. Now Bob needs to determine what message Alice sent him, so he uses a decoder:

Definition 2.2. A decoder is a map D^n that translates output sequences into messages:

$$D^n : \mathcal{Y}^n \to \mathcal{M}. \tag{2.3}$$

We define the rate R of this above coding scheme (i.e. the message set, encoder and decoder) as:

Definition 2.3. The rate R of a coding scheme (M, E^n, D^n) is defined as:

$$R = \frac{\log(|\mathcal{M}|)}{n}. \tag{2.4}$$

Where the logarithm is base 2, so that the rate is measured in bits per channel use. We now define the probability of error for a coding scheme as:

Definition 2.4. The probability of error of a coding scheme is the maximum probability that a message is incorrectly decoded:

$$p_e = \max_{m \in \mathcal{M}} \Pr(D^n(\mathcal{N}^n(E^n(m))) \neq m). \tag{2.5}$$

Thus, the probability of error is simply the maximum probability (over all messages that Alice might want to send) that Bob thinks that Alice sent a different message than she actually did. Although the meaning of the above definition is clear, it is not very rigorous. Indeed, to make it rigorous we note that the probability of error is just equal to:

$$p_e = \max_{m \in \mathcal{M}} \left(1 - \sum_{y_n} p_{Y^n|X^n}(y^n|E^n(m)) I_{\{m\}}(D^n(y^n)) \right), \tag{2.6}$$

Where $I_{\{m\}} : \mathcal{M} \to \{0,1\}$ is the indicator function that indicates whether a message is equal to m (it maps m to 1 and any other message to 0). We say that a coding scheme (consisting of a set of messages, and encoder and a decoder) is an (n, R, ϵ) code if it uses the channel n times, its rate is R and its probability of error is smaller than or equal to ϵ.

2.1.2. Classical Entropy and Mutual Information

In order to understand the statement of the noisy channel coding theorem we have to understand how to quantify ideas such as information content and uncertainty of a random variable. We start with the following definition:

Definition 2.5. Suppose we have a random variable X, whose realisations belong to an alpahbet $\mathcal{X}$, and whose probability density function we denote by $p_X(x)$. Then the information content $i(x)$ of a particular realisation x is:

$$i(x) = -\log(p_X(x)). \tag{2.7}$$

Here the logarithm is base 2.

The fact that we take the logarithm base 2 shows that we measure the information content in units of bits, because a bit can take two values - 0 and 1. We can interpret the information content as a measure of the surprise one has upon learning the outcome of an experiment [Wil19]. The information content is useful for characterising the information of a particular realisation of a random variable, but it is not a useful characterisation of the information of the random variable in general. That is why we define the entropy. It is the expected information content of a random variable:

Definition 2.6. The entropy of a discrete random variable X with probability distribution $P_X(x)$ is:

$$H(X) = -\sum_x p_X(x)\log(p_X(x)) \tag{2.8}$$

It is important to remark that, in this definition, we take $0 \cdot \log(0) = 0$. This can be interpreted as not taking into account events which have no probability of occuring. We now list two important properties of the entropy, as they are presented in section 10.1 of [Wil19].

Property 2.1. The entropy is non-negative for any discrete random variable X, i.e. $H(X) \geq 0$.

Property 2.2. The entropy is concave in the probability density p_X, i.e. if we have two random variables X_1 and X_2 on the same alphabet $\mathcal{X}$, with distributions $p_{X_1}(x)$ and $p_{X_2}(x)$ respectively, then for any $\lambda \in [0,1]$ we have:

$$H(\lambda X_1 + (1-\lambda)X_2) \geq H(X_1) + H(X_2), \tag{2.9}$$

where with $\lambda X_1 + (1-\lambda)X_2$ we mean the random variable with distribution $\lambda p_{X_1}(X) + (1-\lambda)p_{X_2}(x)$.

We now turn to another type of classical entropy, which will play a central role in our study of typicality.

Definition 2.7. Let X and Y be discrete random variables with joint probability distribution given by $p_{X,Y}(x,y)$. The conditional entropy is the expected conditional information content with respect to both X and Y [Wil19]:

$$H(X|Y) = \mathbb{E}(i(X|Y)) = -\sum_{x,y} p_{X,Y}(x,y)\log(p_{X|Y}(x|y)). \tag{2.10}$$

Intuitively, it is immediately clear that the conditional entropy should be less than or equal to the entropy. This is because having information about another random variable Y should only decrease our uncertainty about the random variable X. This intuition turns out to indeed be correct, as proven in section 10.2 of [Wil19]:

Theorem 2.3. For any discrete random variables X and Y we have: $H(X) \geq H(X|Y)$.

We now turn to our last type of classical entropy: the joint entropy. The logic of its definition is similar to that of the conditional entropy:

Definition 2.8. Let X and Y be discrete random variables with joint probability distribution given by $p_{X,Y}(x,y)$. Then the joint entropy is defined as:

$$H(X,Y) = \mathbb{E}(i(X,Y)) = -\sum_{x,y} p_{X,Y}(x,y)\log(p_{X,Y}(x,y)). \tag{2.11}$$

It can be verified by simply wirting out the definitions that the entropy, conditional entropy and joint entropy satisfy the following relation:

Property 2.4. For discrete random variables X and Y we have:

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y). \tag{2.12}$$

We now introduce a measure of the correlation between two random variables. This measure is called the mutual information, and it is absolutely paramount in our study of the classical capacity of a classical channel, because it will turn out that this capacity is actually equal to the mutual information! Given two discrete random variables X and Y, the mutual information basically quantifies how much knowing about one random variable reduces the uncertainty of the other random variable [Wil19]:

Definition 2.9. Let X and Y be discrete random variables with joint probability distribution given by $p_{X,Y}(x,y)$. Then the mutual information of X and Y is the entropy minus the conditional entropy:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X). \tag{2.13}$$

2.1.3. Statement of the Theorem

We now have sufficient background knowledge to state the main result of this chapter: the noisy channel coding theorem. One can define the capacity of a channel as the supremum of all achievable rates, and in that case the noisy channel coding theorem states that the capacity is equal to the maximum mutual information. However, we take a slightly different approach and define the capacity a priori as the maximum mutual information:

$$C(\mathcal{N}) = \max_{p_X} I(X;Y) \tag{2.14}$$

Here $I(X,Y)$ is the mutual information between random variables X and Y. We have, however, not specified the probability distribution of X. The idea is to randomly generate an encoder for Alice. That is, we use a random variable X, which takes values in the input alphabet $\mathcal{X}$ of the channel and which has a distribution p_X, to generate some encoder. For every message $m \in \mathcal{M}$ we independently generate n letters from $\mathcal{X}$ and we form a sequence x^n from these. Thus, we have randomly and independently generated a codeword $x^n(m)$ for every message $m \in \mathcal{M}$. With these ideas in mind we have the following definition for the capacity of a classical channel:

Definition 2.10. Let a classical channel $\mathcal{N}$ be represented by the conditional probability distribution $p_{Y|X}(y|x)$ which gives the probability that the channel outputs the symbol $y \in \mathcal{Y}$ given that the input is $x \in \mathcal{X}$. Moreover, let $p_X(x)$ be the probability distribution of a random variable X which takes values in the input alphabet $\mathcal{X}$. We then have a random variable Y whose distribution is $p_Y(y) = \sum_{x\in\mathcal{X}} p_{Y|X}(y|x)p_X(x)$. We define the ordinary capacity of $\mathcal{N}$ as:

$$C(\mathcal{N}) = \max_{p_X} I(X;Y). \tag{2.15}$$

We note that we have taken the maximum over the probability distributions $p_X(x)$, and not the supremum. This is because the mutual information is a concave function of the probability distribution, such that a maximum actually exists. More precisely, the probability distributions $p_X(x)$ (represented as real functions on $\mathcal{X}$ which can be extended to function on $2^{\mathcal{X}}$ by additivity) form a convex set in the real vector space of functions $f:\mathcal{X}\to\mathbb{R}$. We then have that for a fixed channel the mutual information $I(X;Y)$ can be regarded as a function from this convex set to the real numbers, because given a probability distribution $p_X(x)$ (and keeping the channel fixed) it simply gives the real number:

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y)\log\left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right), \tag{2.16}$$

where we obtain the joint probability distribution $p_{X,Y}(x,y)$ from $p_X(x)$ and $p_{Y|X}(y|x)$ by $p_{X,Y}(x,y) = p_X(x)p_{Y|X}(y|x)$. We can see that the mutual information is indeed a concave function on the convex set of probability distributions on $\mathcal{X}$ by considering two such probability distributions $p_1(x)$ and

$p_2(x)$ corresponding to random variables X_1 and X_2 respectively. Then for $\lambda \in [0,1]$ we define the convex combination $p(x) = (\lambda p_1 + (1-\lambda)p_2)(x) = \lambda p_1(x) + (1-\lambda)p_2(x)$. Let us denote the random variable belonging to this new distribution by X. The mutual information is:

$$I(X;Y) = H(Y) - H(Y|X). \tag{2.17}$$

Remember from property 2.2 that the entropy is concave. Moreover, we can see that the conditional entropy is linear in the probability distribution:

$$\begin{aligned} H(Y|X) &= -\sum_{x,y} p_{X,Y}(x,y)\log(p_{Y|X}(y|x)) = -\sum_{x,y} p_X(x)p_{Y|X}(y|x)\log(p_{Y|X}(y|x)) \\ &= -\sum_{x,y}(\lambda p_1(x) + (1-\lambda)p_2(x))p_{Y|X}(y|x)\log(p_{Y|X}(y|x)) = H(Y|X_1) + H(Y|X_2). \end{aligned} \tag{2.18}$$

Thus, we see that the overall mutual information is concave in the probability distribution $p_X(x)$, since the entropy $H(Y)$ is also concave in the probability distribution, as can be seen by writing out $p_Y(y)$ in terms of $p_{Y|X}(y|X)$ and $p_X(x)$. Now that we have defined the capacity of a classical channel as the maximum mutual information we can state the noisy channel coding theorem.

Theorem 2.5. Let $\mathcal{N}$ be a noisy channel. Then for any rate $R < C$ and for any $\epsilon \in (0,1)$, there exists some $n \in \mathbb{N}$ such that there is an $(n, \tilde{R}, \epsilon)$ coding scheme for this channel, where $\tilde{R} \geq R$. Moreover, for any $R \geq C$ there is some $\epsilon \in (0,1)$ such that there does not exist any $n \in \mathbb{N}$ such that there is an $(n, \tilde{R}, \epsilon)$ coding scheme where $\tilde{R} > R$.

The interpretation of the noisy channel coding theorem is as follows: any rate below the capacity can be achieved (i.e. the probability of error can be made arbitrarily small by using the channel many times), but there are no coding schemes with a rate greater than the capacity such that the probability of error can be made arbitrarily small.

2.1.4. Proof of Direct Coding Part

The proof of the noisy channel coding theorem consists of two parts: the direct coding theorem and the converse theorem. The direct coding theorem shows that rates up to the maximum mutual information are achievable with arbitrarily small error as long as the block length is big enough, which we will now give a simple proof of, following [LF12].

Proof: we begin by again taking some finite set $\mathcal{M}$ to be the message set. We do not specify how big it is yet: we will see at the end of the proof how big we should choose the message set in order to achieve the desired rate. As explained above, we generate a random encoder. That is, for every message $m \in \mathcal{M}$ we generate a sequence of n symbols in the alphabet $\mathcal{X}$, according to the distribution $p_X(x)$. We denote such a sequence as $x^n(m)$, for every $m \in \mathcal{M}$. Again, Bob is aware of the particular encoder that Alice has chosen to use.

Alice chooses some message $m \in \mathcal{M}$, and then she puts the codeword $x^n(m)$ into the noisy channel. The noisy channel will then output a sequence y^n according to the distribution $p_{Y^n|X^n}(y^n|x^n)$, where $\mathcal{Y}$ is the ouput alphabet. Bob receives y^n, and he compares the probabilities $p_{Y^n|X^n}(y^n|x^n(m'))$ for every codeword $x^n(m')$. He chooses the codeword $x^n(m')$ with maximum probability, breaking ties arbitrarily. He will then say that Alice has sent him the message m' which corresponds to $x^n(m')$.

This encoding and decoding scheme could, however, go wrong. That is, Bob could think that Alice has transmitted some message, whereas she actually sent a different message. So say Alice sent message m by putting the codeword $x^n(m)$ into the channel, and Bob received sequence y^n. Then for every $m' \in \mathcal{M}$ we denote the event that the probability of m' is greater than or equal to that of m by $E_{m'}$. That is, $E_{m'}$ is the event that:

$$p_{Y^n|X^n}(y^n|x^n(m') \geq p_{Y^n|X^n}(y^n|x^n(m)). \tag{2.19}$$

We will now invoke the Markov inequality to bound the probability of this event occuring. The Markov inquality states that for a non-negative random variable A and for $t > 0$:

$$\Pr(A \geq t) \leq \frac{\mathbb{E}(A)}{t}. \tag{2.20}$$

The Markov inequality thus gives for the probability of the event $E_{m'}$ occuring (where we use that $p_{Y^n|X^n}(y^n|x^n(m')$ is a non-negative random variable with distribution $p_{X^n}(x^n)$:

$$\Pr(E_{m'}) = \Pr\left(p_{Y^n|X^n}(y^n|x^n(m') \geq p_{Y^n|X^n}(y^n|x^n(m)\right) \leq \frac{\mathbb{E}(p_{Y^n|X^n}(y^n|x^n(m'))}{p_{Y^n|X^n}(y^n|x^n(m))}. \tag{2.21}$$

We observe that the expectation just results in the marginal distribution $p_{Y^n}(y^n)$, so that the above becomes:

$$\Pr(E_{m'}) \leq \frac{p_{Y^n}(y^n)}{p_{Y^n|X^n}(y^n|x^n(m))}. \tag{2.22}$$

It is clear that the total probability of error is bounded by the probability of the union of all such events $E_{m'}$ for all $m' \in \mathcal{M}$. This is because the union can be interpreted as the event that there exists some message $m \neq m'$in $\mathcal{M}$ such that the probability of that message is larger than or equal to the probability of m. Now, if the probability of m' is actually equal to m, than there is a chance that Bob will actually make the right choice. Thus, the union of all events $E_{m'}$ for $m' \in \mathcal{M}$ is not necessarily equal to the total probability of error, but it is a bound. That is:

$$\Pr(E) \leq \Pr\left(\bigcup_{m' \neq m} E_{m'}\right). \tag{2.23}$$

By the union bound, we then get:

$$\Pr(E) \leq \sum_{m' \neq m} \Pr(E_{m'}) \leq \sum_{m' \neq m} \frac{p_{Y^n}(y^n)}{p_{Y^n|X^n}(y^n|x^n(m))} = (|\mathcal{M}|-1)\frac{p_{Y^n}(y^n)}{p_{Y^n|X^n}(y^n|x^n(m))}. \tag{2.24}$$

We will now analyse the behaviour of:

$$\frac{p_{Y^n}(y^n)}{p_{Y^n|X^n}(y^n|x^n(m))} \tag{2.25}$$

for large n. We have:

$$\frac{p_{Y^n}(y^n)}{p_{Y^n|X^n}(y^n|x^n(m))} = \prod_{i=1}^{n} \frac{p_Y(y_i)}{p_{Y|X}(y_i|x_i)}. \tag{2.26}$$

Using this, the law of large numbers tells us that for every $\epsilon > 0$ and $\delta > 0$ there exists a $n \in \mathbb{N}$ such that:

$$\Pr\left(\left|\frac{1}{n}\log\left(\frac{p_{Y^n}(y^n)}{p_{Y^n|X^n}(y^n|x^n(m))}\right) - \mathbb{E}\left(\log\left(\frac{p_Y(Y)}{p_{Y|X}(Y|X)}\right)\right)\right| > \epsilon\right) < \delta, \tag{2.27}$$

where X, Y are random variables such that $p_{X,Y}(x,y) = p_{Y|X}(y|x)p_X(x)$. But if we look at the definition of the mutual information, then we see that:

$$-I(X;Y) = \mathbb{E}\left(\log\left(\frac{p_Y(Y)}{p_{Y|X}(Y|X)}\right)\right). \tag{2.28}$$

Thus, what we really have just found is that for all $\epsilon, \delta > 0$ there exists some $n \in \mathbb{N}$ such that:

$$\Pr\left(\left|\frac{1}{n}\log\left(\frac{p_{Y^n}(y^n)}{p_{Y^n|X^n}(y^n|x^n(m))}\right) + I(X;Y)\right| > \epsilon\right) < \delta. \tag{2.29}$$

But then it follows that for all $\epsilon, \delta > 0$ there exists some $n \in \mathbb{N}$ such with probability at least $1-\delta$:

$$\frac{1}{n}\log\left(\frac{p_{Y^n}(y^n)}{p_{Y^n|X^n}(y^n|x^n(m))}\right) \leq -I(X;Y) + \epsilon. \tag{2.30}$$

We can rewrite this as:

$$\frac{p_{Y^n}(y^n)}{p_{Y^n|X^n}(y^n|x^n(m))} \leq 2^{n(-I(X;Y)+\epsilon)}. \tag{2.31}$$

Thus it follows that for every $\epsilon, \delta > 0$ there exists some $n \in \mathbb{N}$ such that for the probability of error we have:

$$\Pr(E) \leq (|\mathcal{M}| - 1)(\delta + 2^{n(-I(X;Y)+\epsilon)}) \leq |\mathcal{M}|\left(\delta + 2^{n(-I(X;Y)+\epsilon)}\right). \tag{2.32}$$

By recalling the definition of the rate R we immediately see that this can be written as:

$$\Pr(E) \leq \delta + 2^{n(-I(X;Y)+\epsilon+R)}. \tag{2.33}$$

But this proves the direct coding theorem, because we see that for a rate $R < I(X;Y)$ the probability of error can be made arbitrarily small since ϵ and δ can be made arbitrarily small as they were arbitrary. □

2.1.5. Converse Part

Proving the converse part of the noisy channel coding theorem - that is, showing that coding schemes with rates above the capacity can never have a vanishing probability of error - is quite simple: it requires only one lemma: Fano's inequality. We state and prove this result following [Wil19].

Lemma 2.6 (Fano's Inequality)**.** Suppose Alice sends a random variable X through a noisy channel to produce random variable Y. The decoder then gives an estimate $\hat{X}$ of X. Let $p_e = \Pr(\hat{X} \neq X)$ denote the probability of error. Then we have:

$$H(X|Y) \leq H(X|\hat{X}) + h_2(p_e) + p_e\log(|\mathcal{X}| - 1), \tag{2.34}$$

where $h_2(p_e)$ is the binary entropy function $h_2(p_e) = -p_e\log(p_e) - (1-p_e)\log(1-p_e)$.

Proof: let E be the indicator random variable that indicates whether an error occurs, i.e. $E = 0$ if $\hat{X} = X$ and $E = 1$ if $\hat{X} \neq X$. By property 2.4 we have:

$$H(E, X|\hat{X}) = H(X|\hat{X}) + H(E|X, \hat{X}), \tag{2.35}$$

but $H(E|X,\hat{X}) = 0$ since we know E if we know both X and $\hat{X}$. Thus, we get:

$$H(E, X|\hat{X}) = H(X|\hat{X}). \tag{2.36}$$

Now we use another result called the data-processing inequality, which can easily be proven by manipulating the definition of the mutual information. It simply states that post-processing cannot increase the mutual information [Wil19], i.e.

$$I(X;Y) \geq I(X;\hat{X}). \tag{2.37}$$

But since $I(X;Y) = H(X) - H(X|Y)$ and $I(X;\hat{X}) = H(X) - H(X|\hat{X})$ this implies:

$$H(X|\hat{X}) \geq H(X|Y). \tag{2.38}$$

The last result we need is the following:

$$\begin{aligned} H(E,X|\hat{X}) = H(E|\hat{X}) + H(X|E,\hat{X}) \leq H(E) + H(X|E,\hat{X}) \\ = h_2(p_e) + p_e H(X|\hat{X},E=1) + (1-p_e)H(X|\hat{X},E=0) \leq h_2(p_e) + p_e\log(|\mathcal{X}|-1). \end{aligned} \tag{2.39}$$

Here the first inequality follows from theorem 2.3, the second equality follows by simply expanding the two possible values E can take. The last inequality follows form two things: firstly, the fact that $H(X|\hat{X},E=0) = 0$ since we know X if we know $\hat{X}$ and we know that there is no error. Secondly, $H(X|\hat{X},E=1)$ is never greater than the entropy of the uniform distribution on the other possibilities than $\hat{X}$, because the uniform distribution has the most uncertainty of all distributions. The amount of other possibilities is of course equal to $|\mathcal{X}|-1$, and the entropy of that uniform distribution is $\log(|\mathcal{X}|-1)$. Putting equations (2.36), (2.38) and (2.39) together yields:

$$H(X|Y) \leq H(X|\hat{X}) = H(E,X|\hat{X}) \leq h_2(p_e) + p_e\log(|\mathcal{X}|-1), \tag{2.40}$$

which is Fano's inequality. □

With Fano's inequality in hand, we are ready to prove the converse part of the noisy channel coding theorem. It is almost as simple as filling in Fano's inequality above with the uniform choice of message random variable M instead of X and the output codeword random variable Y^n instead of Y.

Proof of Converse Part: we denote by M the uniform random variable corresponding to Alice's uniform message selection from the message set $\mathcal{M}$. We then have:

$$\begin{aligned} \log(\mathcal{M}) - H(M|Y^n) = H(M) - H(M|Y^n) = H(Y^n) - H(Y^n|M) \\ = \sum_{i=1}^{n} H(Y_i) - H(Y_i|M) \leq \sum_{i=1}^{n} H(Y_i) - H(Y_i|X_i) = \sum_{i=1}^{n} I(X_i;Y_i) \leq nC. \end{aligned} \tag{2.41}$$

If we now use M and Y^n in Fano's inequality we get:

$$H(M|Y^n) \leq h_2(p_e) + p_e\log(|\mathcal{M}|-1) \leq 1 + p_e\log(|\mathcal{M}|-1), \tag{2.42}$$

which implies, using the inequality above:

$$p_e \geq \frac{H(M|Y^n)-1}{\log(|\mathcal{M}|)} \geq \frac{\log(|\mathcal{M}|) - nC - 1}{\log(|\mathcal{M}|)} = 1 - \frac{nC}{nR} - \frac{1}{nR} = 1 - \frac{C}{R} - \frac{1}{nR}, \tag{2.43}$$

where we have used the definition of the rate $R = \frac{\log(|\mathcal{M}|)}{n}$. Thus, we see that in the limit where n goes to infinity, the probability of error always remains positive if $R > C$. Thus, no rate above the capacity is achievable. □

2.1.6. Product Channels

We end this section by considering what happens when we put two channels together. Thus, we need to understand what putting two channels means exactly. To the end, we define the following:

Definition 2.11. Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be two noisy channels with input alphabets $\mathcal{X}_1$ and $\mathcal{X}_2$ and output alphabets $\mathcal{Y}_1$ and $\mathcal{Y}_2$ respectively. Let $p_{Y_1|X_1}(y_1|x_1)$ and $p_{Y_2|X_2}(y_2|x_2)$ be the conditional distributions of channels 1 and 2 respectively. Then the product channel $\mathcal{N}_1 \otimes \mathcal{N}_2$ is the channel with input alphabet $\mathcal{X}_1 \times \mathcal{X}_2$, output alphabet $\mathcal{Y}_1 \times \mathcal{Y}_2$ and the following conditional distribution:

$$p_{Y_1\times Y_2|X_1\times X_2}(y_1, y_2|x_1, x_2) = p_{Y_1|X_1}(y_1|x_1) \cdot p_{Y_2|X_2}(y_2|x_2). \tag{2.44}$$

By writing it out, it can be seen that the mutual information of the product of two classical channels is equal to the sum of the individual mutual informations of the channels. Thus, it follows that the ordinary capacity of the product of two channels is equal to the sum of the individual ordinary capacities of the channels. This result is called the additivity of the ordinary capacity. It is quite a unique property, because as we will see in later chapters, all other capacities that we consider in this thesis are not additive. They can be superadditive, which means that the capacity of a product channel can be greater than the sum of the individual capacities. In section 6.2 we will even investigate an extreme form of superadditivity called superactivation.

2.2. Typicality

We now investigate the notion of typicality; both in the classical and quantum case. The basic idea of typicality is to define so-called typical sequences, which are sequences that are highly likely to be generated when one randomly generates sequences like we did in the previous section. When one wished to decode some sequence from a noisy channel, one can check whether this sequence is typical to decide what message it must have come from. We will only discuss strong typicality, since this is the type of typicality that will play a predominant role in our proof of the Holevo-Schumacher-Westmoreland theorem in chapter 5.

A strongly typical subset consists of all sequences which are strongly typical. Now, the idea of a strongly typical sequence is quite intuitive. If we have some random sequence x^n of n symbols in an alphabet $\mathcal{X}$ generated according to the probability distribution $p_X(x)$, then we count the number of times each symbol appears in the sequence. We then say that the sequence is strongly typical if the number of times a symbol occurs in the sequence is close to the number of times we would expect it to occur according to the probability distribution $p_X(x)$.

2.2.1. The Strongly Typical Set

Let us make the above intuition precise, following [Wil19]. Given a sequence $x^n \in \mathcal{X}^n$, we denote by $N(x|x^n)$ the number of times the symbol x occurs in the sequence x^n. If we divide this quantity by n we get an empirical probability distribution $\frac{1}{n}N(x|X^n)$. We then call a sequence δ-typical if this empirical distribution deviates at most by δ from the probability distribution $p_X(x)$ (which was used to randomly generate the sequence in the first place).

Definition 2.12. The δ-strongly typical set $T_\delta^{X^n}$ is the set of all sequences $x^n \in \mathcal{X}^n$ with an empirical distribution $\frac{1}{n}N(x|x^n)$ that has a maximum deviation δ from the true distribution $p_X(x)$. Furthermore, the empirical distribution $\frac{1}{n}N(x|x^n)$ must vanish for any $x \in \mathcal{X}$ for which $p_X(x) = 0$:

$$T_\delta^{X^n} = \left\{ x^n \in \mathcal{X}^n : \forall x \in \mathcal{X} : \left| \frac{1}{n}N(x|x^n) - p_X(x) \right| \leq \delta \text{ if } p_X(x) > 0, \text{ else } \frac{1}{n}N(x|x^n) = 0 \right\}. \tag{2.45}$$

The strongly typical set has three properties which are absolutely paramount in our upcoming proof of the Holevo-Schumacher-Westmoreland theorem. They are called Unit Probability, Exponentially Smaller Cardinality and Equipartition respectively. We will first state these three properties according to [Wil19] and then prove them.

Property 2.7 (Unit Probability)**.** The strongly typical set has unit probability in the limit where n becomes arbitrarily large. That is: for all $\delta > 0$ and $\epsilon \in (0,1)$ there exists some $n \in \mathbb{N}$ such that:

$$\Pr(X^n \in T_\delta^{X^n}) \geq 1 - \epsilon, \tag{2.46}$$

where $\Pr(X^n \in T_\delta^{X^n})$ simply denotes the probability that a sequence randomly generated according to $p_X(x)$ lies in fact in the strongly typical set.

Property 2.8 (Exponentially Smaller Cardinality)**.** The cardinality $|T_\delta^{X^n}|$ of the strongly δ-typical set is exponentially smaller than the total number of sequences $|\mathcal{X}^n|$:

$$|T_\delta^{X^n}| \leq 2^{n(H(X)+c\delta)}, \tag{2.47}$$

where c is some positive constant and X is the random variable with distribution $p_X(x)$. Moreover, we can bound the cardinality of the strongly δ-typical set from below: for all $\delta > 0$ and $\epsilon \in (0,1)$ there exists some $n \in \mathbb{N}$ such that:

$$|T_\delta^{X^n}| \geq (1-\epsilon)2^{n(H(X)-c\delta)}. \tag{2.48}$$

Property 2.9 (Equipartition)**.** The probability of some δ-typical sequence x^n occuring is approximately uniform. That is:

$$2^{-n(H(X)+c\delta)} \leq p_{X^n}(x^n) \leq 2^{-n(H(X)-c\delta)}. \tag{2.49}$$

Now that we have stated the unit probability, exponentially smaller cardinality and equipartition properties of the strongly typical set, we turn to proving them. To this end, we again follow [Wil19].

