Quantum Information
Chapter 10. Quantum Shannon Theory

John Preskill
Institute for Quantum Information and Matter
California Institute of Technology

Updated June 2016

For further updates and additional chapters, see:
http://www.theory.caltech.edu/people/preskill/ph219/

Please send corrections to preskill@caltech.edu

Contents

Preface

This is the 10th and final chapter of my book *Quantum Information*, based on the course I have been teaching at Caltech since 1997. An early version of this chapter (originally Chapter 5) has been available on the course website since 1998, but this version is substantially revised and expanded.

The level of detail is uneven, as I've aimed to provide a gentle introduction, but I've also tried to avoid statements that are incorrect or obscure. Generally speaking, I chose to include topics that are both useful to know and relatively easy to explain; I had to leave out a lot of good stuff, but on the other hand the chapter is already quite long.

My version of Quantum Shannon Theory is no substitute for the more careful treatment in Wilde's book [1], but it may be more suitable for beginners. This chapter contains occasional references to earlier chapters in my book, but I hope it will be intelligible when read independently of other chapters, including the chapter on quantum error-correcting codes.

This is a working draft of Chapter 10, which I will continue to update. See the URL on the title page for further updates and drafts of other chapters. Please send an email to preskill@caltech.edu if you notice errors.

Eventually, the complete book will be published by Cambridge University Press. I hesitate to predict the publication date — they have been far too patient with me.

10
Quantum Shannon Theory

Quantum information science is a synthesis of three great themes of 20th century thought: quantum physics, computer science, and information theory. Up until now, we have given short shrift to the information theory side of this trio, an oversight now to be remedied.

A suitable name for this chapter might have been *Quantum Information Theory*, but I prefer for that term to have a broader meaning, encompassing much that has already been presented in this book. Instead I call it *Quantum Shannon Theory*, to emphasize that we will mostly be occupied with generalizing and applying Claude Shannon's great (classical) contributions to a quantum setting. Quantum Shannon theory has several major thrusts:

1. Compressing quantum information.

2. Transmitting classical and quantum information through noisy quantum channels.

3. Quantifying, characterizing, transforming, and using quantum entanglement.

A recurring theme unites these topics — the properties, interpretation, and applications of Von Neumann entropy.

My goal is to introduce some of the main ideas and tools of quantum Shannon theory, but there is a lot we won't cover. For example, we will mostly consider information theory in an *asymptotic setting*, where the same quantum channel or state is used arbitrarily many times, thus focusing on issues of principle rather than more practical questions about devising efficient protocols.

10.1 Shannon for Dummies

Before we can understand Von Neumann entropy and its relevance to quantum information, we should discuss Shannon entropy and its relevance to classical information.

Claude Shannon established the two core results of classical information theory in his landmark 1948 paper. The two central problems that he solved were:

1. How much can a message be *compressed*; *i.e.*, how redundant is the information? This question is answered by the "source coding theorem," also called the "noiseless coding theorem."

2. At what *rate* can we communicate reliably over a noisy channel; *i.e.*, how much redundancy must be incorporated into a message to protect against errors? This question is answered by the "noisy channel coding theorem."

Both questions concern *redundancy* – how *unexpected* is the next letter of the message, on the average. One of Shannon's key insights was that *entropy* provides a suitable way to quantify redundancy.

I call this section "Shannon for Dummies" because I will try to explain Shannon's ideas quickly, minimizing distracting details. That way, I can compress classical information theory to about 14 pages.

10.1.1 Shannon entropy and data compression

A message is a string of letters, where each letter is chosen from an alphabet of k possible letters. We'll consider an idealized setting in which the message is produced by an "information source" which picks each letter by sampling from a probability distribution

$$X := \{x, p(x)\}; \tag{10.1}$$

that is, the letter has the value

$$x \in \{0, 1, 2, \ldots k-1\} \tag{10.2}$$

with probability $p(x)$. If the source emits an n-letter message the particular string $x = x_1 x_2 \ldots x_n$ occurs with probability

$$p(x_1 x_2 \ldots x_n) = \prod_{i=1}^{n} p(x_i). \tag{10.3}$$

Since the letters are statistically independent, and each is produced by consulting the same probability distribution X, we say that the letters

are *independent and identically distributed*, abbreviated *i.i.d.* We'll use X^n to denote the ensemble of n-letter messages in which each letter is generated independently by sampling from X, and $\vec{x} = (x_1 x_2 \ldots x_n)$ to denote a string of bits.

Now consider long n-letter messages, $n \gg 1$. We ask: is it possible to compress the message to a shorter string of letters that conveys essentially the same information? The answer is: Yes, it's possible, unless the distribution X is uniformly random.

If the alphabet is binary, then each letter is either 0 with probability $1-p$ or 1 with probability p, where $0 \le p \le 1$. For n very large, the law of large numbers tells us that typical strings will contain about $n(1-p)$ 0's and about np 1's. The number of distinct strings of this form is of order the binomial coefficient $\binom{n}{np}$, and from the Stirling approximation $\log n! = n \log n - n + O(\log n)$ we obtain

$$\begin{aligned} \log \binom{n}{np} &= \log \left(\frac{n!}{(np)!\,(n(1-p))!} \right) \\ &\approx n \log n - n - (np \log np - np + n(1-p) \log n(1-p) - n(1-p)) \\ &= nH(p), \end{aligned} \tag{10.4}$$

where

$$H(p) = -p \log p - (1-p) \log(1-p) \tag{10.5}$$

is the *entropy* function.

In this derivation we used the Stirling approximation in the appropriate form for natural logarithms. But from now on we will prefer to use logarithms with base 2, which is more convenient for expressing a quantity of information in bits; thus if no base is indicated, it will be understood that the base is 2 unless otherwise stated. Adopting this convention in the expression for $H(p)$, the number of typical strings is of order $2^{nH(p)}$.

To convey essentially all the information carried by a string of n bits, it suffices to choose a block code that assigns a nonnegative integer to each of the typical strings. This block code needs to distinguish about $2^{nH(p)}$ messages (all occurring with nearly equal *a priori* probability), so we may specify any one of the messages using a binary string with length only slightly longer than $nH(p)$. Since $0 \le H(p) \le 1$ for $0 \le p \le 1$, and $H(p) = 1$ only for $p = \frac{1}{2}$, the block code shortens the message for any $p \ne \frac{1}{2}$ (whenever 0 and 1 are not equally probable). This is Shannon's result. The key idea is that we do not need a codeword for every sequence of letters, only for the *typical* sequences. The probability that the actual message is atypical becomes negligible asymptotically, *i.e.*, in the limit $n \to \infty$.

Similar reasoning applies to the case where X samples from a k-letter alphabet. In a string of n letters, x typically occurs about $np(x)$ times, and the number of typical strings is of order

$$\frac{n!}{\prod_x (np(x))!} \simeq 2^{-nH(X)}, \tag{10.6}$$

where we have again invoked the Stirling approximation and now

$$H(X) = -\sum_x p(x) \log_2 p(x). \tag{10.7}$$

is the *Shannon* entropy (or simply entropy) of the ensemble $X = \{x, p(x)\}$. Adopting a block code that assigns integers to the typical sequences, the information in a string of n letters can be compressed to about $nH(X)$ bits. In this sense a letter x chosen from the ensemble carries, on the average, $H(X)$ bits of information.

It is useful to restate this reasoning more carefully using the *strong law of large numbers*, which asserts that a sample average for a random variable almost certainly converges to its expected value in the limit of many trials. If we sample from the distribution $Y = \{y, p(y)\}$ n times, let $y_i, i \in \{1, 2, \ldots, n\}$ denote the ith sample, and let

$$\mu[Y] = \langle y \rangle = \sum_y y \; p(y) \tag{10.8}$$

denote the expected value of y. Then for any positive ε and δ there is a positive integer N such that

$$\left| \frac{1}{n} \sum_{i=1}^{n} y_i - \mu[Y] \right| \leq \delta \tag{10.9}$$

with probability at least $1 - \varepsilon$ for all $n \geq N$. We can apply this statement to the random variable $\log_2 p(x)$. Let us say that a sequence of n letters is δ-*typical* if

$$H(X) - \delta \leq -\frac{1}{n} \log_2 p(x_1 x_2 \ldots x_n) \leq H(X) + \delta; \tag{10.10}$$

then the strong law of large numbers says that for any $\varepsilon, \delta > 0$ and n sufficiently large, an n-letter sequence will be δ-typical with probability $\geq 1 - \varepsilon$.

Since each δ-typical n-letter sequence $\vec{x}$ occurs with probability $p(\vec{x})$ satisfying

$$p_{\min} = 2^{-n(H+\delta)} \leq p(\vec{x}) \leq 2^{-n(H-\delta)} = p_{\max}, \tag{10.11}$$

we may infer upper and lower bounds on the *number* $N_{\rm typ}(\varepsilon,\delta,n)$ of typical sequences:

$$N_{\rm typ}\, p_{\rm min} \le \sum_{\text{typical } x} p(x) \le 1, \quad N_{\rm typ}\, p_{\rm max} \ge \sum_{\text{typical } x} p(x) \ge 1-\varepsilon, \tag{10.12}$$

implies

$$2^{n(H+\delta)} \ge N_{\rm typ}(\varepsilon,\delta,n) \ge (1-\varepsilon)2^{n(H-\delta)}. \tag{10.13}$$

Therefore, we can encode all typical sequences using a block code with length $n(H+\delta)$ bits. That way, any message emitted by the source can be compressed and decoded successfully as long as the message is typical; the compression procedure achieves a success probability $p_{\rm success} \ge 1-\varepsilon$, no matter how the atypical sequences are decoded.

What if we try to compress the message even further, say to $H(X)-\delta'$ bits per letter, where δ' is a constant independent of the message length n? Then we'll run into trouble, because there won't be enough codewords to cover all the typical messages, and we won't be able to decode the compressed message with negligible probability of error. The probability $p_{\rm success}$ of successfully decoding the message will be bounded above by

$$p_{\rm success} \le 2^{n(H-\delta')}2^{-n(H-\delta)} + \varepsilon = 2^{-n(\delta'-\delta)} + \varepsilon; \tag{10.14}$$

we can correctly decode only $2^{n(H-\delta')}$ typical messages, each occurring with probability no higher than $2^{-n(H-\delta)}$; we add ε, an upper bound on the probability of an atypical message, allowing optimistically for the possibility that we somehow manage to decode the atypical messages correctly. Since we may choose ε and δ as small as we please, this success probability becomes small as $n\to\infty$, if δ' is a positive constant.

