Quantum Shannon Theory Background Chapters

John M. Schanck

May 3, 2011

Contents

List of notations and quantities

Preface

I would like to say that what follows may be read as an introductory text to quantum Shannon theory, but I will not project my aspirations so heavily over what should more properly be seen as a collection of extended notes. My notes from what I can now say, as it draws to an end, has been the most productive and transformative year of my education.

It was less than a year ago that I decided to pursue an understanding of quantum information theory - a decision which I think came as a surprise to many of the people around me. Up until this year I had been working in a very applied fashion in computer science; my interests revolving around distributed systems, cryptography, and anonymity. I was working on privacy enhancing technologies: on my personal project, Anomos, an anonymizing peer-to-peer file-sharing network, and on the Tor project, a more general tool for circumventing traffic analysis on the Internet. The promise of quantum theory to break the encryption on which those tools rely, and to provide a perfectly secure alternative to that encryption, is perhaps what sparked my interest. But other factors certainly came into play. In the spring 2009 semester, Lee Spector's *Unconventional Computing* course introduced me to reversible computation and some basic topics in quantum computation, and Herb Bernstein's *Quantum Mechanics for the Million* served to demystify quantum mechanics and introduced me to its mathematical techniques. These courses, and a fascination with physics going back to my childhood, played a crucial role.

For a year following that spring semester these interests lay dormant while I continued studying classical computer science. Then, in the summer of 2010, I became interested in quantum complexity theory, primarily in how the class **BQP** was related to the polynomial hierarchy, and began perusing the available literature. I was woefully out of my depth with much of what I found, but gleaned what I could from (Ethan) Bernstein and Vazirani's work [BV97], and a few of Scott Aaronson's papers. Around this time Herb suggested I try a gentler introduction to the field and leant me a copy of David Mermin's "Quantum Computer Science," [Mer07] from which I taught myself for the remainder of the summer.

A fortuitous conversation with a friend, just before the start of the fall semester, inspired me to formalize my knowledge of classical information theory. And so I enrolled in a course on the topic at the University of Massachusetts. The topic of the present work would not be quantum Shannon theory if not for this decision. Throughout the semester, while learning the classical theory in class and from [Cov06], I was self-teaching quantum computation and information from Nielsen and Chuang [NC00] and primary sources. In searching for

a Division III topic I meandered through quantum algorithms, quantum complexity, and even back into classical communication complexity; but through all that wandering, the profoundly positive experience I had studying classical information theory kept drawing me back towards Shannon theoretic topics. A discussion with Herb, in December, about the quantum reverse Shannon theorem finally pinned me down.

My concurrent interest in communication complexity lead me to the proof of the "one-shot" classical reverse Shannon theorem in [HJMR07], and so in January my primary focus was on learning the requisite material to prove a similar one-shot version of the quantum reverse Shannon theorem. I don't think I understood at the time the magnitude of the task I was setting out for myself, but I had intentionally picked a lofty target. Without the structure of a traditional course, I've often found that setting seemingly unattainable goals is best way to motivate and direct my learning.

As I explored the literature surrounding topics such as the resource inequality formalism, decoupling theorems, and the known channel coding results, I eventually discovered that Abeyeshinghe *et al.* had proved a one-shot fully quantum reverse Shannon theorem in [ADHW06]. The present work is the result of the research which brought me to that point.

In early discussions about this work with my advisors, the idea was tossed about that it might be written as an introduction to quantum Shannon theory for undergraduates. But as I mentioned at the intro, I prefer to view this as notes which may be more informative to the educator than the student. Perhaps you, as an educator, may view this work as a case study in how a young student might approach quantum information theory. My errors and omissions, from this perspective, indicate the stumbling blocks which must be smoothed over as we work to make this field more comprehensible to a larger audience.

CHAPTER 1

Our Mathematical Toolkit

Of principal importance to the study of quantum Shannon theory is the ability to quantify and compare quantum states in terms of their information content. Much of the development of this field has followed from classical information theory, some familiarity with which is assumed here. We begin by introducing the basic entropic quantities which form the building blocks of the coding theorems and protocols which are to come later. Some classical information theoretic quantities are briefly introduced - primarily for ease of reference. Readers unfamiliar with these quantities are directed to [Cov06] for more rigorous definitions, as well as history and applications.

1.1 Classical information quantities

Entropy Introduced in [Sha48]

$$H(X) = -\mathbb{E}\left[\log p(x)\right] = -\sum_{x\in\mathcal{X}} p(x)\log p(x) \tag{1.1}$$

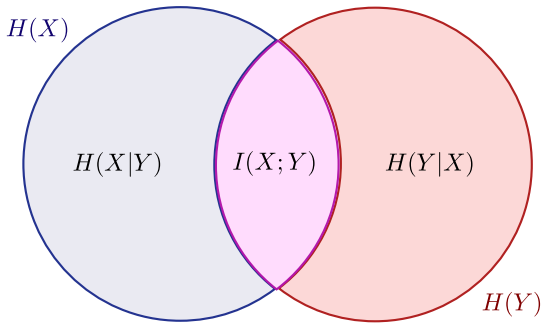

Figure 1.1: An often useful visual aid for remembering the relations between classical entropic quantities. $H(X)$ is the entire left (blue) circle, $H(Y)$ the entire right (red) circle. The union of the two circles is $H(X,Y)$; the intersection $I(X;Y)$; and the difference $H(X|Y)$ or $H(Y|X)$.

Where, by convention, $\log 0$ is taken to be 0 rather than negative infinity. Occasionally one will see $H(X)$ with a subscript, such as $H_2(X)$ or $H_e(X)$, which indicates the base of the logarithms appearing in the expression and subsequently the units of the result. For example, a uniform distribution over 16 values has binary entropy $H_2(X) = -\sum_{x=0}^{15} \frac{1}{16} \log_2 \frac{1}{16} = 4$. As we will see in section 2.1, we can therefore represent samples of this random variable with 4 *b*inary dig*it*s, and so we say that H_2 has units of *bits*. $H_e(X) = \frac{1}{\log_2 e} H_2(X)$ and has units of *nats*; for arbitrary d, $H_d(X)$ has units of *dits*. We will almost always write H without a subscript; logarithms are assumed to be base two unless otherwise noted.

Throughout this text we will use capital letters (X) to denote random variables, calligraphic letters ($\mathcal{X}$) to denote the sample space of a random variable, and lowercase letters (x) to denote individual samples.

Relative Entropy Also known as the Kullback-Leibler- or KL-Divergence, the relative entropy is a distance pseudo-metric[1] between probability distributions. It can be used, for instance, to tell us the amount of uncertainty (in bits) which is introduced by the inappropriate use of one distribution to describe events actually drawn from a different distribution.

$$D(p\|q) = \mathbb{E}\left[\log \frac{p(x)}{q(x)}\right] = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \tag{1.2}$$

Mutual Information is another ubiquitous quantity in classical information theory. It quantifies the amount of "overlap" between random variables, or more explicitly, the amount of information which can be obtained about one random variable from a second. The KL-divergence provides a nice characterization (presented below) of the mutual information as the distance between the joint and factor distributions of X and Y.

$$\begin{aligned} I(X;Y) &= \mathbb{E}\left[\log \frac{p(x,y)}{p(x)p(y)}\right] = D(p(x,y)||p(x)p(y)) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned} \tag{1.3}$$

[1] It fails to be a true metric because it is not symmetric in its arguments $D(p\|q) \neq D(q\|p)$

1.2 Quantum Mechanics

Some background in quantum mechanics is assumed in this text, but we will need surprisingly little of the quantum theory. All that is really necessary is some basic linear algebra, a few rules about the types of objects which constitute valid quantum states, and a few more rules about how these states may be transformed.

1.2.1 Dirac (bra-ket) notation

A pure state is represented by a *ket*, an $n \times 1$ (column) vector of complex numbers. The pure states we will most frequently encounter in this text are those of the computational basis, the *zero* and *one* states of a qubit.

$$|0\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad\qquad |1\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{1.4}$$

The set of all pure quantum states (in a two-dimensional space) is given by the unit circle on the complex plane. In other words,

$$|\psi\rangle = \alpha\,|0\rangle + \beta\,|1\rangle \tag{1.5}$$

is a valid state if and only if $|\alpha|^2 + |\beta|^2 = 1$. Once a basis has been chosen, such as the computational, any pure state which may be written as a linear combination of the basis kets is a *superposition* state. For instance,

$$\frac{(|0\rangle + |1\rangle)}{\sqrt{2}} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \qquad\qquad \frac{(|0\rangle - |1\rangle)}{\sqrt{2}} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix} \tag{1.6}$$

are two possible superpositions of the zero and one states.

To each pure state ϕ-ket, $|\phi\rangle$, corresponds a unique dual state ϕ-*bra*, $\langle\phi|$. A bra is obtained by applying the conjugate transpose operation † to a ket. Suppose $|\phi\rangle = \begin{bmatrix} a+bi \\ c+di \end{bmatrix}$, then

$$|\phi\rangle^\dagger = \langle\phi| = [(a+bi)^*, (c+d)^*] = [a-bi, c-di]\,, \tag{1.7}$$

wherein * denotes complex conjugation.

A bra may be seen as a linear functional taking quantum states to *probability amplitudes*

- complex numbers α for which $0 \leq |\alpha|^2 \leq 1$ may be interpreted as a probability. For a pure state $|\psi\rangle$, $\langle\psi|\psi\rangle = 1$ and $\langle\psi^\perp|\psi\rangle = 0$, where $|\psi^\perp\rangle$ is a state orthogonal to $|\psi\rangle$.

The *bracket*, $\langle\cdot|\cdot\rangle$, and the inner product $\langle\cdot,\cdot\rangle$ found in functional analysis share more than a superficial similarity in notation. A bracket is precisely the inner product defined on the space of quantum states, which is itself a complex Hilbert space.

1.2.2 Kronecker product

We can discuss systems of more than one qubit by taking the direct product of the individual qubits. The matrix direct product goes by the alternative name of the Kronecker product (symbolically, $\otimes$) and takes matrices of arbitrary size to *block matrices*, each element of which is itself a matrix. The block matrix representation is rarely used, as the Kronecker product of an $i \times j$ and a $k \times l$ matrix can also be represented as an $ik \times jl$ matrix.

$$|0\rangle \otimes |0\rangle = \begin{bmatrix} 1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ 0 \begin{bmatrix} 1 \\ 0 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{bmatrix} \equiv \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{1.8}$$

1.2.3 Density matrices

Dirac's bra-ket notation is exceedingly useful in the description of pure states and their evolution. However, we also desire to discuss probabilistic ensembles of pure states, or *mixed states*, and the bra-ket notation leaves something to be desired in such settings. A *density matrix*, on the other hand, gives a complete statistical representation of a quantum state, pure or mixed.

From a pure state, we obtain a density matrix by taking the kronecker product on the right by the state's dual (the kronecker product symbol is rarely written).

$$|0\rangle \otimes \langle 0| \equiv |0\rangle\langle 0| = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \qquad |1\rangle\langle 1| = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \qquad \frac{(|0\rangle + |1\rangle)(\langle 0| + \langle 1|)}{2} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \tag{1.9}$$

A mixed state is simply a probability distribution over a set of pure states. As such a

mixed state may be represented as a convex combination of pure states:

$$\rho = \sum_i p_i \left|x_i\right\rangle \left\langle x_i\right|. \tag{1.10}$$

Since both pure and mixed states may be represented by density matrices, it is useful to have a method for distinguishing them. This is fortunately quite simple, a pure state always satisfies $\mathrm{Tr}\left(\rho^2\right) = 1$ and a mixed state $\mathrm{Tr}\left(\rho^2\right) < 1$.

When we discussed pure states we gave a simple criterion for whether or not a given state was valid. Specifically, if $\left|\psi\right\rangle = \sum_i \mu_i \left|i\right\rangle$, then $\left|\psi\right\rangle$ is a valid state if and only if $\sum_i \left|\mu_i\right|^2 = 1$. Following from this restriction on pure states, and the general construction of density matrices, there are a few criteria by which we may determine if a given matrix is a valid density operator. Density matrices are completely characterized by three properties, they are Hermitian ($\rho = \rho^\dagger$), trace-one ($\mathrm{Tr}\left(\rho\right) = 1$), and positive semi-definite ($\left\langle\phi\right| \rho \left|\phi\right\rangle \geq 0$ for all $\left|\phi\right\rangle$).

The first of these is immediate from the construction of density matrices as $\left(\left|\phi\right\rangle \left\langle\phi\right|\right)^\dagger = \left\langle\phi\right|^\dagger \left|\phi\right\rangle^\dagger = \left|\phi\right\rangle \left\langle\phi\right|$. This generalizes to mixed states since the conjugate transpose operation is distributive.

The matrix trace is a linear function, $\mathrm{Tr}\left(p_1 A + p_2 B\right) = p_1 \mathrm{Tr}\left(A\right) + p_2 \mathrm{Tr}\left(B\right)$, so the trace-one condition simply guarantees that the probability distribution of the mixture is normalized to 1. $\mathrm{Tr}\left(\rho\right) = \sum_i p_i \mathrm{Tr}\left(\rho_i\right) = \sum_i p_i$.

Finally, we note that every density matrix has an eigenvalue decomposition $\rho = \sum_i \lambda_i \left|i\right\rangle \left\langle i\right|$. A matrix is positive semi-definite if its eigenvalues (λ_i) are all greater than or equal to zero. Ensuring that a density matrix is positive semi-definite then simply ensures that it is a valid convex combination of the eigenstates ($\left|i\right\rangle$).

1.2.4 Joint systems

Suppose we have a state distributed across two systems, A and B. This is the case, for instance, when two parties, Alice and Bob, share an entangled state. Alice and Bob's individual particles reside on the Hilbert spaces $\mathcal{H}_A$ and $\mathcal{H}_B$ respectively, and their joint state may be said to reside on the tensor product space $\mathcal{H}_A \otimes \mathcal{H}_B$. We designate with a superscript the systems on which a state resides; Alice and Bob's shared entanglement may be represented as Φ^{AB}.

For concreteness, let us suppose that A and B are both single-qubit systems. Each therefore has a representation in the computational basis, $\{\left|0\right\rangle, \left|1\right\rangle\}$. The joint system is

represented in the basis $\{|0^A\rangle, |1^A\rangle\} \times \{|0^B\rangle, |1^B\rangle\} = \{|0^A 0^B\rangle, |0^A 1^B\rangle, |1^A 0^B\rangle, |1^A 1^B\rangle\}$

1.2.5 Partial Trace

Given a state on multiple systems, we may wish to determine the statistical structure of that state on just one of its subsystems. The action which gives the *reduced density operator* of a multi-partite system is the *partial trace*. It is often referred to casually as "tracing out" or "tracing over" the discarded system(s). For instance, tracing out subsystem B of the bipartite state ρ^{AB} is denoted:

$$\rho^A = \mathrm{Tr}_B\left(\rho^{AB}\right) \tag{1.11}$$

Calculating the partial trace requires our knowing the bases in which each subsystem is represented[2]. It is convenient to treat these as ordered bases, so that we may index a set of n basis vectors by the first n natural numbers. Assuming that the B subsystem of ρ^{AB} has rank m, the value at the ith row and jth column of ρ^A is given by:

$$(\rho^A)_{i,j} = \sum_{k=0}^{m} \left\langle i^A k^B \right| \rho^{AB} \left| j^A k^B \right\rangle \tag{1.12}$$

As a more illustrative example, consider the following. The matrix on the right hand side is a two-qubit state, ρ^{AB} and the left hand side is the reduced density matrix ρ^A.

$$\begin{bmatrix} a_{00,00}+a_{01,01} & a_{00,10}+a_{01,11} \\ a_{10,00}+a_{11,01} & a_{10,10}+a_{11,11} \end{bmatrix} = \mathrm{Tr}_B\left(\begin{bmatrix} a_{00,00} & a_{00,01} & a_{00,10} & a_{00,11} \\ a_{01,00} & a_{01,01} & a_{01,10} & a_{01,11} \\ a_{10,00} & a_{10,01} & a_{10,10} & a_{10,11} \\ a_{11,00} & a_{11,01} & a_{11,10} & a_{11,11} \end{bmatrix}\right) \tag{1.13}$$

1.2.6 Purification

It is a well known and often used fact within quantum information theory that any mixed state may be seen as the reduced density matrix of a pure state on a larger Hilbert space. This is accomplished mathematically by introducing a fictitious *reference* system. A mixed

[2]Unless ρ^{AB} is a product state which factors across the systems, in which case $\mathrm{Tr}_B\left(\rho^A \otimes \rho^B\right) = \rho^A \mathrm{Tr}\left(\rho^B\right) = \rho^A$. The last equality follows because ρ^B, as a density matrix, must have unit trace

state ρ^A is taken to be the reduced density matrix of a pure state on ρ^{AR}, i.e.

$$\rho^A = \mathrm{Tr}_R\left(\rho^{AR}\right) \tag{1.14}$$

ρ^{AR} is obtained from ρ^A in a process called *purification*. A state is purified by first taking its eigenvalue decomposition:

$$\rho^A = \sum_i \lambda_i \left|i\right\rangle \left\langle i\right| \tag{1.15}$$

Then, by choosing any orthonormal basis for the reference system, R (in the following example we use the eigenbasis of ρ^A), the purified state is constructed as

$$\left|\psi\right\rangle = \sum_i \sqrt{\lambda_i}\left|i\right\rangle^A \otimes \left|i\right\rangle^R. \tag{1.16}$$

The density matrix of which is given by

$$\left|\psi\right\rangle\left\langle\psi\right| = \sum_i \sum_j \lambda_i \left|i\right\rangle^A \left|i\right\rangle^R \left\langle j\right|^A \left\langle j\right|^R. \tag{1.17}$$

Purifications of mixed states are not unique, as can easily be seen by considering that we were free to choose the basis in which to represent the reference system in 1.16.