Proof of Unit Probability: this proof is based on the weak law of large numbers, so we start by recalling it. Let $Z_1, ..., Z_n$ denote independent, identically distributed random variables with expectation μ. Then the sample average is:

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i \tag{2.50}$$

The weak law of large numbers then states that for all $\epsilon \in (0,1)$ and $\delta > 0$ there exists some $N \in \mathbb{N}$ such that for all $n > N$ (where $n \in \mathbb{N}$) we have:

$$\Pr(|\bar{Z} - \mu| > \delta) < \epsilon. \tag{2.51}$$

For our proof of the unit probability property we now consider indicator random variables. We denote these by $I(X_i = a)$. These are the random variables which are one when X_i takes the value a, and 0 otherwise. It then is clear that the sample mean of a sequence of these indicator random variables is equal to the empirical distribution random variable:

$$\frac{1}{n}\sum_{i=1}^{n} I(X_i = a) = \frac{1}{n}N(a|X^n). \tag{2.52}$$

Moreover, we make the following trivial observation:

$$\mathbb{E}_X(I(X = a)) = p_X(a). \tag{2.53}$$

Thus, we invoke the weak law of large numbers to obtain the following result: for all $a \in \mathcal{X}$, $\epsilon \in (0,1)$ and $\delta > 0$ there exists some $N_a \in \mathbb{N}$ such that for all $n > N_a$:

$$\Pr\left(\left|\frac{1}{n}N(a|X^n) - p_X(a)\right| > \delta\right) < \frac{\epsilon}{|\mathcal{X}|}. \tag{2.54}$$

Now we choose $N = \max_{a\in\mathcal{X}} N_a$ and we invoke the union bound to conclude that for all $\epsilon \in (0,1)$ and $\delta > 0$ there exists some $N \in \mathbb{N}$ such that for all $n > N$ we have:

$$\begin{aligned}\Pr\left(\text{for some } a \in \mathcal{X} : \left|\frac{1}{n}N(a|X^n) - p_X(a)\right| > \delta\right) &\leq \sum_{a\in\mathcal{X}} \Pr\left(\left|\frac{1}{n}N(a|X^n) - p_X(a)\right| > \delta\right) \\ &< \sum_{a\in\mathcal{X}} \frac{\epsilon}{|\mathcal{X}|} = \epsilon.\end{aligned} \tag{2.55}$$

We can negate this and get that for all $\epsilon \in (0,1)$ and $\delta > 0$ there exists some $N \in \mathbb{N}$ such that for all $n > N$:

$$\Pr\left(\text{for all } a \in \mathcal{X} : \left|\frac{1}{n}N(a|X^n) - p_X(a)\right| \leq \delta\right) \geq 1 - \epsilon. \tag{2.56}$$

Now this is exactly our desired result, if we can show that when $p_X(a) = 0$ we have:

$$\Pr\left(\frac{1}{n}N(a|X^n) = 0\right). \tag{2.57}$$

But this is of course the case, because if $p_X(a) = 0$ then the probability that the random variable X^n takes on the value of a sequence which contains the symbol a is 0, since X^n is generated according to the distribution $p_X(x)$. □

Before we prove the exponentially smaller cardinality, we will first prove the equipartion property because we will need it in the proof of the exponentially smaller cardinality. Again, we follow [Wil19].

Proof of Equipartition: we make the following observation for a strongly typical sequence x^n:

$$p_{X^n}(x^n) = \prod_{i=1}^{n} p_X(x_i) = \prod_{x\in\mathcal{X}^+} p_X(x)^{N(x|x^n)}, \tag{2.58}$$

where we have written $x^n = (x_1, ..., x_n)$ and where $\mathcal{X}^+$ is the set of all $x \in \mathcal{X}$ for which $p_X(x) > 0$. Note that the above expression holds because x^n is strongly typical and therefore none of its symbols has probability zero of occurring (according to the distribution $p_X(x)$). Taking the logarithm and multiplying by $\frac{-1}{n}$ yields:

$$\frac{-1}{n}\log(p_{X^n}(x^n)) = -\sum_{x\in\mathcal{X}^+} \frac{1}{n}N(x|x^n)\log(p_X(x)). \tag{2.59}$$

Now because $x^n \in T_\delta^{X^n}$ we have by definition of the strongly typical set that for all $x \in \mathcal{X}^+$:

$$-\delta + p_X(x) \leq \frac{1}{n}N(x|x^n) \leq \delta + p_X(x). \tag{2.60}$$

We now multiply these inequalities by $-\log(p_X(X))$ (which is of course positive since $x \in \mathcal{X}^+$ and then we sum over all $x \in \mathcal{X}^+$ such that we can use equation (2.59) to write this as:

$$-\sum_{x\in\mathcal{X}^+} (-\delta + p_X(x))\log(p_{X^n}(x^n)) \leq \frac{-1}{n}\log(p_X(x)) \leq -\sum_{x\in\mathcal{X}^+} (\delta + p_X(x))\log(p_X(x)). \tag{2.61}$$

Thus, if we now define the positive constant $c = -\sum_{x\in\mathcal{X}^+}\log(p_X(x))$ and note that the entropy is $H(X) = \sum_{x\in\mathcal{X}^+} p_X(x)\log(p_X(x))$ (because we took $0\cdot\log(0) = 0$ in definition 2.6), we get:

$$-c\delta + H(X) \leq \frac{-1}{n}\log(p_{X^n}(x^n)) \leq c\delta + H(X). \tag{2.62}$$

Or, by multiplying by $-n$ and exponentiating:

$$2^{-n(c\delta+H(X))} \leq p_{X^n}(x^n) \leq 2^{-n(-c\delta+H(X))}, \tag{2.63}$$

which is exactly the result we were looking for. □

Now that we have proven the equipartition property, it is very straightforward to prove exponentially smaller cardinality.

Proof of Exponentially Smaller Cardinality: from the equipartition property we know that for any $x^n \in T_\delta^{X^n}$:

$$2^{-n(c\delta+H(X))} \leq p_{X^n}(x^n) \leq 2^{-n(-c\delta+H(X))}. \tag{2.64}$$

Summing over all typical sequences then gives:

$$\sum_{x^n \in T_\delta^{X^n}} 2^{-n(c\delta+H(X))} \leq \Pr(X^n \in T_\delta^{X^n}) \leq \sum_{x^n \in T_\delta^{X^n}} 2^{-n(-c\delta+H(X))}. \tag{2.65}$$

We recall that X^n denotes the random variable with distribution $p_{X^n}(x^n)$. Noting that $H(X)$ and c do not depend on any specific sequence x^n we can conclude:

$$\left|T_\delta^{X^n}\right| 2^{-n(c\delta+H(X))} \leq \Pr(X^n \in T_\delta^{X^n}) \leq \left|T_\delta^{X^n}\right| 2^{-n(-c\delta+H(X))}. \tag{2.66}$$

But from the unit probability property we know that for any $\delta > 0$ and $\epsilon > 0$ there is some $n \in \mathbb{N}$ such that:

$$1-\epsilon \leq \Pr\left(X^n \in T_\delta^{X^n}\right) \leq 1, \tag{2.67}$$

and thus we combine the above two results to conclude that for all $\delta > 0, \epsilon > 0$ there exists some $n \in \mathbb{N}$ such that:

$$\begin{aligned} \left|T_\delta^{X^n}\right| 2^{-n(c\delta+H(X))} &\leq 1, \\ 1-\epsilon \leq \left|T_\delta^{X^n}\right| 2^{-n(-c\delta+H(X))}&. \end{aligned} \tag{2.68}$$

This can be rewritten as:

$$(1-\epsilon)2^{n(-c\delta+H(X))} \leq \left|T_\delta^{X^n}\right| \leq 2^{n(c\delta+H(X))}, \tag{2.69}$$

which is exactly the statement of the exponentially smaller cardinality propery, with:

$$c = -\sum_{x \in \mathcal{X}^+} \log(p_X(x)), \tag{2.70}$$

concluding the proof. □

2.2.2. Strong Conditional Typicality

Now that we understand the strongly typical set, we consider the so-called conditionally typical set. This concept is basically the same as the typical set, but we now use conditional probability distributions instead of marginal distributions. The conditionally typical set will play a major role in our upcoming exposition of the Holevo-Schumacher-Westmoreland theorem in chapter 5, as well as our second proof of the noisy channel coding theorem in the next section. The general ideas of this section are mostly the same as in the previous section, but some of the proofs are a little more involved, and we will therefore again spell them out in detail. We will still be roughly following the exposition from [Wil19].

In the previous section we used $N(x|x^n)$ to denote the number of times the symbol $x \in \mathcal{X}$ occurs in the sequence $x^n \in \mathcal{X}^n$. We now introduce a similar notation. Suppose we have sequences x^n and y^n which consists of symbols from the alphabets $\mathcal{X}$ and $\mathcal{Y}$ respectively. Then we denote by $N(x,y|x^n,y^n)$ the number of times the symbol x appears in x^n, while at the same time the symbol y appears at the same position in the sequence y^n. That is, if we write x^n and y^n together as $(x_1,y_1),...,(x_n,y_n)$ then $N(x,y|x^n,y^n)$ denotes the number of times the pair (x,y) appears in this combined sequence. We call $\frac{1}{n}N(x,y|x^{\cdot}y^n)$ the joint empirical distribution.

We will now be considering such combined sequences generated by a joint probability distribution $p_{X,Y}(x,y)$ on $\mathcal{X}\times\mathcal{Y}$. This joint distribution can be factored as $p_{X,Y}(x,y) = p_{Y|X}(y|x)p_X(x)$. A conditionally typical sequence y^n conditioned on a strongly typical sequence x^n will then be a sequence in $\mathcal{Y}^n$ which is such that its joint empirical distribution with x^n is close to the product $p_{Y|X}(y|x)\frac{1}{n}N(x|x^n)$, where $\frac{1}{n}N(x|x^n)$ is of course the marginal empirical distribution that we encountered in the previous section. Let us make this intuition precise:

Definition 2.13. Let $p_{X,Y}$ be a joint probability distribution on the product $\mathcal{X}\times\mathcal{Y}$ of alphabets $\mathcal{X}$ and $\mathcal{Y}$ which can be factored as $p_{X,Y}(x,y) = p_{Y|X}(y|x)p_X(x)$, and let $x^n \in T_{\delta'}^{X^n}$. Then the δ-strong conditionally typical set $T_\delta^{Y^n|x^n}$ corresponding to x^n is defined as:

$$T_\delta^{Y^n|x^n} = \left\{ y^n \in \mathcal{Y}^n \mid \forall (x,y)\in\mathcal{X}\times\mathcal{Y}: \begin{cases} N(x,y|x^n,y^n)=0 & \text{if } p(y|x)=0 \\ \left|N(x,y|x^n,y^n) - p(y|x)N(x|x^n)\right| \le n\delta & \text{else} \end{cases} \right\}, \tag{2.71}$$

where we used $p(y|x)$ to denote $p_{Y|X}(y|x)$.

The conditionally typical set has the same three properties as the typical set, which we now state.

Property 2.10 (Unit Probability)**.** Given a strongly typical sequence $x^n \in T_{\delta'}^{X^n}$ the strong conditionally δ-typical set corresponding to x^n has unit probability in the limit where n becomes arbitrarily large. That is: for all $\delta > 0$ and $\epsilon > 0$ there exists some $n \in \mathbb{N}$ such that:

$$\Pr(Y^n \in T_\delta^{Y^n|x^n}) \ge 1-\epsilon, \tag{2.72}$$

where $\Pr(Y^n \in T_\delta^{Y^n|x^n})$ simply denotes the probability that a sequence randomly generated according to $p_Y(y) = \sum_{x\in\mathcal{X}} p(x,y)$ lies in fact in the conditionally typical set.

Property 2.11 (Exponentially Smaller Cardinality)**.** The cardinality $|T_\delta^{Y^n|x^n}|$ of the conditionally δ-typical set corresponding to a typical sequence x^n is exponentially smaller than the total number of sequences $|\mathcal{Y}^n|$:

$$|T_\delta^{Y^n|x^n}| \le 2^{n(H(Y|X)+c(\delta+\delta'))}, \tag{2.73}$$

where c is a positive constant. Moreover, we can bound the cardinality of the conditionally δ-typical set from below: for all $\delta > 0$ and $\epsilon > 0$ there exists some $n \in \mathbb{N}$ such that:

$$|T_\delta^{Y^n|x^n}| \ge (1-\epsilon)2^{n(H(Y|X)-c(\delta+\delta'))}. \tag{2.74}$$

Property 2.12 (Equipartition)**.** Given a typical sequence x^n, the probability of some δ-conditionally typical sequence y^n occuring is approximately uniform. That is:

$$2^{-n(H(Y|X)+c(\delta+\delta'))} \le p_{Y^n|X^n}(y^n|x^n) \le 2^{-n(H(Y|X)-c(\delta+\delta'))}. \tag{2.75}$$

We will now proof the unit probability and equipartition properties. The exponentially smaller cardinality follows from the equipartition property in exactly the same way as in the previous section.

Proof of Unit Probability: we start by labelling the elements of the alphabet $\mathcal{X}$ as $x_1,...,x_{|\mathcal{X}|}$. We can then write the strongly typical sequence $x^n \in \mathcal{X}^n$ as:

$$L(x^n) = \underbrace{x_1 \cdots x_1}_{N(x_1|x^n)} \underbrace{x_2 \cdots x_2}_{N(x_2|x^n)} \cdots \underbrace{x_{|\mathcal{X}|} \cdots x_{|\mathcal{X}|}}_{N(x_{\mathcal{X}}|x^n)}, \tag{2.76}$$

which is called a lexicographic ordering [Wil19], represented by the lexicographic ordering map $L : \mathcal{X}^n \to \mathcal{X}^n$. Now, this lexicographic ordering map also induces a map on the conditionally typical sequences $T_\delta^{Y^n|x^n}$, simply by splitting a sequence y^n into subsequences belonging to $x_1, x_2, ..., x_{|\mathcal{X}|}$. Let us make this precise. Given some $y^n \in T_\delta^{Y^n|x^n}$ we write:

$$y^n = y^1 \cdots y^n, x^n = x^1 \cdots x^n. \tag{2.77}$$

As we saw before, when defining conditional typicality we think of these sequences together as a sequence of pairs $(x^1, y^1)\cdots(x^n, y^n)$. Thus, we can order the letters of y^n according to the lexicographic order of x^n. So we write y^n as the concatenation of subsequences of length $N(x_1|x^n),...,N(x_{|\mathcal{X}|}|x^n)$. Given some $x \in \mathcal{X}$ we can then consider the typical subsequences $y^{N(x|x^n)}$ of length $N(x|x^n)$ whose emperical distribution $\frac{N(y|y^{N(x|x^n)})}{N(x|x^n)}$ is δ-close to the true distribution $p_{Y|X=x}(y)$. To denote these typical subsequences we introduce the following notation:

$$T_\delta^{(Y|X=x)^{N(x|x^n)}} = \left\{ y^{N(x|x^n)} \in \mathcal{Y}^{N(x|x^n)} \mid \forall\, y \in \mathcal{Y} : \left| \frac{N(y|y^{N(x|x^n)})}{N(x|x^n)} - p_{Y|X=x}(y) \right| \leq \delta \right\}. \tag{2.78}$$

We will denote by $L_i(y^n)$ the subsequence of y^n belonging to the i^{th} letter x_i for $i = 1,...,|\mathcal{X}|$. Then we can formulate the following equivalent definition of a conditionally typical sequence $y^n \in T_\delta^{Y^n|x^n}$:

$$y^n \in T_\delta^{Y^n|x^n} \iff \forall\, 1 \leq i \leq |\mathcal{X}| : L_i(y^n) \in T_\delta^{(Y|X=x_i)^{N(x_i|x^n)}}. \tag{2.79}$$

But now we are done, since we can simply use the unit probability property for each typical set $T_\delta^{(Y|X=x)^{N(x|x^n)}}$ to conclude that for every $\epsilon > 0$ and $\delta > 0$ there exists some $n \in \mathbb{N}$ such that:

$$\Pr\left(y^n \in T_\delta^{Y^n|x^n}\right) = \prod_{i=1}^{|\mathcal{X}|} \Pr\left(L_i(y^n) \in T_\delta^{(Y|X=x_i)^{N(x_i|x^n)}}\right) \geq (1-\epsilon)^{|\mathcal{X}|} \geq 1 - |\mathcal{X}|\epsilon \geq 1 - \epsilon, \tag{2.80}$$

concluding the proof. □

Proof of Equipartition: using the same argument as in the previous section we can write for some $y^n \in T_\delta^{Y^n|x^n}$, with $x^n \in T_{\delta'}^{X^n}$ strongly typical:

$$p_{Y^n|X^n}(y^n|x^n) = \prod_{(x,y)\in(\mathcal{X},\mathcal{Y})^+} p_{Y|X}(y|x)^{N(x,y|x^n,y^n)}, \tag{2.81}$$

where $(\mathcal{X},\mathcal{Y})^+$ is the set of all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ with $p(y|x) > 0$. Again taking the logarithm and multiplying by $\frac{-1}{n}$ then gives:

$$\frac{-1}{n}\log\left(p_{Y^n|X^n}(y^n|x^n)\right) = -\sum_{(x,y)\in(\mathcal{X},\mathcal{Y})^+} \frac{1}{n} N(x, y|x^n, y^n)\log\left(p_{Y|X}(y|x)\right). \tag{2.82}$$

Because x^n is strongly δ'typical and y^n is strong conditionally δ-typical we have:

$$\forall\, x \in \mathcal{X}^+ : -\delta' + p_X(x) \leq \frac{1}{n} N(x|x^n) \leq \delta' + p_X(x), \tag{2.83}$$

$$\forall\, (x,y) \in (\mathcal{X},\mathcal{Y})^+ : -\delta + p_{Y|X}(y|x)N(x|x^n) \leq \frac{1}{n} N(x,y|x^n,y^n) \leq \delta + p_{Y|X}(y|x)N(x|x^n). \tag{2.84}$$

We now perform the same steps as in the previous section. That is: we multiply equation (2.84) by $-\log(p_{Y|X}(y|x))$ and then sum over all $(x,y) \in (\mathcal{X}\times\mathcal{Y})^+$. If we then define:

$$c = -\sum_{(x,y)\in(\mathcal{X},\mathcal{Y})^+} \log(p_{Y|X}(y|x), \tag{2.85}$$

which is positive, and combine with equation 2.83 we get:

$$-c(\delta+\delta') + H(Y|X) \leq \frac{-1}{n}\log\left(p_{Y^n|X^n}(y^n|x^n)\right) \leq c(\delta+\delta') + H(Y|X). \tag{2.86}$$

Thus, multiplying by $-n$ and exponentiating gives:

$$2^{-n(H(Y|X)+c(\delta+\delta'))} \leq p_{Y^n|X^n}(y^n|x^n) \leq 2^{-n(H(Y|X)-c(\delta+\delta'))}, \tag{2.87}$$

which is the desired result. □

2.3. Proof of Direct Coding Theorem by Typicality

Now that we understand the definitions and properties of strongly typical and conditional strongly typical sets, we put our knowledge to practice by providing an alternative proof of the direct part of the noisy channel coding theorem. This proof is not simpler than the previous one, but it does generalise to the case of classical communication through quantum channels. We follow [Wil19].

We have the same setup as before, i.e. Alice and Bob randomly generate codewords $x^n(m)$ for every message $m \in \mathcal{M}$ according to the distribution $p_X(x)$, and Alice then randomly selects a message to send to Bob with uniform probability. However, the decoder is different this time. Instead of choosing the message with the highest likelihood, Bob tests whether the sequence he receives is in the strongly δ-typical set, for some fixed δ. If not, he reports an error. After that, Bob checks whether his received sequence is in the conditionally δ-typical set for some codeword that Alice could send. If there is a unique such message, Bob chooses that message. Otherwise, he reports an error. This reporting of an error can be modelled by always choosing a fixed message from the message set. Technically, this means that when Bob reports an error he could still have chosen the correct message, but we will be able to bound the error sufficiently without even taking this into account.

Thus, given that Alice sends the message m using codeword $x^n(m)$, there are three kinds of errors that could occur: $E_1(m)$ denotes the event that when Alice puts in $x^n(m)$, Bob gets a sequence y^n which is not in the typical set $T_\delta^{Y^n}$, $E_2(m)$ denotes the event that $y^n \in T_\delta^{Y^n}$, but $y^n \notin T_\delta^{Y^n|x^n(m)}$, and $E_3(m)$ denotes the event that $y^n \in T_\delta^{Y^n}$, but there is another message $m' \in \mathcal{M}$ such that $m' \in T_\delta^{Y^n|x^n(m')}$.

What we do now is to randomly generate a code (which consists of $|\mathcal{M}|$ codewords) according to the distribution $p_X(x)$. We will then consider the expected average probability of error of such a code, using the decoder described above. After we prove that the average probability of error asymptotically vanishes, we show that this implies that we can find a code such that the maximum probability of error also asymptotically vanishes. Let us denote the random variable corresponding to the random code by $\mathcal{C} = (X^n(1),\ldots,X^n(|\mathcal{M}|))$, where every $X^n(i)$ is a random variable with distribution $p_{X^n}(x^n)$. Thus, the expectation of the average probability of error is:

$$\mathbb{E}_{\mathcal{C}}\left(\frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\Pr\left(E_1(m)\cup E_2(m)\cup E_3(m)\right)\right). \tag{2.88}$$

Here, given a code C and a fixed message m (making the codeword $x^n(m)$ fixed), the distribution of the random variable on the sigma algebra $\Pr:\sigma\left(\{E_1(m),E_2(m),E_3(m)\}\right)\to[0,1]$ is simply induced by the channel distribution $p_{Y^n|X^n}(y^n|x^n)$. That is:

$$\begin{aligned}\Pr(E_1(m)) &= \sum_{y^n\in\mathcal{Y}^n} p_{Y^n|X^n}(y^n|x^n(m))\left(1-I_{T_\delta^{Y^n}}(y^n)\right),\\ \Pr(E_2(m)) &= \sum_{y^n\in\mathcal{Y}^n} p_{Y^n|X^n}(y^n|x^n(m))I_{T_\delta^{Y^n}}(y^n)\left(1-I_\delta^{Y^n|x^n(m)}(y^n)\right),\\ \Pr(E_3(m)) &= \sum_{y^n\in\mathcal{Y}^n} p_{Y^n|X^n}(y^n|x^n(m))\sum_{m'\neq m} I_{T_\delta^{Y^n}}(y^n)I_\delta^{Y^n|x^n(m')}(y^n),\end{aligned} \tag{2.89}$$

where the $I_{T_\delta^{Y^n}}:\mathcal{Y}^n\to\{0,1\}$ and $I_\delta^{Y^n|x^n(m)}:\mathcal{Y}^n\to\{0,1\}$ are the indicator functions that indicate whether a sequence is strongly typical or conditional strongly typical respectively. By linearity, we can exchange the sum and expecation in equation (2.88), which gives:

$$\frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}}\mathbb{E}_{\mathcal{C}}\left(\Pr(E_1(m)\cup E_2(m)\cup E_3(m))\right). \tag{2.90}$$

Now, since the codewords for each message $m\in\mathcal{M}$ are selected in exactly the same way - randomly according to the distribution $p_{X^n}(x^n)$, we can just choose one fixed message, let's say message 1, and write the expected average error as:

$$\mathbb{E}_{\mathcal{C}}\left(\Pr(E_1(1)\cup E_2(1)\cup E_3(1))\right). \tag{2.91}$$

The union bound then gives:

$$\mathbb{E}_{\mathcal{C}}\left(\Pr(E_1(1)\cup E_2(1)\cup E_3(1))\right)\leq\mathbb{E}_{\mathcal{C}}\left(\Pr(E_1(1))\right)+\mathbb{E}_{\mathcal{C}}\left(\Pr(E_2(1))\right)+\mathbb{E}_{\mathcal{C}}\left(\Pr(E_3(1))\right). \tag{2.92}$$

Our strategy now becomes very simple: we bound every one of these three expectations of errors. For the first type of error we have:

$$\begin{aligned}\mathbb{E}_{\mathcal{C}}\left(\Pr(E_1(1))\right) &= \mathbb{E}_{X^n(1)}\left(\Pr(E_1(1))\right) &(2.93)\\ &= \mathbb{E}_{X^n(1)}\left(\sum_{y^n\in\mathcal{Y}^n} p_{Y^n|X^n}(y^n|X^n(1))\left(1-I_{T_\delta^{Y^n}}(y^n)\right)\right) &(2.94)\\ &= 1-\mathbb{E}_{X^n(1)}\left(\sum_{y^n\in\mathcal{Y}^n} p_{Y^n|X^n}(y^n|X^n(1))I_{T_\delta^{Y^n}}(y^n)\right) &(2.95)\\ &= 1-\sum_{x^n\in\mathcal{X}^n} p_{X^n}(x^n)\sum_{y^n\in\mathcal{Y}^n} p_{Y^n|X^n}(y^n|x^n)I_{T_\delta^{Y^n}}(y^n) &(2.96)\\ &= 1-\sum_{y^n\in\mathcal{Y}^n} p_{Y^n}(y^n)I_{T_\delta^{Y^n}}(y^n)=\Pr\left(Y^n\notin T_\delta^{Y^n}\right). &(2.97)\end{aligned}$$

But the unit probability of the strongly typical set tells us that this last probability of a sequence not being strongly typical can become arbitrarily small as n becomes arbitrarily large. Thus, we know that for any $\epsilon>0$ there exists some $n\in\mathbb{N}$ such that:

$$\mathbb{E}_{\mathcal{C}}\left(\Pr(E_1(1))\right)\leq\epsilon. \tag{2.98}$$

Now that we have bounded the expected probability of the first type of error, we move on to the second type. We have:

$$\begin{aligned}\mathbb{E}_{\mathcal{C}}(\Pr(E_2(1)) &= \mathbb{E}_{X^n(1)}\left(\sum_{y^n\in\mathcal{Y}^n} p_{Y^n|X^n}(y^n|X^n(1))I_{T_\delta^{Y^n}}(y^n)\left(1-I_\delta^{Y^n|X^n(1)}(y^n)\right)\right)\\ \le \mathbb{E}_{X^n(1)}&\left(\sum_{y^n\in\mathcal{Y}^n} p_{Y^n|X^n}(y^n|X^n(1))\left(1-I_\delta^{Y^n|X^n(1)}(y^n)\right)\right) = \mathbb{E}_{X^n(1)}\left(\Pr(Y^n\notin T_\delta^{Y^n|X^n(1)})\right).\end{aligned} \tag{2.99}$$

This time, the unit probability of the conditional strongly typical set tells us that for all $\epsilon>0$ there exists some $n\in\mathbb{N}$ such that:

$$\mathbb{E}_{X^n(1)}\left(\Pr(Y^n\notin T_\delta^{Y^n|X^n(1)})\right)\le\mathbb{E}_{X^n(1)}(\epsilon)=\epsilon. \tag{2.100}$$

Now the only type of error that is left for us to bound is the third type:

$$\begin{aligned}&\mathbb{E}_{\mathcal{C}}(\Pr(E_3(1))) = \mathbb{E}_{\mathcal{C}}\left(\sum_{y^n\in\mathcal{Y}^n} p_{Y^n|X^n}(y^n|X^n(1))\sum_{m'\neq 1} I_{T_\delta^{Y^n}}(y^n)I_\delta^{Y^n|X^n(m')}(y^n)\right)\\ &= \sum_{y^n\in\mathcal{Y}^n}\sum_{m'\neq 1}\mathbb{E}_{\mathcal{C}}\left(p_{Y^n|X^n}(y^n|X^n(1))I_{T_\delta^{Y^n}}(y^n)I_\delta^{Y^n|X^n(m')}(y^n)\right)\\ &= \sum_{y^n\in\mathcal{Y}^n}\sum_{m'\neq 1}\mathbb{E}_{X^n(1),X^n(m')}\left(p_{Y^n|X^n}(y^n|X^n(1))I_{T_\delta^{Y^n}}(y^n)I_\delta^{Y^n|X^n(m')}(y^n)\right)\\ &= \sum_{y^n\in\mathcal{Y}^n}\sum_{m'\neq 1}\sum_{x^n(1)\in\mathcal{X}^n} p_{X^n}(x^n(1))\mathbb{E}_{X^n(m')}\left(p_{Y^n|X^n}(y^n|X^n(1))I_{T_\delta^{Y^n}}(y^n)I_\delta^{Y^n|X^n(m')}(y^n)\right)\\ &= \sum_{y^n\in\mathcal{Y}^n}\sum_{m'\neq 1}\mathbb{E}_{X^n(m')}\left(p_{Y^n}(y^n)I_{T_\delta^{Y^n}}(y^n)I_\delta^{Y^n|X^n(m')}(y^n)\right).\end{aligned} \tag{2.101}$$

We now invoke the equipartition property of the strongly typical set, which tells us that for all $y^n\in\mathcal{Y}^n$:

$$p_{Y^n}(y^n)I_{T_\delta^{Y^n}}(y^n)\le 2^{-n(H(Y)-\delta)}. \tag{2.102}$$

Thus, continuing our expression for the third type of error:

$$\sum_{y^n\in\mathcal{Y}^n}\sum_{m'\neq 1}\mathbb{E}_{X^n(m')}\left(p_{Y^n}(y^n)I_{T_\delta^{Y^n}}(y^n)I_\delta^{Y^n|X^n(m')}(y^n)\right) \tag{2.103}$$

$$\le\sum_{y^n\in\mathcal{Y}^n}\sum_{m'\neq 1}\mathbb{E}_{X^n(m')}\left(2^{-n(H(Y)-\delta)}I_\delta^{Y^n|X^n(m')}(y^n)\right) \tag{2.104}$$

$$=2^{-n(H(Y)-\delta)}\sum_{m'\neq 1}\mathbb{E}_{X^n(m')}\left(\sum_{y^n\in\mathcal{Y}^n}I_\delta^{Y^n|X^n(m')}(y^n)\right) \tag{2.105}$$

$$=2^{-n(H(Y)-\delta)}\sum_{m'\neq 1}\mathbb{E}_{X^n(m')}\left(|T_\delta^{Y^n|X^n(m')}|\right) \tag{2.106}$$

$$\le 2^{-n(H(Y)-\delta)}2^{n(H(Y|X+\delta))}\sum_{m'\neq 1}\mathbb{E}_{X^n(m')}(1) \tag{2.107}$$

$$\le|\mathcal{M}|2^{-n(H(Y)-\delta)}2^{n(H(Y|X+\delta))}=|\mathcal{M}|2^{-n(I(X;Y)-2\delta)}. \tag{2.108}$$

Choosing the message set size $|\mathcal{M}|\le 2^{n(I(X;Y)-3\delta)}$ gives:

$$\mathbb{E}_{\mathcal{C}}(\Pr(E_3(1)))\le 2^{-n\delta}. \tag{2.109}$$

Thus, putting the three bounds we have found together, we find that for each $\epsilon > 0, \delta > 0$ there exists an $n \in \mathbb{N}$ such that if we choose $|\mathcal{M}| \leq 2^{n(I(X;Y)-3\delta)}$, we have:

$$\mathbb{E}_{\mathcal{C}}\left(\Pr(E_1(1) \cup E_2(1) \cup E_3(1))\right) \leq 2\epsilon + 2^{-n\delta}. \tag{2.110}$$

Since this is the expectation over all codes of the average probability of error, there definitely exists one code whose average probability of error is bounded by $\epsilon' = 2\epsilon + 2^{-n\delta}$. Now all that is left for us to do is to show that this actually implies the existence of a code whose maximum probability of error can be made arbitrarily small. This step is called the expurgation step [Wil19]. We know that there exists a code such that:

$$\frac{1}{|\mathcal{M}|}\sum_m p_e(m) \leq \epsilon', \tag{2.111}$$

where $p_e(m)$ is the probability of error for message m. Now we consider $p_e : \sigma(M) \to [0,1]$ as a random variable, and use the Markov inequality:

$$\Pr(p_e \geq 2\epsilon') = \frac{1}{|\mathcal{M}|}|\{m \in \mathcal{M} : p_e(m) \geq 2\epsilon'\}| \leq \frac{\mathbb{E}(p_e)}{2\epsilon'} \leq \frac{\epsilon'}{2\epsilon'} = \frac{1}{2}. \tag{2.112}$$

Thus, we see that:

$$|\{m \in \mathcal{M} : p_e(m) \geq 2\epsilon'\}| \leq \frac{|\mathcal{M}|}{2}. \tag{2.113}$$

This means that at least half of the messages has a probability of error less than $2\epsilon'$. Thus, if we simply only use half of the messages, all with probability of error less than $2\epsilon'$, then we have achieved the desired result, because for this code the maximum probability of error is of course bounded by $2\epsilon'$. At the same time, the new rate R' has become:

$$R' = \frac{\log(\frac{|\mathcal{M}|}{2})}{n} = R - \frac{1}{n}, \tag{2.114}$$

which in the limit of large n approaches R. □

3

Classical Zero-Error Communication

In the previous chapter we considered communication through noisy channels, such that the maximum probability of error becomes arbitrarily small as we use the channel arbitrarily many times. We showed that the capacity of a noisy channel for this type of communication is equal to the mutual information of the input and output random variables of the channel. We call this capacity the ordinary capacity. However, there might be situations in which we want to communicate over a noisy channel without any error at all. That is, not even the slightest probability of error is permitted. This leads to the notion of a zero-error capacity - the highest rate at which one can transmit information through a noisy channel without any probability of error. In the present chapter we define and investigate this zero-error capacity. We see that it has a natural formulation in graph theory, which we later generalise to the case of quantum channels in chapter 6.