The number of bits per letter encoding the compressed message is called the *rate* of the compression code, and we say a rate R is *achievable* asymptotically (as $n\to\infty$) if there is a sequence of codes with rate at least R and error probability approaching zero in the limit of large n. To summarize our conclusion, we have found that

$$\begin{aligned} &\text{Compression Rate} = H(X) + o(1) \text{ is } \textit{achievable},\\ &\text{Compression Rate} = H(X) - \Omega(1) \text{ is } \textit{not achievable}, \end{aligned} \tag{10.15}$$

where $o(1)$ denotes a positive quantity which may be chosen as small as we please, and $\Omega(1)$ denotes a positive constant. This is Shannon's source coding theorem.

We have not discussed at all the details of the compression code. We might imagine a huge lookup table which assigns a unique codeword to each message and vice versa, but because such a table has size exponential in n it is quite impractical for compressing and decompressing long

messages. It is fascinating to study how to make the coding and decoding efficient while preserving a near optimal rate of compression, and quite important, too, if we really want to compress something. But this practical aspect of classical compression theory is beyond the scope of this book.

10.1.2 Joint typicality, conditional entropy, and mutual information

The Shannon entropy quantifies my *ignorance* per letter about the output of an information source. If the source X produces an n-letter message, then $n(H(X) + o(1))$ bits suffice to convey the content of the message, while $n(H(X) - \Omega(1))$ bits do not suffice.

Two information sources X and Y can be correlated. Letters drawn from the sources are governed by a joint distribution $XY = \{(x, y), p(x, y)\}$, in which a pair of letters (x, y) appears with probability $p(x, y)$. The sources are independent if $p(x, y) = p(x)p(y)$, but correlated otherwise. If XY is a joint distribution, we use X to denote the marginal distribution, defined as

$$X = \left\{ x, p(x) = \sum_y p(x, y) \right\}, \tag{10.16}$$

and similarly for Y. If X and Y are correlated, then by reading a message generated by Y^n I reduce my ignorance about a message generated by X^n, which should make it possible to compress the output of X further than if I did not have access to Y.

To make this idea more precise, we use the concept of *jointly typical sequences*. Sampling from the distribution X^nY^n, that is, sampling n times from the joint distribution XY, produces a message $(\vec{x}, \vec{y}) = (x_1x_2 \dots x_n, y_1y_2 \dots y_n)$ with probability

$$p(\vec{x}, \vec{y}) = p(x_1, y_1)p(x_2, y_2) \dots p(x_n, y_n). \tag{10.17}$$

Let us say that $(\vec{x}, \vec{y})$ drawn from X^nY^n is *jointly δ-typical* if

$$\begin{aligned} 2^{-n(H(X)+\delta)} &\le p(\vec{x}) \le 2^{-n(H(X)-\delta)}, \\ 2^{-n(H(Y)+\delta)} &\le p(\vec{y}) \le 2^{-n(H(Y)-\delta)}, \\ 2^{-n(H(XY)+\delta)} &\le p(\vec{x}, \vec{y}) \le 2^{-n(H(XY)-\delta)}. \end{aligned} \tag{10.18}$$

Then, applying the strong law of large numbers simultaneously to the three distributions X^n, Y^n, and X^nY^n, we infer that for $\varepsilon, \delta > 0$ and n sufficiently large, a sequence drawn from X^nY^n will be δ-typical with

probability $\geq 1-\varepsilon$. Using Bayes' rule, we can then obtain upper and lower bounds on the *conditional* probability $p(\vec{x}|\vec{y})$ for jointly typical sequences:

$$p(\vec{x}|\vec{y}) = \frac{p(\vec{x},\vec{y})}{p(\vec{y})} \geq \frac{2^{-n(H(XY)+\delta)}}{2^{-n(H(Y)-\delta)}} = 2^{-n(H(X|Y)+2\delta)},$$
$$p(\vec{x}|\vec{y}) = \frac{p(\vec{x},\vec{y})}{p(\vec{y})} \leq \frac{2^{-n(H(XY)-\delta)}}{2^{-n(H(Y)+\delta)}} = 2^{-n(H(X|Y)-2\delta)}. \tag{10.19}$$

Here we have introduced the quantity

$$H(X|Y) = H(XY) - H(Y) = \langle -\log p(x,y) + \log p(y)\rangle = \langle -\log p(x|y)\rangle, \tag{10.20}$$

which is called the *conditional entropy* of X given Y.

The conditional entropy quantifies my *remaining* ignorance about x once I know y. From eq.(10.19) we see that if $(\vec{x},\vec{y})$ is jointly typical (as is the case with high probability for n large), then the number of possible values for $\vec{x}$ compatible with the known value of $\vec{y}$ is no more than $2^{n(H(X|Y)+2\delta)}$; hence we can convey $\vec{x}$ with a high success probability using only $H(X|Y) + o(1)$ bits per letter. On the other hand we can't do much better, because if we use only $2^{n(H(X|Y)-\delta')}$ codewords, we are limited to conveying reliably no more than a fraction $2^{-n(\delta'-2\delta)}$ of all the jointly typical messages. To summarize, $H(X|Y)$ is the number of *additional* bits per letter needed to specify *both* $\vec{x}$ and $\vec{y}$ once $\vec{y}$ is known. Similarly, $H(Y|X)$ is the number of additional bits per letter needed to specify both $\vec{x}$ and $\vec{y}$ when $\vec{x}$ is known.

The information about X that I *gain* when I learn Y is quantified by how much the number of bits per letter needed to specify X is *reduced* when Y is known. Thus is

$$\begin{aligned} I(X;Y) &\equiv H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(XY) \\ &= H(Y) - H(Y|X), \end{aligned} \tag{10.21}$$

which is called the *mutual information*. The mutual information $I(X;Y)$ quantifies how X and Y are correlated, and is symmetric under interchange of X and Y: I find out as much about X by learning Y as about Y by learning X. Learning Y never *reduces* my knowledge of X, so $I(X;Y)$ is obviously nonnegative, and indeed the inequality $H(X) \geq H(X|Y) \geq 0$ follows easily from the convexity of the log function.

Of course, if X and Y are completely uncorrelated, we have $p(x,y) = p(x)p(y)$, and

$$I(X;Y) \equiv \left\langle \log \frac{p(x,y)}{p(x)p(y)} \right\rangle = 0; \tag{10.22}$$

we don't find out anything about X by learning Y if there is no correlation between X and Y.

10.1.3 Distributed source coding

To sharpen our understanding of the operational meaning of conditional entropy, consider this situation: Suppose that the joint distribution XY is sampled n times, where Alice receives the n-letter message $\vec{x}$ and Bob receives the n-letter message $\vec{y}$. Now Alice is to send a message to Bob which will enable Bob to determine $\vec{x}$ with high success probability, and Alice wants to send as few bits to Bob as possible. This task is harder than in the scenario considered in §10.1.2, where we assumed that the encoder and the decoder share full knowledge of $\vec{y}$, and can choose their code for compressing $\vec{x}$ accordingly. It turns out, though, that even in this more challenging setting Alice can compress the message she sends to Bob down to $n\left(H(X|Y) + o(1)\right)$ bits, using a method called *Slepian-Wolf coding*.

Before receiving $(\vec{x}, \vec{y})$, Alice and Bob agree to sort all the possible n-letter messages that Alice might receive into 2^{nR} possible bins of equal size, where the choice of bins is known to both Alice and Bob. When Alice receives $\vec{x}$, she sends nR bits to Bob, identifying the bin that contains $\vec{x}$. After Bob receives this message, he knows both $\vec{y}$ and the bin containing $\vec{x}$. If there is a unique message in that bin which is jointly typical with $\vec{y}$, Bob decodes accordingly. Otherwise, he decodes arbitrarily. This procedure can fail either because $\vec{x}$ and $\vec{y}$ are not jointly typical, or because there is more than one message in the bin which is jointly typical with $\vec{x}$. Otherwise, Bob is sure to decode correctly.

Since $\vec{x}$ and $\vec{y}$ are jointly typical with high probability, the compression scheme works if it is unlikely for a bin to contain an incorrect message which is jointly typical with $\vec{y}$. If $\vec{y}$ is typical, what can we say about the number $N_{\text{typ}|\vec{y}}$ of messages $\vec{x}$ that are jointly typical with $\vec{y}$? Using eq.(10.19), we have

$$1 \geq \sum_{\text{typical } \vec{x}|\vec{y}} p(\vec{x}|\vec{y}) \geq \quad N_{\text{typ}|\vec{y}}\, 2^{-n(H(X|Y)+2\delta)}, \tag{10.23}$$

and thus

$$N_{\text{typ}|\vec{y}} \leq 2^{n(H(X|Y)+2\delta)}. \tag{10.24}$$

Now, to estimate the probability of a decoding error, we need to specify how the bins are chosen. Let's assume the bins are chosen uniformly at random, or equivalently, let's consider averaging uniformly over all codes that divide the length-n strings into 2^{nR} bins of equal size. Then the

probability that a particular bin contains a message jointly typical with a specified $\vec{y}$ purely by accident is bounded above by

$$2^{-nR} N_{\text{typ}|\vec{y}} \geq 2^{-n(R-H(X|Y)-2\delta)}. \tag{10.25}$$

We conclude that if Alice sends R bits to Bob per each letter of the message x, where

$$R = H(X|Y) + o(1), \tag{10.26}$$

then the probability of a decoding error vanishes in the limit $n \to \infty$, at least when we average over uniformly all codes. Surely, then, there must exist a particular sequence of codes Alice and Bob can use to achieve the rate $R = H(X|Y) + o(1)$, as we wanted to show.

In this scenario, Alice and Bob jointly know (x, y), but initially neither Alice nor Bob has access to all their shared information. The goal is to merge all the information on Bob's side with minimal communication from Alice to Bob, and we have found that $H(X|Y) + o(1)$ bits of communication per letter suffice for this purpose. Similarly, the information can be merged on Alice's side using $H(Y|X) + o(1)$ bits of communication per letter from Bob to Alice.

10.1.4 The noisy channel coding theorem

Suppose Alice wants to send a message to Bob, but the communication channel linking Alice and Bob is noisy. Each time they use the channel, Bob receives the letter y with probability $p(y|x)$ if Alice sends the letter x. Using the channel $n \gg 1$ times, Alice hopes to transmit a long message to Bob.