1.2.7 Evolution

Until now we've made no mention of the fact that systems seldom, if ever, stay entirely isolated. They interact with other systems, are acted on by external forces, and otherwise change from moment to moment. In quantum information theory this change is often referred to as the *evolution* of the system, and much like the evolution of biological systems we will not want to attach particularly positive, nor negative, connotations to the term. Evolution is meant to be a fantastically general concept, describing everything from the most purposeful and finely controlled of experimental operations, to the most unpredictable and chaotic of interactions with the environment.

The evolution of closed quantum systems is described by *unitary transformations*. Mathematically, such transformations are given by $n \times n$ complex matrices, U, for which $UU^\dagger = I$. This definition immediately gives rise to some simple observations. First, $U^\dagger$ is also a unitary transformation ($U^\dagger(U^\dagger)^\dagger = U^\dagger U = I$) and as such can be realized by a physical device. Since $U^\dagger = U^{-1}$, U and $U^\dagger$ are both *reversible* operations.

Let us consider some examples of unitary transformations. Among the most commonly encountered in quantum information theory are the Pauli ($\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$), and Hadamard ($\mathbf{H}$) single-qubit gates[3].

$$\mathbf{X} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad \mathbf{Y} = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix} \qquad \mathbf{Z} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \qquad \mathbf{H} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \tag{1.18}$$

The Pauli $\mathbf{X}$ gate can be seen as the quantum analog of the boolean `NOT` operator which takes 0 to 1 and vice versa. Certainly $\mathbf{X}|0\rangle = |1\rangle$ and $\mathbf{X}|1\rangle = |0\rangle$, but as a result of unitarity, $\mathbf{X}$ has a similar action on superpositions of these states. For instance, if $|\phi\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$, then $\mathbf{X}|\phi\rangle = \frac{1}{\sqrt{2}}(|1\rangle - |0\rangle)$. $\mathbf{Z}$ can be seen to act similarly on the phases of a state, in fact the two can be seen to be dual under the change of basis induced by the Hadamard transformation. $\mathbf{Z} = \mathbf{HXH}$.

We will frequently work with density matrices instead of kets. The evolution of a density matrix is given by

$$\rho' = U\rho U^\dagger. \tag{1.19}$$

Which can easily be seen by noting that U is linear and that ρ has a representation as a convex combination of pure states. To simplify the following expression, we drop the probabilities in the sum and write the states $|\phi_i\rangle$ as subnormalized quantities $|\hat{\phi}_i\rangle = \sqrt{p_i}|\phi_i\rangle$.

$$\begin{aligned} U\rho U^\dagger &= U\left(\sum_i |\hat{\phi}_i\rangle\langle\hat{\phi}_i|\right)U^\dagger \\ &= \sum_i U|\hat{\phi}_i\rangle\langle\hat{\phi}_i|U^\dagger \\ &= \sum_i U|\hat{\phi}_i\rangle(U|\hat{\phi}_i\rangle)^\dagger \end{aligned} \tag{1.20}$$

We leave the discussion of open quantum systems and interaction with the environment to chapter 3.

1.2.8 No cloning theorem

One of the major differences between classical and quantum systems is that arbitrary quantum states cannot be copied. This fact was shown in [WZ82] to follow directly from the

[3] In addition to being unitary these matrices are also self-adjoint or *Hermitian*, in other words, $U = U^\dagger$. As such, $\mathbf{X}^2 = \mathbf{Y}^2 = \mathbf{Z}^2 = \mathbf{H}^2 = I$. This is by no means a necessary condition.

linearity of quantum mechanics - it is one of the most fundamental results in quantum information theory, and many important theorems hinge upon it.

We give a brief proof similar to that found in [WZ82], although many others exist.

Suppose there exists a qubit cloning operation U. For such an operation to be physical, i.e. described by quantum mechanics, it must be unitary. Without loss of generality we assume that U acts on two qubits - the first for the to-be-cloned state and the second, initially in the $|0\rangle$ state, for the result.

Suppose further that U acts as expected on two orthogonal states $|\psi\rangle$ and $|\phi\rangle$. That is to say,

$$\begin{aligned} U(|\psi\rangle \otimes |0\rangle) &= |\psi\rangle\,|\psi\rangle\,, \text{ and} \\ U(|\phi\rangle \otimes |0\rangle) &= |\phi\rangle\,|\phi\rangle\,. \end{aligned} \tag{1.21}$$

Before we proceed, note that U which clone orthogonal states do in fact exist. For instance if $|\psi\rangle = |0\rangle$ and $|\phi\rangle = |1\rangle$ then the controlled-not operator (**cX**), given by

$$\mathbf{cX} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{1.22}$$

meets the specifications of 1.21.

To see that no *general* cloning operation exists we need only ask how U acts on the equal superposition of the orthogonal states $|\psi\rangle$ and $|\phi\rangle$, $\frac{1}{\sqrt{2}}(|\psi\rangle + |\phi\rangle)$. If U were truly a cloning operation the result would be two copies of this state, i.e. $\frac{1}{\sqrt{2}}(|\psi\rangle + |\phi\rangle) \otimes \frac{1}{\sqrt{2}}(|\psi\rangle + |\phi\rangle)$.

However, it can be shown that this is not the case:

$$\begin{aligned} U\left(\frac{|\psi\rangle + |\phi\rangle}{\sqrt{2}} \otimes |0\rangle\right) &= U\left(\frac{|\psi\rangle\,|0\rangle + |\phi\rangle\,|0\rangle}{\sqrt{2}}\right) \\ &= \frac{U\,|\psi\rangle\,|0\rangle}{\sqrt{2}} + \frac{U\,|\phi\rangle\,|0\rangle}{\sqrt{2}} \\ &= \frac{1}{\sqrt{2}}(|\psi\rangle\,|\psi\rangle + |\phi\rangle\,|\phi\rangle) \end{aligned} \tag{1.23}$$

Clearly not the result we would expect from a true cloning operation. The second line follows because U is a linear operator. Since the only assumption made was that U *could* clone two states, we've shown that general unitary cloning operations - and therefore physical cloning operations - cannot exist within a universe described by quantum mechanics.

1.3 Quantum information quantities

von Neumann entropy Recall from classical information theory that Shannon's entropy function $H(X) = -\sum_{x \in \mathcal{X}} p_x \log(x)$ gives us, intuitively, a measure of the uncertainty present in the random variable X. Countless uses for this remarkable equation have been found since its exposition in Shannon's seminal 1948 paper [Sha48], and so it is only natural that an analogue suitable for dealing with quantum states would be sought. The measure most frequently used today is the von Neumann entropy, named after John von Neumann[4] who introduced it in 1927. The von Neumann entropy was first given information-theoretic significance by Holevo [**?**], who proved that it places an upper bound on the *accessible information* of a quantum channel, but it was Schumacher [Sch95] who made the analogy with the Shannon entropy complete.

The von Neumann entropy is expressed as, $S(\rho) = -\mathrm{Tr}\left(\rho \log(\rho)\right)$. A superficial relationship with Shannon's entropy emerges when we represent ρ in a basis in which it becomes diagonal. For instance, given the eigenvalue decomposition $\rho = \sum_x \lambda_x |x\rangle \langle x|$, we may reexpress the von Neumann entropy as the Shannon entropy of ρ's eigenvalues: $S(\rho) = -\sum_x \lambda_x \log(\lambda_x)$.

A major contribution of [Sch95] was to show that the von Neumann entropy of an ensemble gives the expected number of qubits needed to encode states from that ensemble. The process for doing so is now known as *Schumacher compression* - it will be introduced in full in section 2.1.

Quantum joint entropy Much like the classical joint entropy, the quantum joint entropy $S(\rho^A, \rho^B)$ is simply given by the von Neumann entropy of the joint system $S(\rho^{AB})$. Frequently one will see such quantities written as $S(A, B)_\rho$. I will continue with this convention for the remainder of this report.

For a pure state $|\psi\rangle^{ABE}$ the following useful equalities hold:

$$\begin{aligned} S(B, E)_\psi &= S(A)_\psi \\ S(A, E)_\psi &= S(B)_\psi \\ S(A, B)_\psi &= S(E)_\psi \end{aligned} \tag{1.24}$$

[4] von Neumann is, incidentally, responsible for suggesting the term "entropy" to Shannon when he was struggling to name his measure of information [TM71].

More generally we may state that bipartitions of any multi-partite pure state have equal von Neumann entropy. Each portion purifies the other which implies (see section 1.2.6) that they have identical eigenvalues.

Quantum mutual information The quantum mutual information is also very similar to its classical counterpart. $S(A:B)_\rho = S(A)_\rho + S(B)_\rho - S(A,B)_\rho$

Quantum conditional entropy The quantum conditional entropy was misunderstood and a subject of debate for much of the early development of quantum Shannon theory. This is because, quite surprisingly, the intuitive definition following from classical information theory of $S(A|B)_\rho = S(AB)_\rho - S(B)_\rho$ can take on negative values. A major conceptual breakthrough came about in [HOW05] in which an operational definition of the negative conditional information was given. [HOW05] introduces the quantum state merging primitive for reasoning about the transfer of partial information from one party to another. In other words, the amount of quantum information Alice must transfer to Bob in order to put her share of bipartite state ρ^{AB} into his hands. [HOW05] lays out three illustrative examples of this which are summarized here.

Case 1: Alice holds a maximally mixed state, $\rho^A = \frac{1}{2}(|0\rangle\langle 0| + |1\rangle\langle 1|)$ and Bob holds $\rho^B = |0\rangle\langle 0|$. $S(A|B)_\rho = S(A,B)_\rho - S(B)_\rho = 1$. And indeed Alice can give Bob her share with 1 qubit of quantum communication.

Case 2: Alice and Bob share the state $\rho^{AB} = \frac{1}{2}(|00\rangle\langle 00|^{AB} + |11\rangle\langle 11|^{AB})$ The conditional entropy of this classically correlated state, $S(A|B)_\rho = 0$. Alice need only measure her particle in the Hadamard basis $\{\frac{1}{\sqrt{2}}(|0\rangle \pm |1\rangle)\}$ and send the result to Bob to enable him to recreate their state locally.

Case 3: Alice and Bob share a fully entangled state $\rho^{AB} = |\Phi^+\rangle\langle\Phi^+|^{AB} = \frac{1}{\sqrt{2}}((|00\rangle + |11\rangle)(\langle 00| + \langle 11|)^{AB})$. We now encounter a negative conditional entropy, $S(A|B)_\rho = -1$, but this is easily understood in the context of state merging. ρ^{AB} actually gives Alice and Bob some capacity for future transmission of quantum information. Bob, knowing that he shares a Φ^+ state with Alice, is free to create an additional Φ^+ locally and keep the one he shares with Alice for future use.

1.4 Measurement

The phases and amplitudes of quantum states are represented mathematically as complex values with unbounded precision, and so it is tempting to believe that, with a delicate enough experimental apparatus, one could store the entire Library of Congress, tweets and all, in but a single relative phase. Indeed, this is not entirely out of the question.

But measurement is not so delicate a process. Our intervention into a quantum system can only reliably distinguish states which are orthogonal to each other. If we limit ourselves for the moment to qubit systems, we may separate $|0\rangle$ from $|1\rangle$ and $|+\rangle$ from $|-\rangle$, but no measurement can draw a definitive line between $|0\rangle$ and $|+\rangle$ - much less for that matter between $|Wikipedia\rangle$ and $|Brittanica\rangle$ states.

Disappointing as this may be, we will find measurement to be an indispensable tool in our exploration of the quantum realm, so we will now consider its various forms.

1.4.1 Projective Measurement

A *projection operator* is any linear transformation, P, such that $P^2 = P$. A simple example with a clear geometric interpretation is the matrix

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{1.25}$$

Which projects vectors in $\mathbb{R}^3$ onto the xy-plane by zeroing out their z component. Applying this matrix twice (in effect zeroing z twice) has no additional effect, as can be verified by checking that $A^2 = A$.

A *projective measurement* is a collection of orthogonal projection operators $\{P_d\}$ such that $\sum_d P_d = I$. The orthogonality constraint here implies that $P_i P_j = \delta_{ij} P_i$. Where δ_{ij} is the Kronecker delta.

We will often refer to measurement in a particular basis, by which we mean that the projection operators are of the form $|\phi_i\rangle \langle\phi_i|$ where ϕ_i are the linearly independent column vectors of the basis matrix in question. The canonical example is measurement in the computational basis, where, for a single qubit system, the projection operators are $\{|0\rangle \langle 0|, |1\rangle \langle 1|\}$. The Bell basis, which distinguishes between the four Bell states, is also common and has projectors $\{|\Phi^+\rangle \langle\Phi^+|, |\Psi^+\rangle \langle\Psi^+|, |\Phi^-\rangle \langle\Phi^-|, |\Psi^-\rangle \langle\Psi^-|\}$

A measurement of a state ρ which yields a result d results in a new density matrix for

the state,

$$\rho' = \frac{P_d \rho P_d}{\mathrm{Tr}\left(P_d \rho\right)}. \tag{1.26}$$

Here $\mathrm{Tr}\left(P_d\rho\right)$ is the probability, prior to measurement, of obtaining d as the measurement result. Dividing the post measurement state by this probability ensures that the result remains a valid density matrix with trace one.

It should be noted that we could have equally well written the denominator as $\mathrm{Tr}\left(P_d\rho\right)$, which may be somewhat clearer depending on ones familiarity with the properties of matrix trace.[5]

Example: Quantum state discrimination Projective measurements are used extensively in quantum algorithms and communication protocols - it has even been shown that measurement is *complete* for quantum computation, by which it is meant that a quantum computer composed entirely out of single-qubit measurements can be made to have the same power as the general quantum circuit model [RB01].

Most often though, measurements are used to distinguish states so as to extract classical information from them. This can be done perfectly, at a rate of one classical bit per measured qubit, assuming the states are orthogonal. And even though one often hears that non-orthogonal states *cannot* be distinguished, what is really meant is that this optimal rate cannot be achieved. Provided we use a clever measurement scheme, and are willing to accept some probability of error, even non-orthogonal states can sometimes be distinguished.

As an example, let's consider a source which prepares an equal mixture of the distinct, yet non-orthogonal, states $|P\rangle = |0\rangle$ and $|Q\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$. The source will be represented as the mixed state $\rho = \frac{1}{2}(|P\rangle\langle P| + |Q\rangle\langle Q|)$.

In general, the best discrimination scheme using projective measurements is to measure in the $\{|P\rangle, |P^\perp\rangle\}$ basis (in which $|P^\perp\rangle$ is a state orthogonal to $|P\rangle$, and therefore $\langle P|P^\perp\rangle = 0$) [Iva87]. In the specific case of our example, this is equivalent to measuring in the computational basis.

A measurement result of $|P^\perp\rangle$ guarantees that the prepared system is in the $|Q\rangle$ state as there is zero probability that $|P\rangle$ would yield $|P^\perp\rangle$. On the other hand, a measurement result of $|P\rangle$ tells us *nothing* about the prepared system as both $|\langle P|P\rangle|^2$ and $|\langle P|Q\rangle|^2$ are nonzero.