3.1. Zero-Error Codes and Capacity

Like in the previous chapter, we will be working with a noisy channel modelled by a conditional probability distribution $p_{Y|X}(y|x)$ on the alphabets $\mathcal{X}$ and $\mathcal{Y}$, such that we can use this channel independently as many times as we wish. We now define our main object of study: the error-free code [GdAM16].

Definition 3.1. An (M, n) error-free code for a noisy channel $\mathcal{N}$ modelled by $p_{Y|X}(y|x)$ consists of a finite message set $\mathcal{M} = \{1, ..., M\}$, an encoder $E^n : \mathcal{M} \to \mathcal{X}^n$ and a decoder $D^n : \mathcal{Y}^n \to \mathcal{M}$, such that for all $m \in \mathcal{M}$:

$$\Pr\left(E^n(\mathcal{N}(D^n(m))) \neq m\right) = 0. \tag{3.1}$$

Like in definition 2.4 we note that:

$$\Pr\left(E^n(\mathcal{N}(D^n(m))) \neq m\right) = 1 - \sum_{y^n \in \mathcal{Y}^n} p_{Y^n|X^n}(y^n|E^n(m)) I_m(D^n(y^n)), \tag{3.2}$$

where I_m is the indicator function on the set $\{m\}$.

For any error-free code $(\mathcal{M}, E^n, D^n)$ for a channel $p_{Y|X}(y|x)$ there cannot be two codewords $x^n(m) = E^n(m)$ and $x^n(m') = E^n(m')$ such that there exists some $y^n \in \mathcal{Y}^n$ for which both probability distributions $p_{Y^n|X^n}(y^n|x^n(m))$ and $p_{Y^n|X^n}(y^n|x^n(m'))$ are nonzero. Indeed, if this would be the case, then if the ouput would be that particular sequence y^n, there would be no way of being sure what message it came from. This gives rise to the following paramount notion [GdAM16]:

Definition 3.2. Given a channel modelled by $p_{Y|X}(y|x)$, two input symbols $x, x' \in \mathcal{X}$ are called adjacent (or indistinguishable) if there exists some $y \in \mathcal{Y}$ such that $p_{Y|X}(y|x) > 0$ and $p_{Y|X}(y|x') > 0$. Otherwise, the symbols x and x' are called non-adjacent (or distinguishable).

We can extend this definition to sequences x^n and $\tilde{x}^n$. These sequences are called indistinguishable if for all $1 \le i \le n$ their i^{th} symbols are adjacent. Otherwise - i.e. if for some i the i^{th} symbols are non-adjacent - the sequences are called distinguishable. We are now in a position to define the zero-error capacity as Shannon originally did [Sha56].

Definition 3.3. Given a noisy channel modelled by $p_{Y|X}(y|x)$, we define $N(n)$ to be the largest number of words $x^n \in \mathcal{X}^n$ such that no two of these words are adjacent. The zero-error capacity C_0 of the channel then is:

$$C_0 = \sup_{n\in\mathbb{N}} \frac{\log(N(n))}{n}. \tag{3.3}$$

Intuitively, we can think of the definition as follows: $N(n)$ corresponds to the maximum number of messages we can send without error by using the channel n times, so $\frac{\log(N(n))}{n}$ is the highest rate at which we can send information when using the channel n times. The capacity is then simply the supremum over these rates, very much like in the definition of the ordinary capacity.

Before we translate the above notions into graph theoretical language, we provide another definition and result that help us better understand the zero-error capacity:

Definition 3.4. A map $f : \mathcal{X} \to \mathcal{X}$ is called adjacency-reducing if for any $x, x' \in \mathcal{X}$ which are non-adjacent we have that $f(x)$ and $f(x')$ are non-adjacent.

An adjacency-reducing mapping can never increase the number of adjacent letters, but it can reduce them. This leads to the following theorem [GdAM16]:

Theorem 3.1. Let $p_{Y|X}(y|x)$ be a noisy channel. If there exists an adjacency-reducing map $f : \mathcal{X} \to \mathcal{X}'$ such that any two symbols in $\mathcal{X}'$ are non-adjacent, then $C_0 = \log(|\mathcal{X}'|)$.

Proof: suppose that such a map $f : \mathcal{X} \to \mathcal{X}'$ exists. Since all symbols in $\mathcal{X}'$ are non-adjacent, there exist at least $|\mathcal{X}'|^n$ distinguishable sequences. Thus, C_0 is at least $\frac{\log(|\mathcal{X}'|^n)}{n} = \log(|\mathcal{X}'|)$. However, to any set of distinguishable codewords $\{x^n(m)\}_{m\in\mathcal{M}}$ for message set $\mathcal{M}$ we can apply the map f to each letter in each codeword individually. Since f is adjacency-reducing, this again yields a set of distinguishable codewords. But all these new codewords are in $\mathcal{X}'^n$, so there can never be more than $|\mathcal{X}'|^n$. That is, the rate of the code cannot exceed $\frac{\log(|\mathcal{X}'|^n)}{n} = \log(|\mathcal{X}'|)$. □

Let us now finish this section by considering a seemingly simple, but actually very profound example.

Example 3.1. The pentagon channel G_5 is a channel with an input and output alphabet consisting of five symbols, which we will simply call 0,1,2,3 and 4. It is represented by the following diagram:

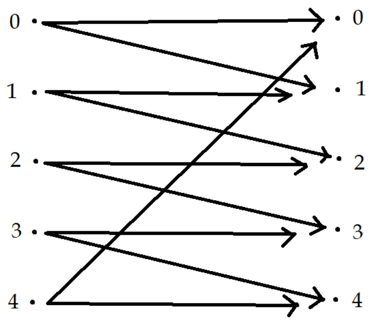

Figure 3.1: The Pentagon channel G_5. An arrow from one symbol to another represents a nonzero probability of the channel outputting the second symbol when the first is the input.

We see that the maximum cardinality of a set of distinguishable symbols is 2 (for example $\{0,2\}$). Thus, using codewords of length 1 would give a maximum rate of $\log(2) = 1$. Using codewords of length 2, however, we might for instance take the set $\{00,12,24,31,43\}$. This set contains only mutually distinguishable codewords, so the maximum rate then becomes $\frac{1}{2}\log(5)$. The reason that the pentagon channel is so interesting is that Shannon was originally able (in [Sha56]) to easily find all capacities of channels with up to five input symbols using theorem 3.1, except for this one. It then remained an open problem to find the zero-errro capacity of the pentagon channel for 20 years [GdAM16], until Lóvasz showed that the lower bound of $\frac{1}{2}\log(5)$ was tight in [Lov79]. This shows that the calculation of the zero-error capacity of even very simple channels can be highly nontrivial. In the next section we will see why the pentagon channel is called the pentagon channel.

3.2. Zero-Error Capacity and Graph Theory

We now reformulate the notions of the previous section in terms of graph theory. This will allow us to prove upper and lower bounds for the zero-error capacity of a noisy channel.

3.2.1. Characteristic Graphs

Associated to every noisy channel $p_{Y|X}(y|x)$ there is a characteristic graph, which, as we will see, contains all information necessary to define the zero-error capacity of that channel [GdAM16]. Let us first provide a proper definition:

Definition 3.5. Given a noisy channel $p_{Y|X}(y|X)$ with input symbol set $\mathcal{X}$, the characteristic graph G is the undirected graph with as its vertices the input symbols $\mathcal{X}$ and edges between the pairs of input symbols which are distinguishable.

Let us now again consider the pentagon channel.

Example 3.2. The characteristic graph of the pentagon channel G_5 is a pentagon, hence the name:

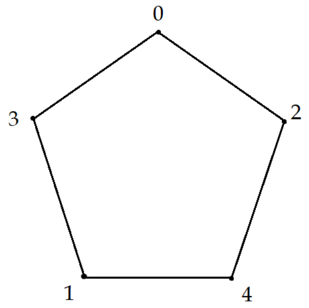

Figure 3.2: The characteristic graph of the pentagon channel.

In order to characterise the zero-error capacity of a channel in terms of its characteristic graph, we make use of the following notion:

Definition 3.6. Let G be an undirected graph. Then a clique in G is a subset of the vertices $V(G)$ such that the induced subgraph is complete. That is, the induced subgraph is such that between every two vertices there is a unique edge.

A clique in a characteristic graph corresponds to a set of mutually distinguishable input symbols - i.e. an error-free code using codewords of length one. We now define the clique number $\omega(G)$ to be the maximal order of a clique in G. It then follows that for a channel with characteristic graph G we have $N(1) = \omega(G)$. In order to define the zero-error capacity in terms of the characteristic graph, however, we need to incorporate input sequences of any length. Thus, we introduce the n-product of a graph [GdAM16]:

Definition 3.7. Given a characteristic graph G with vertices $\mathcal{X}$ we define the n-product G^n of G to be the graph with vertices $\mathcal{X}^n$ and an edge between all vertices $x^n, \tilde{x}^n$ for which there exists some $1 \leq i \leq n$ such that there is an edge in G between x_i and $\tilde{x}_i$.

Intuitively, G^n is the graph whose vertices are all input sequences x^n and which contains an edge between two such input sequences if they are distinguishable. Thus, we see that $N(n) = \omega(G^n)$ and we get the following characterisation of the zero-error capacity of a channel whose characteristic graph is G:

$$C_0 = \sup_{n \in \mathbb{N}} \frac{\log(\omega(G^n))}{n}. \tag{3.4}$$

In graph theory this value is called the Shannon capacity of the graph G [GdAM16]. We will now show that we can also formulate theorem 3.1 in terms of the characteristic graph. To this end, we introduce the notion of a coloring:

Definition 3.8. A colouring of a characteristic graph G is a map $f : V(G) \to K$ from the set of vertices of the graph to some set of colours K such that for any adjacent $x, x' \in V(G)$ we have $f(x) \neq f(x')$.

In other words: a colouring is a designation of a color to every vertex in the graph such that adjacent vertices are not assigned the same colour. This leads to the following:

Definition 3.9. The chromatic number of a characteristic graph G, denoted $\chi(G)$, is the smallest cardinality a colour set K can have such that there exists a colouring $f : V(G) \to K$.

. We are now in a position to formulate theorem 3.1 in terms of the maximal clique number and chromatic number [GdAM16]:

Theorem 3.2. Let $\mathcal{N}$ be a noisy channel with characteristic graph G. If $\omega(G) = \chi(G)$, then $C_0 = \log(\chi(G))$.

Proof: suppose $\omega(G) = \chi(G)$. Take $\mathcal{X}'$ to be the set of vertices of a maximal clique of G, and let $f : V(G) \to K$ be the minimal colouring of G which has $|K| = \chi(G) = \omega(G) = |\mathcal{X}'|$. Since all vertices in $\mathcal{X}'$ are distinguishable they must all bear a different colour. But $\chi(G) = \omega(G)$, so the vertices of G which are outside the maximal clique all have a colour in $f(\mathcal{X}') = K$. Define $\tilde{f} : V(G) \to \mathcal{X}'$ to be the map that takes a vertex to the unique element in $\mathcal{X}'$ that bears the same colour. Then $\tilde{f}$ is adjacency-reducing since it is induced by the colouring f. But we also know that any two symbols in $\mathcal{X}'$ are distinguishable, so $\tilde{f}$ satisfies the properties of theorem 3.1, and we conclude that $C_0 = \log(|\mathcal{X}'|) = \log(\chi(G))$. □

3.2.2. Adjacency Graphs

Now that we have worked with the characteristic graph of a noisy channel we will introduce the complementary graph of the characteristic graph, called the adjacency graph [GdAM16]. First, however, we define the adjacency matrix, originally introduced by Shannon [Sha56].

Definition 3.10. For a noisy channel $p_{Y|X}(y|x)$ with input alphabet $\mathcal{X} = \{x_1, ..., x_{|\mathcal{X}|}\}$ we define the adjacency matrix to be the $|\mathcal{X}| \times |\mathcal{X}|$ matrix A whose entries a_{ij} are 1 if x_i and x_j are adjacent or $i = j$ and zero else.

This then gives rise to the adjacency graph:

Definition 3.11. Given a noisy channel $p_{Y|X}(y|x)$ with input alphabet $\mathcal{X}$ we define the adjacency graph to be the graph with vertex set $\mathcal{X}$ and an edges between adjacent vertices.

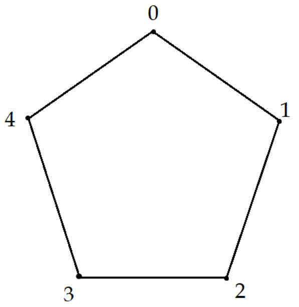

Figure 3.3: The adjacency graph of the pentagon channel.

Example 3.3. The adjacency graph of the pentagon channel G_5 is again a pentagon, but this time with the symbols in increasing order in the clockwise direction.

We are now finally in a position to state and prove the following theorem by Shannon [Sha56]:

Theorem 3.3. Let N be a noisy channel with adjacency matrix $A = [a_{ij}]$, with input alphabet $\mathcal{X} = \{x_1, ..., x_{|\mathcal{X}|}\}$ and output alphabet $\mathcal{Y} = \{y_1, ..., y_{|\mathcal{Y}|}\}$. Then the zero-error capacity C_0 of this channel satisfies:

$$-\log\left(\min_{p_X(x)} \sum_{1\le i,j\le|\mathcal{X}|} a_{ij}\, p_X(x_i)\, p_X(x_j)\right) \le C_0 \le \min_{p_{Y|X}(y|x)} C, \tag{3.5}$$

where $p_X(x)$ is any probability distribution on $\mathcal{X}$, $p_{Y|X}(y|x)$ is any conditional probability distribution on $\mathcal{X}$ and $\mathcal{Y}$ such that its adjacency matrix is A and C is the ordinary capacity of that channel $p_{Y|X}(y|x)$.

Proof: the upper bound follows immediately from the observation that the ordinary capacity of a noisy channel is always greater than or equal to the zero-error capacity, since a code which achieves the ordinary capacity need only have a probability of error which can be made arbitrarily small by using the channel many times, whereas a code which achieves the zero-error capacity must always have zero probability of error. Thus, every error-free code is also a code whose probability of error vanishes in the asymptotic limit, which means that for a specific channel the zero-error capacity cannot achieve the ordinary capacity. However, in the upper bound we minimise over all channels with the same adjacency matrix. It immediately follows that this minimal ordinary capacity is still greater than the zero-error capacity by observing that two channels with the same adjacency matrix have equal zero-error capacities. Moreover, analogously to our argument in equation 2.18, it can be seen that the mutual information and thus the ordinary capacity is convex in the conditional probability distribution $p_{Y|X}(y|x)$, which means we can actually consider the minimum ordinary capacity instead of the infimum [Sha56].

In order to prove the lower bound we follow Shannon's original argument [Sha56]. We again generate a random code of M independent codewords each of length n according to the probability distribution $p_X(x)$ on $\mathcal{X}$, as we have done in the previous chapter. For any two codewords, the probability of their letters at a certain position being adjacent is given by $\sum_{i_j} a_{ij}\, p_X(x_i)\, p_X(x_j)$, because the adjacency matrix is simply 1 when two symbols are adjacent and 0 otherwise. Thus, the probability that two randomly generated codewords are indistinguishable (i.e. all their letters are adjacent) is:

$$\left(\sum_{i_j} a_{ij}\, p_X(x_i)\, p_X(x_j)\right)^n. \tag{3.6}$$

Thus, given that we have generated a codeword, the probability that all other codewords in the code are distinguishable from it is:

$$\begin{aligned}\left(1-\left(\sum_{i_j} a_{ij}p_X(x_i)p_X(x_j)\right)^n\right)^{M-1} &\geq 1-(M-1)\left(\sum_{i_j} a_{ij}p_X(x_i)p_X(x_j)\right)^n \\ &\geq 1-M\left(\sum_{i_j} a_{ij}p_X(x_i)p_X(x_j)\right)^n.\end{aligned} \tag{3.7}$$

Now, for any $\epsilon \in (0,1)$ we can take M such that:

$$(1-\epsilon)^n\left(\sum_{i_j} a_{ij}p_X(x_i)p_X(x_j)\right)^{-n} \leq M \leq \left(\sum_{i_j} a_{ij}p_X(x_i)p_X(x_j)\right)^{-n}. \tag{3.8}$$

Choosing this ϵ arbitrarily small then makes the rate $R=\frac{\log(M)}{n}$, for which we have:

$$-(1-\epsilon)\log\left(\sum_{i_j} a_{ij}p_X(x_i)p_X(x_j)\right) \leq R \leq -\log\left(\sum_{i_j} a_{ij}p_X(x_i)p_X(x_j)\right), \tag{3.9}$$

arbitraily close to $-\log\left(\sum_{i_j} a_{ij}p_X(x_i)p_X(x_j)\right)$. Furthermore, for any $\delta>0$ we can choose some $n\in\mathbb{N}$ such that:

$$(1-\epsilon)^n = M\left(\sum_{i_j} a_{ij}p_X(x_i)p_X(x_j)\right)^n < \delta. \tag{3.10}$$

But this implies that for any codeword the probability of some other codeword being adjacent to it is less than δ. This again implies that for any $\delta>0$ there exists a code with rate arbitrarily close to $-\log\left(\sum_{i_j} a_{ij}p_X(x_i)p_X(x_j)\right)$ whose codewords are length $n\in\mathbb{N}$ such that the fraction of 'undesired' codewords (i.e. codewords which are adjacent to another codeword) to the total number of codewords is less than δ. But if we then omit these undesired codewords the new rate of the error-free code becomes:

$$R' = \frac{\log((1-\delta)M)}{n} = \frac{\log(1-\delta)}{n}+R. \tag{3.11}$$

Thus, we conclude that for every $\epsilon,\delta>0$ we can choose some $n\in\mathbb{N}$ such that there is a zero-error code with whose rate R satisfies:

$$-(1-\epsilon)\log\left(\sum_{i_j} a_{ij}p_X(x_i)p_X(x_j)\right)+\frac{\log(1-\delta)}{n} \leq R \leq \frac{\log(1-\delta)}{n}+-\log\left(\sum_{i_j} a_{ij}p_X(x_i)p_X(x_j)\right). \tag{3.12}$$

By making ϵ and δ arbitrarily small we find the lower bound of the theorem. □

3.3. Product Channels

If we have two noisy channels, we might use them together to transmit information and wonder what the capacity of this resulting channel is. In this section we will first consider Shannon's results and thoughts on this matter, and then we will prove that the zero-error capacity of such a combined channel can actually be greater than the sum of the zero-error capacities of the individual channels - contrary to what Shannon originally conjectured.

3.3.1. Shannon's Results and Conjecture

In order to understand the above ideas we first recall the definition of the product of two classical channels (definition 2.11):

Definition 3.12. Let $\mathcal{N}_1$ and $\mathcal{N}_2$ be two noisy with input alphabets $\mathcal{X}_1$ and $\mathcal{X}_2$ and output alphabets $\mathcal{Y}_1$ and $\mathcal{Y}_2$ respectively. Let $p_{Y_1|X_1}(y_1|x_1)$ and $p_{Y_2|X_2}(y_2|x_2)$ be the conditional distributions of channels 1 and 2 respectively. Then the product channel $\mathcal{N}_1 \otimes \mathcal{N}_2$ is the channel with input alphabet $\mathcal{X}_1 \times \mathcal{X}_2$, output alphabet $\mathcal{Y}_1 \times \mathcal{Y}_2$ and the following conditional distribution:

$$p_{Y_1 \times Y_2|X_1 \times X_2}(y_1, y_2|x_1, x_2) = p_{Y_1|X_1}(y_1|x_1) \cdot p_{Y_2|X_2}(y_2|x_2). \tag{3.13}$$

The following result is due to Shannon [Sha56]:

Theorem 3.4. Let $(\mathcal{X}_1, \mathcal{Y}_1, p_{Y_1|X_1}(y_1|x_1))$ and $(\mathcal{X}_2, \mathcal{Y}_1, p_{Y_2|X_2}(y_2|x_2))$ be two channels with zero-error capacities C_0^1 and C_0^2 respectively. Let C_0 denote the zero-error capacity of their product. Then:

$$C_0 \geq C_0^1 + C_0^2. \tag{3.14}$$

Moreover, the above inequality becomes an equality if the input alphabet of at least one of the channels can be mapped by an adjacency-reducing mapping to a subset of the alphabet such that no two symbols in this subset are adjacent.

Proof: suppose we have two error-free codes for each of the two channels, with message set sizes M_1 and M_2 and codeword lengths n_1 and n_2 respectively. We construct a so-called product code for the product channel. This product code looks as follows: we take codewords of length $\text{lcm}(n_1, n_2)$, where lcm means least common multiple. These codewords consist of $\text{lcm}(n_1, n_2)$ ordered pairs in $\mathcal{X}_1 \times \mathcal{X}_2$ such that the first entries of these pairs are the concatenation of $\frac{\text{lcm}(n_1,n_2)}{n_1}$ codewords from the first code, and the second entries are the concatenation of $\frac{\text{lcm}(n_1,n_2)}{n_2}$ codewords from the second code. This way, we obtain an error-free code for the product channel with the following rate R:

$$\begin{aligned} R = \frac{\log\left(M_1^{\frac{\text{lcm}(n_1,n_2)}{n_1}} M_2^{\frac{\text{lcm}(n_1,n_2)}{n_2}}\right)}{\text{lcm}(n_1,n_2)} &= \frac{\frac{\text{lcm}(n_1,n_2)}{n_1}\log(M_1) + \frac{\text{lcm}(n_1,n_2)}{n_2}\log(M_2)}{\text{lcm}(n_1,n_2)} \\ &= \frac{\log(M_1)}{n_1} + \frac{\log(M_2)}{M_2} = R_1 + R_2. \end{aligned} \tag{3.15}$$

From the construction of this product code whose rate is the sum of the rates of the individual codes it follows that the zero-error capacity of the product channel must at least be the sum of the individual zero-error capacities. We will now prove the second part of the theorem. Suppose, without loss of generality, that there exists an adjacency-reducing mapping $f : \mathcal{X}_1 \to \mathcal{X}_1'$ from the alphabet of the first channel to a distinguishable subset of that alphabet. Suppose we have some error-free code for the product channel which has M codewords of length n. Since each of these codewords consists of n ordered pairs in $\mathcal{X}_1 \times \mathcal{X}_2$ we can map the first entries of these pairs to $\mathcal{X}_1'$ using f. Since f is adjacency-reducing this again gives an error-free code with M codewords of length n. However, this new code can have at most $|\mathcal{X}_1'|^n \cdot 2^{nC_0^2}$ codewords, since there are only a maximum of $|\mathcal{X}_1'|$ and $2^{nC_0^2}$ different symbols for the first and second entries respectively. From theorem 3.1 we know $C_0^1 = \log(|\mathcal{X}_1'|)$, so we see that the rate of the code for the product channel is no greater than:

$$\frac{\log(|\mathcal{X}_1'|^n \cdot 2^{nC_0^2})}{n} = \log(|\mathcal{X}_1'|) + C_0^2 = C_0^1 + C_0^2. \tag{3.16}$$

Combining this with the above result that the zero-error capacity of the product channel can never be less than the sum of the individual zero-error capacities we obtain the statement from the theorem. □

3.3.2. Superadditivity of Zero-Error Capacity

We have now seen that the zero-error capacities of the product of two classical channels is at least the sum of the individual zero-error capacities. We have, however, not shown that there exists channels for which the product zero-error capacity is strictly greater than the sum of the individual capacities. The following theorem was proven in [Alo98]:

Theorem 3.5. There exists a graph G with 27 vertices so that $C^0(G) \leq 7, C^0(\bar{G}) = 3$, whereas $C^0(G + \bar{G}) \geq 2\sqrt{27}$.

Here $\bar{G}$ denotes the complement of G, and $G + \bar{G}$ is the union of G and $\bar{G}$. This union is the adjacency graph of the product of channels whose adjacency graphs are G and $\bar{G}$ [GdAM16]. Thus, this theorem shows that the zero-error capacity of a classical channel can be superadditive, contrary to the ordinary capacity of a classical channel.

II

Quantum Communication

4

Notions of Quantum Information Theory

In this chapter we introduce the basic concepts of quantum information theory. We present the fundamental definitions and theorems which we will make heavy use of later on in this thesis. Since all results in this chapter are basic results of quantum information theory we often refer the reader to the literature for the proofs.

4.1. Density Operators on Hilbert Spaces

We first introduce the density operator formalism, which is ubiquitous in quantum information theory. Note that, although the postulates we now introduce are quite general, we will always be working with finite-dimensional Hilbert spaces.

4.1.1. Postulates of Quantum Mechanics

We start with the fundamental postulates of quantum mechanics in terms of density operators. These are centered around the notion of a Hilbert space. Roughly speaking, this is the space quantum states "live in". We define it according to [Maa]:

Definition 4.1. A Hilbert space is a complex vector space $\mathcal{H}$, endowed with an inner product:

$$\mathcal{H} \times \mathcal{H} \to \mathbb{C}: \quad (\psi, \phi) \mapsto \langle \psi, \phi \rangle \tag{4.1}$$

satifysing the following properties:

(i) $\langle \psi, \phi_1 + \phi_2 \rangle = \langle \psi, \phi_1 \rangle + \langle \psi, \phi_2 \rangle$ for all $\psi, \phi_1, \phi_2 \in \mathcal{H}$;
(ii) $\langle \psi, \lambda\phi \rangle = \lambda \langle \psi, \phi \rangle$ for all $\psi, \phi \in \mathcal{H}$ and all $\lambda \in \mathbb{C}$;
(iii) $\overline{\langle \psi, \phi \rangle} = \langle \phi, \psi \rangle$ for all $\psi, \phi \in \mathcal{H}$;
(iv) $\langle \psi, \psi \rangle \geq 0$ for all $\psi \in \mathcal{H}$;
(v) $\langle \psi, \psi \rangle = 0$ implies $\psi = 0$;
(vi) $\mathcal{H}$ is complete in the norm $||\psi|| = \langle \psi, \psi \rangle^{\frac{1}{2}}$.

The most prominent example of a Hilbert space in our study of quantum information theory is simply that of $\mathbb{C}^n$. A qubit, for example, is described on the Hilbert space $\mathbb{C}^2$. Now that we understand the notion of a Hilbert space, we can introduce the postulates of quantum mechanics in terms of density operators. To this end, we will follow [Hol19] and [Sch].

Postulate 4.1. For every quantum system there is an associated Hilbert space $\mathcal{H}$. The states of the system are all nonnegative trace-class linear maps $\rho : \mathcal{H} \to \mathcal{H}$ for which $\text{Tr}(\rho) = 1$.