Alice and Bob realize that to communicate reliably despite the noise they should use some kind of code. For example, Alice might try sending the same bit k times, with Bob using a majority vote of the k noisy bits he receives to decode what Alice sent. One wonders: for a given channel, is it possible to ensure perfect transmission asymptotically, *i.e.*, in the limit where the number of channel uses $n \to \infty$? And what can be said about the *rate* of the code; that is, how many bits must be sent per letter of the transmitted message?

Shannon answered these questions. He showed that *any* channel can be used for perfectly reliable communication at an asymptotic nonzero rate, as long as there is *some* correlation between the channel's input and its output. Furthermore, he found a useful formula for the optimal rate that can be achieved. These results are the content of the *noisy channel coding theorem*.

Capacity of the binary symmetric channel. To be concrete, suppose we use the binary alphabet $\{0, 1\}$, and the *binary symmetric channel*; this channel acts on each bit independently, flipping its value with probability p, and leaving it intact with probability $1 - p$. Thus the conditional probabilities characterizing the channel are

$$\begin{array}{ll} p(0|0) = 1 - p, & p(0|1) = p, \\ p(1|0) = p, & p(1|1) = 1 - p. \end{array} \tag{10.27}$$

We want to construct a family of codes with increasing block size n, such that the probability of a decoding error goes to zero as $n \to \infty$. For each n, the code contains 2^k *codewords* among the 2^n possible strings of length n. The rate R of the code, the number of encoded data bits transmitted per physical bit carried by the channel, is

$$R = \frac{k}{n}. \tag{10.28}$$

To protect against errors, we should choose the code so that the codewords are as "far apart" as possible. For given values of n and k, we want to maximize the number of bits that must be flipped to change one codeword to another, the *Hamming distance* between the two codewords. For any n-bit input message, we expect about np of the bits to flip — the input diffuses into one of about $2^{nH(p)}$ typical output strings, occupying an "error sphere" of "Hamming radius" np about the input string. To decode reliably, we want to choose our input codewords so that the error spheres of two different codewords do not overlap substantially. Otherwise, two different inputs will sometimes yield the same output, and decoding errors will inevitably occur. To avoid such decoding ambiguities, the total number of strings contained in all $2^k = 2^{nR}$ error spheres should not exceed the total number 2^n of bits in the output message; we therefore require

$$2^{nH(p)} 2^{nR} \leq 2^n \tag{10.29}$$

or

$$R \leq 1 - H(p) := C(p). \tag{10.30}$$

If transmission is highly reliable, we cannot expect the rate of the code to exceed $C(p)$. But is the rate $R = C(p)$ actually *achievable* asymptotically?

In fact transmission with $R = C - o(1)$ and negligible decoding error probability is possible. Perhaps Shannon's most ingenious idea was that this rate can be achieved by an average over "random codes." Though choosing a code at random does not seem like a clever strategy, rather surprisingly it turns out that random coding achieves as high a rate as any other coding scheme in the limit $n \to \infty$. Since C is the optimal rate

for reliable transmission of data over the noisy channel it is called the *channel capacity.*

Suppose that X is the uniformly random ensemble for a single bit (either 0 with $p = \frac{1}{2}$ or 1 with $p = \frac{1}{2}$), and that we sample from X^n a total of 2^{nR} times to generate 2^{nR} "random codewords." The resulting code is known by both Alice and Bob. To send nR bits of information, Alice chooses one of the codewords and sends it to Bob by using the channel n times. To decode the n-bit message he receives, Bob draws a "Hamming sphere" with "radius" slightly large than np, containing

$$2^{n(H(p)+\delta)} \tag{10.31}$$

strings. If this sphere contains a unique codeword, Bob decodes the message accordingly. If the sphere contains more than one codeword, or no codewords, Bob decodes arbitrarily.

How likely is a decoding error? For any positive δ, Bob's decoding sphere is large enough that it is very likely to contain the codeword sent by Alice when n is sufficiently large. Therefore, we need only worry that the sphere might contain another codeword just by accident. Since there are altogether 2^n possible strings, Bob's sphere contains a fraction

$$f = \frac{2^{n(H(p)+\delta)}}{2^n} = 2^{-n(C(p)-\delta)}, \tag{10.32}$$

of all the strings. Because the codewords are uniformly random, the probability that Bob's sphere contains any particular codeword aside from the one sent by Alice is f, and the probability that the sphere contains any one of the $2^{nR} - 1$ invalid codewords is no more than

$$2^{nR} f = 2^{-n(C(p)-R-\delta)}. \tag{10.33}$$

Since δ may be as small as we please, we may choose $R = C(p) - c$ where c is any positive constant, and the decoding error probability will approach zero as $n \to \infty$.

When we speak of codes chosen at random, we really mean that we are averaging over many possible codes. The argument so far has shown that the *average* probability of error is small, where we average over the choice of random code, and for each specified code we also average over all codewords. It follows that there must be a particular sequence of codes such that the average probability of error (when we average over the codewords) vanishes in the limit $n \to \infty$. We would like a stronger result – that the probability of error is small for *every* codeword.

To establish the stronger result, let p_i denote the probability of a decoding error when codeword i is sent. For any positive ε and sufficiently

large n, we have demonstrated the existence of a code such that

$$\frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} p_i \leq \varepsilon. \tag{10.34}$$

Let $N_{2\varepsilon}$ denote the number of codewords with $p_i \geq 2\varepsilon$. Then we infer that

$$\frac{1}{2^{nR}}(N_{2\varepsilon})2\varepsilon \leq \varepsilon \text{ or } N_{2\varepsilon} \leq 2^{nR-1}; \tag{10.35}$$

we see that we can throw away at most half of the codewords, to achieve $p_i \leq 2\varepsilon$ for *every* codeword. The new code we have constructed has

$$\text{Rate} = R - \frac{1}{n}, \tag{10.36}$$

which approaches R as $n \to \infty$. We have seen, then, that the rate $R = C(p) - o(1)$ is asymptotically achievable with negligible probability of error, where $C(p) = 1 - H(p)$.

Mutual information as an achievable rate. Now consider how to apply this random coding argument to more general alphabets and channels. The channel is characterized by $p(y|x)$, the conditional probability that the letter y is received when the letter x is sent. We fix an ensemble $X = \{x, p(x)\}$ for the input letters, and generate the codewords for a length-n code with rate R by sampling 2^{nR} times from the distribution X^n; the code is known by both the sender Alice and the receiver Bob. To convey an encoded nR-bit message, one of the 2^{nR} n-letter codewords is selected and sent by using the channel n times. The channel acts independently on the n letters, governed by the same conditional probability distribution $p(y|x)$ each time it is used. The input ensemble X, together with the conditional probability characterizing the channel, determines the joint ensemble XY for each letter sent, and therefore the joint ensemble X^nY^n for the n uses of the channel.

To define a decoding procedure, we use the notion of joint typicality introduced in §10.1.2. When Bob receives the n-letter output message $\vec{y}$, he determines whether there is an n-letter input codeword $\vec{x}$ jointly typical with $\vec{y}$. If such $\vec{x}$ exists and is unique, Bob decodes accordingly. If there is no $\vec{x}$ jointly typical with $\vec{y}$, or more than one such $\vec{x}$, Bob decodes arbitrarily.

How likely is a decoding error? For any positive ε and δ, the $(\vec{x}, \vec{y})$ drawn from X^nY^n is jointly δ-typical with probability at least $1 - \varepsilon$ if n is sufficiently large. Therefore, we need only worry that there might more than one codeword jointly typical with $\vec{y}$.

Suppose that Alice samples X^n to generate a codeword $\vec{x}$, which she sends to Bob using the channel n times. Then Alice samples X^n a second time, producing another codeword $\vec{x}'$. With probability close to one, both $\vec{y}$ and $\vec{x}'$ are δ-typical. But what is the probability that $\vec{x}'$ is *jointly* δ-typical with $\vec{y}$?

Because the samples are independent, the probability of drawing these two codewords factorizes as $p(\vec{x}', \vec{x}) = p(\vec{x}')p(\vec{x})$, and likewise the channel output $\vec{y}$ when the first codeword is sent is independent of the second channel input $\vec{x}'$, so $p(\vec{x}', \vec{y}) = p(\vec{x}')p(\vec{y})$. From eq.(10.18) we obtain an upper bound on the number $N_{\text{j.t.}}$ of jointly δ-typical $(\vec{x}, \vec{y})$:

$$1 \geq \sum_{\text{j.t. } (\vec{x},\vec{y})} p(\vec{x}, \vec{y}) \geq N_{\text{j.t.}}\, 2^{-n(H(XY)+\delta)} \implies N_{\text{j.t.}} \leq 2^{n(H(XY)+\delta)}. \tag{10.37}$$

We also know that each δ-typical $\vec{x}'$ occurs with probability $p(\vec{x}') \leq 2^{-n(H(X)-\delta)}$ and that each δ-typical $\vec{y}$ occurs with probability $p(\vec{y}) \leq 2^{-n(H(Y)-\delta)}$. Therefore, the probability that $\vec{x}'$ and $\vec{y}$ are jointly δ-typical is bounded above by

$$\begin{aligned} \sum_{\text{j.t. } (\vec{x}',\vec{y})} p(\vec{x}')p(\vec{y}) &\leq N_{\text{j.t.}}\, 2^{-n(H(X)-\delta)} 2^{-n(H(Y)-\delta)} \\ &\leq 2^{n(H(XY)+\delta)} 2^{-n(H(X)-\delta)} 2^{-n(H(Y)-\delta)} \\ &= 2^{-n(I(X;Y)-3\delta)}. \end{aligned} \tag{10.38}$$

If there are 2^{nR} codewords, all generated independently by sampling X^n, then the probability that *any* other codeword besides $\vec{x}$ is jointly typical with $\vec{y}$ is bounded above by

$$2^{nR} 2^{-n(I(X;Y)-3\delta)} = 2^{n(R-I(X;Y)+3\delta)}. \tag{10.39}$$

Since ε and δ are as small as we please, we may choose $R = I(X;Y) - c$, where c is any positive constant, and the decoding error probability will approach zero as $n \to \infty$.

So far we have shown that the error probability is small when we average over codes and over codewords. To complete the argument we use the same reasoning as in our discussion of the capacity of the binary symmetric channel. There must exist a particular sequence of code with zero error probability in the limit $n \to \infty$, when we average over codewords. And by pruning the codewords, reducing the rate by a negligible amount, we can ensure that the error probability is small for *every* codeword. We conclude that the rate

$$R = I(X;Y) - o(1) \tag{10.40}$$

is asymptotically achievable with negligible probability of error. This result provides a concrete operational interpretation for the mutual information $I(X;Y)$; it is the information per letter we can transmit over the channel, supporting the heuristic claim that $I(X;Y)$ quantifies the information we gain about X when we have access to Y.