[5] Recall that the trace is invariant under cyclic permutations i.e. $\mathrm{Tr}\left(ABC\right) = \mathrm{Tr}\left(BCA\right) = \mathrm{Tr}\left(CAB\right)$. By the idempotency of projection operators $\mathrm{Tr}\left(P_d\rho\right) = \mathrm{Tr}\left(P_d^2\rho\right)$; invoking the cyclic property of the trace produces $\mathrm{Tr}\left(P_d\rho P_d\right)$.

Since $|Q\rangle$ is prepared with probability $\frac{1}{2}$, $\left|P^{\perp}\right\rangle$ is detected with probability $\frac{1}{2}|\langle P^{\perp}|Q\rangle|^2 = 0.25$.

This is optimal for schemes which perform single-qubit projective measurements on individual samples from the source. It may come as a surprise, however, that we can do better without considering multi-qubit or aggregate measurements. To understand how we'll need to loosen the constraints and consider measurements in their most general form.

1.4.2 Generalized Measurements

A *positive operator valued measure*, or *POVM*, is any collection of positive-semidefinite matrices which sum to the identity. In other words, $\{E_d\}$ is a POVM if $\sum_i E_d = I$ and each element is a Hermitian matrix with eigenvalues all greater than or equal to zero.

We will find it useful to consider each E_d as being equal to $M_d M_d^{\dagger}$. Such a decomposition is guaranteed to exist by virtue of E_d being positive semidefinite; M_d and $M_d^{\dagger}$ can be obtained by the Cholesky decomposition algorithm, though it should be noted that the decomposition is not unique - there are infinitely many "square roots" of E_d.

The probability of a measurement on ρ yielding outcome d is given by $\mathrm{Tr}\left(E_d\rho\right)$ or equivalently $\mathrm{Tr}\left(M_d\rho M_d\dagger\right)$. The post measurement state is given by:

$$\rho' = \frac{M_d \rho M_d^{\dagger}}{\mathrm{Tr}\left(M_d \rho M_d^{\dagger}\right)}. \tag{1.27}$$

Notice that the post-measurement state given here is entirely equivalent to that following a projective measurement iff the POVM elements are projectors. In that case, $E_d = M_d = M_d^{\dagger}$. In general, POVM measurements result in post-measurement states which are not as well behaved as their projective counterparts. Specifically, there is no guarantee that the measurement is repeatable.

It may not be immediately clear that we've gained anything with this added generality. Indeed, the only liberty we've gained is that the elements of $\{E_d\}$ need not be pairwise orthogonal with each other. To see how this can be used to our advantage, let's consider an example.

Example: Unambiguous quantum state discrimination Returning to our example of quantum state discrimination, we will now see how single-qubit generalized measurements

can outperform projection at a discrimination task.

Our ability to distinguish non-orthogonal states is still conditioned on our willingness to accept some level of error. Using projective measurements, this error was presented as misclassification, $(|0\rangle + |1\rangle)/\sqrt{2}$ states were sometimes labelled as $|0\rangle$. This was because the orthogonality constraint on the projection matrices allowed us to have no more measurement results than the system had dimensions. POVM measurements do not have this restriction and so we can create a POVM which has a measurement result for each state we want to classify and an extra result to indicate that no classification can be made.

A scheme which accomplishes this is known as *unambiguous* quantum state discrimination[6]. And the lack of ambiguity isn't the only improvement over the projective measurement scheme. Recall from the previous example that projective measurements gave us a probability of classification no better than $\frac{1}{2}|\langle P^\perp|Q\rangle|^2$. The use of UQSD [Iva87] improves this to $1 - \langle P|Q\rangle$.

Figure 1.2 depicts the UQSD scheme in question. The POVM elements are built from $\left|P^\perp\right\rangle$, $\left|Q^\perp\right\rangle$, and $|err\rangle$ (the error syndrome). Note that the figure is two-dimensional - the dashed lines indicate possible linear combinations. $|P\rangle$ can be written as a linear combination of $\left|Q^\perp\right\rangle$ and the error syndrome, but $|Q\rangle$ has no $\left|Q^\perp\right\rangle$ component. A measurement result of $\left|Q^\perp\right\rangle$ therefore indicates, with certainty, that the measured state is $|P\rangle$. Similarly, a result of $\left|P^\perp\right\rangle$ signals that the measured state is $|Q\rangle$. No information about the system is gained on measuring the error syndrome.

More formally, the POVM elements are defined as:

$$
\begin{aligned}
E_1 &= k_1(I - |P\rangle\langle P|) = k(|1\rangle\langle 1|) \\
E_2 &= k_2(I - |Q\rangle\langle Q|) = k((|0\rangle - |1\rangle)(\langle 0| - \langle 1|)) \\
E_3 &= I - E_2 - E_3
\end{aligned}
$$

The values of k_1 and k_2 here must be chosen to ensure that E_3 is positive semidefinite, however any choice of k_1 and k_2 which meets this criterion is a valid POVM. Our example valued both non-error results equally, and so k_1 and k_2 were taken to be equal. There is nothing restricting one from choosing POVM elements which are biased towards certain outcomes. In general, there is no closed form solution for determining the optimal POVM for a measurement task, in practice, one

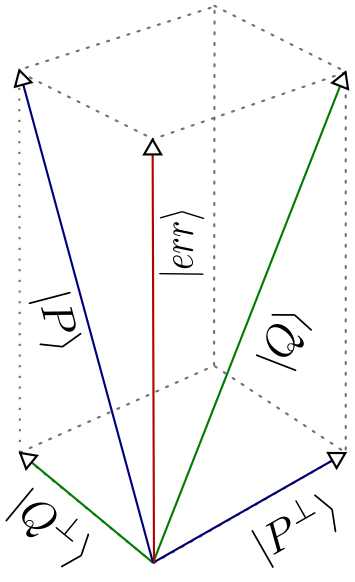

Figure 1.2

[6]UQSD is possible with projective measurements as well but involves the addition of an ancillary system - the use of POVMs is conceptually simpler

is often found numerically using gradient descent or similar techniques [REK05].

In our example, $k_1 = k_2 = \sqrt{2}/(1+\sqrt{2})$. This achieves the optimal probability of correct classification: $1 - \langle P|Q\rangle = 1 - 1/\sqrt{2} \approx 0.293$.

1.5 Distinguishability

Our investigation of measurement has showed us, mathematically at least, how quantum states may be distinguished; it has not, however, shown us how to determine *a priori* whether two states are distinguishable, or for that matter how distinguishable they may be. This problem of distinguishability is intimately related to the very well studied classical problem of distinguishing probability distributions - which is perhaps obvious if we consider that the application of a measurement to density matrix gives rise to a probability distribution on the potential measurement outcomes.

This section describes several distinguishability measures which we will make extensive use of in the chapters to come, for that reason they are presented more rigorously than much of the other material in this text. The reader may choose to read this section lightly and return to it later only to clear up points of confusion. Readers who do not find this section rigorous enough are referred to the excellent explication of these quantities and their applications found in [FvdG97].

1.5.1 Trace Distance

One of the most natural, and simplest to compute, measures of distinguishability is the trace distance:

$$D(\rho, \sigma) = \frac{1}{2}\|\rho - \sigma\|_{tr} \tag{1.28}$$

Where $\|A\|_{tr} = \mathrm{Tr}\left(\sqrt{AA^\dagger}\right)$ is trace norm (also known as the Schatten 1-norm). The factor of $\frac{1}{2}$ constrains the range of D to $[0, 1]$, but is omitted in some texts.

[FvdG97] notes that the trace distance is closely related to the Kolmogorov distance on

classical probability distributions, defined as:

$$K(p,q) = \frac{1}{2}\sum_{x\in\mathcal{X}} |p(x) - q(x)| \tag{1.29}$$

Recalling that the combination of density matrix and measurement yields a probability distribution, the trace distance can be seen as the Kolmogorov distance maximized over all measurements, $\mathcal{E}$, in the set of possible measurements, $\mathcal{M}$. Letting $p_{\mathcal{E}}(\eta)$ be the probability distribution which results from applying $\mathcal{E}$ to η,

$$D(\rho,\sigma) = \max_{\mathcal{E}\in\mathcal{M}} K(p_{\mathcal{E}}(\rho), p_{\mathcal{E}}(\sigma)) \tag{1.30}$$

The optimization here is in general difficult to compute, so 1.28 should be used in practice.

1.5.2 Fidelity

Another frequently used measure of distinguishability is the fidelity. For pure states, the quantum fidelity is a relatively straightforward quantity which can be seen as the "overlap" between the states.

$$F(|\phi\rangle, |\psi\rangle) = |\langle\phi|\psi\rangle|^2 = \cos^2(\theta) \tag{1.31}$$

Where θ is the angle between the vectors. Equation 1.31, henceforth, the pure-state fidelity, generalizes nicely to the fidelity between a pure and a mixed state:

$$F(|\phi\rangle, \sigma) = \langle\phi|\,\sigma\,|\phi\rangle \tag{1.32}$$

Which has a natural interpretation as the average pure-state fidelity between $|\phi\rangle$ and the pure states of the ensemble σ.

Both quantities attain a maximum of 1 when the states are identical, and a minimum of 0 for orthogonal states. Note that this differs from the trace distance which yields 0 for indistinguishable and 1 for perfectly distinguishable states.

The situation is a bit more complex for arbitrary mixed states. Jozsa examined this problem in [Joz94] and set out four reasonable axioms that any mixed-state fidelity should satisfy.

1. The fidelity should be bounded between 0 and 1, and should achieve 1 only when the states are identical. $0 \leq F(\rho,\sigma) \leq 1; F(\rho,\sigma) = 1$ iff $\rho = \sigma$.

2. It should be symmetric in its arguments. $F(\rho, \sigma) = F(\sigma, \rho)$.

3. Equation 1.32 should hold if ρ is pure.

4. The fidelity should be is invariant under unitary transformation, i.e. $F(\rho, \sigma) = F(U\rho U^\dagger, U\sigma U^\dagger)$ for arbitrary unitary transformations U.

He then demonstrated that the following quantity, first introduced by Uhlmann in [Uhl76], satisfies all of these axioms and has several other, noteworthy, properties[7]

$$F(\rho, \sigma) = \left(\mathrm{Tr}\left(\sqrt{\rho^{1/2}\sigma\rho^{1/2}}\right)\right)^2 \tag{1.33}$$

In addition to satisfying Jozsa's axioms, this fidelity measure was shown to be:

a) Multiplicative: $F(\rho_1 \otimes \rho_2, \sigma_1 \otimes \sigma_2) = F(\rho_1, \sigma_1)F(\rho_2, \sigma_2)$

b) Non-decreasing: Following any measurement of ρ and σ with resulting density matrices ρ', σ'; $F(\rho', \sigma') \geq F(\rho, \sigma)$

Equation 1.33 is likely to bewilder the first-time viewer, and though a proof will not be given here, two equivalent statements will be presented which may provide insight. Throughout this work we will use equation 1.33 as the definition of fidelity, as it is easiest to calculate, but the reader is encouraged to think of fidelity as one of the following, more conceptually transparent, quantities.

The first is given by *Uhlmann's theorem*, which states that the fidelity of mixed states is equal to the maximum fidelity between purifications of those states [Uhl76]. As we learned in section 1.2.6, any mixed state may be "purified" - made into a pure state - by introduction of a purely mathematical reference system. Uhlmann's theorem shows that we can use these purifications, and equation 1.31, to evaluate the fidelity of mixed states. Specifically, letting $|\psi_\rho\rangle$ be a fixed purification of ρ and letting $|\psi_\sigma\rangle$ range over all possible purifications of σ,

$$F(\rho, \sigma) = \max_{|\psi_\sigma\rangle} F(|\psi_\rho\rangle, |\psi_\sigma\rangle) = \max_{|\psi_\sigma\rangle} |\langle\psi_\rho|\psi_\sigma\rangle|^2 \tag{1.34}$$

The maximization is necessitated by the fact that purifications are not unique. Without it, various choices of $|\psi_\rho\rangle$ and $|\psi_\sigma\rangle$ would yield different fidelities.

A third, equivalent, definition of fidelity is given by:

$$F(\rho, \sigma) = \min_{\{E_b\}} \left(\sum_b \sqrt{\mathrm{Tr}\left(\rho E_b\right)}\sqrt{\mathrm{Tr}\left(\sigma E_b\right)}\right)^2 \tag{1.35}$$

[7] Some works, notably [NC00], define fidelity as the square root of equation 1.33 - this has caused sufficient confusion to make footnotes such as the current appear almost as frequently as the fidelity.

Wherein the minimization is over all possible measurements (see section 1.4.2) and the term minimized is the square of a measure of the overlap between probability distributions known as the Bhattacharyya coefficient. A proof of this formula's equivalence to 1.33 is given in [FC96]. The Bhattacharyya coefficient can be seen as the dot product of unit-length probability vectors, and as such has a simple geometric interpretation as the cosine of the angle between those vectors. This gives us perhaps the most intuitive definition yet - the fidelity is a measure of how indistinguishable two states remain when the best possible measurement is used to distinguish them.

The optimizations in the later two definitions make them much more difficult to compute than equation 1.33, making 1.33 the preferred definition.

There is a close relationship between fidelity and trace distance. many proofs make use of the following relationship between the two, first introduced in [FvdG97].

$$1 - \sqrt{F(\rho, \sigma)} \leq D(\rho, \sigma) \leq \sqrt{1 - F(\rho, \sigma)} \tag{1.36}$$

CHAPTER 2

Representing Information

Much of what we've covered thus far is common ground for the entire field of quantum information theory. We now begin introducing some uniquely "Shannon theoretic" ideas and techniques - source coding in the present chapter and channel coding in chapter 3. We begin by reviewing the classical cases as they will provide intuition and motivation for their quantum analogs.

2.1 Classical Source Coding

Suppose we want to encode a series of signals for storage on a binary medium - such as a computer's hard disk. For sake of concreteness, we say that the signals are produced by a source X which outputs individual symbols from an alphabet $\mathcal{X}$ according to the probability distribution $p(X)$. For now we limit ourselves to classical information sources with alphabets containing a finite number of completely distinguishable symbols[1]. For example, the source may output decimal digits, the Roman alphabet, or Hangul jamo. This is opposed to quantum information sources, the alphabets of which generally contain indistinguishable, superposed, and unclonable symbols.

Since our disk can only store zeros and ones we need an encoding function mapping the source alphabet to sequences of binary, and a decoding function mapping binary sequences back to the source alphabet. *Source coding* addresses the problem of constructing these functions - often with the goal of representing the signals in a more efficient (i.e. compressed) form than in which they were originally presented.

One very straightforward, but often inefficient, way to encode signals is to assign each element of $\mathcal{X}$ a unique value between 0 and $|\mathcal{X}|-1$. To ensure easy decoding, small values may be padded with zeros so that all of the *codewords* have the same length. For example, the signals from a source with an eight letter alphabet might be encoded as $\{000, 001, 010, 011, \ldots, 111\}$. Using this indexing method, the *expected length* of each codeword is exactly $\lceil \log |\mathcal{X}| \rceil$ bits. Clearly this upper bounds the expected code length of any finite alphabet source - but as

[1] Continuous alphabet sources are commonly studied, but we exclude them here.

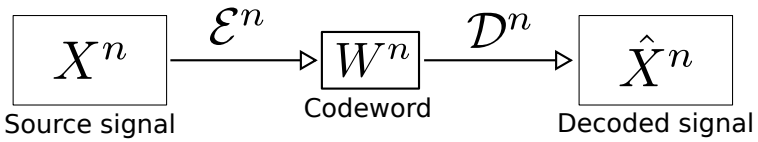

Figure 2.1: Source coding of a classical signal. The encoding ($\mathcal{E}^n$) and decoding ($\mathcal{D}^n$) operators are often inverses, in which case $\hat{X}^n = X^n$.

we will see in following section, we can make use of *block coding* and our knowledge of the probability distribution over source signals to create codes with shorter expected lengths.

To understand how, we need to first introduce the concept of *typical sequences*. The theory of typical sequences will allow us, following Shannon, to demonstrate a fundamental limit on the compressibility of sources and will shed more light on why the entropy function is a good measure of information.