These operators are called density operators, and we denote the set of density operators on $\mathcal{H}$ by $\mathcal{D}(\mathcal{H})$. A trace-class linear map $A : \mathcal{H} \to \mathcal{H}$ is a map for which the trace $\mathrm{Tr}(A) = \sum_k \langle Ae_k, e_k \rangle$ can be defined, where $\{e_k\}$ is a basis of $\mathcal{H}$. In this thesis we will only work with finite-dimensional Hilbert spaces, where every operator is trace-class. An important property of the trace is that it is independent of the basis chosen. This follows immediately from two the fact that $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ for any trace-class operators A, B (this can be checked simply by writing out). To see that basis-indepence follows, recall that a change of basis is represented by a unitary operator U which changes an operator A in one basis to $U^{-1}AU$. We thus have: $\mathrm{Tr}(U^{-1}AU) = \mathrm{Tr}(AUU^{-1}) = \mathrm{Tr}(A)$. Another property of the trace that we will often use is its cyclicity. This can also be checked simply by writing out. It states that for any trace-class operators A, B, C we can cyclically permute: $\mathrm{Tr}(ABC) = \mathrm{Tr}(BCA) = \mathrm{Tr}(CAB)$.

Now that we understand what states are, we need to understand how to define observables.

Postulate 4.2. The observables of a quantum system are the self-adjoint linear maps $A : \mathcal{H} \to \mathcal{H}$. The expectation value of an observable A for a system in state ρ is given by:

$$\langle A \rangle = \mathrm{Tr}(\rho A). \tag{4.2}$$

Another word for self-adjoint is Hermitian. Recall that an operator A is self-adjoint if $A^* = A$, where A^* is the adjoint of A. Recall also that the adjoint A^* is such that $\langle \psi, A^* \phi \rangle = \langle A\psi, \phi \rangle$. Now that we understand states and observables, it is time to consider the two ways in which a quantum system can evolve: unitary dynamics and projective dynamics.

Postulate 4.3. If a system is in a state $\rho(t_1)$ at time t_1, then if no measurement is done on the system, the system will be in the state $\rho(t_2)$ at a later time t_2, given by:

$$\rho(t_2) = U(t_2 - t_1)\rho(t_1)U^{-1}(t_2 - t_1), \tag{4.3}$$

where $U(t_2 - t_1)$ is a unitary operator given by:

$$U(t_2 - t_1) = e^{-\frac{i}{\hbar}H(t_2 - t_1)}. \tag{4.4}$$

Here H is the observable corresponding to the energy of the system (the Hamiltonian).

Now that we understand unitary dynamics, we turn to projective dynamics, which concerns measuring a system:

Postulate 4.4. A measurement on a system in a state ρ is represented by a projection valued measure (PVM), which in the finite-dimensional case is a set of operators $\{M_i\}$ on $\mathcal{H}$ such that:

$$\sum_i M_i^\dagger M_i = \mathbb{1}. \tag{4.5}$$

The probability that we obtain outcome i is given by:

$$\Pr(i) = \mathrm{Tr}(M_i \rho M_i^\dagger). \tag{4.6}$$

Moreover, if the measurement yields outcome i, then the state after the measurement becomes:

$$\frac{M_i \rho M_i^\dagger}{\mathrm{Tr}(M_i \rho M_i^\dagger)}. \tag{4.7}$$

Given a PVM $\{M_i\}$ we can construct a positive-operator valued measure (POVM) $\{F_i\}$ where $F_i = M_i^\dagger M_i$, such that $\sum_i F_i = \mathbb{1}$ and $\Pr(i) = \mathrm{Tr}(\rho F_i)$. Multiple PVM's can correspond to the same POVM, and therefore a POVM does not specify the post-measurement state. In this thesis, however, we often use POVM's to represent measurements, because we often do not need to specify the post-measurement state.

4.1.2. Bras, Kets, Spectral Decompostition, Purity and Projectors

Now that we have presented the postulates of quantum mechanics, we introduce some ubiquitous notation, following [Hol19]: the notation of bras and kets, introduced by Paul Dirac.

Notation 4.5. A ket, denoted by $|\phi\rangle$, represents a unit vector in the Hilbert space $\mathcal{H}$ modulo the action of the circle group.

This modulo the action of the circle group simply means that two unit vectors represent the same physical state if they differ only by a phase factor, i.e. unit vectors $\psi \in \mathcal{H}$ and $\phi \in \mathcal{H}$ are the same state if $\psi = e^{i\theta}\phi$ for some angle $\theta \in [0, 2\pi)$. Now we define a bra, such that together they form the word braket, which is Dirac's wordplay on the word bracket.

Definition 4.2. A bra is a linear functional $\langle\psi| : \mathcal{H} \to \mathbb{C}$, defined by:

$$\langle\psi|(|\phi\rangle) =: \langle\psi|\phi\rangle = \langle\psi, \phi\rangle. \tag{4.8}$$

Given a ket $|\phi\rangle$ we thus have a corresponding bra $\langle\phi|$, which is called the adjoint to $|\phi\rangle$. This notation of bras and kets turns out to be very useful, because it allows for a convenient description of operators [Hol19]. Indeed, given a bra $\langle\psi|$ and a ket $|\phi\rangle$, we can use $A = |\phi\rangle\langle\psi|$ to denote the operator which acts by $A|\chi\rangle = |\phi\rangle\langle\psi|\chi\rangle$. Moreover, given an orthonormal basis $\{|e_i\rangle\}$, we have:

$$\sum_i |e_i\rangle\langle e_i| = \mathbb{1}, \tag{4.9}$$

and thus we can decompose an arbitrary ket $|\phi\rangle$ as:

$$|\phi\rangle = \sum_i |e_i\rangle\langle e_i|\phi\rangle. \tag{4.10}$$

Now, the real usefulness comes from using this notation for operators to describe density operators. It follows from the positive semidefiniteness of density operators that they are also self-adjoint. Thus, the spectral theorem tells us that for every density operator ρ on $\mathcal{H}$ we can find an orthonormal basis $\{|\phi_i\rangle\}$ of $\mathcal{H}$ such that:

$$\rho = \sum_i \lambda_i |\phi_i\rangle\langle\phi_i|, \tag{4.11}$$

where the eigenvalues λ_i are real [Hol19]. Knowing this, we define the following:

Definition 4.3. A pure state is a density operator ρ which can be written as $\rho = |\phi\rangle\langle\phi|$, where $|\phi\rangle$ is some ket.

Continuing this line of thought, we can define the purity of an arbitrary density operator ρ as $\mathrm{Tr}(\rho^2)$. It is not hard to show that the purity is always between 0 and 1, and it is 1 if and only if ρ is a pure state [Wil19].

A class of operators that we will often come across are the projectors. They will be especially important for us in our treatment of typical subspaces, where we will define so-called typical projectors.

Definition 4.4. A projector is a self-adjoint operator P such that $P^2 = P$. For an orthonormal system $\{|e_i\rangle\}$, $P = \sum_i |e_i\rangle\langle e_i|$ is the projector onto the subspace generated by $\{|e_i\rangle\}$.

4.1.3. Composite Systems and Entanglement

We have now treated the most important concepts regarding quantum mechanics in terms of density operators, but there is one more aspect of quantum mechanics that we have not considered, which is actually the source of many of the most useful and interesting quantum mechanical phenomena. It is the composition of quantum mechanical systems, and its consequence: entanglement. Indeed, this aspect is very different from our classical intuition. When we have two quantum mechanical systems and we combine them, the resulting system is not simply described by the direct sum of the underlying Hilbert spaces. Instead, the Hilbert space of the combined system becomes the tensor product of the Hilbert spaces of the subsystems. We define it according to [Hol19].

Definition 4.5. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be Hilbert spaces. Then the tensor product of these spaces $\mathcal{H}_1 \otimes \mathcal{H}_2$ is the vector space which consists of all finite linear combinations:

$$\sum_i c_i\, \phi_1 \otimes \phi_2, \tag{4.12}$$

where $\phi_1 \in \mathcal{H}_1$ and $\phi_2 \in \mathcal{H}_2$. We write $|\phi_1 \otimes \phi_2\rangle = |\phi_1\rangle \otimes |\phi_2\rangle$, and we define an inner product on this vector space by:

$$\langle \psi_1 \otimes \psi_2 | \phi_1 \otimes \phi_2 \rangle = \langle \psi_1 | \phi_1 \rangle \langle \psi_2 | \phi_2 \rangle. \tag{4.13}$$

It can easily be checked, that if $\{|e_i\rangle\}$ and $\{|e_j\rangle\}$ are orthonormal bases of $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively, then $\{|e_i \otimes e_j\rangle\}$ is an orthonormal basis of $\mathcal{H}_1 \otimes \mathcal{H}_2$. Furthermore, $\dim(\mathcal{H}_1 \otimes \mathcal{H}_2) = \dim(\mathcal{H}_1)\dim(\mathcal{H}_2)$ [Hol19]. The tensor product of two operators is defined exactly the way we would expect it to. If A and B are operators on $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively, then $A \otimes B$ acts on $\mathcal{H}_1 \otimes \mathcal{H}_2$ by:

$$A \otimes B |\phi_1 \otimes \phi_2\rangle = A|\phi_1\rangle \otimes B|\phi_2\rangle. \tag{4.14}$$

In this thesis we often talk about the close personal friends Alice and Bob. They both have some quantum system, and we consider how they can communicate with each other using the laws of quantum mechanics. So, say Alice has a system $\mathcal{H}_A$ and Bob has a system $\mathcal{H}_B$. Their combined system is then described by $\mathcal{H}_A \otimes \mathcal{H}_B$. But if Alice acts on her system with some operator, then she only affects her part of the system. That is, Alice only acts *locally*. To capture this notion mathematically, we define the following according to [Wil19]:

Definition 4.6. Let X_{AB} be an operator acting on $\mathcal{H}_A \otimes \mathcal{H}_B$, and let $\{|\phi\rangle_B^i\}$ be an orthonormal basis of $\mathcal{H}_B$. Then the partial trace over $\mathcal{H}_B$ is defined as:

$$\mathrm{Tr}_B(X_{AB}) = \sum_i (\mathbb{1}_A \otimes \langle \phi |_B^i) X_{AB} (\mathbb{1}_A \otimes |\phi\rangle_B^i). \tag{4.15}$$

The partial trace allows us to get back Alice's system from the total composite system. Indeed, if the composite system is in the state ρ_{AB}, then we can get Alice's density operator simply by taking $\rho_A = \mathrm{Tr}_B(\rho_{AB})$. Moreover, if Alice locally performs a measurement described by the POVM Λ_A^j, then the probability that she receives outcome j satisfies the following:

$$\Pr(j) = \mathrm{Tr}(\rho_A \Lambda_A^j) = \mathrm{Tr}(\rho_{AB}(\Lambda_A^j \otimes \mathbb{1}_B)). \tag{4.16}$$

That is: it does not matter whether we consider Alice's system only and perform a local measurement and forget about the outside world, or if we consider the composite system as a whole and perform a measurement which leaves Bob's part of the system invariant (represented by the identity operator $\mathbb{1}_B$).

Now that we have defined composite systems we can indeed turn to the concept that is the source of much of the interesting quantum phenomena we will encounter in this thesis: entanglement. To understand this concept, we define the following:

Definition 4.7. A state ρ_{AB} on the tensor product space $\mathcal{H}_A \otimes \mathcal{H}_B$ is called separable if it can be written as a convex combination of product states:

$$\rho_{AB} = \sum_x p_x \rho_A^x \otimes \rho_B^x. \tag{4.17}$$

That is: $\sum_x p_x = 1$ and $p_x \geq 0$. A state in $\mathcal{H}_A \otimes \mathcal{H}_B$ which is not separable is called entangled.

The canonical example of entanglement is given by the Bell states:

Example 4.1. Let $\mathcal{H}_A$ and $\mathcal{H}_B$ be qubit systems - that is $\mathbb{C}^2$. Then the states $|\Phi^+\rangle\langle\Phi^+|, |\Phi^-\rangle\langle\Phi^-|$, $|\Psi^+\rangle\langle\Psi^+|$ and $|\Psi^-\rangle\langle\Psi^-|$ defined by:

(i) $|\Phi^+\rangle = \frac{1}{\sqrt{2}}(|0\rangle_A \otimes |0\rangle_B + |1\rangle_A \otimes |1\rangle_B)$,
(ii) $|\Phi^-\rangle = \frac{1}{\sqrt{2}}(|0\rangle_A \otimes |0\rangle_B - |1\rangle_A \otimes |1\rangle_B)$,
(iii) $|\Psi^+\rangle = \frac{1}{\sqrt{2}}(|0\rangle_A \otimes |1\rangle_B + |1\rangle_A \otimes |0\rangle_B)$,
(iv) $|\Psi^-\rangle = \frac{1}{\sqrt{2}}(|0\rangle_A \otimes |1\rangle_B - |1\rangle_A \otimes |0\rangle_B)$,

are called the Bell states.

4.2. Quantum Entropy

So far we have introduced the postulates of quantum mechanics, and we have looked at some properties of density operators. We now start our investigation of quantum information theory by generalising the classical entropy and mutual information to the quantum case - resulting in the definition of the quantum entropy. The quantum entropy will be vastly important in the next chapters - especially for proving the HSW theorem in chapter 5.

We define the quantum entropy analogously to the classical entropy, following the notation of [Wil19].

Definition 4.8. Suppose we have a density operator ρ_A acting on the Hilbert space $\mathcal{H}_A$. Then the entropy of the state is defined as:

$$H(A)_\rho = -\mathrm{Tr}(\rho_A \log(\rho_A)). \tag{4.18}$$

We might denote the quantum entropy either by $H(A)_\rho$, by $H(\rho_A)$ or simply by $H(\rho)$. As a reminder: if we write a state ρ in its spectral decomposition $\rho = \sum_x \lambda_x |x\rangle\langle x|$, and we have a function $f: \mathbb{C} \to \mathbb{C}$, then we define $f(\rho)$ by $f(\rho) = \sum_x f(\lambda_x)|x\rangle\langle x|$. The quantum entropy satisfies the same two properties which we previously mentioned for the classical entropy [Wil19]:

Property 4.6. For any density operator $\rho \in \mathcal{D}(\mathcal{H}_A)$ we have $H(\rho_A) \geq 0$.

Property 4.7. The quantum entropy is concave. That is, if we have density operators $\rho_x \in \mathcal{H}$ and a probability distribution $p_X(x)$, then:

$$H(\rho) = H\left(\sum_x p_X(x)\rho_x\right) \geq \sum_x p_X(x) H(\rho_x). \tag{4.19}$$

Before we define the conditional quantum entropy, we will first introduce the joint quantum entropy, since it is very logical:

Definition 4.9. The joint entropy of a state $\rho_{AB} \in \mathcal{H}_A \otimes \mathcal{H}_B$ is:

$$H(AB)_\rho = -\text{Tr}(\rho_{AB}\log(\rho_{AB})). \tag{4.20}$$

We now consider an additivity property not of a capacity, but of the quantum entropy.

Property 4.8. Let $\rho_A \in \mathcal{D}(\mathcal{H}_A)$ and $\sigma_B \in \mathcal{D}(\mathcal{H}_B)$. The quantum entropy is additive for tensor-product states:

$$H(\rho_A \otimes \sigma_B) = H(\rho_A) + H(\sigma_B). \tag{4.21}$$

Considering the way in which we defined the conditional classical entropy, we are now able to similarly define the conditional quantum entropy:

Definition 4.10. Let $\rho_{AB} \in \mathcal{D}(\mathcal{H}_A \otimes \mathcal{H}_B)$. The conditional entropy $H(A|B)_\rho$ of the state ρ_{AB} is equal to the difference between the joint entropy and the marginal entropy:

$$H(A|B)_\rho = H(AB)_\rho - H(B)_\rho. \tag{4.22}$$

It is important to clarify this notation, because one might wonder how $H(B)_\rho$ is defined, since ρ_{AB} really is a state of the composite system $\mathcal{H}_A \otimes \mathcal{H}_B$, and not a state of the subsystem $\mathcal{H}_B$. What we really mean with this, as one might expect, is the entropy of the state ρ_B which is obtained by (partially) tracing out the system A. Thus, we could rewrite the above definition as:

$$H(A|B)_\rho = H(\rho_{AB}) - H(\rho_B). \tag{4.23}$$

In the same fashion as with the classical entropy, conditioning does not increase the quantum entropy [Wil19]:

Theorem 4.9. For any bipartite state $\rho_{AB} \in \mathcal{D}(\mathcal{H}_A \otimes \mathcal{H}_B)$ we have $H(A)_\rho \geq H(A|B)_\rho$.

To finish off this subsection on quantum entropy, we generalise the notion of mutual information to the quantum stage:

Definition 4.11. The quantum mutual information of a state $\rho_{AB} \in \mathcal{D}(\mathcal{H}_A \otimes \mathcal{H}_B)$ is defined as:

$$I(A;B)_\rho = H(A)_\rho + H(B)_\rho - H(AB)_\rho. \tag{4.24}$$

Exactly analogously to the classical case, it satisfies the relations $I(A;B)_\rho = H(A)_\rho - H(A|B)_\rho = H(B)_\rho - H(B|A)_\rho$.

4.3. Quantum Channels

At last, we turn to the fundamental concept in our study of quantum information theory: the quantum channel. It is simply a map which takes quantum states as its input and gives quantum states as its output, satisfying several conditions. The rest of this thesis will be dedicated to investigating how one can send classical information (i.e. bits) over a quantum channel.

4.3.1. Definition of a Quantum Channel

In order to understand the mathematical definition of a quantum channel, we first have to understand the notion of a positive map. To this end, we introduce the following notation:

Notation 4.10. We denote by $\mathcal{L}(\mathcal{H})$ the set of linear operators acting on the Hilbert space $\mathcal{H}$. Moreover, we denote by $\mathcal{L}(\mathcal{H}_A, \mathcal{H}_B)$ the set of linear operators from $\mathcal{H}_A$ to $\mathcal{H}_B$. As we have already seen, we denote the subset of $\mathcal{L}(\mathcal{H})$ which contains the density operators on $\mathcal{H}$ by $\mathcal{D}(\mathcal{H})$.

Now we turn to the notion of positivity:

Definition 4.12. A linear map $\mathcal{M} : \mathcal{L}(\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B)$ is called positive if for all positive semi-definite operators $X_A \in \mathcal{L}(\mathcal{H}_A) : \mathcal{M}(X_A)$ is also a positive semi-definite operator.

Positivity is, however, not strong enough to provide the definition of a quantum channel, because we want to be able to apply a quantum channel to only a part of a larger system and still have it map positive operators to positive operators. This notion is called complete positivity. The intuition is that if Alice has a state $\rho_A \in \mathcal{D}(\mathcal{H}_A)$ which really is her share of a two-party state $\rho_{AR} \in \mathcal{D}(\mathcal{H}_A \otimes \mathcal{H}_R)$, then a quantum channel acting on Alice's system while doing nothing to the reference system should still yield a positive operator [Wil19]. This leads to the following:

Definition 4.13. A linear map $\mathcal{M} : \mathcal{L}(\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B)$ is called completely positive if the map $\mathcal{M} \otimes \mathrm{id}_R : \mathcal{L}(\mathcal{H}_A \otimes \mathcal{H}_R) \to \mathcal{L}(\mathcal{H}_B \otimes \mathcal{H}_R)$ is positive for any reference system R.

One might wonder whether positivity does not just imply complete positivity. Indeed it does not. An elementary example of a positive map which is not completely positive is the transpose map on a qubit, which simply gives the transpose of an operator on $\mathbb{C}^2$ [Wer]. We are now in a position to define quantum channels.

Definition 4.14. A quantum channel $\mathcal{N} : \mathcal{L}(\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B)$ is a linear, completely positive, trace preserving map.

The trace-preserving condition simply means that for all $X_A \in \mathcal{L}(\mathcal{H}_A)$ we have $\mathrm{Tr}(X_A) = \mathrm{Tr}(\mathcal{N}(X_A))$. We remark that we are working in the Schrödinger picture. That is, we consider a quantum channel to be a map that takes a quantum state to another quantum state. The equivalent description in the Heisenberg picture is that a quantum channel is a map from $\mathcal{L}(\mathcal{H}_B) \to \mathcal{L}(\mathcal{H}_A)$ which takes observables to observables. Now, this might be confusing, because one may wonder whether we define a quantum channel to be a map from operators on $\mathcal{H}_A$ to $\mathcal{H}_B$ or vice versa. There is, however, no problem here, because we think of a quantum channel as a map from $\mathcal{L}(\mathcal{H}_A)$ to $\mathcal{L}(\mathcal{H}_B)$ to map *states* to *states*. The reason that we are not restricting the quantum channel to a map from $\mathcal{D}(\mathcal{H}_A)$ to $\mathcal{D}(\mathcal{H}_B)$ is because the density operators do not form a vector space (if $\rho \in \mathcal{D}(\mathcal{H})$ then $-\rho \notin \mathcal{D}(\mathcal{H})$) so we would not be able to talk about linearity if we would do so.

In an analogous fashion to vectors in a Hilbert space, quantum channels also have an adjoint. It is defined as follows:

Definition 4.15. Let $\mathcal{N} : \mathcal{L}(\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B)$ be a quantum channel. Then the adjoint $\mathcal{N}^* : \mathcal{L}(\mathcal{H}_B) \to \mathcal{L}(\mathcal{H}_A)$ of $\mathcal{N}$ is the unique quantum channel satisfying [Wil19]:

$$\langle B, \mathcal{N}(A) \rangle = \langle \mathcal{N}^*(B), A \rangle, \tag{4.25}$$

for any $A \in \mathcal{L}(\mathcal{H}_A), B \in \mathcal{L}(\mathcal{H}_B)$.

We can also straightforwardly define the tensor product of two channels.

Definition 4.16. Let $\mathcal{N}_1, \mathcal{N}_2 : \mathcal{L}(\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B)$ be quantum channels. Then the tensor product $\mathcal{N}_1 \otimes \mathcal{N}_2 : \mathcal{L}(\mathcal{H}_A \otimes \mathcal{L}\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B \otimes \mathcal{L}\mathcal{H}_B)$ is the quantum channel defined by:

$$(\mathcal{N}_1 \otimes \mathcal{N}_2)(X_{AA'}) = \mathcal{N}_1(X_A) \otimes \mathcal{N}_2(X_{A'}), \tag{4.26}$$

where $X_A = \mathrm{Tr}_{A'}(X_{AA'})$ and $X_{A'} = \mathrm{Tr}_A(X_{AA'})$.

We finish this subsection by introducing some useful notation.

Notation 4.11. Let $\mathcal{H}$ be a hilbert space and $\rho \in \mathcal{D}(\mathcal{H})$. Then we write $\mathcal{H}^{\otimes n} = \mathcal{H} \otimes \cdots \otimes \mathcal{H}$ and $\rho^{\otimes n} = \rho \otimes \cdots \otimes \rho$, where these tensor products are performed n times. We do the same for n tensor products $\mathcal{N}^{\otimes n} = \mathcal{N} \otimes \cdots \otimes \mathcal{N}$ of a quantum channel.

4.3.2. Choi-Kraus Decomposition

Now that we know what quantum channels are, we present one of the fundamental theorems of quantum information theory. It tells us that we can always write quantum channels in a certain way which is called the Choi-Kraus decomposition.

Theorem 4.12. A map $\mathcal{N} : \mathcal{L}(\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B)$ is linear, completely positive and trace-preserving (i.e. a quantum channel) if and only if it has a Choi-Kraus decomposition as follows:

$$\mathcal{N}(X_A) = \sum_{l=0}^{d-1} V_l X_A V_l^\dagger, \tag{4.27}$$

for all $X_A \in \mathcal{L}(\mathcal{H}_A)$, and where the $V_l \in \mathcal{L}(\mathcal{H}_A, \mathcal{H}_B)$ are fixed and such that:

$$\sum_{l=0}^{d-1} V_l^\dagger V_l = \mathbb{1}_A, \tag{4.28}$$

and d need not be any larger than $\dim(\mathcal{H}_A)\dim(\mathcal{H}_B)$.

The operators V_l are called Kraus operators. The proof of this theorem can be found in subsection 4.4.1 of [Wil19]. Moreover, if $\{V_l\}$ is a set of Kraus operators of the quantum channel $\mathcal{N}$, then $\{V_l^\dagger\}$ is a set of Kraus operators of the adjoint channel $\mathcal{N}^*$ [CCH11].

4.3.3. Examples of Quantum Channels

We now present three examples of quantum channels, the latter two of which we will meet later on in this thesis. The first of these is the classical-classical channel. It simply corresponds to a classical channel, which can be represented by a conditional probability distribution $p_{Y|X}(y|x)$ on the alphabets $\mathcal{X}$ and $\mathcal{Y}$. The corresponding classical-classical channel then has Kraus operators $\{\sqrt{p_{Y|X}(y|x)}|y\rangle\langle x|\}_{x,y}$ where $|x\rangle$ and $|y\rangle$ are bases of the input and output Hilbert spaces corresponding to the classical input and output letters respectively [Wil19].

The second quantum channel that we present is the classical-quantum channel. This channel first measures the input in some orthonormal basis, and then outputs a density operator [Wil19]. So given an orthonormal basis $|k\rangle_A$ of $\mathcal{H}_A$ and a set of density operators $\{\sigma_B^k\}$ in $\mathcal{D}(\mathcal{H}_B)$, the classical-quantum channel $\mathcal{N}$ has the following acting on a density operator $\rho_A \in \mathcal{L}(\mathcal{H}_A)$:

$$\mathcal{N}(\rho_A) = \sum_k \langle k|_A \rho_A |k\rangle_A \sigma_B^k. \tag{4.29}$$

If we spectrally decompose each density operator as:

$$\sigma_B^k = \sum_i \lambda_i^k |i\rangle_B^k \langle i|_B^k, \tag{4.30}$$

then the Kraus operators are:

$$\{\sqrt{\lambda_i^k}|i\rangle_B^k \langle k|_A\}_{i,k}. \tag{4.31}$$

The last quantum channel which we present is the depolarising channel. Its action on a density operator $\rho \in \mathcal{D}(\mathcal{H}_A)$ is as follows:

$$\mathcal{N}(\rho) = (1-p)\rho + p\pi, \tag{4.32}$$

where $\pi = \frac{1}{d}\mathbb{1}$ is the maximally mixed state with d the dimension of $\mathcal{H}$. If $\{|i\rangle\}_i$ is an orthonormal basis for $\mathcal{H}$, then $\{\sqrt{1-p}\mathbb{1}, \frac{1}{\sqrt{d}}|i\rangle\langle j|\}_{i,j}$ is a set of Kraus operators for the depolarising channel, as can be seen as follows:

$$\sqrt{1-p}\mathbb{1}\rho(\sqrt{1-p}\mathbb{1})^\dagger + \sum_{i,j} \frac{\sqrt{p}}{\sqrt{d}}|i\rangle\langle j|\rho\left(\frac{\sqrt{p}}{\sqrt{d}}|i\rangle\langle j|\right)^\dagger \tag{4.33}$$

$$= (1-p)\rho + \frac{1}{d}\sum_{i,j}|i\rangle\langle j|\rho|i\rangle\langle j| = (1-p)\rho + \frac{p}{d}\sum_i |i\rangle\mathrm{Tr}(\rho)\langle i| = (1-p)\rho + \frac{p}{d}\mathbb{1}, \tag{4.34}$$

where we have used that the trace of a density operator is 1.

4.4. Distance Measures

Now that we understand quantum channels we investigate how to quantify whether two quantum states are close to one another. The most important concept needed in order to do so is the following:

Definition 4.17. Let $\mathcal{H}_A$ and $\mathcal{H}_B$ be Hilbert spaces and let $X \in \mathcal{L}(\mathcal{H}_A, \mathcal{H}_B)$. Then we define the trace norm of X as:

$$||X||_1 = \mathrm{Tr}(|X|), \tag{4.35}$$

where $|X| = \sqrt{X^\dagger X}$.

This norm then induces a distance measure on operators.

Definition 4.18. Let $\mathcal{H}_A$ and $\mathcal{H}_B$ be Hilbert spaces and let $X, Y \in \mathcal{L}(\mathcal{H}_A, \mathcal{H}_B)$. Then we define the trace distance between X and Y to be:

$$||X - Y||_1. \tag{4.36}$$

We now present two lemmas which we will use directly in our proof of the packing lemma in the next chapter. For the proofs we refer the reader to sections 9.1 and 9.4 of [Wil19] respectively.

Lemma 4.13. Let $\mathcal{H}$ be a Hilbert space and let ρ, σ be two hermitian operators on $\mathcal{H}$. Let $\Pi \in \mathcal{L}(\mathcal{H})$ be such that $0 \leq \Pi \leq \mathbb{1}$. Then:

$$\mathrm{Tr}(\Pi\rho) \geq \mathrm{Tr}(\Pi\sigma) - ||\rho - \sigma||_1. \tag{4.37}$$

The next lemma is called the gentle operator lemma, because it concerns the case of a measurement operator which, in a precise sense, gently performs a measurement without changing the state of the system much.

Lemma 4.14. Let $\mathcal{H}$ be a Hilbert space and let $\rho \in \mathcal{D}(\mathcal{H})$. Let $\Lambda \in \mathcal{L}(\mathcal{H})$ be a measurement operator such that $1 \leq \Lambda \leq \mathbb{1}$. Now suppose that Λ has a high probability of detecting ρ in the following sense:

$$\mathrm{Tr}(\Lambda\rho) \geq 1 - \epsilon, \tag{4.38}$$

for some $\epsilon \in [0, 1]$. We then have:

$$\left|\left|\rho - \sqrt{\Lambda}\rho\sqrt{\Lambda}\right|\right|_1 \leq 2\sqrt{\epsilon}. \tag{4.39}$$

5
Quantum Shannon Theory

This chapter is focused on proving a pivotal result in quantum communication theory: the Holevo-Schumacher-Westmoreland theorem, or HSW theorem for short, proven in [SW97]. This theorem provides an expression for the ordinary capacity of a quantum channel to transmit classical information. Thus, it is a direct quantum analog of the noisy channel coding theorem. Indeed, its proof will be based on typicality, like the second proof which we presented of the noisy channel coding theorem in chapter 2. This time, however, it will be the quantum version of typical sets - called typical subspaces - that we will use.