The mutual information $I(X;Y)$ depends not only on the channel's conditional probability $p(y|x)$ but also on the *a priori* probability $p(x)$ defining the codeword ensemble X. The achievability argument for random coding applies for any choice of X, so we have demonstrated that errorless transmission over the noisy channel is possible for any rate R strictly less than

$$C := \max_X I(X;Y). \tag{10.41}$$

This quantity C is called the *channel capacity*; it depends only on the conditional probabilities $p(y|x)$ that define the channel.

Upper bound on the capacity. We have now shown that any rate $R < C$ is achievable, but can R exceed C with the error probability still approaching 0 for large n? To see that a rate for errorless transmission exceeding C is not possible, we reason as follows.

Consider any code with 2^{nR} codewords, and consider the uniform ensemble on the codewords, denoted $\tilde{X}^n$, in which each codeword occurs with probability 2^{-nR}. Evidently, then,

$$H(\tilde{X}^n) = nR. \tag{10.42}$$

Sending the codewords through n uses of the channel we obtain an ensemble $\tilde{Y}^n$ of output states, and a joint ensemble $\tilde{X}^n\tilde{Y}^n$.

Because the channel acts on each letter independently, the conditional probability for n uses of the channel factorizes:

$$p(y_1y_2\cdots y_n|x_1x_2\cdots x_n) = p(y_1|x_1)p(y_2|x_2)\cdots p(y_n|x_n), \tag{10.43}$$

and it follows that the conditional entropy satisfies

$$\begin{aligned} H(\tilde{Y}^n|\tilde{X}^n) = \langle -\log p(\vec{y}|\vec{x})\rangle &= \sum_i \langle -\log p(y_i|x_i)\rangle \\ &= \sum_i H(\tilde{Y}_i|\tilde{X}_i), \end{aligned} \tag{10.44}$$

where $\tilde{X}_i$ and $\tilde{Y}_i$ are the marginal probability distributions for the ith letter determined by our distribution on the codewords. Because Shannon entropy is subadditive, $H(XY) \le H(X) + H(Y)$, we have

$$H(\tilde{Y}^n) \le \sum_i H(\tilde{Y}_i), \tag{10.45}$$

and therefore

$$\begin{aligned} I(\tilde{Y}^n;\tilde{X}^n) &= H(\tilde{Y}^n) - H(\tilde{Y}^n|\tilde{X}^n) \\ &\leq \sum_i (H(\tilde{Y}_i) - H(\tilde{Y}_i|\tilde{X}_i)) \\ &= \sum_i I(\tilde{Y}_i;\tilde{X}_i) \leq nC. \end{aligned} \tag{10.46}$$

The mutual information of the messages sent and received is bounded above by the sum of the mutual information per letter, and the mutual information for each letter is bounded above by the capacity, because C is defined as the maximum of $I(X;Y)$ over all input ensembles.

Recalling the symmetry of mutual information, we have

$$\begin{aligned} I(\tilde{X}^n;\tilde{Y}^n) &= H(\tilde{X}^n) - H(\tilde{X}^n|\tilde{Y}^n) \\ &= nR - H(\tilde{X}^n|\tilde{Y}^n) \leq nC. \end{aligned} \tag{10.47}$$

Now, if we can decode reliably as $n \to \infty$, this means that the input codeword is completely determined by the signal received, or that the conditional entropy of the input (per letter) must get small

$$\frac{1}{n} H(\tilde{X}^n|\tilde{Y}^n) \to 0. \tag{10.48}$$

If errorless transmission is possible, then, eq. (10.47) becomes

$$R \leq C + o(1), \tag{10.49}$$

in the limit $n \to \infty$. The asymptotic rate cannot exceed the capacity. In Exercise 10.8, you will sharpen the statement eq.(10.48), showing that

$$\frac{1}{n} H(\tilde{X}^n|\tilde{Y}^n) \leq \frac{1}{n} H_2(p_e) + p_e R, \tag{10.50}$$

where p_e denotes the decoding error probability, and $H_2(p_e) = -p_e \log_2 p_e - (1-p_e)\log_2(1-p_e)$.

We have now seen that the capacity C is the highest achievable rate of communication through the noisy channel, where the probability of error goes to zero as the number of letters in the message goes to infinity. This is Shannon's noisy channel coding theorem. What is particularly remarkable is that, although the capacity is achieved by messages that are many letters in length, we have obtained a *single-letter formula* for the capacity, expressed in terms of the optimal mutual information $I(X;Y)$ for just a single use of the channel.

The method we used to show that $R = C - o(1)$ is achievable, averaging over random codes, is not constructive. Since a random code has no

structure or pattern, encoding and decoding are unwieldy, requiring an exponentially large code book. Nevertheless, the theorem is important and useful, because it tells us what is achievable, and not achievable, in principle. Furthermore, since $I(X;Y)$ is a concave function of $X = \{x, p(x)\}$ (with $\{p(y|x)\}$ fixed), it has a unique local maximum, and C can often be computed (at least numerically) for channels of interest. Finding codes which can be efficiently encoded and decoded, and come close to achieving the capacity, is a very interesting pursuit, but beyond the scope of our lightning introduction to Shannon theory.

10.2 Von Neumann Entropy

In classical information theory, we often consider a source that prepares messages of n letters ($n \gg 1$), where each letter is drawn independently from an ensemble $X = \{x, p(x)\}$. We have seen that the Shannon entropy $H(X)$ is the number of incompressible bits of information carried per letter (asymptotically as $n \to \infty$).

We may also be interested in correlations among messages. The correlations between two ensembles of letters X and Y are characterized by conditional probabilities $p(y|x)$. We have seen that the mutual information

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X), \tag{10.51}$$

is the number of bits of information per letter about X that we can acquire by reading Y (or vice versa). If the $p(y|x)$'s characterize a noisy channel, then, $I(X;Y)$ is the amount of information per letter that can be transmitted through the channel (given the *a priori* distribution X for the channel inputs).

We would like to generalize these considerations to *quantum* information. We may imagine a source that prepares messages of n letters, but where each letter is chosen from an ensemble of quantum states. The signal alphabet consists of a set of quantum states $\{\boldsymbol{\rho}(x)\}$, each occurring with a specified *a priori* probability $p(x)$.

As we discussed at length in Chapter 2, the probability of any outcome of any measurement of a letter chosen from this ensemble, if the observer has no knowledge about which letter was prepared, can be completely characterized by the density operator

$$\boldsymbol{\rho} = \sum_x p(x)\boldsymbol{\rho}(x); \tag{10.52}$$

for a POVM $\boldsymbol{E} = \{\boldsymbol{E}_a\}$, the probability of outcome a is

$$\text{Prob}(a) = \text{tr}(\boldsymbol{E}_a\boldsymbol{\rho}). \tag{10.53}$$

For this (or any) density operator, we may define the Von Neumann entropy

$$H(\boldsymbol{\rho}) = -\mathrm{tr}(\boldsymbol{\rho} \log \boldsymbol{\rho}). \tag{10.54}$$

Of course, we may choose an orthonormal basis $\{|a\rangle\}$ that diagonalizes $\boldsymbol{\rho}$,

$$\boldsymbol{\rho} = \sum_a \lambda_a |a\rangle\langle a|; \tag{10.55}$$

the vector of eigenvalues $\lambda(\boldsymbol{\rho})$ is a probability distribution, and the Von Neumann entropy of $\boldsymbol{\rho}$ is just the Shannon entropy of this distribution,

$$H(\boldsymbol{\rho}) = H(\lambda(\boldsymbol{\rho})). \tag{10.56}$$

If $\boldsymbol{\rho}_A$ is the density operator of system A, we will sometimes use the notation

$$H(A) := H(\boldsymbol{\rho}_A). \tag{10.57}$$

Our convention is to denote quantum systems with $A, B, C, \ldots$ and classical probability distributions with $X, Y, Z, \ldots$.

In the case where the signal alphabet $\{|\varphi(x)\rangle, p(x)\}$ consists of mutually orthogonal pure states, the quantum source reduces to a classical one; all of the signal states can be perfectly distinguished, and $H(\boldsymbol{\rho}) = H(X)$, where X is the classical ensemble $\{x, p(x)\}$. The quantum source is more interesting when the signal states $\{\boldsymbol{\rho}(x)\}$ are not mutually commuting. We will argue that the Von Neumann entropy quantifies the incompressible information content of the quantum source (in the case where the signal states are pure) much as the Shannon entropy quantifies the information content of a classical source.

Indeed, we will find that Von Neumann entropy plays multiple roles. It quantifies not only the *quantum* information content per letter of the pure-state ensemble (the minimum number of qubits per letter needed to reliably encode the information) but also its *classical* information content (the maximum amount of information per letter—in bits, not qubits—that we can gain about the preparation by making the best possible measurement). And we will see that Von Neumann information enters quantum information in yet other ways — for example, quantifying the entanglement of a bipartite pure state. Thus quantum information theory is largely concerned with the interpretation and uses of Von Neumann entropy, much as classical information theory is largely concerned with the interpretation and uses of Shannon entropy.

In fact, the mathematical machinery we need to develop quantum information theory is very similar to Shannon's mathematics (typical sequences, random coding, ...); so similar as to sometimes obscure that the conceptual context is really quite different. The central issue in quantum information theory is that nonorthogonal quantum states cannot be perfectly distinguished, a feature with no classical analog.

10.2.1 Mathematical properties of $H(\boldsymbol{\rho})$

There are a handful of properties of the Von Neumann entropy $H(\boldsymbol{\rho})$ which are frequently useful, many of which are closely analogous to corresponding properties of the Shannon entropy $H(X)$. Proofs of some of these are Exercises 10.1, 10.2, 10.3.

1. **Pure states**. A pure state $\boldsymbol{\rho} = |\varphi\rangle\langle\varphi|$ has $H(\boldsymbol{\rho}) = 0$.

2. **Unitary invariance**. The entropy is unchanged by a unitary change of basis,
$$H(\boldsymbol{U}\boldsymbol{\rho}\boldsymbol{U}^{-1}) = H(\boldsymbol{\rho}), \tag{10.58}$$
because $H(\boldsymbol{\rho})$ depends only on the eigenvalues of $\boldsymbol{\rho}$.