2.1.1 Typical sequences

To gain intuition, let's introduce a fictional unit, the JLB, defined as a book from Jorge Luis Borges' short story *The Library of Babel*. These books are four hundred and ten pages long, forty lines to the page, and eighty characters to the line [Bor98]. Let's say the alphabet contains 32 characters, including spaces and punctuation. That makes $32^{(410*40*80)}$ unique JLBs each requiring something on the order of 6.56 million bits, or about 800 kilobytes of ASCII text to uniquely index. Almost all such JLBs are gibberish, but of course all of last week's best sellers, every book you've ever read, and everything you've ever written would also make an appearance somewhere in the mix (perhaps as multiple volumes, or with some special characters stripped out).

Suppose we're trying to compress a source, X, which outputs random JLBs. If $p(X)$ is uniform over all JLBs, then we can do no better than the indexing method described above. But if $p(X)$ is instead a suitable function of how many copies of the book were sold online last week, we can uniquely identify each JLB we're likely to encounter by its title, author, and date of publication - no more than a few hundred bits. Even better, we might use the book's ISBN, only 34 bits!

In this example, the top sellers, all of the classics, and likely every book you've ever read, are members of what we'll soon come to know as the *typical set*, whereas the overwhelming

majority of all possible JLBs belong to the *atypical set*.

An important thing to keep in mind is that the typical set is always defined with respect to a probability distribution. The choice of number-of-copies-sold was not arbitrary, it was an essential choice for us to be able to use the ISBN as a good code.

Formally, given a probability distribution, $p(x)$, the ϵ-typical set of length n sequences, $\mathcal{A}_\epsilon^{(n)}$, is:

$$\mathcal{A}_\epsilon^{(n)} = \left\{ (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n : \left| -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i) - H(X) \right| < \epsilon \right\}. \tag{2.1}$$

The *typicality* of this set is due to the information theoretic analog of the weak law of large numbers: the asymptotic equipartition property[2]. The AEP states that as n increases, the average log probability of n independent and identically distributed (i.i.d.) symbols, $-\frac{1}{n}\sum_{i=1}^{n} \log p(x_i)$, converges (due to the weak law of large numbers) to the expected log probability, $-\mathbb{E}\left[\log p(X)\right]$, which is exactly the entropy, $H(X)$. This fact will form the backbone of our proof of Shannon's source coding theorem, as it effectively states that approximately $nH(X)$ bits are sufficient to represent any typical sequence.

The AEP allows us to make some other general statements about the properties of typical sets. Most importantly we can bound both the size and the probability mass of a typical set. We state these facts briefly here, without formal proof. Detailed explanations may be found in [Cov06].

First we note that the probability mass of the typical set (the sum probability of its members) can be made arbitrarily close to 1. Fix any $\delta > 0$, the convergence guaranteed by the AEP ensures that for sufficiently large n, the probability that an i.i.d. length n sequence is δ-typical is greater than $1 - \delta$.

$$\Pr\left\{ \left| -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i) - H(X) \right| < \delta \right\} \geq 1 - \delta \tag{2.2}$$

Which implies that block coding can be used to push the probability of encountering an atypical sequence arbitrarily close to zero.

The definition of typicality gives upper and lower bounds for the probabilities of individual typical set elements. No sequence has probability less than $2^{-n(H(X)+\epsilon)}$, nor greater than $2^{-n(H(X)-\epsilon)}$. These element-wise probabilities can be used to determine bounds on the size of the typical set. An upper bound is obtained by assuming that every element has minimal probability, and that the sum of these probabilities is close to 1. Similarly, a lower bound is

[2]The presentation given here is very informal, see [Cov06] for a very readable, yet rigorous, treatment of the AEP and typicality

given by assuming that each element has maximal probability and that their sum probability is only $(1-\epsilon)$. Therefore,

$$(1-\epsilon)2^{n(H(X)-\epsilon)} \leq |\mathcal{A}_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)} \tag{2.3}$$

2.1.2 Shannon's source coding theorem

Shannon used the theory of typical sequences in his groundbreaking work "A Mathematical Theory of Communication" [Sha48] to place a fundamental limit on the compressibility of sources. The theorem he derived is now known as Shannon's source coding theorem (or alternatively, for reasons which will be explained in chapter 3, as his noiseless channel coding theorem).

One goal of source coding is to decrease the expected length of messages transmitted by the source. The key word here is expected - the reason we're able to compress the source at all is because some messages are more likely than others. If we assign the shortest codes to the most likely messages, then the expected length of a message should decrease. However, the unlikely messages must either remain lengthy or they must fail to be decompressable[3].

The source coding theorem states that for any $\delta > 0$, $H(X) + \delta$ bits per symbol is an achievable compression rate. This can be proved as follows.

Put members of the typical set in a one-to-one correspondence with the integers between zero and $2^{nH(X)}$. Assign each member of the atypical set a value in $\left[2^{nH(X)}, 2^{n|\mathcal{X}|}\right]$. The codewords for typical (atypical) set members are obtained by prepending a 0 (1) to their assigned value. The length of codewords is therefore $nH(X)+1$ for typical sequences, and $n|\mathcal{X}|+1$ for atypical sequences. The expected length per symbol is obtained by dividing these quantities by the block length, n, and taking the expectation with respect to the probability that a sequence is typical:

$$(1-\epsilon)\left(H(X)+\tfrac{1}{n}\right) + \epsilon\left(|\mathcal{X}|+\tfrac{1}{n}\right) = H(X) + \epsilon\left(|\mathcal{X}| - H(X)\right) + \tfrac{1}{n} \tag{2.4}$$

For sufficiently small ϵ and large n (both of which are free parameters) the above quantity is less than $H(X)+\delta$.

Shannon also proved the converse theorem - that for any $\delta > 0$, the compression rate of $H(X) - \delta$ bits per symbol is not attainable. This can be seen by noting that any attempt to

[3]This is proved by invoking the pigeonhole principle. Specifically if we have n messages (pigeons), and only $m < n$ indices by which to address them (pigeonholes), then some indices must correspond to multiple indices and decoding fails to be a proper function.

index the typical set with fewer than $nH(X)$ bits will leave some portion of typical sequences uncoded. These uncoded sequences will receive code lengths longer than $n(H(X) - \delta)$ bits and, unlike atypical sequences, their probability and contribution to the expected code length will not be diminished by increasing n or decreasing ϵ.

The source coding theorem and its converse firmly establish $H(X)$ as an appropriate measure of the information content of classical sources. In the coming sections we will show, by similar arguments, that $S(X)$ plays a similar role for quantum sources.

2.2 Quantum source coding

Following closely in Shannon's footsteps, Benjamin Schumacher introduced a theory of typical subspaces, a quantum source coding theorem, and the term "qubit" in [Sch95]. This work opened the door for an entirely new perspective on how the quantum theory of information should be structured. Rather than applying classical information theory to the probability distributions described by quantum mechanics, Schumacher invented distinctly quantum mechanical measures of information. These measures have allowed for the formulation of a quantum theory of information distinct from, yet capable of subsuming, the classical theory.

2.2.1 Typical subspaces

With the theory of typical sequences in hand we can generalize fairly easily to their quantum analog, *typical subspaces*.

Typicality for classical sources was defined for length n blocks of bits. Typicality for quantum sources will similarly be defined for blocks of qubits - but we first need to specify exactly what this means.

One of the major conceptual innovations of [Sch95] was to realize that arbitrary quantum states could be encoded on sequences of qubits in much the same way which any classical message can be encoded on bits. We can therefore assume that a quantum source, $\mathcal{M}$, only outputs qubit states. Though, unlike a classical source, it is not constrained to orthogonal states. $\mathcal{M}$ is assumed to output any one of k signal states, $|a_i\rangle$, according to the probability distribution $p(a_i)$. This is described by the density matrix $\rho = \sum_{i=0}^{k} p(a_i) \, |a_i\rangle \, \langle a_i|$.

Each signal from the source is assumed to be independent and identically distributed. This implies that a block of n qubits - a sample from $\mathcal{M}^n$ - may be represented as a prod-

uct state $|\alpha\rangle = |\alpha_1\rangle |\alpha_2\rangle \dots |\alpha_n\rangle$, and the probability of α is given by the product of the probabilities of its subsystems, $p(\alpha) = p(\alpha_1)p(\alpha_2)\dots p(\alpha_n)$. We describe samples from $\mathcal{M}^n$ by a density matrix, which is conveniently given by the product of the single signal density matrices, $\rho^n = \rho_1 \otimes \rho_2 \otimes \dots \otimes \rho_n$.

ρ^n gives a complete statistical description of samples from $\mathcal{M}^n$, but it is still a bit more unwieldy than the simple collection of alphabet-and-probability-distribution which describe classical signal blocks. There is, however, a nice reduction from ρ^n to something very similar to this classical object. We note that the eigenvalues of ρ^n form a probability distribution on the eigenstates of ρ^n. This is plainly seen by considering an eigenvalue decomposition of the state: $\rho^n = \sum_i \lambda_i |i\rangle \langle i|$. Density matrices are required to be postive semi-definite, and to have unit trace, so their eigenvalues must be positive real numbers which sum to 1; $\sum_i \lambda_i = 1$. This convex set satisfies all of the formal requirements of a probability distribution, and we can see, from our discussion of measurement, that λ_i has an operational definition as the probability of obtaining result $|i\rangle$ after a measurement in the eigenbasis.

Assigning a classical letter to each eigenstate, we can treat the quantum system as if it were a classical source. A quantum state will be considered typical if the sequence of eigenstates describing it is typical.

Schumacher notes [Sch95] that due to the product nature of ρ^n, the eigenstates are given by the Kronecker product of the eigenstates of each subsystem. Similarly, the eigenvalues are the scalar product of the eigenvalues of the subsystems.

As we saw in section 1.3 the von Neumann entropy of a state, $S(\rho) = \mathrm{Tr}\,(\rho \log \rho)$, is most easily calculated when we know an eigenvalue decomposition of the state, $\rho = \sum_x \lambda_x |x\rangle \langle x|$. In this case, $S(\rho) = H(\{\lambda_x\}) = -\sum_x \lambda_x \log \lambda_x$.

With the equivalence between Shannon and von Neumann entropies in mind, we may now give a definition of quantum typicality almost identical to the classical case. Here, $\lambda(\rho, i)$ is the ith largest eigenvalue of ρ.

$$\mathcal{T}_\epsilon^{(n)} = \left\{ \rho : \left| -\frac{1}{n} \sum_{i=1}^{n} \log \lambda(\rho, i) - S(\rho) \right| < \epsilon \right\} \tag{2.5}$$

This definition is technically correct, but leaves something to be desired if it is to be used in practice. Consider, for instance, that while superposition states may be typical, we cannot directly test for their typicality (which would involve measuring each subsystem) without destroying the state. For this reason we need our typicality test to be *gentle*.

In all measurements there is a trade-off between the amount of information gained and the level of disturbance caused to the system - the greater the information gain, the greater the disturbance. This has been proven formally in [Win99]. The gentle operator lemma

states that given a state ρ and operator Λ if $\mathrm{Tr}\,(\rho\Lambda) \geq 1 - \lambda$, then

$$\|\rho - \sqrt{\Lambda}\rho\sqrt{\Lambda}\|_1 \leq \sqrt{8\lambda} \tag{2.6}$$

Fortunately for us *almost all* i.i.d. product states are typical, and so we gain very little information by determining typicality. For large enough n we are able to perform typicality measurements which cause little disturbance to the measured system.

To perform such a measurement we need to first identify the set of ϵ-typical quantum states with the subspace in which they lie: the *ϵ-typical subspace*. A projector into this subspace is constructed by summing over the density matrices of all ϵ-typical product states:

$$\Lambda_\epsilon^{(n)} = \sum_{\rho \in \mathcal{T}_\epsilon^{(n)}} \rho = \sum_{\rho \in \mathcal{T}_\epsilon^{(n)}} |\alpha_1\rangle\langle\alpha_1| \otimes |\alpha_2\rangle\langle\alpha_2| \otimes \cdots \otimes |\alpha_n\rangle \tag{2.7}$$

We can now make some general claims about the properties of typical subspaces. Proofs are omitted but may be found in [NC00].

1. As n increases, the probability that a i.i.d. product state is typical converges to 1. Specifically, for all $\epsilon, \delta > 0$ there exists an n_0 such that for all $n \geq n_0$:

$$\mathrm{Tr}\left(\Lambda_\epsilon^{(n)} \rho \Lambda_\epsilon^{(n)}\right) \geq 1 - \delta. \tag{2.8}$$

   By the completeness relation, the orthogonal (atypical) subspace, has vanishing weight.

$$\mathrm{Tr}\left((I - \Lambda_\epsilon^{(n)})\rho(I - \Lambda_\epsilon^{(n)})\right) < \delta. \tag{2.9}$$

2. The dimension of the typical subspace is approximately $2^{n(S(\rho)\pm\epsilon)}$.

2.2.2 The quantum noiseless coding theorem

The quantum noiseless coding theorem generalizes Shannon's noiseless channel coding theorem (section 2.1.2) to quantum sources. Like Shannon's theorem, it places upper and lower bounds on the compressibility of a source, but there are some fairly substantial differences between classical and quantum sources which complicate the task. For one, as we saw in 1.2.8, quantum sources cannot be copied - they can't even necessarily be distinguished. So we will not be able to say that a state after compression and subsequent decompression is

exactly the same as the original. Instead we will bound the *fidelity* between initial and final states.

Doing so introduces an element of error which was not present in the classical case. Note, however, that many proofs of Shannon's source coding theorem (specifically when the theorem is referred to as the noiseless channel coding theorem) bound an error rate instead of the expected code length. These are entirely equivalent - one can enforce a code length of $nH(X)$ bits, allowing for only $2^{nH(X)}$ sequences to be properly decoded. The remaining sequences are mapped to arbitrary code words of length $nH(X)$, preserving the expected code length but introducing erroneous decodings. A proof of this form will show that the error rate may be made arbitrarily small in the limit of large block lengths.

Schumacher's proof of the quantum source coding theorem is more along these lines. Consider a source, M, which outputs independent and identically distributed signals according to a distribution $p(M)$. The goal is to block encode these signals using as few qubits per signal as possible - we are trying to compress the source in such a way that it can be decompressed with high probability. Encoding and decoding will be represented by unitary operators $\mathcal{E} : \mathcal{H}_Q^{\otimes n} \to \mathcal{H}_Q^{\otimes n}$ and $\mathcal{D} : \mathcal{H}_Q^{\otimes n} \to \mathcal{H}_Q^{\otimes n}$. Notice that these functions return states encoded on Hilbert spaces of the same dimension as their inputs. While the information content of the signals will be stored on some number of qubits $m < n$, unitarity restricts us from "throwing out" qubit systems. To circumvent this, we require that $\mathcal{E}^n$ and $\mathcal{D}^n$ leave the last $n - m$ qubits in a standard state such as $|0\rangle$. The assumption is then that these qubits are discarded after encoding and that ancilla qubits in the $|0\rangle$ state are added prior to decoding. With this assumption, it makes sense to speak of $\mathcal{E}^n$ and $\mathcal{D}^n$ as compression and decompression functions.

Let ρ' be Bob's estimate of Alice's input; $\rho' = \mathcal{D}^n(\mathcal{E}^n(\rho))$. This representation of ρ is ϵ-faithful if the fidelity[4] is greater that $(1 - \epsilon)$; $F(\rho, \rho') > 1 - \epsilon$.

Fix $\delta > 0$. Provided that $S(\rho) + \delta$ qubits are available to encode each signal, then for any $\epsilon > 0$ there exists an N_0 such that for all $N \geq N_0$ it is possible to transmit blocks of N signals with fidelity $F > 1 - \epsilon$. This is justified by the fact that, as N becomes large, the probability mass concentrated on the typical subspace approaches 1, i.e. for some N_0 and all N greater, the probability mass concentrated outside the typical subspace drops below ϵ

The converse proof was also given in [Sch95]. It states that the bound given above is in fact tight. Given only $S(\rho) - \delta$ qubits to encode each signal the achievable fidelity for large N is $F < \epsilon$.

[4]The definition of fidelity given in section 1.5.2 actually predates Schumacher's noiseless coding theorem. Schumacher's fidelity, for an initial state π_a and final state w_a, is given by $\sum_a p(a)\mathrm{Tr}\,(\pi_a w_a)$. But our fidelity measure is also sufficient.