There is, however, a complication that arises when transitioning from communication over classical channels to communication over quantum channels. This complication is the fact that measurement works so differently in the quantum world. Indeed, where in the classical case we simply received an output sequence of our classical channel which we could then decode to decide what message was sent, in the quantum case we receive an output sequence of quantum states, i.e. density operators. These density operators must first be measured according to a POVM before they can be translated into classical information. This complication of measurement is dealt with by using the packing lemma. The packing lemma tells us how we can encode classical information into an ensemble of quantum states [Wil19].

Besides this complication of measurement, there is another effect which makes quantum communication radically different from classical communication: the existence of entanglement. Indeed, we will see that it is entanglement which really sets apart quantum communication from classical communication, and because of entanglement we will be unable to characterise the classical capacity of a quantum channel as cleanly as the capacity of a classical channel. This is manifested by the fact that the HSW theorem provides an expression of the classical capacity in terms of the regularised Holevo information, whereas in the classical case such a notion of regularisation playes no role. This shows that we do not know how to approximate the classical capacity of a quantum channel.

5.1. Preliminary Notions

In this section we treat the two techniques used in our proof of the HSW theorem: the notion of the typical subspace and the packing lemma.

5.1.1. Quantum Typicality

The definition of the typical subspace is very similar to that of the typical subset. Indeed, the typical subspace shares the same three properties of the typical subset: unit probability, exponentially smaller cardinality and equipartition. However, the typical subspace is defined with respect to a certain quantum state, i.e. a density operator. The spectral decomposition of this density operator gives a probability distribution (namely the eigenvalues of the density operator), and the typical subspace is defined analogously to the typical subset with respect to this induced probability distribution [Wil19]. Let us now properly define this.

Definition 5.1. Let $\mathcal{H}_A$ be a Hilbert space with a density operator ρ_A on it. We write ρ_A in its spectral decomposition:

$$\rho_A = \sum_{x \in \mathcal{X}} p_X(x) |x\rangle\langle x|_A, \tag{5.1}$$

where $\{|x\rangle_A\}_{x\in\mathcal{X}}$ forms a complete orthonormal basis of $\mathcal{H}_A$. Then $\{|x^n\rangle\}_{x^n\in\mathcal{X}^n}$ is a complete orthonormal basis of $\mathcal{H}_{A^n} = \mathcal{H}_{A_1} \otimes ... \otimes \mathcal{H}_{A_n}$. We define the typical subspace $T^{\delta}_{A^n}$ with respect to the state $\rho_{A^n} = (\rho_A)^{\otimes n}$ as:

$$T^{\delta}_{A^n} = \text{span}\{|x^n\rangle_{A^n} \mid x^n \in T^{X^n}_{\delta}\}, \tag{5.2}$$

where the typical set $T^{X^n}_{\delta}$ is with respect to the distribution $p_X(x)$ consisting of the eigenvalues of ρ_A.

The typical subspace will play the same role in the proof of the HSW theorem as the typical set in the proof of the noisy channel coding theorem. However, in the HSW theorem we will be working with quantum states and of course, these need to be measured using a POVM. To this end, we define the so-called typical projector. It is the projector that projects onto the typical subspace, and it will play a pivotal role in the construction of a POVM in the HSW theorem.

Definition 5.2. Given a density operator ρ_A on $\mathcal{H}_A$ with corresponding typical subspace $T^{\delta}_{A^n}$ and eigenvalues $\{p_X(x)\}_{x\in\mathcal{X}}$, we define the typical projector as:

$$\Pi^{\delta}_{A^n} = \sum_{x^n \in T^{X^n}_{\delta}} |x^n\rangle\langle x^n|_{A^n}. \tag{5.3}$$

With these definitions in hand we can now state the three most important properties of typical subspaces, completely analogously to the properties of the typical set.

Property 5.1 (Unit Probability)**.** The typical subspace $T^{\delta}_{A^n}$ corresponding to $\rho_A \in \mathcal{D}(\mathcal{H})$ has unit probability in the limit where n becomes arbitrarily large. That is: for all $\delta > 0$ and $\epsilon \in (0,1)$ there exists some $n \in \mathbb{N}$ such that:

$$\text{Tr}\left(\Pi^{\delta}_{A^n}\rho_{A^n}\right) \geq 1 - \epsilon. \tag{5.4}$$

Property 5.2 (Exponentially Smaller Dimension)**.** The dimension $\dim(T^{\delta}_{A^n})$ of the typical subspace is exponentially smaller than dimension of the total space $\dim(\mathcal{H}_A)^n$:

$$\text{Tr}\left(\Pi^{\delta}_{A^n}\right) \leq 2^{n(H(A)+c\delta)}, \tag{5.5}$$

where c is some positive constant. Moreover, we can bound the dimension of the typical subspace from below: for all $\delta > 0$ and $\epsilon >\in (0,1)$ there exists some $n \in \mathbb{N}$ such that:

$$\text{Tr}\left(\Pi^{\delta}_{A^n}\right) \geq (1-\epsilon)2^{n(H(A)-c\delta)}. \tag{5.6}$$

Property 5.3 (Equipartition)**.** We have the following operator inequality:

$$2^{-n(H(A)+c\delta)}\Pi^{\delta}_{A^n} \leq \Pi^{\delta}_{A^n}\rho_{A^n}\Pi^{\delta}_{A^n} \leq 2^{-n(H(A)-c\delta)}\Pi^{\delta}_{A^n}. \tag{5.7}$$

This is a statement about the eigenvalues of the operators $\Pi^{\delta}_{A^n}\rho_{A^n}\Pi^{\delta}_{A^n}$ and $\Pi^{\delta}_{A^n}$, which have the same eigenvectors because they commute [Wil19]. We can think of the operator $\Pi^{\delta}_{A^n}\rho_{A^n}\Pi^{\delta}_{A^n}$ as the typical part of ρ_{A^n}. If we think this way, we can also write the above operator inequality as follows:

$$\forall x^n \in T^{X^n}_{\delta} : 2^{-n(H(A)+c\delta)} \leq p_{X^n}(x^n) \leq 2^{-n(H(A)-c\delta)}, \tag{5.8}$$

where the $\{p_{X^n}(x^n)\}_{x^n\in\mathcal{X}^n}$ are of course the eigenvalues of ρ_{A^n}.

We now prove the above three properties. This is actually quite simple: it just amounts to reducing to the case of the typical set.

Proof of Unit Probability: we have the following:

$$\begin{aligned}\operatorname{Tr}\left(\Pi^{\delta}_{A^n}\rho_{A^n}\right) &= \operatorname{Tr}\left(\sum_{x^n\in T^{X^n}_{\delta}}|x^n\rangle\langle x^n|_{A^n}\sum_{x^n\in\mathcal{X}^n}p_{X^n}(x^n)|x^n\rangle\langle x^n|_{A^n}\right)\\ = \operatorname{Tr}\left(\sum_{x^n\in T^{X^n}_{\delta}}p_{X^n}(x^n)|x^n\rangle\langle x^n|_{A^n}\right) &= \sum_{x^n\in T^{X^n}_{\delta}}p_{X^n}(x^n) = \Pr\left(X^n\in T^{X^n}_{\delta}\right).\end{aligned} \tag{5.9}$$

Now the result follows immediately from property 2.7. □

Proof of Exponentially Smaller Dimension: note that, by definition of the typical subspace, we have:

$$\operatorname{Tr}\left(\Pi^{\delta}_{A^n}\right) = \dim\left(T^{\delta}_{A^n}\right) = |T^{X^n}_{\delta}|. \tag{5.10}$$

Moreover, note that:

$$\begin{aligned}H(A) = -\operatorname{Tr}\left(\rho_A\log(\rho_A)\right) &= -\operatorname{Tr}\left(\sum_{x\in\mathcal{X}}p_X(x)|x\rangle\langle x|_A\log\left(\sum_{x'\in\mathcal{X}}p_X(x')|x'\rangle\langle x'|_A\right)\right)\\ &= -\operatorname{Tr}\left(\sum_{x\in\mathcal{X}}p_X(x)|x\rangle\langle x|_A\sum_{x'\in\mathcal{X}}\log(p_X(x'))|x'\rangle\langle x'|_A\right)\\ &= -\operatorname{Tr}\left(\sum_{x\in\mathcal{X}}p_X(x)\log(p_X(x))|x\rangle\langle x|_A\right) = \sum_{x\in\mathcal{X}}p_X(x)\log(p_X(x)) = H(X),\end{aligned} \tag{5.11}$$

where H(A) is the quantum entropy of the density operator ρ_A and $H(X)$ is the classical entropy of the random variable X with distribution $p_X(x)$. With these results, the statement of the property follows immediately from its classical counterpart: property 2.8. □

Proof of Equipartition: we have the following:

$$\begin{aligned}\Pi^{\delta}_{A^n}\rho_{A^n}\Pi^{\delta}_{A^n} = \sum_{x^n\in T^{X^n}_{\delta}}|x^n\rangle\langle x^n|_{A^n}\sum_{x'^n\in T^{X^n}_{\delta}}p_{X^n}(x'^n)|x'^n\rangle\langle x'^n|_{A^n}\sum_{\tilde{x}^n\in T^{X^n}_{\delta}}|\tilde{x}^n\rangle\langle\tilde{x}^n|_{A^n}\\ = \sum_{x^n\in T^{X^n}_{\delta}}p_{X^n}(x^n)|x^n\rangle\langle x^n|_{A^n}.\end{aligned} \tag{5.12}$$

Thus, we see that we can indeed write the equipartition property in its operator inequality form equivalently as:

$$\forall x^n\in T^{X^n}_{\delta}: 2^{-n(H(A)+c\delta)} \le p_{X^n}(x^n) \le 2^{-n(H(A)-c\delta)}. \tag{5.13}$$

But in the previous proof we showed that $H(A) = H(X)$, so we can again equivalently write this as:

$$\forall x^n\in T^{X^n}_{\delta}: 2^{-n(H(X)+c\delta)} \le p_{X^n}(x^n) \le 2^{-n(H(X)-c\delta)}, \tag{5.14}$$

and we know that this is true from property 2.9. □

5.1.2. Conditional Quantum Typicality

When we investigated classical typicality in chapter 2, we defined two types of typical set: the strongly typical set and the conditional strongly typical set. Both of these incarnations of typicality were necessary to prove the noisy channel coding theorem. This is no different in the proof of the

HSW theorem: we also need a conditionally typical subspace. In order to explain these concepts we follow [Wil19].

The conditionally typical subspace is defined for a certain ensemble $\{p_x(x), \rho_B^x\}$ consisting of a probability distribution on the finite set $\mathcal{X}$ and for each $x \in \mathcal{X}$ a corresponding density operator ρ_B^x on the Hilbert space $\mathcal{H}_B$. We can then write the spectral decomposition of each ρ_B^x as follows:

$$\rho_B^x = \sum_{y \in \mathcal{Y}} p_{Y|X}(y|x) |y_x\rangle\langle y_x|_B. \tag{5.15}$$

Here every $\{|y_x\rangle\}_{y\in\mathcal{Y}}$ is an orthonormal basis of B. Although it might not seem so at first sight, we have done nothing new. We have simply written the eigenvalues of each density operator as a conditional probability distribution $p_{Y|X}(y|x)$, conditioned on $x \in \mathcal{X}$. Moreover, we define the classical-quantum state:

$$\rho_{XB} = \sum_{x \in \mathcal{X}} p_X(x) |x\rangle\langle x|_X \otimes \rho_B^x, \tag{5.16}$$

such that the conditional entropy of this state becomes:

$$H(B|X)_\rho = \sum_{x \in \mathcal{X}} p_X(x) H(\rho_B^x). \tag{5.17}$$

With these notations in place we define the conditionally typical subspace.

Definition 5.3. The conditionally typical subspace corresponding to a stronlgy typical sequence $x^n \in \mathcal{X}^n$ and to an ensemble $\{p_X(x), \rho_B^x\}$ whose density operators have eigenvalues $p_{Y|X}(y|x)$ and corresponding eigenvectors $|y_x\rangle_B$ is:

$$T^\delta_{B^n|x^n} = \text{span}\left(\bigotimes_{x\in\mathcal{X}} |y_x^{I_x}\rangle_{B^{I_x}} : \forall x \in \mathcal{X} : y^{I_x} \in T_\delta^{Y^{|I_x|}|x^{|I_x|}} \right). \tag{5.18}$$

Here $I_x = \{i \mid x_i = x\}$ is an indicator set which selects the indices i for which $x_i = x$. Similarly, B^{I_x} selects the subsystems of $B^n = B_1 \otimes ... \otimes B_n$ for which the entry of x^n is equal to x. Moreover, $|y_x^{I_x}\rangle$ is some string of states from $\{|y_x\rangle\}$ and y^{I_x} is the corresponding classical sequence. Lastly, $T_\delta^{Y^{|I_x|}|x^{|I_x|}}$ is just a conditional strongly typical set as defined in definition 2.13.

We also define the conditionally typical subspace projector.

Definition 5.4. The conditionally typical subspace projector corresponding to a strongly typical sequence x^n and an ensemble $\{p_X(x), \rho_B^x\}$ is:

$$\Pi^\delta_{B^n|x^n} = \bigotimes_{x\in\mathcal{X}} \Pi^{\rho_x,\delta}_{B^{I_x}}, \tag{5.19}$$

where $\Pi^{\rho_x,\delta}_{B^{I_x}}$ is simply the subspace projector corresponding to the state ρ_B^x, and where B^{I_x} again selects the subsystems of B^n onto which the typical subspace projector for ρ_B^x projects.

The conditionally typical subspace again satisfies the same three properties that we have come across several times before. However, the proofs are more involved this time. We will now state and then proof each of them.

Property 5.4 (Unit Probability)**.** For all $\delta > 0$ and $\epsilon \in (0,1)$ there exists some $n \in \mathbb{N}$ such that the probability that we measure a state $\rho_{B^n}^{x^n}$ to be in the conditionally typical subspace $T^\delta_{B^n|x^n}$ satisfies:

$$\text{Tr}\left(\Pi^\delta_{B^n|x^n} \rho^{x^n}_{B^n}\right) \geq 1 - \epsilon. \tag{5.20}$$

Property 5.5 (Exponentially Smaller Dimension)**.** The dimension $\dim(T^{\delta}_{B^n|x^n})$ of the conditionally typical subspace is exponentially smaller than dimension of the total space $\dim(\mathcal{H}_B)^n$:

$$\mathrm{Tr}\left(\Pi^{\delta}_{B^n|x^n}\right) \leq 2^{n(H(B|X)+\delta'')}, \tag{5.21}$$

where δ'' is a positive constant. Moreover, we can bound the dimension of the conditionally typical subspace from below: for all $\delta > 0$ and $\epsilon >\in (0,1)$ there exists some $n \in \mathbb{N}$ such that:

$$\mathrm{Tr}\left(\Pi^{\delta}_{B^n|x^n}\right) \geq (1-\epsilon)2^{n(H(B|X)-\delta'')}. \tag{5.22}$$

Property 5.6 (Equipartition)**.** We have the following operator inequality:

$$2^{-n(H(B|X)+\delta'')}\Pi^{\delta}_{B^n|x^n} \leq \Pi^{\delta}_{B^n|x^n}\rho^{x^n}_{B^n}\Pi^{\delta}_{B^n|x^n} \leq 2^{-n(H(B|X)-\delta'')}\Pi^{\delta}_{B^n|x^n}. \tag{5.23}$$

We now prove the unit probability and equipartition properties. The exponentially smaller dimension property follows from the equipartition property the same way it did in the classical case.

Proof of Unit Probability: like in the case of the conditionally typical set we denote $\mathcal{X} = \{a_1, ..., a_{|\mathcal{X}|}\}$ so that we can lexicographically order a quantum state ρ_{x^n}:

$$\rho_{x^n} = \underbrace{\rho_{a_1} \otimes \cdots \otimes \rho_{a_1}}_{N(a_1|x^n)} \otimes \cdots \otimes \underbrace{\rho_{a_{|\mathcal{X}|}} \otimes \cdots \otimes \rho_{a_{|\mathcal{X}|}}}_{N(a_{|\mathcal{X}|}|x^n)}. \tag{5.24}$$

The conditionally strong typical projector can then be written as:

$$\Pi^{\delta}_{B^n|x^n} = \bigotimes_{x\in\mathcal{X}} \Pi^{\rho_x,\delta}_{B^{N(x|x^n)}}. \tag{5.25}$$

Now we use property 5.1 to conclude that for every $\delta > 0$ and $\epsilon \in (0,1)$ there exists an $n \in \mathbb{N}$ such that:

$$\begin{aligned}\mathrm{Tr}\left(\Pi^{\delta}_{B^n|x^n}\rho^{x^n}_{B^n}\right) = \mathrm{Tr}\left(\bigotimes_{x\in\mathcal{X}}\left(\Pi^{\rho_x,\delta}_{B^{N(x|x^n)}}\rho_x^{\otimes N(x|x^n)}\right)\right) &= \prod_{x\in\mathcal{X}} \mathrm{Tr}\left(\Pi^{\rho_x,\delta}_{B^{N(x|x^n)}}\rho_x^{\otimes N(x|x^n)}\right) \\ &\geq (1-\epsilon)^{|\mathcal{X}|} \geq 1-|\mathcal{X}|\epsilon \geq 1-\epsilon,\end{aligned} \tag{5.26}$$

which is the desired result. □

Proof of Equipartition: we again write $\rho^{x^n}_{B^n}$ in lexicographical order, such that we can write:

$$\Pi^{\delta}_{B^n|x^n}\rho^{x^n}_{B^n}\Pi^{\delta}_{B^n|x^n} = \bigotimes_{x\in\mathcal{X}} \Pi^{\rho_x,\delta}_{B^{N(x|x^n)}}\rho_x^{\otimes N(x|x^n)}\Pi^{\rho_x,\delta}_{B^{N(x|x^n)}}. \tag{5.27}$$

The equipartition property for each typical subspace projector individually (property 5.3) then gives:

$$\bigotimes_{x\in\mathcal{X}} \Pi^{\rho_x,\delta}_{B^{N(x|x^n)}}2^{-N(x|x^n)(H(\rho_x)+c\delta)} \leq \bigotimes_{x\in\mathcal{X}} \Pi^{\rho_x,\delta}_{B^{N(x|x^n)}}\rho_x^{\otimes N(x|x^n)}\Pi^{\rho_x,\delta}_{B^{N(x|x^n)}} \tag{5.28}$$

$$\leq \bigotimes_{x\in\mathcal{X}} \Pi^{\rho_x,\delta}_{B^{N(x|x^n)}}2^{-N(x|x^n)(H(\rho_x)-c\delta)}. \tag{5.29}$$

Now x^n is strongly typical, so $|\frac{N(x|x^n)}{n} - p_X(x)| < \delta'$ for some $\delta' > 0$. The above equation thus implies:

$$\bigotimes_{x\in\mathcal{X}} \Pi^{\rho_x,\delta}_{B^{N(x|x^n)}}2^{-n(p_X(x)+\delta')(H(\rho_x)+c\delta)} \leq \bigotimes_{x\in\mathcal{X}} \Pi^{\rho_x,\delta}_{B^{N(x|x^n)}}\rho_x^{\otimes N(x|x^n)}\Pi^{\rho_x,\delta}_{B^{N(x|x^n)}} \tag{5.30}$$

$$\leq \bigotimes_{x\in\mathcal{X}} \Pi^{\rho_x,\delta}_{B^{-n(p_X(x)-\delta')}}2^{-n(p_X(x)+\delta')(H(\rho_x)-c\delta)}. \tag{5.31}$$

If we now take the exponentials out of the tensor products and put them in front, we get:

$$\prod_{x\in\mathcal{X}} 2^{-n(p_X(x)+\delta')(H(\rho_x)+c\delta)} \bigotimes_{x\in\mathcal{X}} \Pi^{\rho_x,\delta}_{B^{N(x|x^n)}} \leq \bigotimes_{x\in\mathcal{X}} \Pi^{\rho_x,\delta}_{B^{N(x|x^n)}} \rho_x^{\otimes N(x|x^n)} \Pi^{\rho_x,\delta}_{B^{N(x|x^n)}} \tag{5.32}$$

$$\leq \prod_{x\in\mathcal{X}} 2^{-n(p_X(x)+\delta')(H(\rho_x)-c\delta)} \bigotimes_{x\in\mathcal{X}} \Pi^{\rho_x,\delta}_{B^{N(x|x^n)}}. \tag{5.33}$$

Now using $\Pi^{\delta}_{B^n|x^n} = \bigotimes_{x\in\mathcal{X}} \Pi^{\rho_x,\delta}_{B^{N(x|x^n)}}$ we can write this as:

$$\prod_{x\in\mathcal{X}} 2^{-n(p_X(x)+\delta')(H(\rho_x)+c\delta)} \Pi^{\delta}_{B^n|x^n} \leq \Pi^{\delta}_{B^n|x^n} \rho_{x^n} \Pi^{\delta}_{B^n|x^n} \tag{5.34}$$

$$\leq \prod_{x\in\mathcal{X}} 2^{-n(p_X(x)+\delta')(H(\rho_x)-c\delta)} \Pi^{\delta}_{B^n|x^n}. \tag{5.35}$$

Distributing over the brackets in the exponential and writing out the products gives:

$$\begin{aligned} 2^{-n\big(H(B|X)+\sum_x(H(\rho_x\delta'+cp_X(x)\delta+c\delta\delta'))\big)} \Pi^{\delta}_{B^n|x^n} &\leq \Pi^{\delta}_{B^n|x^n} \rho_{x^n} \Pi^{\delta}_{B^n|x^n} \\ &\leq 2^{-n\big(H(B|X)+\sum_x(-H(\rho_x\delta'-cp_X(x)\delta+c\delta\delta'))\big)} \Pi^{\delta}_{B^n|x^n}. \end{aligned} \tag{5.36}$$

Since $\sum_x p_X(x) = 1$ and $\sum_x H(\rho_x) \leq |\mathcal{X}|\log(d)$ where d is the dimension of ρ_x, we can write this as:

$$\begin{aligned} 2^{-n\big(H(B|X)+\delta'|\mathcal{X}|\log(d)+c\delta+|\mathcal{X}|c\delta\delta'\big)} \Pi^{\delta}_{B^n|x^n} &\leq \Pi^{\delta}_{B^n|x^n} \rho_{x^n} \Pi^{\delta}_{B^n|x^n} \\ &\leq 2^{-n\big(H(B|X)-\delta'|\mathcal{X}|\log(d)+c\delta+|\mathcal{X}|c\delta\delta'\big)} \Pi^{\delta}_{B^n|x^n}. \end{aligned} \tag{5.37}$$

Or, by defining $\delta'' = \delta'|\mathcal{X}|\log(d) + c\delta + |\mathcal{X}|c\delta\delta'$:

$$2^{-n\big(H(B|X)+\delta''\big)} \Pi^{\delta}_{B^n|x^n} \leq \Pi^{\delta}_{B^n|x^n} \rho_{x^n} \Pi^{\delta}_{B^n|x^n} \leq 2^{-n\big(H(B|X)-\delta''\big)} \Pi^{\delta}_{B^n|x^n}. \tag{5.38}$$

This is exactly the statement of the equipartition property. □

5.1.3. Packing Lemma

The second ingredient - besides typical subspaces - of our proof of the HSW theorem is the so-called packing lemma. It is called so because it tells us how many classical messages we can pack into a Hilbert space. We will again follow [Wil19] for the exposition of the packing lemma and its proof.

Theorem 5.7 (Packing Lemma)**.** Let $\{p_X(x), \sigma_x\}_{x\in\mathcal{X}}$ be an ensemble, where each $\sigma_x \in \mathcal{D}(\mathcal{H})$. We denote the expected density operator of the ensemble by $\sigma = \sum_x p_X(x)\sigma_x$. Suppose we have a projector Π and a set of projectors $\{\Pi_x\}_{x\in\mathcal{X}}$ which project onto subspaces of $\mathcal{H}$. We call these projectors the code subspace projector and codeword subspace projectors respectively. Now suppose the following conditions are satisfied for all $x \in \mathcal{X}$:

$$\mathrm{Tr}(\Pi\sigma_x) \geq 1-\epsilon, \tag{5.39}$$

$$\mathrm{Tr}(\Pi_x\sigma_x) \geq 1-\epsilon, \tag{5.40}$$

$$\mathrm{Tr}(\Pi_x) \leq d, \tag{5.41}$$

$$\Pi\sigma\Pi \leq \frac{1}{D}\Pi, \tag{5.42}$$

where $\epsilon \in (0,1), D > 0$ and $d \in (0,D)$. Now let $\mathcal{M} = \{1,\ldots,|\mathcal{M}|\}$ be a finite set. We now generate a code $\mathcal{C} = \{C_m\}_{m\in\mathcal{M}}$, consisting of a set of random variables which each take values in $\mathcal{X}$ according to the distribution $p_X(x)$. Then there exsists a POVM $\{\Lambda_m\}_{m\in\mathcal{M}}$ such that:

$$\mathbb{E}_{\mathcal{C}}\left(\frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \mathrm{Tr}\left(\Lambda_m\sigma_{C_m}\right)\right) \geq 1-2(\epsilon-2\sqrt{\epsilon})-4|\mathcal{M}|\frac{d}{D}. \tag{5.43}$$

Proof: as explained in the definition, we randomly generate a code as a set of random variables taking values in $\mathcal{X}$. So $\mathcal{C} = \{c_m\}_{m\in\mathcal{M}}$ is a realisation of such a code. The probability that a particular code is generated is given by:

$$p(\mathcal{C}) = \prod_{m\in\mathcal{M}} p_X(c_m). \tag{5.44}$$

We now define the following operators:

$$\Upsilon_x = \Pi\Pi_x\Pi. \tag{5.45}$$

Using these operators and given a code $\mathcal{C} = \{c_m\}_{m\in\mathcal{M}}$, we construct a so-called square-root measurement:

$$\Lambda_m = \left(\sum_{m'\in\mathcal{M}} \Upsilon_{m'}\right)^{-\frac{1}{2}} \Upsilon_{c_m} \left(\sum_{m'\in\mathcal{M}} \Upsilon_{m'}\right)^{-\frac{1}{2}}. \tag{5.46}$$

We might have $\sum_{m\in\mathcal{M}} \Lambda_m < \mathbb{1}$, so we add the operator $\Lambda_0 = \mathbb{1} - \sum_{m\in\mathcal{M}} \Lambda_m$ so that $\{\Lambda_0\} \cup \{\Lambda_m\}_{m\in\mathcal{M}}$ becomes a POVM. The intuition behind all this is that the elements Λ_m are used to distinguish a particular message m, whereas Λ_0 corresponds to an error.

Now that we have our code generation and POVM we are able to start the error analysis. To this end, we use a result by Hayashi and Nagaoka, the proof of which can be found in appendix A.