3. **Maximum**. If $\boldsymbol{\rho}$ has d nonvanishing eigenvalues, then
$$H(\boldsymbol{\rho}) \leq \log d, \tag{10.59}$$
with equality when all the nonzero eigenvalues are equal. The entropy is maximized when the quantum state is maximally mixed.

4. **Concavity**. For $\lambda_1, \lambda_2, \cdots, \lambda_n \geq 0$ and $\lambda_1 + \lambda_2 + \cdots + \lambda_n = 1$,
$$H(\lambda_1\boldsymbol{\rho}_1 + \cdots + \lambda_n\boldsymbol{\rho}_n) \geq \lambda_1 H(\boldsymbol{\rho}_1) + \cdots + \lambda_n H(\boldsymbol{\rho}_n). \tag{10.60}$$
The Von Neumann entropy is larger if we are *more ignorant* about how the state was prepared. This property is a consequence of the convexity of the log function.

5. **Subadditivity**. Consider a bipartite system AB in the state $\boldsymbol{\rho}_{AB}$. Then
$$H(AB) \leq H(A) + H(B) \tag{10.61}$$
(where $\boldsymbol{\rho}_A = \mathrm{tr}_B(\boldsymbol{\rho}_{AB})$ and $\boldsymbol{\rho}_B = \mathrm{tr}_A(\boldsymbol{\rho}_{AB})$), with equality for $\boldsymbol{\rho}_{AB} = \boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B$. Thus, entropy is *additive* for uncorrelated systems, but otherwise the entropy of the whole is less than the sum of the entropy of the parts. This property is the quantum generalization of subadditivity of Shannon entropy:
$$H(XY) \leq H(X) + H(Y). \tag{10.62}$$

6. **Bipartite pure states**. If the state $\boldsymbol{\rho}_{AB}$ of the bipartite system AB is pure, then
$$H(A) = H(B), \tag{10.63}$$
because $\boldsymbol{\rho}_A$ and $\boldsymbol{\rho}_B$ have the same nonzero eigenvalues.

7. **Quantum mutual information**. As in the classical case, we define the mutual information of two quantum systems as

$$I(A;B) = H(A) + H(B) - H(AB), \tag{10.64}$$

which is nonnegative because of the subadditivity of Von Neumann entropy, and zero only for a product state $\boldsymbol{\rho}_{AB} = \boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B$.

8. **Triangle inequality (Araki-Lieb inequality)**. For a bipartite system,

$$H(AB) \geq |H(A) - H(B)|. \tag{10.65}$$

To derive the triangle inequality, consider the tripartite pure state $|\psi\rangle_{ABC}$ which purifies $\boldsymbol{\rho}_{AB} = \mathrm{tr}_C\left(|\psi\rangle\langle\psi|\right)$. Since $|\psi\rangle$ is pure, $H(A) = H(BC)$ and $H(C) = H(AB)$; applying subadditivity to BC yields $H(A) \leq H(B) + H(C) = H(B) + H(AB)$. The same inequality applies with A and B interchanged, from which we obtain eq.(10.65).

The triangle inequality contrasts sharply with the analogous property of Shannon entropy,

$$H(XY) \geq H(X), H(Y). \tag{10.66}$$

The Shannon entropy of just part of a classical bipartite system cannot be greater than the Shannon entropy of the whole system. Not so for the Von Neumann entropy! For example, in the case of an entangled bipartite pure quantum state, we have $H(A) = H(B) > 0$, while $H(AB) = 0$. The entropy of the global system vanishes because our ignorance is minimal — we know as much about AB as the laws of quantum physics will allow. But we have incomplete knowledge of the parts A and B, with our ignorance quantified by $H(A) = H(B)$. For a quantum system, but not for a classical one, information can be encoded in the correlations among the parts of the system, yet be invisible when we look at the parts one at a time.

Equivalently, a property that holds classically but not quantumly is

$$H(X|Y) = H(XY) - H(Y) \geq 0. \tag{10.67}$$

The Shannon conditional entropy $H(X|Y)$ quantifies our remaining ignorance about X when we know Y, and equals zero when knowing Y makes us certain about X. On the other hand, the Von Neumann conditional entropy,

$$H(A|B) = H(AB) - H(B), \tag{10.68}$$

can be negative; in particular we have $H(A|B) = -H(A) = -H(B) < 0$ if $\boldsymbol{\rho}_{AB}$ is an entangled pure state. How can it make sense that "knowing"

the subsystem B makes us "more than certain" about the subsystem A? We'll return to this intriguing question in §10.8.2.

When X and Y are perfectly correlated, then $H(XY) = H(X) = H(Y)$; the conditional entropy is $H(X|Y) = H(Y|X) = 0$ and the mutual information is $I(X;Y) = H(X)$. In contrast, for a bipartite pure state of AB, the quantum state for which we may regard A and B as perfectly correlated, the mutual information is $I(A;B) = 2H(A) = 2H(B)$. In this sense the quantum correlations are stronger than classical correlations.

10.2.2 Mixing, measurement, and entropy

The Shannon entropy also has a property called *Schur concavity*, which means that if $X = \{x, p(x)\}$ and $Y = \{y, q(y)\}$ are two ensembles such that $p \prec q$, then $H(X) \geq H(Y)$. In fact, any function on probability vectors is Schur concave if it is invariant under permutations of its arguments and also concave in each argument. Recall that $p \prec q$ (q *majorizes* p) means that "p is at least as random as q" in the sense that $p = Dq$ for some doubly stochastic matrix D. Thus Schur concavity of H says that an ensemble with more randomness has higher entropy.

The Von Neumann entropy $H(\boldsymbol{\rho})$ of a density operator is the Shannon entropy of its vector of eigenvalues $\lambda(\boldsymbol{\rho})$. Furthermore, we showed in Exercise 2.6 that if the quantum state ensemble $\{|\varphi(x)\rangle, p(x)\}$ realizes $\boldsymbol{\rho}$, then $p \prec \lambda(\boldsymbol{\rho})$; therefore $H(\boldsymbol{\rho}) \leq H(X)$, where equality holds only for an ensemble of mutually orthogonal states. The decrease in entropy $H(X) - H(\boldsymbol{\rho})$ quantifies how *distinguishability is lost* when we mix nonorthogonal pure states. As we will soon see, the amount of information we can gain by measuring $\boldsymbol{\rho}$ is no more than $H(\boldsymbol{\rho})$ bits, so some of the information about which state was prepared has been irretrievably lost if $H(\boldsymbol{\rho}) < H(X)$.

If we perform an orthogonal measurement on $\boldsymbol{\rho}$ by projecting onto the basis $\{|y\rangle\}$, then outcome y occurs with probability

$$q(y) = \langle y|\boldsymbol{\rho}|y\rangle = \sum_a |\langle y|a\rangle|^2 \lambda_a, \quad \text{where} \quad \boldsymbol{\rho} = \sum_a \lambda_a |a\rangle\langle a| \tag{10.69}$$

and $\{|a\rangle\}$ is the basis in which $\boldsymbol{\rho}$ is diagonal. Since $D_{ya} = |\langle y|a\rangle|^2$ is a doubly stochastic matrix, $q \prec \lambda(\boldsymbol{\rho})$ and therefore $H(Y) \geq H(\boldsymbol{\rho})$, where equality holds only if the measurement is in the basis $\{|a\rangle\}$. Mathematically, the conclusion is that for a nondiagonal and nonnegative Hermitian matrix, the diagonal elements are more random than the eigenvalues. Speaking more physically, the outcome of an orthogonal measurement is easiest to predict if we measure an observable which commutes with the density operator, and becomes less predictable if we measure in a different basis.

This majorization property has a further consequence, which will be useful for our discussion of quantum compression. Suppose that $\boldsymbol{\rho}$ is a density operator of a d-dimensional system, with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ and that $\boldsymbol{E}' = \sum_{i=1}^{d'} |e_i\rangle\langle e_i|$ is a projector onto a subspace Λ of dimension $d' \leq d$ with orthonormal basis $\{|e_i\rangle\}$. Then

$$\mathrm{tr}\left(\boldsymbol{\rho}\boldsymbol{E}'\right) = \sum_{i=1}^{d'} \langle e_i|\boldsymbol{\rho}|e_i\rangle \leq \sum_{i=1}^{d'} \lambda_i, \tag{10.70}$$

where the inequality follows because the diagonal elements of $\boldsymbol{\rho}$ in the basis $\{|e_i\rangle\}$ are majorized by the eigenvalues of $\boldsymbol{\rho}$. In other words, if we perform a two-outcome orthogonal measurement, projecting onto either Λ or its orthogonal complement $\Lambda^\perp$, the probability of projecting onto Λ is no larger than the sum of the d' largest eigenvalues of $\boldsymbol{\rho}$ (the *Ky Fan dominance principle*).

10.2.3 Strong subadditivity

In addition to the subadditivity property $I(X;Y) \geq 0$, correlations of classical random variables obey a further property called *strong subadditivity*:

$$I(X;YZ) \geq I(X;Y). \tag{10.71}$$

This is the eminently reasonable statement that the correlations of X with YZ are at least as strong as the correlations of X with Y alone.

There is another useful way to think about (classical) strong subadditivity. Recalling the definition of mutual information we have

$$\begin{aligned} I(X;YZ) - I(X;Y) &= -\left\langle \log \frac{p(x)p(y,z)}{p(x,y,z)} + \log \frac{p(x,y)}{p(x)p(y)} \right\rangle \\ &= -\left\langle \log \frac{p(x,y)}{p(y)} \, \frac{p(y,z)}{p(y)} \, \frac{p(y)}{p(x,y,z)} \right\rangle \\ &= -\left\langle \log \frac{p(x|y)p(z|y)}{p(x,z|y)} \right\rangle = \sum_y p(y) I(X;Z|y) \geq 0. \end{aligned} \tag{10.72}$$

For each fixed y, $p(x,z|y)$ is a normalized probability distribution with nonnegative mutual information; hence $I(X;YZ) - I(X;Y)$ is a convex combination of nonnegative terms and therefore nonnegative. The quantity $I(X;Z|Y) := I(X;YZ) - I(X;Y)$ is called the *conditional mutual information*, because it quantifies how strongly X and Z are correlated

when Y is known; strong subadditivity can be restated as the nonnegativity of conditional mutual information,

$$I(X;Z|Y) \geq 0. \tag{10.73}$$

One might ask under what conditions strong subadditivity is satisfied as an equality; that is, when does the conditional mutual information vanish? Since $I(X;Z|Y)$ is sum of nonnegative terms, each of these terms must vanish if $I(X;Z|Y) = 0$. Therefore for each y with $p(y) > 0$, we have $I(X;Z|y) = 0$. The mutual information vanishes only for a product distribution, therefore

$$p(x,z|y) = p(x|y)p(z|y) \implies p(x,y,z) = p(x|y)p(z|y)p(y). \tag{10.74}$$

This means that the correlations between x and z arise solely from their shared correlation with y, in which case we say that x and z are *conditionally independent*.