Schumacher's source coding theorem gives a wonderful operational definition to the von Neumann entropy $S(\rho)$. Like its classical counterpart, the Shannon entropy, S gives a tight classification of the number of qubits needed to transmit a state from one party to another. If more than $S(\rho)$ qubits are available per state, then the state may be transmitted with fidelity arbitrarily close to 1. However, if less than $S(\rho)$ qubits are available, that fidelity will inevitably approach zero.

CHAPTER 3

Channels

The previous chapters have demonstrated some general techniques for quantifying and representing information, but this presentation has been largely idealized and has ignored dynamic systems. Ultimately, we want to understand how quantum information changes over time, and we want a single mechanism capable of describing any evolution. We may wish to describe the free evolution of a closed or open quantum system, the action of gates or measurement apparatuses, or the transmission of quantum information in a communication setting. In analogy with concepts of classical information theory, we will call this mechanism a *quantum channel*.

In the classical theory a channel is simply any medium through which information may be transmitted - a fiber-optic line or radio frequency for transmission through space; an optical disc or photographic film for storage over time. A quantum channel may be very similar; there will be a sender and a receiver (who we typically call Alice and Bob), and there will be some model of how the information changes during transmission. The later might be intentional evolution, as in the case of a gate in a quantum computer; unintentional noise, as when the principal system interacts with an environment; or even adversarial, in the case of a third party who attempts to read or disrupt the information.

The intricacies of quantum mechanics, such as unitary evolution and its consequences like the no-cloning theorem, cause noise to be more nuanced in the quantum domain than in the classical. Classical noise can often be thought of as being independent from the transmission and the impact which classical signals have on their environment may often be completely ignored. In contrast, quantum signals and their environments evolve together and can be conceptualized independently only in the most idealized settings. Schumacher, in [Sch96], states that "[t]his illustrates very clearly a general principle: In quantum information theory, noise is exactly information exchange with an external system."

We begin now by introducing the concept of *quantum operations*, which will allow us to describe the evolution of open systems (ones in which an environment is present).

3.1 Quantum Operations

In section 1.2.7 we briefly examined the evolution of closed systems. Such a discussion is vital in developing an understanding of how it is possible to encode information, and even perform computation, entirely on quantum states. Yet closed systems rarely, if ever, exist in the real world - in order to understand how quantum information processing can be made practical, we will need to introduce some new mathematics for describing the evolution of *open systems*.

Within the context of quantum Shannon theory we are primarily concerned with how the introduction of the environment affects our ability to reliably communicate information (quantum or classical). In general, these interactions will manifest as *noise* and will appear as non-unitary perturbations of the principal system. The focus of this section is to define a way in which these interactions may be modelled.

One might recall from basic quantum mechanics that the time evolution of all quantum systems is unitary. As such, the principal system only seems to undergo non-unitary evolution in the presence of noise because we have only considered its *reduced dynamics* - we have ignored the evolution of the environment. Therein lines an immensely important idea; the evolution of any system, no matter how chaotic, may be described as the reduced dynamics of the unitary evolution of a larger system. Invoking such a larger system is referred to, colloquially, as "going to the church of the larger Hilbert space." More formally the act is known as a Stinespring dilation, or Stinespring extension.

3.1.1 Stinespring's dilation theorem

Stinespring's theorem is a result from functional analysis which characterizes what are known as *completely positive* maps. As it turns out, all quantum channels are described by completely positive trace preserving (CPTP) maps which is really just to say that if the map's input is a density matrix its output will be as well. The map must be completely positive in order to ensure that the acted upon density matrices' eigenvalues remain positive; and the map must be trace preserving in order to preserve probabilities[1].

[1]The CPTP restriction has been questioned on several occasions. Pechukas, in [Pec94], notes that the map describing the reduced dynamics of a system, ρ^S, need only be completely positive if the system is initially decoupled from the environment (i.e. $\rho^{SE} = \rho^S \otimes \omega^E$). While such decoupling is a reasonable in the idealized quantum information setting, it is not easily accomplished experimentally. As such, Pechukas suggests that the positivity constraint be dropped. Alicki notes, [Ali95], that doing so introduces additional difficulties, and that positivity can be preserved by carefully modeling the process which prepares ρ^{SE}. We will assume that the systems we study are initially decoupled from the environment - the question of how

The theorem [Sti55] states exactly what we have already covered informally: if a map $\mu(\rho)$ is completely positive for states ρ on a Hilbert space $\mathcal{H}$, then it has a representation as $\mu(\rho) = Vk(\rho)V^\dagger$. The matrix V is a unitary operation on the larger Hilbert space $\mathcal{K}$; $k(\rho)$ lifts ρ from $\mathcal{H}$ to $\mathcal{K}$. $k(\rho)$ might be, for instance, $k(\rho) = \rho \otimes I$, where the dimension of I is dim $\mathcal{K}$/dim $\mathcal{H}$.

We can now describe the action of an arbitrary channel $\mathcal{E}$. Consider a state ρ on system S which is coupled to an environment, R, the action of $\mathcal{E}$ on ρ is given by:

$$\rho^S \to \mathcal{E}(\rho^S) = \mathrm{Tr}_R \left(V^{SR}(\rho^S \otimes \omega^R)V^{\dagger SR} \right) \tag{3.1}$$

Wherein the quantity traced over is a Stinespring dilation of $\mathcal{E}$.

This technique merely shows the existence of V^{SE}, there still seem to be some obstacles to still overcome before we can actually put it to use. For one, we have not yet bound the size of the larger Hilbert space on which V will act. Immediately it seems that doing so may be difficult, as the environment can be effectively infinite dimensional - do we therefore need to leave behind the simple finite structures we have used up until now? Fortunately, the answer is no. A description of the environment of no more than d dimensions (i.e. the joint system $\mathcal{H}_S \otimes \mathcal{H}_R$ has d^2 dimensions) is sufficient to model any possible evolution of the principal system. We will easily see why if we consider the *operator sum representation* of quantum operations.

3.1.2 Operator sum representation

Building on the result of Stinespring, Choi gave an alternate characterization of the completely positive maps, [Cho75], which states that $\mathcal{E}$ is completely positive if and only if it admits a representation as:

$$\mathcal{E}(\rho) = \sum_k^M A_k \rho A_k^\dagger. \tag{3.2}$$

This is known as the *operator sum* or *Kraus* representation provided that the *Kraus operators* (the A_k) satisfy

$$\sum_k^M A_k A_k^\dagger = I. \tag{3.3}$$

This last condition ensures that the map is trace preserving. Any physical evolution of a principal system which is initially decoupled from its environment may be written in this

best to model systems for which this is not the case is still the subject of current research.

form.

The operator sum representation is closely related to the representation given by dilation. In fact we can derive the operator sum representation by explicitly calculating[2] the partial trace in 3.1. For now we assume that ρ^S is pure, $\rho^S = \left|\phi^S\right\rangle\left\langle\phi^S\right|$. We are free to purify the environment, and choose its basis, so we let $\omega^R = \left|0^R\right\rangle\left\langle 0^R\right|$.

$$\begin{aligned}\mathcal{E}(\rho) &= \mathrm{Tr}_R\left(V^{SR}(\rho^S \otimes \omega^R)V^{SR\dagger}\right) \\ &= \mathrm{Tr}_R\left(V^{SR}(\left|\phi^S\right\rangle\left\langle\phi^S\right| \otimes \left|0^R\right\rangle\left\langle 0^R\right|)V^{SR\dagger}\right) \\ &= \sum_k^M \left\langle k^R\right| V^{SR}(\left|\phi^S\right\rangle\left\langle\phi^S\right| \otimes \left|0^R\right\rangle\left\langle 0^R\right|)V^{SR\dagger}\left|k^R\right\rangle\end{aligned} \tag{3.4}$$

Letting $A_k = \left\langle k^R\right| V^{SR}(\left|\phi^S\right\rangle \otimes \left|0^R\right\rangle)$ we recover the operator sum representation for pure states. Since ρ may be written as a convex combination of pure states, $\sum_i p_i \left|\phi_i\right\rangle\left\langle\phi_i\right|$, and $\mathcal{E}$ is a linear operator, we recover the operator sum representation for arbitrary mixed states by $\mathcal{E}(\rho) = \mathcal{E}\left(\sum_i p_i \left|\phi_i\right\rangle\left\langle\phi_i\right|\right) = \sum_i p_i \mathcal{E}(\left|\phi_i\right\rangle\left\langle\phi_i\right|)$.

A treatment similar to the above as well as a more formal proof may be found in [Sch96].

For a d dimensional system ρ^S the number of Kraus operators, M, specifying its evolution need be no more than d^2 - intuitively, this is a sufficient number to alter each element of the $d \times d$ matrix representation of ρ^S individually.

3.2 Types of channels

The noiseless channel $\mathbf{id}^{A\to B}$ takes states on system A to system B with perfect fidelity.

We will sometimes denote a general noisy channel by $\mathcal{N}^{A\to B}$, however since noise can only come through interaction with an environment, it is often more appropriate to deal with the channel's Stinespring extension $\mathcal{U}_{\mathcal{N}}^{A\to BE}$. The later makes it clear that the information which was not conveyed to Bob has been retained by the environment (or perhaps by an adversary, Eve), not destroyed.

[2]Some notational liberty is taken in equation 3.4, see section 1.2.5 for the actual explicit calculation of the partial trace.

CHAPTER 4

Channel Capacity Results

In section 2.1 we were concerned in some sense with how "densely" we could represent messages from a source, or equivalently, with the amount of redundancy we could remove from messages while retaining the ability to read them later. The results were given under the assumption that the data would be stored in a completely noiseless form in between its initial encoding and eventual decoding. Source coding results can also be seen through the lens of chapter 3 - they are statements about the capacity of noiseless channels. In the present chapter we drop the assumption that channels and storage devices are noiseless so as to get one step closer towards examining the communication capacity of real devices.

In the classical information theory, Shannon's noisy channel coding theorem more or less opens and closes the book on what can be said about the capacity of a noisy channel shared between two parties. No similar claim can be made about the existing theorems for quantum channels. There are a plethora of capacity results and as of yet no single formula has been able to wrangle these together. We will encounter a few - though by no means all - of these results in this chapter, specifically: the classical capacity of a quantum channel (C), the entanglement assisted capacity (C_E), and the quantum capacity (Q).

C gives the asymptotic rate at which classical information may be transmitted over a noisy quantum channel. One might expect this to be relatively straightforward to classify, yet it is still not fully understood, and a recent *non-additivity* result from Hastings [Has09] shows that we may still have a long way to go before it is.

On the other hand, the examination of C_E - the rate at which classical information may be communicated between parties who are assisted by an unlimited share of entanglement - has given some of the most definitive results in the field. The presence of unlimited entanglement so dramatically simplifies the relations between channels that C_E can be seen as the only parameter of interest [BSST01]. The so-called quantum reverse Shannon theorem states, among other things, that any two channels with equivalent C_E may simulate each other's action on i.i.d. inputs with perfect asymptotic efficiency [BDH$^+$09].

The quantum capacity, Q, gives the asymptotic rate at which quantum information may be transmitted over a channel. Much progress has been made towards understanding this quantity, but the amount which is still unknown only underlines the fact that quantum

information theory is still in its formative years - there is much we still do not know about even the most basic properties of quantum channels. Recently [SY08] showed, quite counter intuitively, that there exist channels which have zero capacity when used on their own, but which have positive capacity when combined with each other. Such results show that the quantum capacity cannot be said to characterize a channel outright - it must be considered with respect to the context in which the channel is to be used.

With the exception of the entanglement-assisted capacity, the results we present here apply only to the capacity of a channel to transmit information from i.i.d. sources. The effect which entangling the inputs to a channel may have is still very poorly understood. To clarify the types of inputs to which a capacity result applies we will mark each capacity with a superscript. $C^{(1)}$ and $Q^{(1)}$ indicate the classical and quantum capacities for a channel given i.i.d. inputs. By extension $C^{(n)}$ is the classical capacity of a channel which allows arbitrary entanglement within length-n input blocks, but for which each separate length-n block must be i.i.d. A major open problem is the additivity of the classical capacity, in other words, is $C^{(n)}(\mathcal{N}^{\otimes n}) = nC^{(1)}(\mathcal{N})$. The previously mentioned result by Hastings, shows that our best known measure for the classical capacity is *non-additive*, but it is still unknown whether an additive measure might be found.

So, with the caveat in mind that these are likely not the final word on the subject, we now present some of the known capacity results.

4.1 Classical capacity of a noiseless quantum channel

The field got a bit ahead of itself with Schumacher's noiseless channel coding theorem (section 2.2.2), because while it gives the *quantum* capacity of the noiseless channel, it leaves open the seemingly simpler problem of that channel's *classical capacity*. Hausladen, Jozsa, Schumacher, Westmoreland and Wootters got around to proving this result in [HJS+96]. We present this here, rather than in the source coding section, as it was this result which lead to the proof of the Holevo-Schumacher-Westmoreland theorem which gives the $C^{(1)}$ capacity of a noisy quantum channel.

One of the niceties we take for granted when analyzing classical channels is that the symbols of their alphabets are perfectly distinguishable, and for that matter, that any uncertainty about the message is introduced by noise during transmission. These assumptions reduce the problem of determining the channel's classical capacity to determining how much one can compress its input.

Schumacher's source coding scheme allows us to do something similar for arbitrary quan-

tum states - it tells us how much we can compress these states, and therefore tells us the rate at which we can send them over a noiseless channel. But it is not immediately obvious what this result says about the ability of a quantum channel to communicate purely classical information. How much classical information can we encode on a single qubit?

Recall that $\overline{C}$, the classical capacity of a *classical* channel, is given by Shannon's noisy channel coding theorem which states that $\overline{C}$ is equal to the mutual information between the channel's inputs and its outputs after an optimization over input distributions.

$$\overline{C} = \max_{p(X)} I(X;Y) = \max_{p(X)} \left(H(X) - H(X|Y)\right). \tag{4.1}$$

Our goal is to define a similar term for quantum channels. Alice is still trying to send classical information, but now she and Bob are connected by a quantum channel, and they have a quantum encoder and decoder respectively.

Alice will use her encoder to assign some set of classical signals to the states $\{\rho_x\}$. With probability $p(X)$ she will attempt to send the classical message x by inputting $\{\rho_x\}$ to the channel. We therefore represent the channel's inputs as the mixed state $\rho = \sum_x p(x)\rho_x$. Each use of the channel, like the classical case, is independent and identically distributed - $p(X)$ does not change, and we do not allow multiple inputs to the channel to be entangled with each other.

We can no longer assume that individual letters of Alice's channel alphabet ($\{\rho_x\}$) are distinguishable, and therefore we cannot assume that there exist decoding operations capable of extracting n bits of information from a sample with von Neumann entropy n. The amount of information which Bob can extract is termed the *accessible information*. If Bob uses the best possible decoding scheme, the accessible information is equal to the mutual information between Alice's classical inputs and Bob's decoded outputs, $I(X;\hat{X})$.

Certainly the accessible information will never exceed the von Neumann entropy of the source, and we will shortly see that for a noiseless channel there are input distributions and channel alphabets for which $S(\rho)$ is attainable. However, a celebrated theorem by Holevo [**?**] gives an even better upper bound which will also be useful in the noisy channel scenario: $I(X;\hat{X}) \leq S(\rho) - \sum_x p(x)S(\rho_x)$.

The term on the right hand side is referred to as χ, or the *Holevo information*.

$$\chi = S\left(\sum_x p(x)\rho_x\right) - \sum_x p(x)S\left(\rho_x\right) \tag{4.2}$$

The result of Hausladen *et al.* , [HJS+96], shows that the Holevo information is an attainable communication rate per use of the noiseless channel. The Holevo information itself

is maximized by a uniform distribution over an orthonormal set of pure states. For example, for a qubit channel, Alice's best encoding scheme is to send $|0\rangle$ and $|1\rangle$ each with probability 0.5. Bob's optimal decoding operator is a projective measurement in the standard basis. The proof in [HJS+96] shows that the Holevo information gives the correct capacity in the more general case where Alice is not free to choose the alphabet (i.e. for channel alphabets of pure, yet non-orthogonal states). We have covered most of the mathematical machinery that is needed to prove this (typical subspaces, block encoding/decoding operations), but the proof technique itself (random coding) has been superseded by newer techniques so we will not review it here.