Lemma 5.8 (Hayashi-Nagaoka)**.** Let $S, T \in \mathcal{L}(\mathcal{H})$ be positive semi-definite operators such that $\mathbb{1} - S$ is also postive semi-definite. Then for any $c > 0$ we have:

$$\mathbb{1} - (S+T)^{-\frac{1}{2}} S (S+T)^{-\frac{1}{2}} \leq (1+c)(\mathbb{1} - S) + (2 + c + \frac{1}{c})T. \tag{5.47}$$

Suppose we have generated a particular code $\mathcal{C}$. If we take $S = \Upsilon_{c_m}$ and $T = \sum_{m'\neq m} \Upsilon_{c_{m'}}$, then with $c = 1$ the lemma gives:

$$\mathbb{1} - \left(\sum_{m'} \Upsilon_{c_{m'}}\right)^{-\frac{1}{2}} \Upsilon_{c_m} \left(\sum_{m'} \Upsilon_{c_{m'}}\right)^{-\frac{1}{2}} \leq (1+1)(\mathbb{1} - \Upsilon_{c_m}) + (2+1+\frac{1}{1}) \sum_{m'\neq m} \Upsilon_{c_{m'}}. \tag{5.48}$$

Or by using our definition of Λ_m:

$$\mathbb{1} - \Lambda_m \leq 2(\mathbb{1} - \Upsilon_{c_m}) + 4 \sum_{m'\neq m} \Upsilon_{c_{m'}}. \tag{5.49}$$

If a particular message m is sent, then the probability that it is decoded incorrectly is given by:

$$p_e(m,\mathcal{C}) = \mathrm{Tr}\left((\mathbb{1} - \Lambda_m)\sigma_{c_m}\right). \tag{5.50}$$

Combining the above two results gives:

$$\begin{aligned} p_e(m,\mathcal{C}) &\leq \mathrm{Tr}\left(\left((2(\mathbb{1} - \Upsilon_{c_m}) + 4 \sum_{m'\neq m} \Upsilon_{c_{m'}}\right)\sigma_{c_m}\right) \\ &= 2\mathrm{Tr}\left((\mathbb{1} - \Upsilon_{c_m})\sigma_{c_m}\right) + 4 \sum_{m'\neq m} \mathrm{Tr}\left(\Upsilon_{c_{m'}}\sigma_{c_m}\right). \end{aligned} \tag{5.51}$$

Using the gentle operator lemma 4.14 and lemma 4.13 we can derive the following:

$$\begin{aligned} \mathrm{Tr}\left(\Upsilon_{c_m}\sigma_{c_m}\right) &= \mathrm{Tr}\left(\Pi\Pi_{c_m}\Pi\sigma_{c_m}\right) = \mathrm{Tr}\left(\Pi_{c_m}\Pi\sigma_{c_m}\Pi\right) \\ &\geq \mathrm{Tr}\left(\Pi_{c_m}\sigma_{c_m}\right) - \left|\left|\Pi\sigma_{c_m}\Pi - \sigma_{c_m}\right|\right|_1 \geq 1 - \epsilon - 2\sqrt{\epsilon}, \end{aligned} \tag{5.52}$$

and thus we get:

$$\mathrm{Tr}\left((\mathbb{1} - \Upsilon_{c_m})\sigma_{c_m}\right) = 1 - \mathrm{Tr}\left(\Upsilon_{c_m}\sigma_{c_m}\right) \leq \epsilon + 2\sqrt{\epsilon}. \tag{5.53}$$

For the probability of error we now have:

$$p_e(m,\mathcal{C}) \leq 2(\epsilon + 2\sqrt{\epsilon}) + 4 \sum_{m' \neq m} \mathrm{Tr}\left(\Upsilon_{c_{m'}} \sigma_{c_m}\right). \tag{5.54}$$

Assuming that the message to be sent is selected randomly and uniformly, we get the following average probability of error for a certain code $\mathcal{C}$:

$$\bar{p}_e(\mathcal{C}) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} p_e(m,\mathcal{C}) \leq 2(\epsilon + 2\sqrt{\epsilon}) + \frac{4}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{m' \neq m} \mathrm{Tr}\left(\Upsilon_{c_{m'}} \sigma_{c_m}\right). \tag{5.55}$$

In Shannon like fashion, we will now consider the expectation over all randomly generated codes:

$$\begin{aligned} \mathbb{E}_{\mathcal{C}}(\bar{p}_e(\mathcal{C})) &\leq \mathbb{E}_{\mathcal{C}} \left(2(\epsilon + 2\sqrt{\epsilon}) + \frac{4}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{m' \neq m} \mathrm{Tr}\left(\Upsilon_{c_{m'}} \sigma_{c_m}\right) \right) \\ &= 2(\epsilon + 2\sqrt{\epsilon}) + \frac{4}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{m' \neq m} \mathbb{E}_{\mathcal{C}}\left(\mathrm{Tr}\left(\Upsilon_{c_{m'}} \sigma_{c_m}\right)\right). \end{aligned} \tag{5.56}$$

Our task now amounts to bounding $\mathbb{E}_{\mathcal{C}}\left(\mathrm{Tr}\left(\Upsilon_{c_{m'}} \sigma_{c_m}\right)\right)$. To do so, we first note that this quantity depends only on the specific codewords C_m and $C_{m'}$. Thus, we find:

$$\mathbb{E}_{\mathcal{C}}\left(\mathrm{Tr}\left(\Upsilon_{c_{m'}} \sigma_{c_m}\right)\right) = \mathbb{E}_{\mathcal{C}}\left(\mathrm{Tr}\left(\Pi\Pi_{C_{m'}}\Pi\sigma_{C_m}\right)\right) \mathbb{E}_{\mathcal{C}}\left(\mathrm{Tr}\left(\Pi_{C_{m'}}\Pi\sigma_{C_m}\Pi\right)\right) \tag{5.57}$$

$$= \mathbb{E}_{C_{m'}}\mathbb{E}_{C_m}\left(\mathrm{Tr}\left(\Pi_{C_{m'}}\Pi\sigma_{C_m}\Pi\right)\right) = \mathrm{Tr}\left(\mathbb{E}_{C_{m'}}\left(\Pi_{C_{m'}}\right)\Pi\mathbb{E}_{C_m}\left(\sigma_{C_m}\right)\Pi\right) \tag{5.58}$$

$$\leq \mathrm{Tr}\left(\mathbb{E}_{C_{m'}}\left(\Pi_{C_{m'}}\right)\frac{1}{D}\Pi\right) = \frac{1}{D}\mathrm{Tr}\left(\mathbb{E}_{C_{m'}}\left(\Pi_{C_{m'}}\right)\Pi\right) \leq \frac{1}{D}\mathrm{Tr}\left(\mathbb{E}_{C_{m'}}\left(\Pi_{C_{m'}}\right)\right) \tag{5.59}$$

$$= \frac{1}{D}\mathbb{E}_{C_{m'}}\left(\mathrm{Tr}\left(\Pi_{C_{m'}}\right)\right) \leq \frac{1}{D}\mathbb{E}_{C_{m'}}\left(\mathrm{Tr}\left(d\right)\right) = \frac{d}{D}. \tag{5.60}$$

Thus, for the expectation of the average error we get:

$$\bar{p}_e(\mathcal{C}) \leq 2(\epsilon + 2\sqrt{\epsilon}) + \frac{4}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \sum_{m' \neq m} \frac{d}{D} \leq 2(\epsilon + 2\sqrt{\epsilon}) + 4|\mathcal{M}|\frac{d}{D}, \tag{5.61}$$

which is the statement of the packing lemma. □

To apply the packing lemma in the proof of the HSW theorem, however, we do not wish to have a statement about the expectation of the average error over all codes. Instead, we wish to have a similar statement but then about the existence of one code. Thus, we perform the same expurgation argument as we did in the proof of the noisy channel coding theorem in section 2.3: we first note that there must exist at least one code satisfying the same error bound as in the packing lemma and we then again omit the worse half of the codewords. This doubles the error bound and leaves half the number of messages and leads to the following corollary, which will play a pivotal role in our upcoming proof of the HSW theorem:

Corollary 5.9. Let $\{p_X(x), \sigma_x\}_{x \in \mathcal{X}}$ be an ensemble, where each $\sigma_x \in \mathcal{D}(\mathcal{H})$. We denote the expected density operator of the ensemble by $\sigma = \sum_x p_X(x)\sigma_x$. Suppose we have a projector Π and a set of projectors $\{\Pi_x\}_{x \in \mathcal{X}}$ which project onto subspaces of $\mathcal{H}$. We call these projectors the code subspace projector and codeword subspace projectors respectively. Now suppose the following conditions are satisfied:

$$\begin{aligned} \mathrm{Tr}(\Pi\sigma_x) &\geq 1 - \epsilon, \\ \mathrm{Tr}(\Pi_x\sigma_x) &\geq 1 - \epsilon, \\ \mathrm{Tr}(\Pi_x) &\leq d, \\ \Pi\sigma\Pi &\leq \frac{1}{D}\Pi, \end{aligned} \tag{5.62}$$

where $\epsilon \in (0,1), D > 0$ and $d \in (0, D)$. Then there exist a code $\mathcal{C} = \{c_m\}_{m\in\mathcal{M}}$ with codewords taking values in $\mathcal{X}$ and a POVM $\{\Lambda_m\}_{m\in\mathcal{M}}$ such that for all $m \in \mathcal{M}$:

$$\mathrm{Tr}(\Lambda_m \sigma_{c_m}) \geq 1 - 4(\epsilon + 2\sqrt{\epsilon}) - 16|\mathcal{M}|\frac{d}{D}. \tag{5.63}$$

5.2. Holevo-Schumacher-Westmoreland Theorem

We are now finally in a position to state and prove one of the highlights of this thesis: the Holevo-Schumacher-Westmoreland theorem. It is the quantum analog of the noisy channel coding theorem, but there is one main difference in its statement: the appearance of the so-called regularisation of the Holevo information. This regularisation does not appear in the noisy channel coding theorem, and it is due to the fact that when dealing with quantum channels, we are also dealing with entanglement. In this section we will again roughly follow [Wil19].

5.2.1. Statement of the Theorem

Let us now define the Holevo information:

Definition 5.5. The Holevo information $\chi(\mathcal{N})$ of a quantum channel $\mathcal{N} : \mathcal{L}(\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B)$ is:

$$\chi(\mathcal{N}) = \max_{\rho_{XB}} I(X;B)_\rho, \tag{5.64}$$

where ρ_{XB} is a classical-quantum state of the form:

$$\rho_{XB} = \sum_{x\in\mathcal{X}} p_X(x)|x\rangle\langle x|_X \otimes \mathcal{N}(\rho_A^x), \tag{5.65}$$

where $\rho_A^x \in \mathcal{D}(\mathcal{H}_A)$.

The regularisation of the Holevo information is then:

$$\chi_{\text{reg}}(\mathcal{N}) = \lim_{k\to\infty} \frac{1}{k}\chi(\mathcal{N}^{\otimes k}). \tag{5.66}$$

We will see shortly that this regularised Holevo information is the highest achievable rate for a quantum channel. Analogously to the case of a classical noisy channel, a code for a quantum channel $\mathcal{N} : \mathcal{L}(\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B)$ consists of a finite message set $\mathcal{M}$, a set of $|\mathcal{M}|$ quantum states $\rho_{A^n}^m$ which can be put into n uses of the channel (similar to an encoder) and a POVM $\{\Lambda_m\}_{m\in\mathcal{M}}$ (similar to a decoder). We again think of two friends Alice and Bob who want to communicate through the channel. If Alice sends message m, then the probability that Bob correctly decodes it is:

$$\mathrm{Tr}(\Lambda_m \mathcal{N}^{\otimes n}(\rho_{A^n}^m)). \tag{5.67}$$

Thus, the probability of error when a particular message m is sent is given by:

$$p_e(m) = \mathrm{Tr}((\mathbb{1} - \Lambda_m)\mathcal{N}^{\otimes n}(\rho_{A^n}^m)). \tag{5.68}$$

If the maximum probability of error $\max_{m\in\mathcal{M}} p_e(m)$ is no greater than ϵ, and if the rate of the code is $R = \frac{|\mathcal{M}|}{n}$ then we again say that we have a (n, R, ϵ) code. We now state the HSW theorem:

Theorem 5.10 (Holevo-Schumacher-Westmoreland)**.** We define the capacity of a quantum channel $\mathcal{N}$ as:

$$C(\mathcal{N}) = \chi_{\text{reg}}(\mathcal{N}). \tag{5.69}$$

Then any rate $R < C$ is achievable, that is: for any $\epsilon \in (0,1)$ there exists some $n \in \mathbb{N}$ such that there is an $(n, \tilde{R}, \epsilon)$ code, where $\tilde{R} \geq R$. Moreover, for any $R > C$ there exists some $\epsilon \in (0,1)$ such that there is no $n \in \mathbb{N}$ for which an $(n, \tilde{R}, \epsilon)$ code exists with $\tilde{R} \geq R$.

5.2.2. Direct Coding Part

We will now prove the direct coding part (which states that rates below the capacity are achievable) using typical subspaces and the packing lemma. Let $\{p_X(x), \rho_A^x\}_{x\in\mathcal{X}}$ be any ensemble. We will show that $I(X;B)_\rho$ with $\rho_{XB} = \sum_x p_X(x)|x\rangle\langle x|_X \otimes \mathcal{N}(\rho_A^x)$ is an achievable rate. Since we picked a random ensemble, it will then follow that the Holevo information of the channel can be achieved.

First of all, Alice selects $|\mathcal{M}|$ codewords $\{x^n(m)\}_{m\in\mathcal{M}}$ randomly and independently according to the following distribution:

$$p_{X'^n}(x^n) = \begin{cases} p_{X^n}(x^n)\left(\sum_{x^n\in T_\delta^{X^n}} p_{X^n}(x^n)\right)^{-1} & \text{if } x^n \in T_{X^n}^\delta \\ 0 & \text{else} \end{cases}. \tag{5.70}$$

These classical codewords $x^n(m) = x_1(m)\cdots x_n(m)$ give rise to quantum codewords:

$$\rho_{A^n}^{x^n(m)} = \rho_A^{x_1(m)} \otimes \cdots \otimes \rho_A^{x_n(m)}. \tag{5.71}$$

When Alice puts such a quantum codeword into the channel the output is:

$$\sigma_{B^n}^{x^n(m)} = \sigma_B^{x_1(m)} \otimes \cdots \otimes \sigma_B^{x_n(m)} = \mathcal{N}(\rho_A^{x_1(m)}) \otimes \cdots \otimes \mathcal{N}(\rho_A^{x_n(m)}). \tag{5.72}$$

We will now use the derandomised version of the packing lemma (corollary 5.9) and in order to do so we need several objects satisfying the correct properties. The first of these objects will be the ensemble $\{p'_{X'^n}(x^n), \sigma_{B^n}^{x^n}\}_{x^n\in\mathcal{X}^n}$ deduced as explained above from our original ensemble $\{p_X(x), \rho_A^x\}_{x\in\mathcal{X}}$. The second object is the expected density operator of this ensemble:

$$\mathbb{E}_{X'^n}(\sigma^{X'^n}) = \sum_{x^n\in\mathcal{X}^n} p'_{X'^n}(x^n)\sigma_{B^n}^{x^n}. \tag{5.73}$$

The last two objects we need are a code subspace projector and a set of codeword subspace projects for each message. We take the code subspace projector to be the typical subspace projector $\Pi_{B^n}^\delta$ for the state $\sigma^{\otimes n}$, where:

$$\sigma_{B^n} = \sum_{x\in\mathcal{X}} p_X(x)\sigma_B^x = \sum_{x\in\mathcal{X}} p_X(x)\mathcal{N}(\rho_A^x). \tag{5.74}$$

Lastly, we take the codeword subspace projectors to be the conditionally typical subspace projectors $\Pi_{B^n|x^n}^\delta$ for the ensemble $\{p_X(x_i), \sigma^{x_i}\}_{1\le i\le|\mathcal{M}|}$ belonging to the codeword $\sigma_{B^n}^{x^n}$. Now, the first three properties needed to use the packing lemma are immediately satisfied by the properties of typical subspaces: for every $\delta > 0$ and $\epsilon \in (0,1)$ there exists an $n\in\mathbb{N}$ such that:

$$\mathrm{Tr}\left(\Pi_{B^n}^\delta \sigma_{B^n}^{x^n}\right) \ge 1-\epsilon, \tag{5.75}$$

$$\mathrm{Tr}\left(\Pi_{B^n|x^n}^\delta \sigma_{B^n}^{x^n}\right) \ge 1-\epsilon, \tag{5.76}$$

$$\mathrm{Tr}\left(\Pi_{B^n|x^n}\right) \le 2^{n(H(B|X)+c\delta)}. \tag{5.77}$$

In this last property we have used the constant c to absorb the δ'' in property 5.5. Now all we need to do is show that the last property needed for the packing lemma holds. Indeed, by using that $\mathbb{E}_{X'^n}\left(\sigma_{B^n}^{X'^n}\right) \le \frac{1}{1-\epsilon}\sigma_{B^n}$ and the equipartition property of the typical subspace projector $\Pi_{B^n}^\delta$ we find that:

$$\Pi_{B^n}^\delta \mathbb{E}_{X'^n}\left(\sigma_{B^n}^{X'^n}\right)\Pi_{B^n}^\delta \le \frac{1}{1-\epsilon}2^{-n(H(B)-c'\delta)}\Pi_{B^n}^\delta, \tag{5.78}$$

where c' is also a positive constant. Thus, all conditions necessary to use the derandomised packing lemma (corollary 5.9) are satisfied, and we find that the maximal probability of error p_e^* satisfies:

$$p_e^* = \max_m \mathrm{Tr}\left((\mathbb{1} - \Lambda_m)\mathcal{N}^{\otimes n}(\rho^{x^n(m)})\right) \tag{5.79}$$

$$\le 4(\epsilon + 2\sqrt{\epsilon}) + \frac{16}{1-\epsilon}2^{-n(H(B)-H(B|X)-(c+c')\delta)}|\mathcal{M}| \tag{5.80}$$

$$\le 4(\epsilon + 2\sqrt{\epsilon}) + \frac{16}{1-\epsilon}2^{-n(I(X;B)-(c+c')\delta)}|\mathcal{M}|. \tag{5.81}$$

We see that if we choose $|\mathcal{M}| = 2^{n(I(X;B)-(c+c'+1)\delta)}$ we get a rate of:

$$\frac{1}{n}\log(|\mathcal{M}|) = I(X;B) - (c+c'+1)\delta, \tag{5.82}$$

and a maximal probability of error of:

$$p_e^* \leq 4(\epsilon + 2\sqrt{\epsilon}) + \frac{16}{1-\epsilon}2^{-n\delta}, \tag{5.83}$$

which can be made arbitrarily small when n can become arbitrarily large. Thus, we have seen that the Holevo information $I(X;B)_\rho$ with respect to the following classical-quantum state:

$$\rho_{XB} = \sum_{x\in\mathcal{X}} p_X(x)|x\rangle\langle x|_X \otimes \mathcal{N}(\rho^x), \tag{5.84}$$

is an achievable rate [Wil19]. If we now instead use a code for the tensor product channel $\mathcal{N}^{\otimes k}$ then the rate $\frac{1}{k}\chi(\mathcal{N}^{\otimes k})$ can be achieved. Thus, by taking k arbitrarily large, the regularised Holevo information $\chi_{\text{reg}}(\mathcal{N})$ can be achieved [Wil19]. This concludes the proof of the direct coding part. □

We now sketch the proof of converse part of the HSW theorem, which states that rates above the regularised Holevo information are not achievable. The details can be found in subsection 20.3.2 of [Wil19]. The idea is to consider a different kind of task than classical communication over a quantum channel $\mathcal{N} : \mathcal{L}(\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B)$. This different task is called randomness distribution.

The protocol for randomness distribution is as follows: Alice has a message set $\mathcal{M}$ and uniformly randomly generates a classical state of the following form:

$$\Phi = \sum_{m\in\mathcal{M}} \frac{1}{|\mathcal{M}|}|m\rangle\langle m|_M \otimes |m\rangle\langle m|_{M'}. \tag{5.85}$$

She then encodes half of the state into quantum codewords, giving her the state:

$$\sum_{m\in\mathcal{M}} \frac{1}{|\mathcal{M}|}|m\rangle\langle m|_M \otimes \rho^m_{A'^n}, \tag{5.86}$$

and transmits these quantum codewords over n uses of the channel, giving:

$$\sum_{m\in\mathcal{M}} \frac{1}{|\mathcal{M}|}|m\rangle\langle m|_M \otimes \mathcal{N}^{\otimes n}(\rho^m_{A'^n}). \tag{5.87}$$

Bob then performs a POVM $\{\Lambda_m\}$ on the system that he received through the channel. If the state

$$\omega_{MM'} = \sum_{m,m'\in\mathcal{M}} \frac{1}{|\mathcal{M}|} \operatorname{Tr}\left\{\Lambda_{m'}\mathcal{N}^{\otimes n}\left(\rho^m_{A'^n}\right)\right\}|m\rangle\left\langle m\right|_M \otimes \left| m'\right\rangle\left\langle m'\right|_{M'} \tag{5.88}$$

is then close to the orginial state in the following sense:

$$\frac{1}{2}||\Phi_{MM'} - \omega_{MM'}||_1 \leq \epsilon, \tag{5.89}$$

then we call this protocol an (N,R,ϵ) protocol, where $R = \frac{1}{n}\log(|\mathcal{M}|)$. In an analogous fashion to the classical capacity, the capacity for randomness distribution is then the supremum of achievable rates. As shown in [Wil19], the capacity for randomness distribution is at least as great as the classical capacity of a quantum channel. One can show that the regularised Holevo information bounds the capacity for randomness distribution from above, and must therefore also bound the classical capacity from above. Since we showed in the direct coding part that the regularised Holevo information bounds the classical capacity from below, we can thus conclude that the classical capacity is equal to the regularised Holevo information.

Example 5.1. The capacity of the depolarising channel with parameter p on a Hilbert space $\mathcal{H}$ is the following [Wil19]:

$$\log(d) + (1-p+\frac{p}{d})\log\left(1-p+\frac{p}{d}\right) + (d-1)\frac{p}{d}\log\left(\frac{p}{d}\right), \tag{5.90}$$

where d is the dimension of $\mathcal{H}$.

5.2.3. Superadditivty of Holevo Information

We finish this chapter with a consideration of the (non-)additivity of the Holevo information. Additivity of the Holevo information would mean that:

$$\chi(\mathcal{N} \otimes \mathcal{M}) = \chi(\mathcal{N}) + \chi(\mathcal{M}), \tag{5.91}$$

for any two quantum channels $\mathcal{N}$ and $\mathcal{M}$. If the Holevo information would be additive, then it would be a good characterisation of the classical capacity of a quantum channel, because it would imply that the regularisation of the Holevo information is unnecessary. Indeed, if the Holevo information were additive, then for any quantum channel $\mathcal{N}$ we would have $\chi_{\text{reg}}(\mathcal{N}) = \chi(\mathcal{N})$.

It was first conjectured that the Holevo information is indeed additive, because several quantum channels were discovered for which it was the case. In 2009, however, Hastings showed that this conjecture is false in [Has09]. He showed that a different quantity called the minimum output entropy is non-additive, and it was already known that non-additivity of the minimum output entropy implies non-additivity of the Holevo information.

This means that using entangled states as codewords can increase the capacity of a quantum channel. It also means that the most basic question of the classical capacity of a quantum channel remains open and that further research is needed to determine in what situations entanglement can increase the classical capacity of a quantum channel [Has09].

6

Quantum Zero-Error Communication

In chapter three we considered zero-error communication through classical channels. We showed that a classical channel can be represented by a graph which contains all information needed to express the zero-error capacity of that channel. Moreover, we provided a lower and upper bound for the zero-error capacity and we showed that the zero-error capacity is non-additive. That is, the zero-error capacity of two channels combined can be greater than the sum of the zero-error capacities of the individual channels.

In the present chapter we will extend these concepts to the case of zero-error communication through quantum channels. Moreover, we will see that in the quantum case there exists an even more surprising phenomenon than superadditivity, called superactivation. Indeed, we will show that there exist two quantum channels which individually have no zero-error classical capacity, but whose combined channel does have zero-error classical capacity. This result is the most advanced so far and can be regarded as the climax of this thesis.

6.1. Zero-Error Classical Capacity of a Quantum Channel

In order to investigate the zero-error classical capacity of a quantum channel we first need to define it. We will then translate this definition into graph-theoretical language, analogously to what we did in chapter 3. This graph-theoretical language will allow us to further study the properties of the zero-error classical capacity. We then finish the section by relating the zero-error classical capacity to the Holevo-Schumacher-Westmoreland capacity from the previous chapter.

6.1.1. Zero-Error Quantum Codes and Capacity

We begin our study of the zero-error classical capacity of quantum channels by generalising the error-free code that we defined in chapter 3 to the case of quantum channels. We present this generalisation according to [GdAM16].

Definition 6.1. Let $\mathcal{N}:\mathcal{L}(\mathcal{H}_A)\to\mathcal{L}(\mathcal{H}_B)$ be a quantum channel. An (M,n) error-free quantum code over $\mathcal{N}$ consists of a finite set $\mathcal{M}=\{1,\dots,M\}$ of messages, an encoder $E^n:\mathcal{M}\to\mathcal{D}(\mathcal{H}_A)^{\otimes n}$ and a decoding POVM $\{\Lambda_m\}_{m\in\mathcal{M}}$ such that the probability of error is zero for all $m\in\mathcal{M}$:

$$\operatorname{Tr}\left((\mathbb{1}-\Lambda_m)\mathcal{N}(E(m))\right)=0. \tag{6.1}$$

We can now define the zero-error classical capacity over a quantum channel, which is really simply a straightforward generalisation of the zero-error capacity of a classical channel.

Definition 6.2. Let $\mathcal{N}$ be a quantum channel. Then the zero-error classical capacity $C^0(\mathcal{N})$ of this channel is defined as:

$$C^0(\mathcal{N})=\sup_{n\in\mathbb{N}}\frac{1}{n}\log(N(n)), \tag{6.2}$$

where $N(n)$ is the maximum number M of messages that can be sent over $\mathcal{N}$ using an (M,n) error-free quantum code.

This definition is exactly the same as in the classical case. It will, however, be useful for us to formulate it slightly differently because it will then be easier to understand the graph-theoretical formulation in the next subsection. We thus get [GdAM16]:

Definition 6.3. Let $\mathcal{N}:\mathcal{L}(\mathcal{H}_A)\to\mathcal{L}(\mathcal{H}_B)$ be a quantum channel. Then the zero-error classical capacity $C^0(\mathcal{N})$ of this channel is defined as:

$$C^0(\mathcal{N}) = \sup_{S}\sup_{n\in\mathbb{N}}\frac{1}{n}\log(\alpha_n(\mathcal{N})), \tag{6.3}$$

where $S\subset\mathcal{D}(\mathcal{H}_A)$ is a finite set of input states and $\alpha_n(\mathcal{N})$ is the maximum number M of messages which can be transmitted over $\mathcal{N}$ with an (M,n) error-free quantum code using only input states in S.

The fact that these definitions are in fact equivalent follows from $\sup_S \alpha_n(\mathcal{N}) = N(n)$ for any $n\in\mathbb{N}$, because we can write $\sup_S\sup_{n\in\mathbb{N}}\frac{1}{n}\log(\alpha_n(\mathcal{N})) = \sup_{n\in\mathbb{N}}\frac{1}{n}\log(\sup_S\alpha_n(\mathcal{N}))$ since the logarithm is a continuous and increasing function.

In order to study the zero-error classical capacity of quantum channels we will - analogously to the classical case - introduce the notion of adjacency between states. This notion will allow us generalise the concept of the characteristic graph from chapter 3 to the quantum case.

Definition 6.4. Let $\mathcal{N}:\mathcal{L}(\mathcal{H}_A)\to\mathcal{L}(\mathcal{H}_B)$ be a quantum channel. Then two states $\rho_1,\rho_2\in\mathcal{D}(\mathcal{H}_A)$ are said to be adjacent (or indistinguishable) if the supports of their output states are not orthogonal. That is:

$$\mathrm{Tr}\left(\mathcal{N}(\rho_1)\mathcal{N}(\rho_2)\right) > 0. \tag{6.4}$$

Otherwise, we call ρ_1 and ρ_2 non-adjacent (or distinguishable).

The above definition can immediately be generalised to tensor product states. Indeed, if $\rho_1 = \rho_1^1\otimes\ldots\otimes\rho_1^n$ and $\rho_2=\rho_2^1\otimes\ldots\otimes\rho_2^n$ are density operators in $\mathcal{H}_A^{\otimes n}$, then we call them distinguishable if there is some $1\le i\le n$ such that ρ_1^i and ρ_2^i are non-adjacent. If all entries are adjacent, then we call the tensor product states indistinguishable. Let us now end this subsection by considering a simple example: the depolarising channel.

Example 6.1. Let $\mathcal{N}:\mathcal{L}(\mathcal{H})\to\mathcal{L}(\mathcal{H})$ be a depolarising channel, where $d=\dim(\mathcal{H})$. Recall that the action of the depolarising channel on a density operator $\rho\in\mathcal{D}(\mathcal{H})$ is as follows:

$$\mathcal{N}(\rho) = (1-p)\rho + \frac{p}{d}\mathbb{1}, \tag{6.5}$$

for some $p\in(0,1)$. Now let $\rho_1,\rho_2\in\mathcal{D}(\mathcal{H})$. Then we have:

$$\mathrm{Tr}\left(\mathcal{N}(\rho_1)\mathcal{N}(\rho_2)\right) = \mathrm{Tr}\left(\left((1-p)\rho_1+\frac{p}{d}\mathbb{1}\right)\left((1-p)\rho_2+\frac{p}{d}\mathbb{1}\right)\right) \tag{6.6}$$

$$= \mathrm{Tr}\left((1-p)^2\rho_1\rho_2 + \frac{(1-p)p}{d}(\rho_1+\rho_2)+\frac{p^2}{d^2}\mathbb{1}\right) > \frac{p^2}{d^2} > 0. \tag{6.7}$$

Here we have used that the trace of the product of two density operators ρ_1 and ρ_2 is non-negative. To see that this is true, consider the following:

$$\mathrm{Tr}(\rho_1\rho_2) = \mathrm{Tr}(\sqrt{\rho_1}\sqrt{\rho_1}\rho_2) = \mathrm{Tr}(\sqrt{\rho_1}\rho_2\sqrt{\rho_1}^{\dagger}) \ge 0. \tag{6.8}$$

We see that any two input states ρ_1,ρ_2 are indistinguishable. Therefore, no POVM could completely reliably distinguish between them, meaning that the depolarising channel has a zero-error classical capacity equal to zero.

6.1.2. Graph-Theoretical Formulation

In this subsection we define the characteristic graph of a quantum channel in exactly the same way as for a classical channel. We then translate the expression for the zero-error classical capacity into an expression about the characteristic graph. This formulation will be useful in proving a result on the zero-error classical capacity in the next subsection.

Definition 6.5. Let $\mathcal{N} : \mathcal{L}(\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B)$ be a quantum channel and let S be a finite set of states in $\mathcal{D}(\mathcal{H}_A)$. The characteristic graph G corresponding to $\mathcal{N}$ and the set S is the undirected graph whose vertices are the elements of S and which has edges between the states in S which are distinguishable. We can generalise this to the tensor product channel $\mathcal{N}^{\otimes n}$, where the vertices are the elements of $S^{\otimes n}$ and the edges are still between distinguishable tensor product states.

With this definition we can reformulate the expression for the zero-error classical capacity of a quantum channel. Indeed, given a quantum channel $\mathcal{N} : \mathcal{L}(\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B)$ and some finite subset $S \subset \mathcal{D}(\mathcal{H}_A)$ of input states, the maximum number of messages M which can be transmitted by an (M, n) error-free quantum code over the channel using states from S is equal to the clique number $\omega(G)$ (defintion 3.6) of the corresponding characteristic graph G. Thus, we find an alternative expression for the zero-error classical capacity:

$$C^0(\mathcal{N}) = \sup_{S} \sup_{n\in\mathbb{N}} \frac{1}{n} \log(\omega(G)), \tag{6.9}$$

where the suprema are taken over all input sets S and code lengths n.