Correlations of quantum systems also obey strong subadditivity:

$$I(A;BC) - I(A;B) := I(A;C|B) \geq 0. \tag{10.75}$$

But while the proof is elementary in the classical case, in the quantum setting strong subadditivity is a rather deep result with many important consequences. We will postpone the proof until §10.8.3, where we will be able to justify the quantum statement by giving it a clear operational meaning. We'll also see in Exercise 10.3 that strong subadditivity follows easily from another deep property, the monotonicity of relative entropy:

$$D(\boldsymbol{\rho}_A \| \boldsymbol{\sigma}_A) \leq D(\boldsymbol{\rho}_{AB} \| \boldsymbol{\sigma}_{AB}), \tag{10.76}$$

where

$$D(\boldsymbol{\rho} \| \boldsymbol{\sigma}) := \operatorname{tr} \boldsymbol{\rho} \left(\log \boldsymbol{\rho} - \log \boldsymbol{\sigma} \right) . \tag{10.77}$$

The relative entropy of two density operators on a system AB cannot be less than the induced relative entropy on the subsystem A. Insofar as we can regard the relative entropy as a measure of the "distance" between density operators, monotonicity is the reasonable statement that quantum states become no easier to distinguish when we look at the subsystem A than when we look at the full system AB. It also follows (Exercise 10.3), that the action of a quantum channel $\mathcal{N}$ cannot increase relative entropy:

$$D(\mathcal{N}(\boldsymbol{\rho}) \| \mathcal{N}(\boldsymbol{\sigma})) \leq D(\boldsymbol{\rho} \| \boldsymbol{\sigma}) \tag{10.78}$$

There are a few other ways of formulating strong subadditivity which are helpful to keep in mind. By expressing the quantum mutual information in terms of the Von Neumann entropy we find

$$H(ABC) + H(B) \leq H(AB) + H(BC). \tag{10.79}$$

While A, B, C are three disjoint quantum systems, we may view AB and BC as overlapping systems with intersection B and union ABC; then strong subadditivity says that the sum of the entropies of two overlapping systems is at least as large as the sum of the entropies of their union and their intersection. In terms of conditional entropy, strong subadditivity becomes

$$H(A|B) \geq H(A|BC); \tag{10.80}$$

loosely speaking, our ignorance about A when we know only B is no smaller than our ignorance about A when we know both B and C, but with the proviso that for quantum information "ignorance" can sometimes be negative!

As in the classical case, it is instructive to consider the condition for equality in strong subadditivity. What does it mean for systems to have *quantum conditional independence*, $I(A;C|B) = 0$? It is easy to formulate a sufficient condition. Suppose that system B has a decomposition as a direct sum of tensor products of Hilbert spaces

$$\mathcal{H}_B = \bigoplus_j \mathcal{H}_{B_j} = \bigoplus_j \mathcal{H}_{B_j^L} \otimes \mathcal{H}_{B_j^R}, \tag{10.81}$$

and that the state of ABC has the block diagonal form

$$\boldsymbol{\rho}_{ABC} = \bigoplus_j p_j \; \boldsymbol{\rho}_{AB_j^L} \otimes \boldsymbol{\rho}_{B_j^R C}. \tag{10.82}$$

In each block labeled by j the state is a tensor product, with conditional mutual information

$$I(A;C|B_j) = I(A;B_jC) - I(A;B_j) = I(A;B_j^L) - I(A;B_j^L) = 0; \tag{10.83}$$

What is less obvious is that the converse is also true — any state with $I(A;C|B) = 0$ has a decomposition as in eq.(10.82). This is a useful fact, though we will not give the proof here.

10.2.4 Monotonicity of mutual information

Strong subadditivity implies another important property of quantum mutual information, its *monotonicity* — a quantum channel acting on system B cannot increase the mutual information of A and B. To derive monotonicity, suppose that a quantum channel $\mathcal{N}^{B\to B'}$ maps B to B'. Like any quantum channel, $\mathcal{N}$ has an isometric extension, its Stinespring dilation $\boldsymbol{U}^{B\to B'E}$, mapping B to B' and a suitable environment system E. Since

the isometry $\boldsymbol{U}$ does not change the eigenvalues of the density operator, it preserves the entropy of B and of AB,

$$H(B) = H(B'E), \quad H(AB) = H(AB'E), \tag{10.84}$$

which implies

$$\begin{aligned} I(A;B) &= H(A) + H(B) - H(AB) \\ &= H(A) + H(B'E) - H(ABE') = I(A;B'E). \end{aligned} \tag{10.85}$$

From strong subadditivity, we obtain

$$I(A;B) = I(A;B'E) \geq I(A,B') \tag{10.86}$$

the desired statement of monotonicity.

10.2.5 Entropy and thermodynamics

The concept of entropy first entered science through the study of thermodynamics, and the mathematical properties of entropy we have enumerated have many interesting thermodynamic implications. Here we will just mention a few ways in which the nonnegativity and monotonicity of quantum relative entropy relate to ideas encountered in thermodynamics.

There are two distinct ways to approach the foundations of quantum statistical physics. In one, we consider the evolution of an isolated closed quantum system, but ask what we will observe if we have access to only a portion of the full system. Even though the evolution of the full system is unitary, the evolution of a subsystem is not, and the subsystem may be accurately described by a thermal ensemble at late times. Information which is initially encoded locally in an out-of-equilibrium state becomes encoded more and more nonlocally as the system evolves, eventually becoming invisible to an observer confined to the subsystem.

In the other approach, we consider the evolution of an open system A, in contact with an unobserved environment E, and track the evolution of A only. From a fundamental perspective this second approach may be regarded as a special case of the first, since AE is closed, with A as a privileged subsystem. In practice, though, it is often more convenient to describe the evolution of an open system using a master equation as in Chapter 3, and to analyze evolution toward thermal equilibrium without explicit reference to the environment.

Free energy and the second law. Tools of quantum Shannon theory can help us understand why the state of an open system with Hamiltonian $\boldsymbol{H}$

might be expected to be close to the thermal *Gibbs state*

$$\boldsymbol{\rho}_\beta = \frac{e^{-\beta \boldsymbol{H}}}{\mathrm{tr}\left(e^{-\beta \boldsymbol{H}}\right)}, \tag{10.87}$$

where $kT = \beta^{-1}$ is the temperature. Here let's observe one noteworthy feature of this state. For an arbitrary density operator $\boldsymbol{\rho}$, consider its *free energy*

$$F(\boldsymbol{\rho}) = E(\boldsymbol{\rho}) - \beta^{-1} S(\boldsymbol{\rho}) \tag{10.88}$$

where $E(\boldsymbol{\rho}) = \langle \boldsymbol{H} \rangle_{\boldsymbol{\rho}}$ denotes the expectation value of the Hamiltonian in this state; for this subsection we respect the conventions of thermodynamics by denoting Von Neumann entropy by $S(\boldsymbol{\rho})$ rather than $H(\boldsymbol{\rho})$ (lest H be confused with the Hamiltonian $\boldsymbol{H}$), and by using natural logarithms. Expressing $F(\boldsymbol{\rho})$ and the free energy $F(\boldsymbol{\rho}_\beta)$ of the Gibbs state as

$$\begin{aligned} F(\boldsymbol{\rho}) &= \mathrm{tr}\left(\boldsymbol{\rho}\boldsymbol{H}\right) - \beta^{-1} S(\boldsymbol{\rho}) = \beta^{-1} \mathrm{tr}\ \boldsymbol{\rho}\left(\ln \boldsymbol{\rho} + \beta \boldsymbol{H}\right), \\ F(\boldsymbol{\rho}_\beta) &= -\beta^{-1} \ln\left(\mathrm{tr}\ e^{-\beta \boldsymbol{H}}\right), \end{aligned} \tag{10.89}$$

we see that the relative entropy of $\boldsymbol{\rho}$ and $\boldsymbol{\rho}_\beta$ is

$$\begin{aligned} D(\boldsymbol{\rho} \| \boldsymbol{\rho}_\beta) &= \mathrm{tr}\left(\boldsymbol{\rho} \ln \boldsymbol{\rho}\right) - \mathrm{tr}\left(\boldsymbol{\rho} \ln \boldsymbol{\rho}_\beta\right) \\ &= \beta\left(F(\boldsymbol{\rho}) - F(\boldsymbol{\rho}_\beta)\right) \geq 0, \end{aligned} \tag{10.90}$$

with equality only for $\boldsymbol{\rho} = \boldsymbol{\rho}_\beta$. The nonnegativity of relative entropy implies that at a given temperature β^{-1}, the Gibbs state $\boldsymbol{\rho}_\beta$ has the lowest possible free energy. Our open system, in contact with a thermal reservoir at temperature β^{-1}, will prefer the Gibbs state if it wishes to minimize its free energy.

What can we say about the *approach* to thermal equilibrium of an open system? We may anticipate that the joint unitary evolution of system and reservoir induces a quantum channel $\mathcal{N}$ acting on the system alone, and we know that relative entropy is monotonic — if

$$\mathcal{N} : \boldsymbol{\rho} \mapsto \boldsymbol{\rho}', \quad \mathcal{N} : \boldsymbol{\sigma} \mapsto \boldsymbol{\sigma}', \tag{10.91}$$

then

$$D(\boldsymbol{\rho}' \| \boldsymbol{\sigma}') \leq D(\boldsymbol{\rho} \| \boldsymbol{\sigma}). \tag{10.92}$$

Furthermore, if the Gibbs state is an *equilibrium* state, we expect this channel to preserve the Gibbs state

$$\mathcal{N} : \boldsymbol{\rho}_\beta \mapsto \boldsymbol{\rho}_\beta; \tag{10.93}$$

therefore,

$$D(\boldsymbol{\rho}'\|\boldsymbol{\rho}_\beta) = \beta\left(F(\boldsymbol{\rho}') - F(\boldsymbol{\rho}_\beta)\right) \leq \beta\left(F(\boldsymbol{\rho}) - F(\boldsymbol{\rho}_\beta)\right) = D(\boldsymbol{\rho}\|\boldsymbol{\rho}_\beta), \tag{10.94}$$

and hence

$$F(\boldsymbol{\rho}') \leq F(\boldsymbol{\rho}). \tag{10.95}$$

Any channel that preserves the Gibbs state cannot increase the free energy; instead, free energy of an out-of-equilibrium state is monotonically decreasing under open-state evolution. This statement is a version of the second law of thermodynamics.