We can now more formally state the $C^{(1)}$ capacity of a noiseless qudit channel for which Alice chooses the channel alphabet.

$$\begin{aligned} C^{(1)}(\mathbf{id}_D^{A\to B}) &= \chi_{max}(\mathbf{id}_D^{A\to B}) \\ &= \max_{\{p(a),\rho_a\}} S(\rho) \quad \text{where } \rho = \sum_a p(a)\,|\psi_a\rangle\langle\psi_a| \\ &= S\left(\frac{1}{D}\sum_{d=0}^{D}|\psi_d\rangle\langle\psi_d|\right) \\ &= \log D \text{ bits} \end{aligned} \tag{4.3}$$

This is, perhaps, an unsurprising result, but it is an important one nonetheless. One-way noiseless quantum channels provide no improvement over noiseless classical channels for the transmission of classical information.

Hausladen *et al.* conjectured that the classical capacity of a noisy quantum channel would also be given by the χ quantity. The proof of which is now known as the Holevo-Schumacher-Westmoreland (HSW) theorem.

4.2 Holevo-Schumacher-Westmoreland Theorem

Building off the conjecture of [HJS+96], proofs of the classical capacity of a *noisy* quantum channel were given independently by Holevo in [Hol98], as well as Schumacher and Westmoreland in [SW97]. Both use notions of typicality and random coding techniques to answer the conjecture in the affirmative - the Holevo information gives the $C^{(1)}$ capacity for a noisy quantum channel.

The proofs of this theorem are somewhat too involved to present here, they both involve rigorous definitions of typicality and an application of random coding. Intuitively, random

coding bounds Bob's expected error rate over all coding schemes. If a random code of rate R is expected to have an error rate of p_e, then a simple modification of any code, namely one in which the highest error rate codewords are deleted, can be made to have lower error rate. These random coding proofs show that in the limit of large block lengths, p_e can be pushed to zero for all codes with rates bounded above by the Holevo information.

The end result is a characterization of the $C^{(1)}$ capacity in terms of the Holevo information (we will also refer to this as the Holevo capacity).

$$C^{(1)}(\mathcal{N}) = \chi_{max}(\mathcal{N}) \tag{4.4}$$

It was an open question until very recently whether or not the Holevo χ quantity was additive, in other words, whether $\chi_{max}(\mathcal{N} \otimes \mathcal{N}) = \chi_{max}(\mathcal{N}) + \chi_{max}(\mathcal{N})$. If this were the case, then the χ quantity would give a complete description of the classical capacity; it would settle the conjecture $C \stackrel{?}{=} C^{(1)}$ in the affirmative.

The recent result by Hastings shows that the χ quantity and two other quantities known as the entanglement of formation and the minimum output entropy are not additive.

As a result we find ourselves without a general formula for the classical capacity of a quantum channel. The best known formula is given by the regularized Holevo capacity.

$$C = \lim_{n\to\infty} \frac{1}{n} C^{(n)}(\mathcal{N}^{\otimes n}) \tag{4.5}$$

4.2.1 Example: Classical Capacity of the Depolarizing Channel

The regularized Holevo capacity cannot generally be computed. That said, the Holevo capacity has been shown to be additive for a limited set of channels, and as such we can compute exact classical capacities for those channels.

One nice example of such a channel is one which "depolarizes" its input. With probability p the depolarizing channel produces a random output, and with probability $(1-p)$ it transmits the input state with perfect fidelity. Equivalently, one may think of this channel as randomly applying one of the four Pauli operators $\{\mathbf{I}, \mathbf{X}, \mathbf{Y}, \mathbf{Z}\}$, with probability p.

As given in [NC00] and elsewhere, this can be conveniently written in the operator-sum form,

$$\mathcal{E}(\rho) = (1 - \frac{3p}{4})\rho + \frac{p}{4}(\mathbf{X}\rho\mathbf{X} + \mathbf{Y}\rho\mathbf{Y} + \mathbf{Z}\rho\mathbf{Z}) \tag{4.6}$$

If we abandon the operator sum representation (which is mostly useful to show that this is a valid physical channel), we can recover the original interpretation of the depolarizing

channel as one which probabilistically randomizes its output. We note that

$$\frac{1}{4}\left(\rho + \mathbf{X}\rho\mathbf{X} + \mathbf{Y}\rho\mathbf{Y} + \mathbf{Z}\rho\mathbf{Z}\right) = \frac{I}{2}. \tag{4.7}$$

From which we may rewrite equation 4.6 as

$$\mathcal{E}(\rho) = (1-p)\rho + \frac{p}{2}I. \tag{4.8}$$

The HSW theorem states that the $C^{(1)}$ capacity of the depolarizing channel should be given by a maximization over the Holevo information of the output states.

$$C^{(1)}(\mathcal{E}) = \chi_{max}(\mathcal{E}) = S\left(\mathcal{E}(\rho)\right) - \sum_i p_i S(\mathcal{E}(\rho_i)) \tag{4.9}$$

We can assume that Alice inputs pure states to her channel, so the ρ_i are equivalently represented by $|\psi_i\rangle\langle\psi_i|$. Consider the action of $\mathcal{E}$ on a pure state $|\psi_i\rangle$,

$$\mathcal{E}(|\psi_i\rangle\langle\psi_i|) = p\,|\psi_i\rangle\langle\psi_i| + (1-p)\frac{I}{2} \tag{4.10}$$

The eigenvalues of the two matrices are respectively $\{p, 0\}$ and $\{\frac{(1-p)}{2}, \frac{(1-p)}{2}\}$. Since the two matrices commute, the eigenvalues of their sum are $\{\frac{(1+p)}{2}, \frac{(1-p)}{2}\}$. Note that this is entirely independent of the choice of $|\psi_i\rangle$.

The entropy of $\mathcal{E}(\rho_i)$ is therefore[1] $H\left(\frac{1+p}{2}\right)$ for all pure ρ_i. As such we've reduced the problem of determining the depolarizing channel's capacity to calculating:

$$\chi_{max}(\mathcal{E}) = \max_{\rho}\ S\left(\mathcal{E}(\rho)\right) - H\left(\frac{1+p}{2}\right) \tag{4.11}$$

A uniform distribution on any orthonormal set of pure states (such as $|0\rangle, |1\rangle$) will serve to maximize this quantity, as $\mathcal{E}(\frac{I}{2}) = \frac{I}{2}$, and $S(\frac{I}{2}) = 1$. Therefore, the classical capacity of the depolarizing channel with parameter p is

$$C^{(1)}(\mathcal{E}) = 1 - H\left(\frac{1+p}{2}\right) \tag{4.12}$$

Since this depends only on p and not the input state, this is the general classical capacity, C, as well.

[1] We adopt the convention here that the Shannon entropy of a real number, $H(p)$, is defined to be the Shannon entropy of a Bernoulli random variable with weight p, i.e. $H(\{p, 1-p\})$

4.3 Classical Capacity of an Entanglement-Assisted Quantum Channel

What if, in addition to a noisy channel, Alice and Bob share a pool of entangled qubits. Can they use these to enhance the capacity of their channel? How might such an *entanglement-assisted* capacity relate to the unassisted capacity? These questions, and their answers, have lead to some of the most definitive results in the field of quantum Shannon theory.

It was conjectured in [BSST01], and subsequently proved in [BDH+09], that in the presence of unlimited shared entanglement, the only parameter needed to completely classify a quantum channel is its entanglement assisted classical capacity - if $C_E(\mathcal{N}_1) \geq C_E(\mathcal{N}_2)$, then $\mathcal{N}_1$ may be used to simulate $\mathcal{N}_2$. This result is known as the quantum reverse Shannon theorem.

The entanglement assisted classical capacity of a quantum channel is given by a maximization of the quantum mutual information between that channel's inputs and outputs [BSST01].

$$\begin{aligned} C_E &= \max_{\rho} \; S(\rho : \mathcal{N}(\rho)) \\ &= \max_{\rho} \; H(\rho) + H(\mathcal{N}(\rho)) - H((\mathcal{N}^{Q\to Q'} \otimes I^R)\Phi_\rho^{QR}) \end{aligned} \tag{4.13}$$

Remarkably, this formula is additive $C_E(\mathcal{N}^{\otimes n}) = nC_E(\mathcal{N})$, and therefore completely characterizes the entanglement assisted capacity of any channel - there is no need to regularize the expression.

The presence of entanglement so dramatically simplifies the relationships between the capacities that we can now even state a precise bound on the *quantum* capacity of an entanglement-assisted channel. The proof relies on the teleportation and superdense coding protocols which we will encounter more formally in chapter 6. For now it suffices to mention that quantum teleportation [BBC+93] allows for the noiseless transmission of an arbitrary qubit in exchange for one qubit of entanglement and two bits of classical communication, and that superdense coding [BW92] allows for two classical bits to be communicated via the noiseless transfer of a single maximally entangled state. Both protocols are known to be optimal, which implies that the entanglement assisted classical capacity must be exactly twice the entanglement assisted quantum capacity, $C_E = 2Q_E$.

4.4 Quantum Capacity

The quantum capacity of a noisy quantum channel is another area in which there is still room for considerable improvements over known results. Lloyd [Llo97], Shor, and Devetak [Dev03], have all given proofs of the $Q^{(1)}$ capacity with increasing levels of rigor. All of these find that the $Q^{(1)}$ capacity is given by the *coherent information*, a term introduced by Schumacher and Nielsen in [SN96].

The coherent information has quickly become one of the most important quantities in quantum information theory. It shares many similarities with the classical mutual information, and appears in similar circumstances. For instance, it was shown in [SN96] that the coherent information obeys a data processing inequality.

In some sense, the coherent information measures of how well the distinctly quantum attributes of a system are preserved under evolution. Formally, the coherent information is given by

$$I(A\rangle B)_\rho = -S(A|B) = S(B) - S(A,B) = S(B) - S(E). \tag{4.14}$$

In order to apply this to a quantum channel, it is most helpful to express the action of the channel, on an input ρ, as a unitary transformation, U, on the purified state $\rho^A \otimes |0\rangle\langle 0|^E$. The coherent information can then be evaluated on the reduced states of $\sigma^{BE} = U\rho^A \otimes |0\rangle\langle 0|^E U^\dagger$ as $S(B)_\sigma - S(E)_\sigma$.

Maximizing the coherent information in this form over all input states, we obtain an expression for the $Q^{(1)}$ capacity as

$$Q^{(1)} = \max_\rho \left(S(B)_\sigma - S(E)_\sigma\right). \tag{4.15}$$

It has been known for some time that the coherent information is not additive [DiV98], except for in the case of a limited set of channels known as *degradable* quantum channels [DS05] [CRS08].

As with the classical capacity, we can define a regularized formula for the quantum capacity,

$$Q = \lim_{n\to\infty} \frac{1}{n} Q^{(n)}(\mathcal{N}^{\otimes n}), \tag{4.16}$$

which *is* additive, i.e. $Q(\mathcal{N}^{\otimes k}) = kQ(\mathcal{N})$. However, there is no efficient way to calculate Q for general channels.

For some time researchers believed that any gap between $Q^{(1)}$ and Q was likely to be small, but Smith and Yard have shown [SY08] that this gap may be made arbitrarily large via the combination of *zero capacity* channels. There exist channels for which $Q(\mathcal{N}) = Q(\mathcal{M}) = 0$

but for which $Q(\mathcal{N} \otimes \mathcal{M}) > 0$, and for which $Q(\mathcal{N}^{\otimes n} \otimes \mathcal{M}^{\otimes n})$ increases with n. Studying this bizarre "superactiviation" effect may lead to profound new insights, but as of yet this effect only makes it all the more clear that we still have much to learn about the quantum capacity.

4.4.1 Example: Quantum capacity of the erasure channel

The ϵ-quantum erasure channel (ϵ-QEC) takes qubits as input and noiselessly transmits them with probability $(1-\epsilon)$. With probability ϵ it outputs an *erasure state* which indicates to the recipient that the input was "lost" along the way. The erasure state $|2\rangle$ is orthogonal to both $|0\rangle$ and $|1\rangle$, so it is very easy to calculate the density matrices for both Bob's system and the environment following Alice's use of an ϵ-QEC.

$$\rho^B = (1-\epsilon)\rho^A + \epsilon\,|2\rangle\langle 2| \qquad\qquad \rho^E = \epsilon\rho^A + (1-\epsilon)\,|2\rangle\langle 2| \tag{4.17}$$

Where ρ^A is Alice's input.

Assuming that the environment starts in the pure state $|2\rangle$ the joint ABE system can be seen to be pure which allows us to compute the coherent information as:

$$I(A\rangle B) = -S(A|B)_\rho = S(B)_\rho - S(E)_\rho \tag{4.18}$$

The quantum capacity of the ϵ-QEC is given by maximizing the coherent information over all input states. In this case the maximum is given by sending a completely mixed state through the channel. Note that this is equivalent to sending one member of a fully entangled pair through the channel.

Alice generates a Bell state locally, $\rho^{RA} = |\Phi^+\rangle\langle\Phi^+|^{RA}$. She will hold on to the R system and input $\mathrm{Tr}_R\left(\rho^{RA}\right) = \rho^A = I/2$ to the channel.

Following Equation 4.17, this gives us

$$\rho^B = (1-\epsilon)(I/2) + \epsilon\,|2\rangle\langle 2| \qquad\qquad \rho^E = \epsilon(I/2) + (1-\epsilon)\,|2\rangle\langle 2|\,, \tag{4.19}$$

or in matrix form

$$\rho^B = (1-\epsilon)\begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix} + \epsilon\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \rho^E = \epsilon\begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix} + (1-\epsilon)\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.20}$$

The von Neumann entropy, like the classical Shannon entropy, is a convex function, therefore,

$$S(\rho^B) = (1-\epsilon)S(I/2) + \epsilon S(|2\rangle\langle 2|) \tag{4.21}$$

Pure states have zero entropy, so we only need to calculate $S(I/2)$. Since this matrix is diagonal, the von Neumann entropy is equal to the Shannon entropy of its diagonal entries. $H(\{\frac{1}{2},\frac{1}{2}\}) = 1$, so we have our result, $\rho^B = (1-\epsilon)$. An identical argument shows that $\rho^E = \epsilon$.

With these quantities in hand we can now calculate the coherent information and the quantum capacity.

$$Q(\epsilon - \text{QEC}) = S(A\rangle B)_\rho = S(B)_\rho - S(E)_\rho = 1 - 2\epsilon \tag{4.22}$$

Which corresponds exactly with the result given by Bennett *et al.* in [BDS97]. Those researchers gave an alternative description of the ϵ-QEC's quantum capacity based on arguments from [BDSW96]. Imagine that for $\epsilon \geq \frac{1}{2}$ the QEC has a positive quantum capacity. This implies that Bob is able to reconstruct, with high fidelity, Alice's original input state from fewer than half of the transmitted qubits. In fact, if the channel really has positive capacity, it makes no difference which portion of the transmitted qubits he receives. By claiming that the ϵ-QEC has positive capacity, we are implicitly claiming that there exists a scheme by which Bob can recover Alice's n qubit message from any subset of her message of size approximately $(1-\epsilon)n$.

This cannot be true as it immediately gives rise to a scheme for cloning arbitrary quantum information. We imagine that the ϵn "erased" qubits were actually intercepted by a third party, Charles. Assuming that Bob can reconstruct Alice's message from $(1-\epsilon)n$ qubits implies that, for $\epsilon \geq \frac{1}{2}$, Charles can as well - violating the no-cloning principle of section 1.2.8.

4.4.2 The effect of a classical side channel

Can Alice and Bob boost the quantum capacity of their channel through the use of a classical side channel?

First the bad news, in [BDSW96] it was shown that a one-way classical side channel from Alice to Bob *does not* increase the quantum capacity of the quantum channel; $Q_1 = Q$. However, there are instances in which two-way classical communication can enhance the quantum capacity. The previous example of the ϵ-QEC is one such case.

Notice that Bob always knows whether or not the source signal has been erased - if it has, he receives a $|2\rangle$ state which is orthogonal to both of the source letters. So, if Alice and Bob

share a two-way classical channel, Bob may signal to Alice which qubits he received. Alice then ignores the ϵn lost qubits, and is left with a pool of $(1-\epsilon)n$ qubits which she knows are maximally entangled with Bob's system. With a classical side channel at her disposal, Alice has everything she needs to perform the teleportation protocol, and as such, she can use each of her known-to-be-good qubits as single-serving noiseless channels. This implies that the Q_2 capacity of the ϵ-QEC is $(1-\epsilon)$ - a result which perfectly matches the classical result for the classical erasure channel.