6.1.3. Properties of Zero-Error Classical Capacity

Now that we understand how the zero-error classical capacity can be expressed in terms of characteristic graphs we will prove two properties of the zero-error classical capacity. The first one states that the zero-error classical capacity can always be achieved by using a set of pure input states:

Theorem 6.1. Let $\mathcal{N} : \mathcal{L}(\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B)$ be a quantum channel. Then the zero-error capacity of this channel is equal to:

$$C^0(\mathcal{N}) = \sup_{S'} \sup_{n\in\mathbb{N}} \frac{1}{n} \log(\omega(G')), \tag{6.10}$$

where S' is a set of pure input states and G' is the corresponding characteristic graph.

Proof: we prove that for any set S of input states we can construct a set S' of pure input states such that:

$$\sup_{n\in\mathbb{N}} \frac{1}{n} \log(\omega(G)) = \sup_{n\in\mathbb{N}} \frac{1}{n} \log(\omega(G')), \tag{6.11}$$

where G and G' are the characteristic graphs corresponding to the input sets S and S' respectively. If this statement is indeed true, then the theorem immediately follows from the characterisation of the zero-error capacity in equation 6.9. Thus, let S be any set of input states. We then define S' to be a set of pure states $\{|\phi_i\rangle\langle\phi_i|_A\}$ such that $|\phi_i\rangle_A$ is an eigenvector of $\rho_i \in S$, whose corresponding eigenvalue is nonzero. That is, for every density operator in S we pick one of its eigenvectors and add its corresponding pure density operator to S'.

We will now show that a pair of pure states in S' is distinguishable when the corresponding density operators in S are distinguishable. Let $\{N_k\}$ be the Kraus operators of $\mathcal{N}$, and let $\rho_1, \rho_2 \in S$. We write these two states in their spectral decompositions: $\rho_1 = \sum_i \lambda_1^i |\psi_1^i\rangle\langle\psi_1^i|_A$ and $\rho_2 = \sum_i \lambda_2^i |\psi_2^i\rangle\langle\psi_2^i|_A$.

Suppose ρ_1 and ρ_2 are distinguishable. This implies:

$$0 = \mathrm{Tr}\left(\mathcal{N}(\rho_2)\mathcal{N}(\rho_2)\right) = \mathrm{Tr}\left(\sum_k N_k\rho_1 N_k^\dagger \sum_l N_l\rho_2 N_l^\dagger\right) \tag{6.12}$$

$$= \mathrm{Tr}\left(\sum_k N_k \sum_i \lambda_1^i |\psi_1^i\rangle\langle\psi_1^i|_A N_k^\dagger \sum_l N_l \sum_j \lambda_2^j |\psi_2^j\rangle\langle\psi_2^j|_A N_l^\dagger\right) \tag{6.13}$$

$$= \mathrm{Tr}\left(\sum_{i,j,k,l} \lambda_1^i\lambda_2^j N_k|\psi_1^i\rangle\langle\psi_1^i|_A N_k^\dagger N_l|\psi_2^j\rangle\langle\psi_2^j|_A N_l^\dagger\right) = \sum_{i,j,k,l}\lambda_1^i\lambda_2^j\left|\langle\psi_1^i|N_k^\dagger N_l|\psi_2^j\rangle\right|^2. \tag{6.14}$$

Thus, we conclude that for all i, j, k, l we have $\langle\psi_1^i|N_k^\dagger N_l|\psi_2^j\rangle = 0$. We will now use this fact to show that any eigenvector of ρ_1 is distinguishable from any eigenvector of ρ_2:

$$\mathrm{Tr}\left(\mathcal{N}(|\psi_1^i\rangle\langle\psi_1^i|_A)\mathcal{N}(|\psi_2^j\rangle\langle\psi_2^j|_A)\right) = \mathrm{Tr}\left(\sum_k N_k|\psi_1^i\rangle\langle\psi_1^i|_A N_k^\dagger \sum_l N_l|\psi_2^j\rangle\langle\psi_2^j|_A N_l^\dagger\right) \tag{6.15}$$

$$= \mathrm{Tr}\left(\sum_{k,l} N_k|\psi_1^i\rangle\langle\psi_1^i|_A N_k^\dagger N_l|\psi_2^j\rangle\langle\psi_2^j|_A N_l^\dagger\right) = \sum_{k,l}\left|\langle\psi_1^i|N_k^\dagger N_l|\psi_2^j\rangle\right|^2 = 0. \tag{6.16}$$

But this means that the graph G' contains all edges that G contains. G' can only differ from G in the sense that it can have more edges, i.e. more distinguishable input states. But the clique number of a graph can only increase when one adds edges to it, so we conclude that the clique number $\omega(G')$ is at least as great as $\omega(G)$. This means that $\sup_S \sup_n \frac{1}{n}\log(\omega(G)) \leq \sup_{S'} \sup_n \frac{1}{n}\log(\omega(G'))$. But because using only sets of pure input states instead of using any set of input states is a restriction we trivially have $\sup_S \sup_n \frac{1}{n}\log(\omega(G)) \geq \sup_{S'} \sup_n \frac{1}{n}\log(\omega(G'))$. Thus, the statement of the theorem follows. □

We will now state and prove the last result of this section, which relates the zero-error classical capacity to the HSW capacity:

Theorem 6.2. Let $\mathcal{N}$ be a quantum channel and let $C(\mathcal{N})$ denote its HSW capacity. Then:

$$C^0(\mathcal{N}) \leq C(\mathcal{N}). \tag{6.17}$$

Proof: in chapter 5 we saw that the HSW capacity is the supremum of all achievable rates over $\mathcal{N}$. In this case, achievable rate means that for every $\epsilon > 0$ there exists a code with at least that rate and with probability of error no greater than ϵ. Now let $n \in \mathbb{N}$. Then the greatest zero-error rate R which can be achieved using the channel n times is $\frac{1}{n}\log(N(n))$, where $N(n)$ denotes the maximum number M of messages for which there exists an (M, n) error-free quantum code. But R is then clearly achievable in the sense of the HSW capacity. Thus we find that $R < C(\mathcal{N})$. But since this hold for all $n \in \mathbb{N}$ we conclude that:

$$\sup_{n\in\mathbb{N}} \frac{1}{n}\log(N(n)) \leq C(\mathcal{N}), \tag{6.18}$$

which is just the statement that the zero-error classical capacity is no greater than the HSW capacity. □

6.2. Superactivation of Zero-Error Classical Capacity

Now that we understand the main definitions and results concerning the zero-error classical capacity of a quantum channel, we investigate an interesting and surprising phenomenon: the superactivation of the zero-error classical capacity, which was proven to exist by Cubitt, Chen and Harrow. This phenomenon entails the following: there exist two quantum channels which each have no zero-error classical capacity at all, whereas their combined (tensor product) channel does have positive zero-error classical capacity. It is called superactivation because we think of the two channels as activating each other, thus yielding a positive zero-error capacity whereas individually they do not have any.

In this section we present the main idea of the proof which shows that the zero-error classical capacity can be superactivated. For the complete proof and more details we refer the reader to the original paper [CCH11]. Before we present the main idea, however, we develop some preliminary notions and results on so-called conjugate-divisible maps.

6.2.1. Conjugate-Divisble Maps

Let us start by stating the definition of a conjugate-divisible map.

Definition 6.6. Let $\mathcal{H}_A$ be a Hilbert space. Then a map $\mathcal{N} : \mathcal{H}_A \to \mathcal{H}_A$ is called conjugate-divisible if it can be written:

$$\mathcal{N} = \mathcal{E}^* \circ \mathcal{E}, \tag{6.19}$$

where $\mathcal{E} : \mathcal{H}_A \to \mathcal{H}_B$ is a quantum channel and $\mathcal{E}^*$ is its adjoint.

Conjugate-divisible maps will play a major role in our proof that superactivated channels exist, because studying the map $\mathcal{E}^* \circ \mathcal{E}$ for a quantum channel $\mathcal{E}$ will turn out to carry information about the zero-error classical capacity of $\mathcal{E}$. It will be useful for us to study these maps via their Choi-Jamiolkowski operators, provided by the Choi-Jamiolkowski isomorphism. More on the Choi-Jamiolkowski operator and isomorphism can be found in section 4.4.1 of [Wil19]. We will first use the standard Choi-Jamiolkowski operator, which is defined as follows [Wil19]:

Definition 6.7. Let $\mathcal{H}_R$ and $\mathcal{H}_A$ be isomorphic Hilbert spaces and let $\{|i\rangle_R\}$ and $\{|i\rangle_A\}$ be orthonormal basis for them respectively. Let $\mathcal{H}_B$ be another Hilbert space and let $\mathcal{N} : \mathcal{L}(\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B)$ be any linear map. Then a Choi-Jamiolkowski operator corresponding to this map is given by:

$$\sigma_{RB} = (\mathbb{1}_R \otimes \mathcal{N})(|\omega\rangle\langle\omega|_{RA}), \tag{6.20}$$

where $|\omega\rangle_{RA}$ is an unnormalised vector given by:

$$|\omega\rangle_{RA} = \sum_{i=0}^{d_A} \lambda_i |\psi_i\rangle_R \otimes |\phi_i\rangle_A, \tag{6.21}$$

with $\{|\psi\rangle_R\}$ and $\{|\phi\rangle_A\}$ orthonormal bases and no λ_i equal to zero.

Since $\mathcal{H}_A$ and $\mathcal{H}_R$ are isomorphic, we will from now on write A instead of R, since this makes our equations more orderly. Introducing the basis change $U|\psi_i\rangle = |\phi_i\rangle$ (with U unitary) we can also get back the action of the map $\mathcal{N}$ in the above definition by [CCH11]:

$$\mathcal{N}(\rho_A) = \mathrm{Tr}_A(U\sigma_A^{-\frac{1}{2}} \otimes \mathbb{1}_B \cdot \sigma_{AB} \cdot \sigma_A^{-\frac{1}{2}} U^\dagger \otimes \mathbb{1}_B \cdot \rho_A^T \otimes \mathbb{1}_B), \tag{6.22}$$

where $\sigma_A = \mathrm{Tr}_B(\sigma_{AB})$. The standard Choi-Jamiolkowski operator $\tilde{\sigma}_{AB}$ is obtained by setting $|\omega\rangle = \sum_i |i\rangle|i\rangle$ where $|i\rangle$ is a basis of $\mathcal{H}_A$, and the above equation simplifies to:

$$\mathcal{N}(\rho_A) = \mathrm{Tr}_A(\tilde{\sigma}_{AB} \cdot \rho_A^T \otimes \mathbb{1}_B). \tag{6.23}$$

Moreover, we see that we can write:

$$\tilde{\sigma}_{AB} = U\sigma_A^{-\frac{1}{2}} \otimes \mathbb{1}_B \cdot \sigma_{AB} \cdot \sigma_A^{-\frac{1}{2}} U^\dagger, \tag{6.24}$$

where U is unitary [CCH11].

In order to prove our first result about conjugate divisible maps we need a lemma, and this lemma makes use of the following operation:

Definition 6.8. Let $|\psi\rangle_{AB}$ be a bipartite state in $\mathcal{H}_A \otimes \mathcal{H}_B$, where $|i\rangle_A$ and $|j\rangle_B$ are orthonormal bases of the isomorphic Hilbert spaces $\mathcal{H}_A$ and $\mathcal{H}_B$ respectively. Writing $|\psi\rangle_{AB} = \sum_{i,j} c_{ij}|i\rangle_A \otimes |j\rangle_B$ the flip operation is defined as follows:

$$\mathbb{F}(|\psi\rangle_{AB}) = \mathbb{F}\left(\sum_{i,j} c_{ij}|i\rangle_A \otimes |j\rangle_B\right) = \sum_{i,j} \bar{c}_{ij}|j\rangle_A \otimes |i\rangle_B, \tag{6.25}$$

where $\bar{c}_{ij}$ denotes the complex conjugate of c_{ij}. This definition can be extended to a density operator $\rho_{AB} = \sum_{i,j} \lambda_{ij}|i\rangle_A|j\rangle_B\langle i|_A\langle j|_B$ by:

$$\mathbb{F}(\rho_{AB}) = \sum_{i,j} \bar{\lambda}_{ij}|j\rangle_A|i\rangle_B\langle j|_A\langle i|_B. \tag{6.26}$$

We call a bipartite state or operator conjugate-symmetric if it is invariant under the flip operation. Similarly, we call a subspace of $\mathcal{H}_A \otimes \mathcal{H}_B$ conjugate-symmetric if it is invariant under the flip operation.

Using this flip operator we get the following lemma:

Lemma 6.3. Let $\mathcal{E} : \mathcal{L}(\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B)$ be a quantum channel with standard Choi-Jamiolkowski operator ρ_{AB}. Then the standard Choi-Jamiolkowski operator of $\mathcal{E}^*$ is given by:

$$\mathbb{F}(\rho_{AB}) = \bar{\rho}_{BA}. \tag{6.27}$$

Proof: by definition, the adjoint $\mathcal{E}^* : \mathcal{L}(\mathcal{H}_B) \to \mathcal{L}(\mathcal{H}_A)$ of $\mathcal{E} : \mathcal{L}(\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B)$ is defined to be the unique map which satisfies:

$$\mathrm{Tr}(\psi \cdot \mathcal{E}(\phi)) = \mathrm{Tr}(\mathcal{E}^*(\psi) \cdot \phi), \tag{6.28}$$

for any $\phi \in \mathcal{L}(\mathcal{H}_A)$ and $\psi \in \mathcal{L}(\mathcal{H}_B)$. Thus, we get the following:

$$\begin{aligned}
\mathrm{Tr}(\mathcal{E}^*(\psi)^\dagger \cdot \phi) = \mathrm{Tr}(\mathcal{E}^*(\psi^\dagger) \cdot \phi) = \mathrm{Tr}(\psi^\dagger \cdot \mathcal{E}(\phi)) &= \mathrm{Tr}\left(\psi^\dagger \cdot \mathrm{Tr}_A(\rho_{AB} \cdot (\phi^T \otimes \mathbb{1}_\psi))\right) && (6.29)\\
&= \mathrm{Tr}\left(\mathbb{1}_A \otimes \psi^\dagger \cdot \rho_{AB} \cdot \phi^T \otimes \mathbb{1}_B\right) = \mathrm{Tr}\left(\mathrm{Tr}_B(\mathbb{1}_A \otimes \psi^\dagger)\rho_{AB})\phi^T\right) && (6.30)\\
&= \mathrm{Tr}\left(\mathrm{Tr}_B(\mathbb{1}_A \otimes \psi^\dagger \cdot \rho_{AB})^T\phi\right) = \mathrm{Tr}\left(\mathrm{Tr}_B(\mathbb{1}_A \otimes \bar{\psi} \cdot \rho_{AB}^{T_B})\phi\right) && (6.31)\\
&= \mathrm{Tr}\left(\mathrm{Tr}_B(\rho_{AB}^{T_B} \cdot \mathbb{1}_A \otimes \bar{\psi})\phi\right) = \mathrm{Tr}\left(\mathrm{Tr}_B(\rho_{BA}^{T_B} \cdot \bar{\psi} \otimes \mathbb{1}_A)\phi\right) && (6.32)\\
&= \mathrm{Tr}\left(\mathrm{Tr}_B(\bar{\rho}_{BA} \cdot \psi \otimes \mathbb{1}_A)^\dagger\phi\right) = \mathrm{Tr}\left(\mathrm{Tr}_B(\mathbb{F}(\rho_{AB}) \cdot \psi^T \otimes \mathbb{1}_A)^\dagger\phi\right), && (6.33)
\end{aligned}$$

where T_B denotes the transpose with respect to the subsystem B only.

From this we recognise that:

$$\mathcal{E}^*(\psi) = \mathrm{Tr}_B(\mathbb{F}(\rho_{AB}) \cdot \psi^T \otimes \mathbb{1}_A), \tag{6.34}$$

and thus we conclude that the standard Choi-Jamiolkowski operator of $\mathcal{E}$ is indeed $\mathbb{F}(\rho_{AB})$. □

We can now state and prove our first result on conjugate-divisble maps.

Lemma 6.4. If $\mathcal{E} : \mathcal{L}(\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B)$ is a quantum channel with standard Choi-Jamiolkowski operator ρ_{AB}, then the standard Choi-Jamiolkowski matrix of the map $\mathcal{N} = \mathcal{E}^* \circ \mathcal{E}$ is given by:

$$\sigma_{AA'} = \mathrm{Tr}_B\left(\rho_{AB} \otimes \mathbb{1}_{A'} \cdot \mathbb{1}_A \otimes \mathbb{F}(\rho_{A'B})^{T_B}\right). \tag{6.35}$$

Proof: for any $\phi \in \mathcal{L}(\mathcal{H}_A)$ we have:

$$\mathcal{N}(\phi) = \mathrm{Tr}_B\left(\mathbb{F}(\rho_{A'B}) \cdot \left(\mathcal{E}(\phi)\right)^T \otimes \mathbb{1}'_A\right) = \mathrm{Tr}_B\left(\mathbb{F}(\rho_{A'B}) \cdot \left(\mathrm{Tr}_A(\rho_{AB} \cdot \phi^T \otimes \mathbb{1}_B)\right)^T \otimes \mathbb{1}'_A\right) \tag{6.36}$$

$$= \mathrm{Tr}_B\left(\left(\mathrm{Tr}_A(\rho_{AB} \cdot \phi^T \otimes \mathbb{1}_B)\right)^T \otimes \mathbb{1}'_A \cdot \mathbb{F}(\rho_{A'B})\right) \tag{6.37}$$

$$= \mathrm{Tr}_B\left(\mathrm{Tr}_A(\rho_{AB} \cdot \phi^T \otimes \mathbb{1}_B) \otimes \mathbb{1}'_A \cdot \mathbb{F}(\rho_{A'B})^{T_B}\right) \tag{6.38}$$

$$= \mathrm{Tr}_A\left(\mathrm{Tr}_B(\rho_{AB} \otimes \mathbb{1}_{A'} \cdot \mathbb{1}_A \otimes \mathbb{F}(\rho_{A'B})^{T_B}) \cdot \phi^T \otimes \mathbb{1}_B \otimes \mathbb{1}'_A\right). \tag{6.39}$$

We indeed recognise that $\sigma_{AA'} = \mathrm{Tr}_B\left(\rho_{AB} \otimes \mathbb{1}_{A'} \cdot \mathbb{1}_A \otimes \mathbb{F}(\rho_{A'B})^{T_B}\right)$. □

As explained in [CCH11], we can generalise this result to non-standard Choi-Jamiolkowski operators by recognising that the standard operator is related to any non-standard one by equation (6.24). This yields the following result:

Corollary 6.5. Let $\mathcal{E}$ be a channel with Choi-Jamiolkowski operator (standard or non-standard) ρ_{AB}. Then

$$\sigma_{AA'} = \mathrm{Tr}_B\left(\rho_{AB} \otimes \mathbb{1}_{A'} \cdot \mathbb{1}_A \otimes \mathbb{F}(\rho_{A'B})^{T_B}\right) \tag{6.40}$$

is a Choi-Jamiolkowski operator of $\mathcal{N} = \mathcal{E}^* \circ \mathcal{E}$.

Before we further investigate the properties of Choi-Jamiolkowski operators of conjugate-divisible maps we introduce some new notation.

Notation 6.6. Let $|\psi\rangle_{AB}$ be a bipartite state in $\mathcal{H}_A \otimes \mathcal{H}_B = \mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B}$. Let us write it out in the product basis:

$$|\psi\rangle_{AB} = \sum_{i,j} M_{ij} |i\rangle|j\rangle. \tag{6.41}$$

Then we define the isomorphism $\mathbb{M}$ to take a bipartite state to the $d_A \times d_B$ matrix with coefficients M_{ij}. Similarly, given a subspace $S \subset \mathcal{H}_A \otimes \mathcal{H}_B$ we write $\mathbb{M}(S)$ for the corresponding matrix subspace. A state $|\psi\rangle$ is then called conjugate-symmetric (invariant under the flip operation) if its matrix $\mathbb{M}(|\psi\rangle)$ is hermitian, and a subspace S is called conjugate-symmetric if $\mathbb{M}(S)$ is spanned by a basis of hermitian matrices.

We now use this notation to define the following [CCH11]:

Definition 6.9. Let $|\psi\rangle_{AB}$ be a bipartite state in $\mathcal{H}_A \otimes \mathcal{H}_B = \mathbb{C}^{d_A} \otimes \mathbb{C}^{d_B}$. We call $|\psi\rangle_{AB}$ positive semidefinite if $\mathbb{M}(|\psi\rangle)$ is a positive semidefinite matrix (and therefore also hermitian). Similarly, we call a subspace S positive semidefinite if $\mathbb{M}(S)$ is spanned by a basis of positive semidefinite matrices.

We remark that a positive semi-definite subspace need not contain only positive semidefinite elements. It only needs to have a positive semidefinite basis. Moreover, we remark that a subspace is already positive semi-definite if it contains one positive definite element, because then we can just add this element many times to any basis and turn it into a positive semidefinite basis [CCH11]. Lastly, we remark that since a positive semidefinite matrix is in particular hermitian, any positive semidefinite subspace is also a conjugate-symmetric subspace.

In order to prove the main theorem of this subsection we need two lemmas, of which we now present and prove the first.

Lemma 6.7. The support of the Choi-Jamiolkowski operator (the span of its column vectors) of a conjugate-divisible map is a positive semidefinite subspace.

Proof: let $\mathcal{N} = \mathcal{E}^* \circ \mathcal{E}$ be a conjugate divisible map, where $\mathcal{E} : \mathcal{L}(\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B)$ is a quantum channel with Choi-Jamiolkowski operator ρ_{AB}. From corollary 6.5 we know the Choi-Jamiolkowski operator of $\mathcal{N}$ is given by:

$$\sigma_{AA'} = \mathrm{Tr}_B\left(\rho_{AB} \otimes \mathbb{1}_{A'} \cdot \mathbb{1}_A \otimes \mathbb{F}(\rho_{A'B})^{T_B}\right). \tag{6.42}$$

If we now write $\rho_{AB} = \sum_k |\phi_k\rangle\langle\phi_k|_{AB}$ with $|\phi_k\rangle_{AB} = \sum_i |\psi_i^k\rangle_A |i\rangle_B$, where eigenvalues and coefficients have been absorbed into the unnormalised basis vectors $|\phi_k\rangle_{AB}$ and $|\psi_i^k\rangle_A$, we get the following:

$$\begin{aligned}
\sigma_{AA'} &= \mathrm{Tr}_B\left(\rho_{AB} \otimes \mathbb{1}_{A'} \cdot \mathbb{1}_A \otimes \mathbb{F}(\rho_{A'B})^{T_B}\right) && (6.43)\\
&= \mathrm{Tr}_B\left(\sum_{i,j,k} |\psi_i^k\rangle_A |i\rangle_B \langle\psi_j^k|_A \langle j|_B \otimes \mathbb{1}_{A'} \cdot \mathbb{1}_A \otimes \mathbb{F}\left(\sum_{l,m,n} |\psi_l^n\rangle_{A'} |l\rangle_B \langle\psi_m^n|_{A'} \langle m|_B\right)^{T_B}\right) && (6.44)\\
&= \mathrm{Tr}_B\left(\sum_{i,j,k} |\psi_i^k\rangle_A |i\rangle_B \langle\psi_j^k|_A \langle j|_B \otimes \mathbb{1}_{A'} \cdot \mathbb{1}_A \otimes \left(\sum_{l,m,n} |l\rangle_B |\bar{\psi}_l^n\rangle_{A'} \langle m|_B \langle\bar{\psi}_m^n|_{A'}\right)^{T_B}\right) && (6.45)\\
&= \mathrm{Tr}_B\left(\sum_{i,j,k} |\psi_i^k\rangle_A |i\rangle_B \langle\psi_j^k|_A \langle j|_B \otimes \mathbb{1}_{A'} \cdot \mathbb{1}_A \otimes \sum_{l,m,n} |m\rangle_B |\bar{\psi}_l^n\rangle_{A'} \langle l|_B \langle\bar{\psi}_m^n|_{A'}\right) && (6.46)\\
&= \sum_{i,j,k,l} |\psi_i^k\rangle_A |\bar{\psi}_i^l\rangle_{A'} \langle\psi_j^k|_A \rangle\bar{\psi}_j^l|_{A'}. && (6.47)
\end{aligned}$$

Thus, if we use $S_{AA'}$ to denote the support of $\sigma_{AA'}$ we find:

$$S_{AA'} = \mathrm{span}\left(\sum_i |\psi_i^k\rangle_A |\bar{\psi}_i^l\rangle_{A'}\right)_{k,l}. \tag{6.48}$$

But now we see that $\mathbb{M}(S_{AA'})$ contains the following matrix:

$$\mathbb{M}\left(\sum_{i,k} |\psi_i^k\rangle_A |\bar{\psi}_i^k\rangle_{A'}\right) = \sum_{i,k} |\psi_i^k\rangle\langle\psi_i^k|, \tag{6.49}$$

which is a positive definite matrix. We conclude that $S_{AA'}$ is a positive semidefinite subspace. □

The second lemma that we need is the following [CCH11]:

Lemma 6.8. For any conjugate symmetric subspace $S_{AA'} \subset \mathcal{H}_A \otimes \mathcal{H}_{A'}$ such that $\mathrm{supp}(\mathrm{Tr}_{A'}(S_{AA'})) = \bigcup_{|\psi\rangle \in S_{AA'}} \mathrm{supp}(\mathrm{Tr}_{A'}(|\psi\rangle\langle\psi|)) = \mathcal{H}_A$ we can construct a Choi-Jamiolkowski operator $\sigma_{AA'}$ of a conjugate-divisible map $\mathcal{N} = \mathcal{E}^* \circ \mathcal{E}$ such that $\mathrm{supp}(\sigma_{AA'}) = S_{AA'}$. The quantum channel $\mathcal{E} : \mathcal{L}(\mathbb{H}_A) \to \mathcal{L}(\mathcal{H}_B)$ has input dimension $d_A = \mathcal{H}_A$, rank $d_E = \dim(S_{AA'})$ and output dimension $d_B = d_A d_E$.

Proof: since $S_{AA'}$ is positive semidefinite, $\mathbb{M}(S_{AA'})$ has a basis $\{M_k\}$ of matrices $M_k \geq 0$ (which are therefore also Hermitian). Thus, we can spectrally decompose and absorb the positive eigenvalues into the eigenstates for every such matrix:

$$M_k = \sum_i |\psi_i^k\rangle\langle\psi_i^k|. \tag{6.50}$$

We thus have:

$$S_{AA'} = \mathrm{span}\left(\sum_i |\psi_i^k\rangle|\bar{\psi}_i^k\rangle\right)_k. \tag{6.51}$$

We now consider the following operator:

$$\rho_{AB} = \sum_{i,j,k} |\psi_i^k\rangle_A |k,i\rangle_B \langle\psi_j^k|_A \langle k,j|_B, \tag{6.52}$$

which is positive semidefinite. Moreover, we have that the rank of $\mathrm{Tr}(\rho_{AB})$ is d_A. Thus, by the Choi-Jamiolkowski isomorphism it is the Choi-Jamiolkowski operator of a quantum channel $\mathcal{E} : \mathcal{L}(H_A) \to \mathcal{L}(H)_B$ which has rank $d_E = \dim(S_{AA'})$ and where $d_B = d_A d_E$ [CCH11]. We can now use corollary 6.5 to find the Choi-Jamiolkowski operator $\sigma_{AA'}$ of the conjugate-divisble map $\mathcal{N} = \mathcal{E}^* \circ \mathcal{E}$:

$$\sigma_{AA'} = \mathrm{Tr}_B \left(\rho_{AB} \otimes \mathbb{1}_{A'} \cdot \mathbb{1}_A \otimes \mathbb{F}(\rho_{A'B})^{T_B}\right) = \mathrm{Tr}_B \Big(\sum_{i,j,k} |\psi_i^k\rangle_A |k,i\rangle_B \langle\psi_j^k|_A \langle k,j|_B \otimes \mathbb{1}_{A'} \tag{6.53}$$

$$\cdot \mathbb{1}_A \otimes \mathbb{F}\left(\sum_{l,m,n} |\psi_l^n\rangle_{A'} |n,l\rangle_B \langle\psi_m^n|_{A'} \langle n,m|_B \right)^{T_B} \Big) \tag{6.54}$$

$$= \mathrm{Tr}_B \left(\sum_{i,j,k} |\psi_i^k\rangle_A |k,i\rangle_B \langle\psi_j^k|_A \langle k,j|_B \otimes \mathbb{1}_{A'} \cdot \mathbb{1}_A \otimes \left(\sum_{l,m,n} |n,l\rangle_B |\bar{\psi}_l^n\rangle_{A'} \langle n,m|_B \langle\bar{\psi}_m^n|_{A'} \right)^{T_B} \right) \tag{6.55}$$

$$= \mathrm{Tr}_B \left(\sum_{i,j,k} |\psi_i^k\rangle_A |k,i\rangle_B \langle\psi_j^k|_A \langle k,j|_B \otimes \mathbb{1}_{A'} \cdot \mathbb{1}_A \otimes \sum_{l,m,n} |n,m\rangle_B |\bar{\psi}_l^n \langle_{A'} \langle n,l|_B \langle\bar{\psi}_m^n|_{A'} \right) \tag{6.56}$$

$$= \sum_{i,j,k} |\psi_i^k\rangle_A |\bar{\psi}_i^k\rangle_{A'} \langle\psi_j^k|_A \langle\bar{\psi}_j^k|_{A'}. \tag{6.57}$$

The support of this operator is $\mathrm{span}\left(\sum_i |\psi_i^k\rangle_A |\bar{\psi}_i^k\rangle_{A'}\right)_k$, which we saw is equal to $S_{AA'}$ at the beginning of this proof. □

We are now in a position to state the main result of this section. This theorem will allow us to study the zero-error classical capacity of a quantum channel by studying its corresponding conjugate divisible map. It is directly implied by lemmas 6.7 and 6.8.