10.2.6 Bekenstein's entropy bound.

Similar ideas lead to Bekenstein's bound on entropy in quantum field theory. The field-theoretic details, though interesting, would lead us far afield. The gist is that Bekenstein proposed an inequality relating the energy and the entropy in a bounded spatial region. This bound was motivated by gravitational physics, but can be formulated without reference to gravitation, and follows from properties of relative entropy.

A subtlety is that entropy of a region is infinite in quantum field theory, because of contributions coming from arbitrarily short-wavelength quantum fluctuations near the boundary of the region. Therefore we have to make a subtraction to define a finite quantity. The natural way to do this is to subtract away the entropy of the same region in the vacuum state of the theory, as any finite energy state in a finite volume has the same structure as the vacuum at very short distances. Although the vacuum is a pure state, it, and any other reasonable state, has a marginal state in a finite region which is highly mixed, because of entanglement between the region and its complement.

For the purpose of our discussion here, we may designate any mixed state $\boldsymbol{\rho}_0$ we choose supported in the bounded region as the "vacuum," and define a corresponding "modular Hamiltonian" $\boldsymbol{K}$ by

$$\boldsymbol{\rho}_0 = \frac{e^{-\boldsymbol{K}}}{\operatorname{tr}\left(e^{-\boldsymbol{K}}\right)}. \tag{10.96}$$

That is, we regard the state as the thermal mixed state of $\boldsymbol{K}$, with the temperature arbitrarily set to unity (which is just a normalization convention for $\boldsymbol{K}$). Then by rewriting eq.(10.90) we see that, for any state $\boldsymbol{\rho}$, $D(\boldsymbol{\rho}\|\boldsymbol{\rho}_0) \geq 0$ implies

$$S(\boldsymbol{\rho}) - S(\boldsymbol{\rho}_0) \leq \operatorname{tr}\left(\boldsymbol{\rho}\boldsymbol{K}\right) - \operatorname{tr}\left(\boldsymbol{\rho}_0\boldsymbol{K}\right) \tag{10.97}$$

The left-hand side, the entropy with vacuum entropy subtracted, is not larger than the right-hand side, the (modular) energy with vacuum energy subtracted. This is one version of Bekenstein's bound. Here $\boldsymbol{K}$, which is dimensionless, can be loosely interpreted as ER, where E is the energy contained in the region and R is its linear size.

While the bound follows easily from nonnegativity of relative entropy, the subtle part of the argument is recognizing that the (suitably subtracted) expectation value of the modular Hamiltonian is a reasonable way to define ER. The detailed justification for this involves properties of relativistic quantum field theory that we won't go into here. Suffice it to say that, because we constructed $\boldsymbol{K}$ by regarding the marginal state of the vacuum as the Gibbs state associated with the Hamiltonian $\boldsymbol{K}$, we expect $\boldsymbol{K}$ to be linear in the energy, and dimensional analysis then requires inclusion of the factor of R (in units with $\hbar = c = 1$).

Bekenstein was led to conjecture such a bound by thinking about black hole thermodynamics. Leaving out numerical factors, just to get a feel for the orders of magnitude of things, the entropy of a black hole with circumference $\sim R$ is $S \sim R^2/G$, and its mass (energy) is $E \sim R/G$, where G is Newton's gravitational constant; hence $S \sim ER$ for a black hole. Bekenstein realized that unless $S = O(ER)$ for arbitrary states and regions, we could throw extra stuff into the region, making a black hole with lower entropy than the initial state, thus violating the (generalized) second law of thermodynamics. Though black holes provided the motivation, G drops out of the inequality, which holds even in nongravitational relativistic quantum field theories.

10.2.7 Entropic uncertainty relations

The uncertainty principle asserts that noncommuting observables cannot simultaneously have definite values. To translate this statement into mathematics, recall that a Hermitian observable $\boldsymbol{A}$ has spectral representation

$$\boldsymbol{A} = \sum_x |x\rangle a(x) \langle x| \tag{10.98}$$

where $\{|x\rangle\}$ is the orthonormal basis of eigenvectors of $\boldsymbol{A}$ and $\{a(x)\}$ is the corresponding vector of eigenvalues; if $\boldsymbol{A}$ is measured in the state $\boldsymbol{\rho}$, the outcome $a(x)$ occurs with probability $p(x) = \langle x|\boldsymbol{\rho}|x\rangle$. Thus $\boldsymbol{A}$ has expectation value $\mathrm{tr}(\boldsymbol{\rho A})$ and variance

$$(\Delta A)^2 = \mathrm{tr}\left(\boldsymbol{\rho A}^2\right) - (\mathrm{tr}\boldsymbol{\rho A})^2 \,. \tag{10.99}$$

Using the Cauchy-Schwarz inequality, we can show that if $\boldsymbol{A}$ and $\boldsymbol{B}$ are two Hermitian observables and $\boldsymbol{\rho} = |\psi\rangle\langle\psi|$ is a pure state, then

$$\Delta A \Delta B \geq \frac{1}{2}|\langle\psi|[\boldsymbol{A}, \boldsymbol{B}]|\psi\rangle|. \tag{10.100}$$

Eq.(10.100) is a useful statement of the uncertainty principle, but has drawbacks. It depends on the state $|\psi\rangle$ and for that reason does not fully capture the incompatibility of the two observables. Furthermore, the variance does not characterize very well the unpredictability of the measurement outcomes; entropy would be a more informative measure.

In fact there are *entropic uncertainty relations* which do not suffer from these deficiencies. If we measure a state $\boldsymbol{\rho}$ by projecting onto the orthonormal basis $\{|x\rangle\}$, the outcomes define a classical ensemble

$$X = \{x, p(x) = \langle x|\boldsymbol{\rho}|x\rangle\}; \tag{10.101}$$

that is, a probability vector whose entries are the diagonal elements of $\boldsymbol{\rho}$ in the x-basis. The Shannon entropy $H(X)$ quantifies how uncertain we are about the outcome before we perform the measurement. If $\{|z\rangle\}$ is another orthonormal basis, there is a corresponding classical ensemble Z describing the probability distribution of outcomes when we measure the same state $\boldsymbol{\rho}$ in the z-basis. If the two bases are incompatible, there is a tradeoff between our uncertainty about X and about Z, captured by the inequality

$$H(X) + H(Z) \geq \log\left(\frac{1}{c}\right) + H(\boldsymbol{\rho}), \tag{10.102}$$

where

$$c = \max_{x,z} |\langle x|z\rangle|^2. \tag{10.103}$$

The second term on the right-hand side, which vanishes if $\boldsymbol{\rho}$ is a pure state, reminds us that our uncertainty increases when the state is mixed. Like many good things in quantum information theory, this entropic uncertainty relation follows from the monotonicity of the quantum relative entropy.

For each measurement there is a corresponding quantum channel, realized by performing the measurement and printing the outcome in a classical register,

$$\begin{aligned} \mathcal{M}_X &: \boldsymbol{\rho} \mapsto \sum_x |x\rangle\langle x|\boldsymbol{\rho}|x\rangle\langle x| =: \boldsymbol{\rho}_X, \\ \mathcal{M}_Z &: \boldsymbol{\rho} \mapsto \sum_z |z\rangle\langle z|\boldsymbol{\rho}|z\rangle\langle z| =: \boldsymbol{\rho}_Z. \end{aligned} \tag{10.104}$$

The Shannon entropy of the measurement outcome distribution is also the Von Neumann entropy of the corresponding channel's output state,

$$H(X) = H(\boldsymbol{\rho}_X), \quad H(Z) = H(\boldsymbol{\rho}_Z); \tag{10.105}$$

the entropy of this output state can be expressed in terms of the relative entropy of input and output, and the entropy of the channel input, as in

$$H(X) = -\mathrm{tr}\boldsymbol{\rho}_X \log \boldsymbol{\rho}_X = -\mathrm{tr}\boldsymbol{\rho} \log \boldsymbol{\rho}_X = D(\boldsymbol{\rho}\|\boldsymbol{\rho}_X) + H(\boldsymbol{\rho}). \tag{10.106}$$

Using the monotonicity of relative entropy under the action of the channel $\mathcal{M}_Z$, we have

$$D(\boldsymbol{\rho}\|\boldsymbol{\rho}_X) \geq D(\boldsymbol{\rho}_Z\|\mathcal{M}_Z(\boldsymbol{\rho}_X)), \tag{10.107}$$

where

$$D(\boldsymbol{\rho}_Z\|\mathcal{M}_Z(\boldsymbol{\rho}_X)) = -H(\boldsymbol{\rho}_Z) - \mathrm{tr}\boldsymbol{\rho}_Z \log \mathcal{M}_Z(\boldsymbol{\rho}_X), \tag{10.108}$$

and

$$\mathcal{M}_Z(\boldsymbol{\rho}_X) = \sum_{x,z} |z\rangle\langle z|x\rangle\langle x|\boldsymbol{\rho}|x\rangle\langle x|z\rangle\langle z|. \tag{10.109}$$

Writing

$$\log \mathcal{M}_Z(\boldsymbol{\rho}_X) = \sum_z |z\rangle \log \left(\sum_x \langle z|x\rangle\langle x|\boldsymbol{\rho}|x\rangle\langle x|z\rangle \right) \langle z|, \tag{10.110}$$

we see that

$$-\mathrm{tr}\boldsymbol{\rho}_Z \log \mathcal{M}_Z(\boldsymbol{\rho}_X) = -\sum_z \langle z|\boldsymbol{\rho}|z\rangle \log \left(\sum_x \langle z|x\rangle\langle x|\boldsymbol{\rho}|x\rangle\langle x|z\rangle \right). \tag{10.111}$$

Now, because $-\log(\cdot)$ is a monotonically decreasing function, we have

$$\begin{aligned} -\log \left(\sum_x \langle z|x\rangle\langle x|\boldsymbol{\rho}|x\rangle\langle x|z\rangle \right) &\geq -\log \left(\max_{x,z} |\langle x|z\rangle|^2 \sum_x \langle x|\boldsymbol{\rho}|x\rangle \right) \\ &= \log \left(\frac{1}{c} \right), \end{aligned} \tag{10.112}$$

and therefore

$$-\mathrm{tr}\boldsymbol{\rho}_Z \log \mathcal{M}_Z(\boldsymbol{\rho}_X) \geq \log \left(\frac{1}{c} \right). \tag{10.113}$$

Finally, putting together eq.(10.106), (10.107) (10.108), (10.113), we find

$$H(X) - H(\boldsymbol{\rho}) = D(\boldsymbol{\rho}\|\boldsymbol{\rho}_X) \geq D(\boldsymbol{\rho}_Z\|\mathcal{M}_Z(\boldsymbol{\rho}_X))$$
$$= -H(Z) - \mathrm{tr}\boldsymbol{\rho}_Z \log \mathcal{M}_Z(\boldsymbol{\rho}_X) \geq -H(Z) + \log\left(\frac{1}{c}\right), \tag{10.114}$$

which is equivalent to eq.(10.102).