4.5 Quantum data processing inequality

Regardless of the doubt the last sections may have just cast on the accuracy of the coherent information as a measure of quantum capacity, it has several properties that make it attractive none the less. Notably, it allows for the statement of a quantum *data processing inequality* analogous to that found in classical information theory.

In the classical theory, the data processing inequality states that if

$$X \to Y \to Z \tag{4.23}$$

is a Markov chain (i.e. each state is conditionally independent of all but the previous state so that the joint distribution may be written $p(X,Y,Z) = p(X)p(Y|X)p(Z|Y)$), then the mutual information across operations must be non-increasing

$$H(X) \geq I(X;Y) \geq I(X;Z). \tag{4.24}$$

Intuitively, one may think of this as stating that post-processing of Y may not increase the information Y contains about X. This is not to say that post-processing cannot make it "easier," in terms of say, the number of CPU cycles required to extract the information contained in Y. The data processing inequality simply states that post-processing may not increase the correlation between X, and Y.

An analogous claim about post-processing of quantum states sent through a channel. Consider two successive evolutions, $\mathcal{E}^{A\to B}$ and $\mathcal{F}^{B\to C}$, so that,

$$\rho^A \xrightarrow{\mathcal{E}} \rho^B \xrightarrow{\mathcal{F}} \rho^C \tag{4.25}$$

it was shown in [SN96] that the coherent information is non-increasing under these operations

$$S(A) \geq I(A\rangle B) \geq I(A\rangle C) \tag{4.26}$$

CHAPTER 5

The Resource Inequality Formalism

Introduced in [DW03a] and formally developed in [DHW08], the resource inequality formalism provides a simple notation for representing coding theorems and studying their relations. Its inventors noted that the known coding results in quantum Shannon theory could be seen as inter-conversions between information processing resources, and suggested that a unified treatment of these inter-conversions, or *inequalities*, might unveil structure amongst the coding theorems which had previously gone unseen.

The protocols of quantum Shannon theory all describe situations in which one or more parties begin a process with one resource at their disposal, such as a noisy quantum channel, and end the process with another resource - pure entanglement, perhaps, or a noiseless channel. The protocols themselves are in someways immaterial, serving to prove that an interconversion is possible, sometimes, but often arriving on the scene long after the inequality itself has been proven. This is the case with the classical Shannon theory as well. The canonical example being Shannon's proof of channel coding theorem which gave no means of constructing good codes while showing that they had to exist. Protocols for actually achieving capacity have been notoriously elusive; provable solutions for even the simplest classes of channels (such as Arikan's *polar codes* which are capacity achieving on binary discrete memoryless channels [Ari08]) have emerged only in recent years.

And so the resource inequality formalism strips away the baggage of protocols and leaves us with simple statements of input and output resources. Provided that the rules developed in [DHW08] are adhered to in constructing resource inequalities, we may construct new coding results by merely chaining statements together.

Claims to the utility of this approach are sure to inspire skepticism in some, and it is not altogether unlikely that this framework will be soon forgotten in this rapidly evolving field. For now, however, it appears in sufficiently many papers as to warrant mention.

5.1 Resources and Resource Inequalities

Within quantum Shannon theory we find at least three ways to divide resources: classical/quantum, noisy/noiseless, and static/dynamic. The last of which is hopefully the only requiring some explanation. Static resources are the objects of the theory, such as bits and qubits and their multipartite counterparts shared random bits and entangled qubits. Dynamic resources are processes which act on static resources such as channels and measurements. These latter resources are always accompanied by a notion of directionality, and we denote this with an arrow.

The noiseless classical bit and qubit channels are written as $[c \to c]$ and $[q \to q]$ respectively. The square brackets indicate that the resource is noiseless, curly braces in their place would imply that the channels are noisy. Similarly, the static resources shared random bits and entangled qubits are written as $[cc]$ and $[qq]$.

For instance, $\{c \to q\}$, indicates the noisy preparation of a quantum state from classical information, and $\{q \to c\}$, a measurement yielding classical information.

Addition and scalar multiplication of resources holds in the straightforward way: two entangled qubits are denoted as $2\,[qq]$, and access to both a quantum and a classical channel as $[q \to q] + [c \to c]$.

Using just the above portions of the resource framework, we can already express some powerful ideas and protocols. Consider, from our discussion of the HSW theorem, the inequality representing the fact that a noiseless qubit channel may be used to send at most one classical bit of information,

$$[q \to q] \geq [c \to c]\,. \tag{5.1}$$

The inequality here gives the sense that the process is "one-way", that while the qubit channel is as powerful as the classical bit channel, the converse is not true.

We also know that one may use a noiseless qubit channel to share entanglement with another party. One simply generates an entangled pair locally, and then communicates one member of the pair through the channel. This gives us

$$[q \to q] \geq [qq]\,. \tag{5.2}$$

From these two inequalities we can immediately derive the famous superdense coding conversion of [BW92]. It states that two parties may communicate at the "super dense" rate of two bits per qubit by spending one qubit of entanglement per transmission. The details of the protocol will be covered in section 6.1, but for now we consider only the interconversion it implies.

Using a qubit channel twice we have $[q \to q] + [q \to q] \geq 2\,[c \to c]$, and by application of 5.2 to one instance of $[q \to q]$ on the left hand side, we obtain superdense coding

$$[q \to q] + [qq] \geq 2\,[c \to c]\,. \tag{5.3}$$

The notation introduced thus far is lacking in many respects. For instance, we have no means of indicating whether or not the parties know which entangled state they hold. It is essential in superdense coding that the party initiating the communication have this information. In general, we will not want to talk about just any noisy channel, or just any entangled state, so additional notation is introduced to handle specific resources.

The inequality,

$$\langle \mathcal{N}^{A \to B} \rangle \geq R\,[c \to c]\,, \tag{5.4}$$

states that a particular noisy quantum channel, $\mathcal{N}$, which takes states from system A to system B, is equivalent to at least R uses of a classical noiseless bit channel. Following this example, we will often write dynamic resources with the name of the function to which they correspond. $\langle \mathbf{id_2} \rangle$ is a noiseless qubit channel, and $\langle \overline{\mathbf{id}}_\mathbf{2} \rangle$, a noiseless bit channel.

For entanglement we may write $\langle \Phi_2^+ \rangle$ to denote a standard Bell Φ^+ state; a value d in the subscript would denote a generalized Bell state on a d level system.

There are still many nuances of the resource framework which have not been covered here. One which we will need in the next section is the notion of a *relative resource*. The behavior of these resources is guaranteed only when applied to a particular input. For example if $\tau_2 = \frac{1}{2}(|0\rangle\langle 0| + |1\rangle\langle 1|) = I_2^A$ (a fully mixed state), then the resource $\langle \mathbf{id_2} : \tau_2 \rangle$ is a qubit channel which is noiseless for τ_2 states, but with behavior is undefined for other states such as $|0\rangle$. We might use this resource to say that a channel which distributes τ_2 states can be used to generate shared classical randomness: $\langle \mathbf{id_2} : \tau \rangle \geq [cc]$.

CHAPTER 6

Noiseless protocols

6.1 Superdense coding

One of the earliest and simplest communication protocols, superdense coding, shows us that the qubit channel can be used to send 2 classical bits per transmission in the presence of previously shared entanglement. This feat is possible because local operations on Alice's (or for the matter, Bob's) half of a $|\Phi^+\rangle^{AB}$ state are sufficient to transform the shared state into any of the four Bell states - even when the particles are spatially separated.

$$\begin{aligned}(\mathbf{I}\otimes\mathbf{I})\left|\Phi^+\right\rangle^{AB} &= \left|\Phi^+\right\rangle^{AB}\\ (\mathbf{X}\otimes\mathbf{I})\left|\Phi^+\right\rangle^{AB} &= \left|\Psi^+\right\rangle^{AB}\\ (\mathbf{Z}\otimes\mathbf{I})\left|\Phi^+\right\rangle^{AB} &= \left|\Phi^-\right\rangle^{AB}\\ (\mathbf{ZX}\otimes\mathbf{I})\left|\Phi^+\right\rangle^{AB} &= \left|\Psi^-\right\rangle^{AB}\end{aligned} \tag{6.1}$$

Having performed one of these four local transformations, Alice sends her qubit to Bob who measures the pair in the Bell basis. His measurement informs him of which of the four operations Alice applied, and thus yields two classical bits of information (the maximum he could possibly attain, by the HSW theorem, from measurement of a two-qubit system).

The protocol readily generalizes to the N dimensional qudit case. Alice replaces her $\mathbf{X}$ gate with a cyclic permutation, $\mathbf{S}$, and her $\mathbf{Z}$ gate with a phase shift by an Nth root of unity, $\mathbf{D}$ (three dimensional versions of these matrices are given below). In N dimensions there are N^2 generalized Bell states obtainable from the generalized Φ^+ state by combinations of cyclic permutations and phase shifts.

$$\mathbf{S}_3 = \begin{bmatrix}0 & 0 & 1\\ 1 & 0 & 0\\ 0 & 1 & 0\end{bmatrix} \qquad \mathbf{D}_3 = \begin{bmatrix}1 & 0 & 0\\ 0 & \omega & 0\\ 0 & 0 & \omega^2\end{bmatrix};\quad \omega = e^{2\pi i/N} \tag{6.2}$$

Counting resources, the qubit case yields 2 classical bits per spent ebit and use of a

noiseless qubit channel, as we saw previously:

$$[q \to q] + [qq] \geq 2\,[c \to c]\,. \tag{6.3}$$

And the qudit case gives the more general trade-off:

$$\langle \Phi_N^+ \rangle^{AB} + \langle \mathbf{id}_N^{A \to B} \rangle \geq 2 \log N\,[c \to c] \tag{6.4}$$

Both cases were originally presented in [BW92].

6.2 Quantum Teleportation

Quantum teleportation (QT) [BBC+93] can be expressed in the resource framework as:

$$2\,[c \to c] + [qq] \geq [q \to q]\,. \tag{6.5}$$

Although in this section we'll drop the descriptive shorthand and write it as:

$$2\langle \overline{\mathbf{id}}_\mathbf{2} \rangle + \langle \Phi_2 \rangle \geq \langle \mathbf{id}_\mathbf{2} \rangle. \tag{6.6}$$

In which $\langle \overline{\mathbf{id}}_\mathbf{2} \rangle$ is a noiseless classical bit channel; $\langle \Phi_2 \rangle$ is a fully entangled state on the 2-dimensional systems A and B (held by Alice and Bob respectively); and $\langle \mathbf{id}_\mathbf{2} \rangle$ is a noiseless qubit channel.

This generalizes quite easily to the N-dimensional case [**?**][1]:

$$(2 \log(N))\langle \overline{\mathbf{id}}_\mathbf{2} \rangle + \langle \Phi_N \rangle \geq \langle \mathbf{id}_\mathbf{N} \rangle. \tag{6.7}$$

In which the bipartite state and the quantum channel are both of dimension N. In N dimensions the QT protocol can communicate $2(N-1)$ parameters, the relative amplitudes and phases.

[1] All logarithms are taken to be base 2

6.3 Superdense Teleportation

H. J. Bernstein has developed a protocol, SuperDense Teleportation (henceforth SDT), whereby the $(N-1)$ relative phases encoded on an N-level system, may be teleported using $\log(N)$ bits of classical communication [Ber10].

Alice and Bob begin SDT by sharing a fully entangled state $\left|\Phi_N^+\right\rangle = \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} |j\rangle^A |j\rangle^B$. A third party, Charles, chooses the state $|\psi\rangle = \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} \exp(i\theta_j) |j\rangle$ to be teleported - $\{\theta_1, \theta_2, \ldots, \theta_{N-1}\}$ are defined relative to θ_0 which is constrained to zero by the equivalence of states modulo a global phase factor. Charles applies the $N-1$ relative phases to Alice's particle, which takes Alice and Bob's shared state to

$$|\Phi_N\rangle = \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} \exp(i\theta_j) |j\rangle^A |j\rangle^B. \tag{6.8}$$

Alice then performs the discrete Fourier transformation given by the $N \times N$ matrix $\mathbf{T}_{jk} = \frac{1}{\sqrt{N}} \omega^{jk}$, where ω is a primitive Nth root of unity.

At this point a von Neumann measurement by Alice yields each of the N possible outcomes with equal probability - a fact which will be especially important later when we consider the coherent version of SDT. Alice classically communicates the $\log(N)$ bit result of her measurement to Bob, who can then apply the appropriate decoding operation to recover $|\psi\rangle$. Bob's operations are given by powers of the diagonal matrix $\mathbf{D}_{jj} = \frac{1}{\sqrt{N}} \omega^{j(N-1)}$; to recover $|\psi\rangle$ he applies $\mathbf{D}^x$ where x is Alice's message taken as an integer.

Provided that high dimensional entanglement is available, SDT requires as little as half the classical channel capacity that a QT protocol teleporting the same number of parameters would require[2]. More importantly though, the operations which Bob must perform to correct his state are simpler and less numerous than those of QT. In SDT the receiving party need only be capable of performing a number of relative phase shifts linearly related to the dimension of the shared state. The experimental appeal of SDT becomes clear when we consider that QT holds a quadratic relation between these quantities.

Stated as a resource inequality, SDT may be written:

$$\log(N) \langle \overline{\mathbf{id_2}} \rangle + \langle \Phi_N \rangle \geq \langle \mathbf{id_N} : \Psi_N \rangle \tag{6.9}$$

In which the relative resource $\langle \mathbf{id_N} : \Psi_N \rangle$ is a channel which reliably transmits states of the

[2]To teleport $2N-2$ parameters, QT requires a N-dimensional shared state and $2\log N$ bits of classical communication; SDT requires a $(2N-1)$-dimensional shared state and $\log N$ bits

form:

$$|\Psi_N\rangle = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \exp(i\theta_k) |k\rangle. \tag{6.10}$$

6.4 Coherent protocols

The coherent, or *cobit*, channel $\Delta_d : A' \to AB$ is given by the isometry $\sum_x |x\rangle^A |x\rangle^B \langle x|^{A'}$, and may be thought of as the quantum analogue of a noiseless classical channel with feedback [Har03] [DHW08]. The inequality $\langle \Delta_d \rangle \geq \langle \overline{\mathbf{id}}_\mathbf{d} \rangle$ is trivially attained by having Alice discard her system after using the cobit channel. Additionally, entanglement is generated when Alice inputs a superposition of states. In general this will be partial entanglement, but full entanglement is generated when Alice inputs an equimodular state, therefore $\langle \Delta_d \rangle \geq \langle \Phi_d^+ \rangle$. Of much interest, as recognized in [Har03], is the ability of the cobit channel to communicate and generate entanglement simultaneously provided certain conditions are met. **TODO:** conditions

6.4.1 Coherent Teleportation

A coherent version of qubit QT was given in [Har03], although a circuit implementing it had previously been discussed in [BBC98]. Alice and Bob begin by sharing a $|\Phi_2\rangle$ state; Alice holds an additional state, $|\psi\rangle$, which is to be teleported. Rather than performing a measurement in the Bell basis on her two particles as she normally would to initiate teleportation, Alice unitarily rotates her system into the computational basis by applying $(\mathbf{H} \otimes \mathbf{I})\mathbf{cX}$. $\mathbf{cX}$ here is the controlled-not gate. The resulting joint state is $\frac{1}{2} \sum_{ij} |ij\rangle^A (\mathbf{X}^i \mathbf{Z}^j |\psi\rangle^B)$. She can then use the cobit channel to send a copy of her state, $\frac{1}{2} \sum_{ij} |ij\rangle^A |ij\rangle^B (X^i Z^j |\psi\rangle^B)$. And finally Bob can perform a $\mathbf{cX}$ followed by a $\mathbf{cZ}$ (controlled-Z) to recover $|\psi\rangle$. However that's not all they have accomplished, the final state of their joint system is $(|\Phi_2\rangle |\Phi_2\rangle)^{AB} |\psi\rangle^B$. In addition to teleporting $|\psi\rangle$ they've generated two bits of entanglement. This leads us to the resource inequality: $2\,[q \to qq] + [qq] \geq [q \to q] + 2\,[qq]$.