Theorem 6.9. Let $S_{AA'} \subset \mathcal{H}_A \otimes \mathcal{H}_{A'}$ be a subspace which is such that $\mathrm{supp}(\mathrm{Tr}_{A'}(S_{AA'})) = \mathcal{H}_A$. Then there exists a conjugate divisible map with Choi-Jamiolkowski matrix $\sigma_{AA'}$ satisfying $\mathrm{supp}(\sigma_{AA'}) = S_{AA'}$ if and only if $S_{AA'}$ is a positive semidefinite subspace.

6.2.2. Existence of Superactivated Channels

In this section we will explain the proof that there exist two quantum channels whose zero-error classical capacity is superactivated. In order to do so, we first translate this statement about the existence of two quantum channels into a statement about the existence of a subspace $S \subset \mathcal{H}_A \otimes \mathcal{H}_A$ and two unitary operators U, V on $\mathcal{H}_A$ satisfying certain conditions. We then explain how one can show that such a subspace and unitary operators exist, thereby proving the existence of superactivation of the zero-error classical capacity.

Now, a quantum channel $\mathcal{E} : \mathcal{L}(\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B)$ has nonzero zero-error classical capacity if and only if there exist two distinguishable states, i.e. for all $n \in \mathbb{N}$ there exist $|\psi\rangle, |\phi\rangle \in \mathcal{H}_A^{\otimes n}$ such that:

$$\mathrm{Tr}\left(\mathcal{E}^{\otimes n}(|\psi\rangle\langle\psi|)\mathcal{E}^{\otimes n}(|\phi\rangle\langle\phi|)\right) = 0. \tag{6.58}$$

Since a quantum channel maps density operators to density operators, and density operators are self-adjoint, we can write this equivalently as:

$$\mathrm{Tr}\left(\mathcal{E}^{\otimes n}(|\psi\rangle\langle\psi|)^\dagger \mathcal{E}^{\otimes n}(|\phi\rangle\langle\phi|)\right) = 0. \tag{6.59}$$

Let $\{N_k\}$ be the Kraus operators of $\mathcal{E}$. Then we see, using the cyclicity of the trace:

$$\mathrm{Tr}\left(\mathcal{E}^{\otimes n}(|\psi\rangle\langle\psi|)\mathcal{E}^{\otimes n}(|\phi\rangle\langle\phi|)\right) = \mathrm{Tr}\left(\sum_k N_k^{\otimes n}|\psi\rangle\langle\psi|N_k^{\dagger^{\otimes n}} \sum_l N_l^{\otimes n}|\phi\rangle\langle\phi|N_l^{\dagger^{\otimes n}}\right) \tag{6.60}$$

$$= \sum_{k,l} \mathrm{Tr}\left(N_k^{\otimes n}|\psi\rangle\langle\psi|N_k^{\dagger^{\otimes n}} N_l^{\otimes n}|\phi\rangle\langle\phi|N_l^{\dagger^{\otimes n}}\right) = \sum_{k,l} \mathrm{Tr}\left(|\psi\rangle\langle\psi|N_k^{\dagger^{\otimes n}} N_l^{\otimes n}|\phi\rangle\langle\phi|N_l^{\dagger^{\otimes n}} N_k^{\otimes n}\right) \tag{6.61}$$

$$= \mathrm{Tr}\left(\sum_{k,l} |\psi\rangle\langle\psi|N_k^{\dagger^{\otimes n}} N_l^{\otimes n}|\phi\rangle\langle\phi|N_l^{\dagger^{\otimes n}} N_k^{\otimes n}\right) = \mathrm{Tr}\left(|\psi\rangle\langle\psi| \cdot \mathcal{E}^{*^{\otimes n}}\left(\mathcal{E}^{\otimes n}(|\phi\rangle\langle\phi|)\right)\right). \tag{6.62}$$

The above equations immediately imply the following result, which is the first step in our translation of the existence of superactivated channels into a statement about a subspace and unitary operators:

Lemma 6.10. Let $\mathcal{E}_1, \mathcal{E}_2 : \mathcal{L}(\mathcal{H}_A) \to \mathcal{L}(\mathcal{H}_B)$ be two quantum channels. Then the statement that $\mathcal{E}_1, \mathcal{E}_2$ have no zero-error classical capacity whereas $\mathcal{E}_1 \otimes \mathcal{E}_2$ does have zero-error classical capacity is equivalent to:

$$\forall i = 1,2 : \forall n \in \mathbb{N} : \forall |\psi\rangle, |\phi\rangle \in \mathcal{H}_A^{\otimes n} : \mathrm{Tr}\left(\psi \cdot \mathcal{E}_i^{*^{\otimes n}} \circ \mathcal{E}_i^{\otimes n}(\phi)\right) \neq 0, \tag{6.63}$$

$$\exists |\psi\rangle, |\phi\rangle \in \mathcal{H}_A^{\otimes 2n} : \mathrm{Tr}\left(\psi \cdot (\mathcal{E}_1^* \otimes \mathcal{E}_2^*)^{\otimes n} \circ (\mathcal{E}_1 \otimes \mathcal{E}_2)^{\otimes n}(\phi)\right) = 0, \tag{6.64}$$

where we have used the notation $\psi = |\psi\rangle\langle\psi|, \phi = |\phi\rangle\langle\phi|$.

We will now find equivalent conditions for the above two equations. To do so for the first, let $\sigma_{1,2}$ denote Choi-Jamiolkowski operators of the conjugate divisible maps $\mathcal{N}_{1,2} = \mathcal{E}_{1,2}^* \circ \mathcal{E}_{1,2}$. Then the first condition simply states:

$$\forall n \in \mathbb{N} : \forall |\psi\rangle, |\phi\rangle \in \mathcal{H}_A^{\otimes n} : \mathrm{Tr}\left(\psi_{A'^{\otimes n}} \cdot \mathrm{Tr}_{A^{\otimes n}}(\sigma_{1,2}^{\otimes n} \cdot \phi_A^T \otimes \mathbb{1}_{A'})\right) \tag{6.65}$$

$$= \mathrm{Tr}\left(\sigma_{1,2}^{\otimes n} \cdot \phi_A^T \otimes \psi_{A'}\right) \neq 0. \tag{6.66}$$

From corrolary 6.5 we know that this holds for any Choi-Jamiolkowski operator and not only the standard one, because the rescaling unitaries can be absorbed into ψ and ϕ [CCH11]. Alternatively, we can interpret the above equation as stating that the orthogonal complements of the supports of the operators $\sigma_{1,2}$ contain no product states:

$$\forall n \in \mathbb{N} : \nexists |\psi\rangle, |\phi\rangle \in \mathcal{H}_A^{\otimes n} : |\psi\rangle \otimes |\phi\rangle \in S_{1,2}^{\otimes n^{\perp}}. \tag{6.67}$$

We now turn our attention to the second condition of lemma 6.10. Since this conditions requires the existence of two vectors $|\psi\rangle, |\phi\rangle \in \mathcal{H}_A^{\otimes 2n}$ satisfying a certain property, we can restrain these vectors to be of the following form:

$$|\psi\rangle = U_{A_1}^{\otimes n} \otimes V_{A_2}^{\otimes n} |\omega\rangle^{\otimes n}, \tag{6.68}$$

$$|\phi\rangle = W_{A_1'}^{\otimes n} \otimes X_{A_2'}^{\otimes n} |\omega\rangle^{\otimes n}, \tag{6.69}$$

where $|\omega\rangle_{A_1 A_2} = \sum_i |ii\rangle_{A_1 A_2}$ is the maximally entangled, unnormalised vector and U, V, W, X are unitary. Thus, if we again use $\sigma_{1,2}$ to denote the Choi-Jamiolkowski operators of $\mathcal{N}_{1,2} = \mathcal{E}_{1,2}^* \circ \mathcal{E}_{1,2}$, the second condition of 6.10 for $|\psi\rangle$ and $|\phi\rangle$ as just defined translates to [CCH11][GdAM16]:

$$\mathrm{Tr}\left(\psi \cdot (\mathcal{N}_1 \otimes \mathcal{N}_2)^{\otimes n}\right) = \mathrm{Tr}\left(\psi \cdot \mathrm{Tr}_A\left((\sigma_1 \otimes \sigma_2)^{\otimes n} \cdot \phi^T \otimes \mathbb{1}\right)\right) = \mathrm{Tr}\left((\sigma_1 \otimes \sigma_2)^{\otimes n} \cdot \phi^T \otimes \psi\right) \tag{6.70}$$

$$= \mathrm{Tr}\left((\sigma_1 \otimes \sigma_2)^{\otimes n} \cdot (\bar{U}^{\otimes n} \otimes \bar{V}^{\otimes n} \omega^{T^{\otimes n}} U^{T^{\otimes n}} \otimes V^{T^{\otimes n}}) \otimes (W^{\otimes n} \otimes X^{\otimes n} \omega^{\otimes n} W^{\dagger^{\otimes n}} \otimes X^{\dagger^{\otimes n}})\right) \tag{6.71}$$

$$= \mathrm{Tr}\Big(\left((\bar{U}^{\otimes n} \otimes W^{\otimes n}) \cdot \sigma_1^{\otimes n} \cdot (U^{T^{\otimes n}} \otimes W^{\dagger^{\otimes n}})\right)^T \tag{6.72}$$

$$\cdot \left((\bar{V}^{\otimes n} \otimes X^{\otimes n}) \cdot \sigma_2^{\otimes n} \cdot (V^{T^{\otimes n}} \otimes X^{\dagger^{\otimes n}})\right)\Big) \tag{6.73}$$

$$= \mathrm{Tr}\left(\sigma_1^{T^{\otimes n}} \cdot (U'^{\otimes n} \otimes V'^{\otimes n}) \sigma_2^{\otimes n} (U'^{\dagger^{\otimes n}} \otimes V'^{\otimes^{\otimes n}})\right) = 0. \tag{6.74}$$

From this it follows that a sufficient condition for $\mathcal{E}_{1,2}$ to satisfy the second condition of 6.10 is that there exist two unitary operators U, V such that:

$$S_2^T = U \otimes V \cdot S_1^{\perp}, \tag{6.75}$$

where $S_{1,2}$ denote the supports of $\sigma_{1,2}$ [CCH11]. Redefining, for easy of notation, $S_2 = \text{supp}(\sigma_2^T)$, we now present the following lemma:

Lemma 6.11. If there exists subspaces $S_{1,2} \subset \mathcal{H}_A \otimes \mathcal{H}_A$ and unitary operators U, V on $\mathcal{H}_A$ satisfying:

$$\forall n \in \mathbb{N} : \nexists |\psi\rangle, |\phi\rangle \in \mathcal{H}_A^{\otimes n} : |\psi\rangle \otimes |\phi\rangle \in S_{1,2}^{\otimes n^{\perp}}, \tag{6.76}$$

$$S_2 = U \otimes V \cdot S_1^{\perp}, \tag{6.77}$$

$$\mathbb{F}(S_{1,2}) = S_{1,2}, \tag{6.78}$$

$$\exists \{M_i^{1,2} \geq 0\} : \mathbb{M}(S_{1,2}) = \text{span}\{M_i^{1,2}\}, \tag{6.79}$$

then there exist quantum channels $\mathcal{E}_{1,2}$ which individually have no zero-error classical capacity but whose tensor product channel has positive zero-error classical capacity.

Proof: the first two conditions are simply the conditions which we established above, which guarantee two quantum channels $\mathcal{E}_{1,2}$ to individually have no zero-error classical capacity but to together have positive zero-error classical capacity. The existence of these channels, however, is guaranteed by lemma 6.8 because the fourth property of this lemma is simply positive semidefiniteness. The property $\text{supp}(\text{Tr}_{A'}(S_{AA'}))$ in lemma 6.8 can be neglected because we can always shrink $\mathcal{H}_A$ such that this property is indeed satisfied, without changing the requirements of the lemma [CCH11]. The third property (conjugate-symmetry) is redundant because it is implied by the fourth (positive semidefiniteness). □

The redundant requirement of conjugate-symmetry is useful in proving that the appropriate subspaces and unitary operators exist.

Now, the orthogonal complement of a conjugate-symmetric subspace is again conjugate- symmetric [CCH11]. So the second condition $S_2 = U \otimes V \cdot S_1^{\perp}$ of the above lemma implies the following if S_1 and S_2 are conjugate-symmetric:

$$U \otimes V \cdot S_1 = S_2^{\perp} = \mathbb{F}(S_2^{\perp}) = \mathbb{F}(U \otimes V \cdot S_1). \tag{6.80}$$

Conversly, if S_1 is conjugate-symmetric and the above equation holds, then we have $\mathbb{F}(S_{1,2}) = S_{1,2}$ [CCH11]. Thus, we see that for any conjugate-symmetric subspace S, setting $S_1 = S$ and $S_2 = U \otimes V \cdot S^{\perp}$ gives an alternative characterisation of the above lemma in terms of only one subspace:

Theorem 6.12. Let $\mathcal{H}_A$ be a Hilbert space. If there exist a subspace $S \subset \mathcal{H}_A \otimes \mathcal{H}_A$ and unitary operators U, V on $\mathcal{H}_A$ satisfying the following conditions:

$$\forall n \in \mathbb{N} : \nexists |\psi\rangle, |\phi\rangle \in \mathcal{H}_A^{\otimes n} : |\psi\rangle \otimes |\phi\rangle \in S^{\otimes n^{\perp}}, \tag{6.81}$$

$$\forall n \in \mathbb{N} : \nexists |\psi\rangle, |\phi\rangle \in \mathcal{H}_A^{\otimes n} : |\psi\rangle \otimes |\phi\rangle \in S^{\perp^{\otimes n^{\perp}}}, \tag{6.82}$$

$$\mathbb{F}(S) = S \tag{6.83}$$

$$\mathbb{F}(U \otimes V \cdot S) = U \otimes V \cdot S, \tag{6.84}$$

$$\exists \{M_i \geq 0\} : \mathbb{M}(S) = \text{span}\{M_i\}, \tag{6.85}$$

$$\exists \{M_j \geq 0\} : \mathbb{M}(U \otimes V \cdot S^{\perp}) = \text{span}\{M_j\}, \tag{6.86}$$

then there exist two quantum channels which individually have no zero-error classical capacity but whose tensor product channel has positive zero-error classical capacity.

All that is left to do now is to prove that a subspace and unitary operators satisfying the above conditions actually exist. This proof is carried out in [CCH11], using some concepts from algebraic geometry. We finish this chapter by briefly sketching the proof.

One considers the set of subspaces which satisfy the third and fourth conditions. Let us denote it by F. A measure can be defined on this set. This measure is the restriction of the measure on the Grassmannian (the set of subspaces of a certain dimension), which is in turn induced from the Haar measure on the unitary group. More details on the construction of this measure can be found in [CS12] and in chapter three of [KP08].

First, the two unitary operators are chosen in a fixed way. It can then be shown that, for these unitaries, the subset of F consisting of all subspaces which also satisfy the first and second conditions is full measure in F. Moreover, it can be shown that the subset of F consisting of all subspaces which also satisfy the fifth and sixth conditions is nonzero measure in F. Lastly, it can be shown that F is not empty. We can interpret these results as follows: the probability that a random subspace in F (i.e. satisfying the third and fourth conditions) also satisfies the first and second conditions is 1, and the probability that it also satisfies the fifth and sixth conditions is nonzero. Since F is non-empty there must therefore exist a subspace satisfying all conditions, and therefore there exist superactivated channels.

The proof makes sure of algebraic geometric notions such as projective varieties, the Zariski topology, the Plücker embedding and complete varieties. This is once again a testament to the wide use of mathematical concepts in quantum information theory.

7

Conclusion

In this thesis two different classical capacities of both classical and quantum channels were studied, thus resulting in a total of four different settings. A classical capacity is a maximum rate at which information can be transmitted over a communication channel per use of the channel, subject to certain conditions concerning the reliability of this transmission. Relations between and properties of these capacities were highlighted and proven.

The first kind of classical capacity that was investigated was the capacity of a channel to transmit information with a probability of error that becomes arbitrarily small when the channel can be used, called the ordinary capacity. This capacity was studied both in the setting of classical and quantum channels. The second kind of classical capacity that was studied is the capacity of a channel to transmit information with zero probability of error. This capacity was also studied for both classical and quantum channels.

The first setting that was studied was the ordinary capacity of classical channels. The pivotal theorem concerning this type of capacity is the noisy channel coding theorem, orginally proven by Claude Shannon in [Sha48]. Two different proofs of this theorem were presented. The first was based on the Markov inequality and the Law of Large Numbers. The second was based on the notion of a typical set. The first of these proofs was simpler and required less prerequisites. However, the second proof could be easily generalised to the case of quantum channels, which was done in chapter 5. Lastly, it was shown that the ordinary capacity of a classical channel is additive. That is, the ordinary capacity of the product of two channels is equal to the sum of the individual ordinary capacities of those channels.

The second setting that was investigated is the zero-error capacity of classical channels. Both a lower and an upper bound of this capacity were proven. The upper bound was given in terms of the ordinary capacity. In order to prove the lower bound, the zero-error capacity of a classical channel was studied by considering the adjacency graph of that channel. This is a graph which contains the information about the distinguishability of symbols which are sent through the channel. It was shown that the zero-error capacity of a classical channel can be characterised completely in terms of its adjacency graph. Lastly, it was explained that the zero-error capacity of a classical channel is superadditive. That is, there exist classical channels for which the zero-error capacity of their product channel is greater than the sum of their individual zero-error capacities.

The third setting that was investigated was the ordinary classical capacity of quantum channels. The main theorem on this capacity was the Holevo-Schumacher-Westmoreland theorem, which can be considered to be a generalisation of the noisy channel coding theorem that was studied in chapter two. This theorem was proven using the notion of a typical subspace, following a similar approach as in the proof of the noisy channel coding theorem. However, it was found that there are several fundamental complications that arise when considering quantum channels rather than classical channels. First and foremost, quantum channels exhibit entanglement. This requires the introduction of the regularisation of the Holevo information in the expression of the ordinary classical capacity of a quantum channel. This regularisation reflects the fact that we are unable to approximate the ordinary classical capacity of a quantum channel well. Instead, we must resort to an expression containing the limit of infinitely many channel uses. This is different than the case of classical channels, where the ordinary capacity could simply be expressed in terms of the mutual information of the random variables representing the input and output of the channel. The need for regularisation shows that further work is needed to determine in what situations the use of entanglement can increase the ordinary classical capacity of a quantum channel.

Another complication which arose in the study of the ordinary capacity of quantum channels was the phenomenon of collective measurement. When several quantum states are transmitted over several copies of a quantum channel, one measures these states collectively, using a collective POVM. This is radically different for classical channels, where there is no difference between individual and collective measurement. In order to account for collective measurement when coding for quantum channels, the packing lemma was proven and used. This lemma describes how classical information can be stored into quantum states, allowing for retrieval of this classical information using a POVM with a relatively small probability of error. The packing lemma was subsequently used to prove the Holevo-Schumacher-Westmoreland theorem.

The fourth and last setting that was studied was the zero-error classical capacity of quantum channels. This zero-error classical capacity was, in a similar fashion to the case of classical channels in chapter three, characterised equivalently in terms of the characteristic graph of a quantum channel. This graph-theoretical formulation was then used to prove that the zero-error classical capacity of a quantum channel can be achieved using only pure quantum states. Furthermore, it was shown that the zero-error classical capacity is bounded from above by the ordinary classical capacity.

The investigation of the zero-error classical capacity of quantum channels was then concluded with a study of a phenomenon unique to this specific classical capacity: superactivation. Two quantum channels are said to be superactivated if they individually have no zero-error classical capacity, but together have positive zero-error classical capacity. It was proven that two superactivated quantum channels exist if there exists a subspace of a product Hilbert space and two unitary operators satisfying several conditions. This proof made heavy use of the theory of conjugate-divisible maps, which were introduced in the same chapter. Finally, the proof showing that the appropriate subspace and unitary operators do indeed exist was sketched.

Besides the theorem on the different classical capacities that were presented in this thesis, a less explicit result was found: the diversity of the mathematics of quantum information theory. Indeed, many different areas of mathematics were found to have important applications in studying classical capacities, most notably probability theory, graph theory, abstract algebra and algebraic geometry. The application of this wide range of mathematical disciplines was a welcome surprise, because it only added to the intrigue of quantum information theory.

A straightforward topic for further research could be the generalisation of the results in this thesis to quantum capacities. Indeed, quantum channels can be used to transmit quantum information instead of classical information, thus allowing one to define quantum capacities. It would be interesting to see what definitions and results from this thesis have a quantum capacity-counterpart. Furthermore, a possible topic for further research is an explicit construction of two quantum channels whose zero-error classical capacity is superactivated. For as far as we know, there is no such an explicit description - the proof of the existence of superactivated channels presented in chapter 6 was implicit. Lastly, further research can be conducted on determining in what situations entanglement increases the ordinary classical capacity of a quantum channel. This remains one of the major open problems in quantum information theory.

7.1. Acknowledgements

First of all, I would like to thank my supervisors Dr. D. Elkouss and Dr. B. Janssens for their guidance. They gave me useful feedback on all parts of my thesis and helped me in understanding difficult concepts. Moreover, they allowed me the freedom to pursue the topics that I found most interesting and gave me the time to properly investigate these. Furthermore, I wish to thank my family for supporting and hosting me during these strange last months. I want to specifically thank my mother for providing an endless supply of amazing coffee. Lastly, I wish to thank my friends and fellow students Pepijn Klooster and Uki Ognjanovic for pre-reading my thesis and giving valuable feedback.

Bibliography

[Alo98] Noga Alon. The shannon capacity of a union. *Combinatorica*, 18:301–310, 1998.

[CCH11] T.S. Cubitt, Jianxin Chen, and A.W. Harrow. Superactivation of the asymptotic zero-error classical capacity of a quantum channel. *Information Theory, IEEE Transactions on*, 57:8114–8126, 12 2011.

[CS12] T. S. Cubitt and G. Smith. An extreme form of superactivation for quantum zero-error capacities. *IEEE Transactions on Information Theory*, 58(3):1953–1961, 2012.

[GdAM16] Elloá B. Guedes, Francisco Marcos de Assis, and Rex A. C. Medeiros. Quantum zero-error information theory. In *Springer International Publishing*, 2016.

[Has09] Matthew B. Hastings. Superadditivity of communication capacity using entangled inputs. 2009.

[Hol19] Alexander S. Holevo. *Quantum Systems, Channels and Information.* Walter de Gruyter, Berlin/Boston, 2019.

[KP08] Steven G. Krantz and Harold R. Parks. Geometric integration theory. 2008.

[LF12] Yuval Lomnitz and Meir Feder. A simpler derivation of the coding theorem. 2012.

[Lov79] László Lovász. On the shannon capacity of a graph. *IEEE Trans. Inf. Theory*, 25:1–7, 1979.

[Maa] Hans Maassen. Quantum probability, quantum information theory, quantum computing.

[Sch] Frederic Schuller. Lectures on quantum theory.

[Sha48] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[Sha56] C. Shannon. The zero error capacity of a noisy channel. *IRE Transactions on Information Theory*, 2(3):8–19, 1956.

[SW97] Benjamin Schumacher and Michael D. Westmoreland. Sending classical information via noisy quantum channels. *Phys. Rev. A*, 56:131–138, Jul 1997.

[Wer] Reinhard Werner. Mathematical methods of quantum information theory.

[Wil19] Mark M. Wilde. *From Classical to Quantum Shannon Theory.* Cambridge University Press, Baton Rouge, Louisiana, 2019.

A

Proof of the Hayashi-Nagaoka Lemma

The Hayashi-Nagaoka lemma states:

Lemma A.1 (Hayashi-Nagaoka)**.** Let $S, T \in \mathcal{L}(\mathcal{H})$ be positive semi-definite operators such that $\mathbb{1} - S$ is also postive semi-definite. Then for any $c > 0$ we have:

$$\mathbb{1} - (S+T)^{-\frac{1}{2}} S (S+T)^{-\frac{1}{2}} \leq (1+c)(\mathbb{1} - S) + (2 + c + \frac{1}{c})T. \tag{A.1}$$

Proof: we follow [Wil19]. For any operators $A, B \in \mathcal{L}(\mathcal{H})$ and constant $c > 0$ we have:

$$(A - cB)^\dagger (A - cB) \geq 0. \tag{A.2}$$

This follows directly from the fact that $X^\dagger X \geq 0$ for any $X \in \mathcal{L}(\mathcal{H})$, since $\langle X^\dagger X\psi, \psi\rangle = \langle X\psi, X\psi\rangle \geq 0$. We can equivalently write the above equation as:

$$\frac{1}{c} A^\dagger A + cB^\dagger B \geq A^\dagger B + B^\dagger A. \tag{A.3}$$

Let $R \in \mathcal{L}(\mathcal{H})$ and pick $A = \sqrt{T}R, B = \sqrt{T}(\mathbb{1} - R)$. The above equation then yields:

$$\frac{1}{c} R^\dagger T R + c(\mathbb{1} - R^\dagger) T (\mathbb{1} - R) \geq R^\dagger T (\mathbb{1} - R) + (\mathbb{1} - R^\dagger) T R. \tag{A.4}$$

Here we have used that T is a positive semidefinite operator (and therefore Hermitian). Now consider the following:

$$T = 2R^\dagger T R - 2R^\dagger T R + R^\dagger T - R^\dagger T + TR - TR + T \tag{A.5}$$

$$= R^\dagger T R + R^\dagger T(\mathbb{1} - R) + (\mathbb{1} - R^\dagger) T R + (\mathbb{1} - R^\dagger) T (\mathbb{1} - R) \tag{A.6}$$

$$\leq R^\dagger T R + \frac{1}{c} R^\dagger T R + c(\mathbb{1} - R^\dagger) T (\mathbb{1} - R) + (\mathbb{1} - R^\dagger) T (\mathbb{1} - R) \tag{A.7}$$

$$= (1 + \frac{1}{c}) R^\dagger T R + (1+c)(\mathbb{1} - R^\dagger) T (\mathbb{1} - R). \tag{A.8}$$

Since we let R be arbitrary, we can take $R = (S+T)^{\frac{1}{2}}$. The above equation then becomes:

$$T \leq (1 + \frac{1}{c})(S+T)^{\frac{1}{2}} T (S+T)^{\frac{1}{2}} + (1+c)(\mathbb{1} - (S+T)^{\frac{1}{2}}) T (\mathbb{1} - (S+T)^{\frac{1}{2}}), \tag{A.9}$$

where we have again used that S and T are positive semidefinite and therefore Hermitian. Using that $T \leq S+T$ since $S \geq 0$ and that $S \leq S^{\frac{1}{2}} \leq (S+T)^{\frac{1}{2}}$ since $S \leq \mathbb{1}$, we get:

$$\begin{aligned}
T \leq& \left(1+c^{-1}\right)(S+T)^{1/2}T(S+T)^{1/2} + (1+c)\left(\mathbb{1}-(S+T)^{1/2}\right)(S+T)\left(\mathbb{1}-(S+T)^{1/2}\right) && \text{(A.10)}\\
=&(S+T)^{1/2}\left[\left(1+c^{-1}\right)T+(1+c)\left(\mathbb{1}+S+T-2(S+T)^{1/2}\right)\right](S+T)^{1/2} && \text{(A.11)}\\
=&(S+T)^{1/2}\left[\left(2+c+c^{-1}\right)T+(1+c)\left(\mathbb{1}+S-2(S+T)^{1/2}\right)\right](S+T)^{1/2} && \text{(A.12)}\\
\leq&(S+T)^{1/2}\left[\left(2+c+c^{-1}\right)T+(1+c)(\mathbb{1}+S-2S)\right](S+T)^{1/2} && \text{(A.13)}\\
=&(S+T)^{1/2}\left[\left(2+c+c^{-1}\right)T+(1+c)(\mathbb{1}-S)\right](S+T)^{1/2}. && \text{(A.14)}
\end{aligned}$$

Multiplying this inequality both from the left and the right by $(S+T)^{-\frac{1}{2}}$ gives:

$$(S+T)^{-\frac{1}{2}}T(S+T)^{-\frac{1}{2}} \leq (2+c+c^{-1})T+(1+c)(\Pi_{S+T}-S), \tag{A.15}$$

where Π_{S+T} is the projector onto the supports of S and T [Wil19]. For this projector we have that $\Pi_{S+T}S\Pi_{S+T} = S$ and $\Pi_{S+T}T\Pi_{S+T} = T$. Thus, the following holds:

$$\begin{aligned}
&\mathbb{1}-(S+T)^{-1/2}S(S+T)^{-1/2} && \text{(A.16)}\\
&=\mathbb{1}-(S+T)^{-1/2}(S+T)(S+T)^{-1/2}+(S+T)^{-1/2}T(S+T)^{-1/2} && \text{(A.17)}\\
&=\mathbb{1}-\Pi_{S+T}+(S+T)^{-1/2}T(S+T)^{-1/2} && \text{(A.18)}\\
&\leq(1+c)(\mathbb{1}-\Pi_{S+T})+\left(2+c+c^{-1}\right)T+(1+c)(\Pi_{S+T}-S) && \text{(A.19)}\\
&=\left(2+c+c^{-1}\right)T+(1+c)(\mathbb{1}-S). && \text{(A.20)}
\end{aligned}$$

This is exactly the statment of the lemma. □