6.4.2 Coherent SDT[3]

Recall that Alice and Bob initially share a $\left|\Phi_N^+\right\rangle$ state which is altered by Charles to produce

$$|\Phi_N\rangle = \frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} \exp(i\theta_j) \, |j\rangle^A \, |j\rangle^B \tag{6.11}$$

Alice performs the discrete Fourier transformation ($\mathbf{T}$) as usual, yielding

$$\frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} |j\rangle^A \otimes \left(\exp(i\theta_j)\mathbf{T} \, |j\rangle^B\right) \tag{6.12}$$

Now instead of measuring her particle, Alice sends it through the coherent channel.

$$\frac{1}{N} \sum_{j=0}^{N-1} |j\rangle^A \, |j\rangle^B \otimes \left(\exp(i\theta_j)\mathbf{T} \, |j\rangle^B\right) \tag{6.13}$$

Following [DHW04], we replace Bob's decoding operation, $\mathbf{D}$, with a controlled operation $\mathbf{cD} = \sum_{j=0}^{N-1} |j\rangle \langle j|^B \otimes \mathbf{D}^j$. Applying this to his system transforms the joint state into

$$\left(\frac{1}{\sqrt{N}} \sum_{j=0}^{N-1} |j\rangle^A \, |j\rangle^B\right) \otimes |\psi\rangle^B \tag{6.14}$$

Or simply $\left|\Phi_N^+\right\rangle^{AB} |\psi\rangle^B$. Just as in coherent QT, the parties have managed to recover the entanglement spent during teleportation.

[3]This section is original work although it follows very closely from presentation of the coherent teleportation in [DHW04].

CHAPTER 7

Noisy Protocols, A Family Tree

We have encountered a number of noisy protocols already - in the form of channel coding results. In this chapter we will examine an attempt that was made to tie the disparate coding results together into a "family tree" of quantum protocols. Its success is perhaps the strongest reason to believe that the resource inequality formalism holds some utility, and the relationships discovered are interesting in their own right.

The mother and father resource inequalities as well as their relationships with their "children" were introduced in [DHW04].

7.1 The mother protocol

The mother resource inequality embodies an achievability bound on entanglement purification schemes. Given n instances of a noisy bipartite state ρ^{AB} and access to a noiseless quantum channel, each instance of ρ may be converted into $\frac{1}{2}I(A;B)$ qubits of pure entanglement at the cost of $\frac{1}{2}I(A;E)$ qubits of noiseless communication. As a resource inequality we have

$$\frac{1}{2}I(A;E)_{\psi}\left[q \to q\right] + \left\{qq\right\} \geq \frac{1}{2}I(A;B)_{\psi}\left[qq\right], \tag{7.1}$$

where $|\psi\rangle^{ABE}$ is taken to be a purification of ρ^{AB} with the environment. To sanity check this statement, note that if the input entanglement is pure (thereby requiring no purification), then $I(A;E) = 0$ and $\frac{1}{2}I(A;B)$ is equal to n - unsurprisingly, using zero communication we can turn n entangled qubits into n entangled qubits. If ρ is a product state we find, as we should expect, that $I(A;B) = 0$ since product states can not be distilled into pure entanglement.

By combining the mother with the noiseless superdense coding and teleportation protocols, Devetak, Harrow and Winter showed [DHW04] how to derive three other noisy protocols.

The first, a form of noisy teleportation is obtained by appending teleportation to the mother. Teleportation allows us to convert the mother's $\frac{1}{2}I(A;B)_{\psi}\left[qq\right]$ into qubit commu-

nication at a rate of two classical bits per qubit.

$$I(A;E)_\psi\left[q\to q\right]+\{qq\}+I(A;B)_\psi\left[c\to c\right]\geq\tfrac{1}{2}I(A;B)_\psi\left[q\to q\right] \tag{7.2}$$

We can simplify this further, although as we do it becomes less clear that there exists a protocol satisfying the inequality. Since ψ is pure,

$$\begin{aligned} I(A;B)-I(A;E) &= S(A)+S(B)-S(A,B)-S(A)-S(E)+S(A,E)\\ &= S(A)+S(B)-S(A,B)-S(A)-S(A,B)+S(B)\\ &= 2\ (S(B)-S(A,B))\\ &= 2\ I(A\rangle B). \end{aligned} \tag{7.3}$$

So, subtracting the quantum communication used in the input to the mother from both sides of equation 7.2 we obtain

$$\{qq\}+I(A;B)_\psi\left[c\to c\right]\geq I(A\rangle B)_\psi\left[q\to q\right]. \tag{7.4}$$

The next inequality easily derived from the mother is gotten by prepending teleportation. The resulting inequality happens to correspond to a bound (the hashing inequality [DW03b]) on the one-way distillable entanglement which is known to be optimal.

$$I(A;E)\left[c\to c\right]+\{qq\}\geq I(A\rangle B)\left[qq\right] \tag{7.5}$$

Here the communicating parties have spent $\frac{1}{2}I(A;E)$ qubits of pure entanglement and $I(A;E)$ bits of classical communication in place of the $\frac{1}{2}I(A;E)$ qubits of noiseless communication required by the mother. The input entanglement has been subtracted from the result.

A third inequality follows from appending superdense coding to the mother.

$$\tfrac{1}{2}(I(A;B)+I(A;E))\left[q\to q\right]+\{qq\}\geq I(A;B)\left[c\to c\right] \tag{7.6}$$

The mother's pure entangled output is spent, along with $\frac{1}{2}I(A;B)$ qubits of quantum communication, for a number of classical bits achieving the superdense limit. The qubit com-

munication cost is simplified to $S(A)$ by the following equivalence:

$$\begin{aligned}\tfrac{1}{2}(I(A;B)+I(A;E)) &= \tfrac{1}{2}(S(A)+S(B)-S(A,B)+S(A)+S(E)-S(A,E)) \\ &= \tfrac{1}{2}(S(A)+S(B)-S(E)+S(A)+S(E)-S(B)) \\ &= S(A)\end{aligned} \tag{7.7}$$

7.2 The father protocol

A "father" inequality was also introduced by Devetak, Harrow and Winter in [DHW04]. It is dual to the mother in the sense that it takes an identical form with dynamic resources replacing the mother's static resources and vice versa.

The father is stated as

$$\tfrac{1}{2}I(A;E)\,[qq]+\{q\to q\}\geq\tfrac{1}{2}I(A;B)\,[q\to q]\,. \tag{7.8}$$

Playing a similar game as they did with the mother protocol, Devetak *et al.* were able to re-derive two well known results from their father inequality.

First, by appending superdense coding, a statement of the entanglement-assisted classical capacity is found.

$$S(A)\,[qq]+\{q\to q\}\geq I(A;B)\,[c\to c] \tag{7.9}$$

Again, using equation 7.7 to simplify the input resources.

One final result concludes our traversal of the family tree. By devoting $\frac{1}{2}I(A;B)$ qubits of the father's output to the restoration of the pure entangled input, a channel coding result matching the discussion of section 4.4 is obtained.

$$\{q\to q\}\geq I(A\rangle B)\,[q\to q] \tag{7.10}$$

7.3 More fundamental protocols

Devetak *et al.* failed to find a relationship between their mother and father protocols beyond that induced by swapping dynamic and static resources[1]. The continued search for structure

[1]Some researchers have noted that this leaves us with more of a Brady Bunch than a genealogical family tree.

underlying these relationships has, however, been fruitful. Abeyeshinghe *et al.* [ADHW06] introduced a new resource inequality corresponding to the "mother of all" protocols, the Fully Quantum Slepian-Wolf (FQSW) protocol. They showed how to derive the mother and father protocols of the last section, as well as the state-merging primative of Horodecki *et al.* , from this one inequality.

$$\langle W^{S\to AB} : \varphi^S \rangle + \tfrac{1}{2} I(A;R)_\varphi \left[q \to q\right] \geq \tfrac{1}{2} I(A;B)_\varphi \left[qq\right] + \langle \mathbf{id}^{S\to B} : \varphi^S \rangle \tag{7.11}$$

The proof that protocols exist attaining this trade-off is based on a powerful new technique called *decoupling*.

Bibliography

[ADHW06] Anura Abeyesinghe, Igor Devetak, Patrick Hayden, and Andreas Winter. The mother of all protocols: Restructuring quantum information's family tree. *quant-ph/0606225*, June 2006. Proc. R. Soc. A 465(2108):2537-2563, 2009. 5, 60

[Ali95] Robert Alicki. Comment on "Reduced dynamics need not be completely positive". *Phys. Rev. Lett.*, 75(16):3020, October 1995. 35

[Ari08] Erdal Arikan. Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *0807.3917*, July 2008. 49

[BBC+93] Charles H. Bennett, Gilles Brassard, Claude Crépeau, Richard Jozsa, Asher Peres, and William K. Wootters. Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Physical Review Letters*, 70(13):1895, March 1993. 44, 53

[BBC98] Gilles Brassard, Samuel L. Braunstein, and Richard Cleve. Teleportation as a quantum computation. *Physica D: Nonlinear Phenomena*, 120(1-2):43–47, September 1998. 55

[BDH+09] Charles H Bennett, Igor Devetak, Aram W Harrow, Peter W Shor, and Andreas Winter. Quantum reverse shannon theorem. *0912.5537*, December 2009. 38, 44

[BDS97] Charles H Bennett, David P DiVincenzo, and John A Smolin. Capacities of quantum erasure channels. *Phys. Rev. Lett.*, 78(16):3217–3220, April 1997. 47

[BDSW96] Charles H Bennett, David P DiVincenzo, John A Smolin, and William K Wootters. Mixed state entanglement and quantum error correction. *quant-ph/9604024*, April 1996. Phys.Rev.A54:3824-3851,1996. 47

[Ber10] Herbert J Bernstein. Remote Semi-State preparation as SuperDense quantum teleportation. 2010. 54

[Bor98] Jorge Borges. *Collected fictions*. Viking, New York N.Y. U.S.A., 1998. 26

[BSST01] Charles H Bennett, Peter W Shor, John A Smolin, and Ashish V Thapliyal. Entanglement-assisted capacity of a quantum channel and the reverse shannon theorem. *quant-ph/0106052*, June 2001. 38, 44

[BV97] E. Bernstein and U. Vazirani. Quantum complexity theory. *SIAM J. Comput.*, 26(5):1411–1473, 1997. 4

[BW92] Charles H. Bennett and Stephen J. Wiesner. Communication via one- and two-particle operators on Einstein-Podolsky-Rosen states. *Physical Review Letters*, 69(20):2881, November 1992. 44, 50, 53

[Cho75] M. D Choi. Completely positive linear maps on complex matrices. *Linear Algebra Appl.*, 10:285–290, 1975. 36

[Cov06] T Cover. *Elements of information theory*. Wiley-Interscience, Hoboken N.J., 2nd ed. edition, 2006. 4, 6, 27

[CRS08] Toby S Cubitt, Mary-Beth Ruskai, and Graeme Smith. The structure of degradable quantum channels. *0802.1360*, February 2008. J. Math. Phys. 49, 102104 (2008). 45

[Dev03] I. Devetak. The private classical capacity and quantum capacity of a quantum channel. *quant-ph/0304127*, April 2003. 44

[DHW04] Igor Devetak, Aram W. Harrow, and Andreas Winter. A family of quantum protocols. *Physical Review Letters*, 93(23):230504, December 2004. 56, 57, 59

[DHW08] I. Devetak, A. W Harrow, and A. J Winter. A resource framework for quantum shannon theory. *Information Theory, IEEE Transactions on*, 54(10):4587 –4618, October 2008. 49, 55

[DiV98] David DiVincenzo. Quantum-channel capacity of very noisy channels. *Physical Review A*, 57(2):830–839, 1998. 45

[DS05] I. Devetak and P. W. Shor. The capacity of a quantum channel for simultaneous transmission of classical and quantum information. *Communications in Mathematical Physics*, 256(2):287–303, 2005. 45

[DW03a] I. Devetak and A. Winter. Distilling common randomness from bipartite quantum states. *quant-ph/0304196*, April 2003. 49

[DW03b] Igor Devetak and Andreas Winter. Distillation of secret key and entanglement from quantum states. *quant-ph/0306078*, June 2003. Proc. R. Soc. Lond. A, vol 461, pp 207–235, 2005. 58

[FC96] Christopher A Fuchs and Carlton M Caves. Mathematical techniques for quantum communication theory. *quant-ph/9604001*, April 1996. Open Systems & Information Dynamics 3 (1995) 1. 24

[FvdG97] Christopher A Fuchs and Jeroen van de Graaf. Cryptographic distinguishability measures for quantum mechanical states. *quant-ph/9712042*, December 1997. 21, 24

[Har03] Aram W Harrow. Coherent communication of classical messages. *quant-ph/0307091*, July 2003. Phys. Rev. Lett. 92, 097902 (2004). 55

[Has09] M. B. Hastings. Superadditivity of communication capacity using entangled inputs. *Nat Phys*, 5(4):255–257, April 2009. 38

[HJMR07] Prahladh Harsha, Rahul Jain, David McAllester, and Jaikumar Radhakrishnan. The communication complexity of correlation. *Computational Complexity, Annual IEEE Conference on*, 0:10–23, 2007. 5

[HJS+96] Paul Hausladen, Richard Jozsa, Benjamin Schumacher, Michael Westmoreland, and William K. Wootters. Classical information capacity of a quantum channel. *Physical Review A*, 54(3):1869, 1996. 39, 40, 41

[Hol98] A. S Holevo. The capacity of the quantum channel with general signal states. *Information Theory, IEEE Transactions on*, 44(1):269 – 273, January 1998. 41

[HOW05] Michal Horodecki, Jonathan Oppenheim, and Andreas Winter. Quantum information can be negative. *quant-ph/0505062*, May 2005. Nature 436:673-676 (2005) as "Partial quantum information". 16

[Iva87] I. D. Ivanovic. How to differentiate between non-orthogonal states. *Physics Letters A*, 123(6):257–259, August 1987. 18, 20

[Joz94] R. Jozsa. Fidelity for mixed quantum states. *Journal of Modern Optics*, 41(12):2315–2323, 1994. 22

[Llo97] Seth Lloyd. Capacity of the noisy quantum channel. *Physical Review A*, 55(3):1613, March 1997. 44

[Mer07] N. David Mermin. *Quantum Computer Science: An Introduction*. Cambridge University Press, 1 edition, September 2007. 4

[NC00] Michael Nielsen and Isaac Chuang. *Quantum computation and quantum information*. Cambridge University Press, Cambridge ;;New York, 2000. 4, 23, 31, 42

[Pec94] Philip Pechukas. Reduced dynamics need not be completely positive. *Phys. Rev. Lett.*, 73(8):1060–1062, August 1994. 35

[RB01] Robert Raussendorf and Hans J. Briegel. A One-Way quantum computer. *Physical Review Letters*, 86(22):5188, May 2001. 18

[REK05] Jaroslav Rcaroneháccaronek, Berthold-Georg Englert, and Dagomir Kaszlikowski. Iterative procedure for computing accessible information in quantum communication. *Physical Review A*, 71(5):054303, May 2005. 21

[Sch95] Benjamin Schumacher. Quantum coding. *Physical Review A*, 51(4):2738, April 1995. 15, 29, 30, 32

[Sch96] Benjamin Schumacher. Sending quantum entanglement through noisy channels. *quant-ph/9604023*, April 1996. 34, 37

[Sha48] Claude Shannon. A mathematical theory of communication. *The Bell Systems Technical Journal*, 27:379–423,623–656, 1948. 6, 15, 28

[SN96] Benjamin Schumacher and M. A. Nielsen. Quantum data processing and error correction. *Physical Review A*, 54(4):2629, October 1996. 45, 48

[Sti55] W. F Stinespring. Positive functions on c*-algebras. In *Proc. Amer. Math. Soc*, volume 6, page 3, 1955. 36

[SW97] Benjamin Schumacher and Michael D. Westmoreland. Sending classical information via noisy quantum channels. *Physical Review A*, 56(1):131, July 1997. 41

[SY08] Graeme Smith and Jon Yard. Quantum communication with Zero-Capacity channels. *0807.4935*, July 2008. Science 321, 1812 - 1815 (2008). 39, 45

[TM71] M. Tribus and E.C. McIrvine. Energy and information. *Scientific American*, (224):178–184, September 1971. 15

[Uhl76] A. Uhlmann. The 'transition probability' in the state space of a *-algebra. *Reports on Mathematical Physics*, 9(2):273 – 279, 1976. 23

[Win99] A. Winter. Coding theorem and strong converse for quantum channels. *Information Theory, IEEE Transactions on*, 45(7):2481 –2485, November 1999. 30

[WZ82] WK Wootters and WH Zurek. A single quantum cannot be cloned. *Nature*, 299(5886):802–803, October 1982. 13, 14