We say that two different bases $\{|x\rangle\}$, $\{|z\rangle\}$ for a d-dimensional Hilbert space are *mutually unbiased* if for all x, z

$$|\langle x|z\rangle|^2 = \frac{1}{d}; \tag{10.115}$$

thus, if we measure any x-basis state $|x\rangle$ in the z-basis, all d outcomes are equally probable. For measurements in two mutually unbiased bases performed on a pure state, the entropic uncertainty relation becomes

$$H(X) + H(Z) \geq \log d. \tag{10.116}$$

Clearly this inequality is tight, as it is saturated by x-basis (or z-basis) states, for which $H(X) = 0$ and $H(Z) = \log d$.

10.3 Quantum Source Coding

What is the quantum analog of Shannon's source coding theorem?

Let's consider a long message consisting of n letters, where each letter is a pure quantum state chosen by sampling from the ensemble

$$\{|\varphi(x)\rangle, p(x)\}. \tag{10.117}$$

If the states of this ensemble are mutually orthogonal, then the message might as well be classical; the interesting quantum case is where the states are not orthogonal and therefore not perfectly distinguishable. The density operator realized by this ensemble is

$$\boldsymbol{\rho} = \sum_x p(x)|\varphi(x)\rangle\langle\varphi(x)|, \tag{10.118}$$

and the entire n-letter message has the density operator

$$\boldsymbol{\rho}^{\otimes n} = \boldsymbol{\rho} \otimes \cdots \otimes \boldsymbol{\rho}. \tag{10.119}$$

How *redundant* is the quantum information in this message? We would like to devise a *quantum code* allowing us to compress the message to a smaller Hilbert space, but without much compromising the fidelity of the message. Perhaps we have a quantum memory device, and we know the

statistical properties of the recorded data; specifically, we know $\boldsymbol{\rho}$. We want to conserve space on our (very expensive) quantum hard drive by compressing the data.

The optimal compression that can be achieved was found by Schumacher. As you might guess, the message can be compressed to a Hilbert space $\mathcal{H}$ with

$$\dim \mathcal{H} = 2^{n(H(\boldsymbol{\rho})+o(1))} \tag{10.120}$$

with negligible loss of fidelity as $n \to \infty$, while errorless compression to dimension $2^{n(H(\boldsymbol{\rho})-\Omega(1))}$ is not possible. In this sense, the Von Neumann entropy is the number of *qubits* of quantum information carried per letter of the message. Compression is always possible unless $\boldsymbol{\rho}$ is maximally mixed, just as we can always compress a classical message unless the information source is uniformly random. This result provides a precise operational interpretation for Von Neumann entropy.

Once Shannon's results are known and understood, the proof of Schumacher's compression theorem is not difficult, as the mathematical ideas needed are very similar to those used by Shannon. But conceptually quantum compression is very different from its classical counterpart, as the imperfect distinguishability of nonorthogonal quantum states is the central idea.

10.3.1 Quantum compression: an example

Before discussing Schumacher's quantum compression protocol in full generality, it is helpful to consider a simple example. Suppose that each letter is a single qubit drawn from the ensemble

$$|\uparrow_z\rangle \quad = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad p = \frac{1}{2}, \tag{10.121}$$

$$|\uparrow_x\rangle \quad = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \qquad p = \frac{1}{2}, \tag{10.122}$$

so that the density operator of each letter is

$$\begin{aligned} \boldsymbol{\rho} &= \frac{1}{2}|\uparrow_z\rangle\langle\uparrow_z| + \frac{1}{2}|\uparrow_x\rangle\langle\uparrow_x| \\ &= \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix}. \end{aligned} \tag{10.123}$$

As is obvious from symmetry, the eigenstates of $\boldsymbol{\rho}$ are qubits oriented up and down along the axis $\hat{n} = \frac{1}{\sqrt{2}}(\hat{x} + \hat{z})$,

$$\begin{aligned} |0'\rangle \equiv |\uparrow_{\hat{n}}\rangle &= \begin{pmatrix} \cos\frac{\pi}{8} \\ \sin\frac{\pi}{8} \end{pmatrix}, \\ |1'\rangle \equiv |\downarrow_{\hat{n}}\rangle &= \begin{pmatrix} \sin\frac{\pi}{8} \\ -\cos\frac{\pi}{8} \end{pmatrix}; \end{aligned} \tag{10.124}$$

the eigenvalues are

$$\begin{aligned} \lambda(0') &= \frac{1}{2} + \frac{1}{2\sqrt{2}} = \cos^2\frac{\pi}{8}, \\ \lambda(1') &= \frac{1}{2} - \frac{1}{2\sqrt{2}} = \sin^2\frac{\pi}{8}; \end{aligned} \tag{10.125}$$

evidently $\lambda(0') + \lambda(1') = 1$ and $\lambda(0')\lambda(1') = \frac{1}{8} = \det\boldsymbol{\rho}$. The eigenstate $|0'\rangle$ has equal (and relatively large) overlap with both signal states

$$|\langle 0'|\uparrow_z\rangle|^2 = |\langle 0'|\uparrow_x\rangle|^2 = \cos^2\frac{\pi}{8} = .8535, \tag{10.126}$$

while $|1'\rangle$ has equal (and relatively small) overlap with both,

$$|\langle 1'|\uparrow_z\rangle|^2 = |\langle 1'|\uparrow_x\rangle|^2 = \sin^2\frac{\pi}{8} = .1465. \tag{10.127}$$

Thus if we don't know whether $|\uparrow_z\rangle$ or $|\uparrow_x\rangle$ was sent, the best guess we can make is $|\psi\rangle = |0'\rangle$. This guess has the maximal *fidelity* with $\boldsymbol{\rho}$

$$F = \frac{1}{2}|\langle\uparrow_z|\psi\rangle|^2 + \frac{1}{2}|\langle\uparrow_x|\psi\rangle|^2, \tag{10.128}$$

among all possible single-qubit states $|\psi\rangle$ ($F = .8535$).

Now imagine that Alice needs to send three letters to Bob, but she can afford to send only two qubits. Still, she wants Bob to reconstruct her state with the highest possible fidelity. She could send Bob two of her three letters, and ask Bob to guess $|0'\rangle$ for the third. Then Bob receives two letters with perfect fidelity, and his guess has $F = .8535$ for the third; hence $F = .8535$ overall. But is there a more clever procedure that achieves higher fidelity?

Yes, there is. By diagonalizing $\boldsymbol{\rho}$, we decomposed the Hilbert space of a single qubit into a "likely" one-dimensional subspace (spanned by $|0'\rangle$) and an "unlikely" one-dimensional subspace (spanned by $|1'\rangle$). In a similar way we can decompose the Hilbert space of three qubits into likely and unlikely subspaces. If $|\psi\rangle = |\psi_1\rangle \otimes |\psi_2\rangle \otimes |\psi_3\rangle$ is any signal state,

where the state of each qubit is either $|\uparrow_z\rangle$ or $|\uparrow_x\rangle$, we have

$$\begin{aligned}
|\langle 0'0'0'|\psi\rangle|^2 &= \cos^6\left(\frac{\pi}{8}\right) = .6219, \\
|\langle 0'0'1'|\psi\rangle|^2 &= |\langle 0'1'0'|\psi\rangle|^2 = |\langle 1'0'0'|\psi\rangle|^2 = \cos^4\left(\frac{\pi}{8}\right)\sin^2\left(\frac{\pi}{8}\right) = .1067, \\
|\langle 0'1'1'|\psi\rangle|^2 &= |\langle 1'0'1'|\psi\rangle|^2 = |\langle 1'1'0'|\psi\rangle|^2 = \cos^2\left(\frac{\pi}{8}\right)\sin^4\left(\frac{\pi}{8}\right) = .0183, \\
|\langle 1'1'1'|\psi\rangle|^2 &= \sin^6\left(\frac{\pi}{8}\right) = .0031. \qquad (10.129)
\end{aligned}$$

Thus, we may decompose the space into the likely subspace Λ spanned by $\{|0'0'0'\rangle, |0'0'1'\rangle, |0'1'0'\rangle, |1'0'0'\rangle\}$, and its orthogonal complement $\Lambda^\perp$. If we make an incomplete orthogonal measurement that projects a signal state onto Λ or $\Lambda^\perp$, the probability of projecting onto the likely subspace Λ is

$$p_{\text{likely}} = .6219 + 3(.1067) = .9419, \qquad (10.130)$$

while the probability of projecting onto the unlikely subspace is

$$p_{\text{unlikely}} = 3(.0183) + .0031 = .0581. \qquad (10.131)$$

To perform this measurement, Alice could, for example, first apply a unitary transformation $\boldsymbol{U}$ that rotates the four high-probability basis states to

$$|\cdot\rangle \otimes |\cdot\rangle \otimes |0\rangle, \qquad (10.132)$$

and the four low-probability basis states to

$$|\cdot\rangle \otimes |\cdot\rangle \otimes |1\rangle; \qquad (10.133)$$

then Alice measures the third qubit to perform the projection. If the outcome is $|0\rangle$, then Alice's input state has in effect been projected onto Λ. She sends the remaining two unmeasured qubits to Bob. When Bob receives this compressed two-qubit state $|\psi_{\text{comp}}\rangle$, he decompresses it by appending $|0\rangle$ and applying $\boldsymbol{U}^{-1}$, obtaining

$$|\psi'\rangle = \boldsymbol{U}^{-1}(|\psi_{\text{comp}}\rangle \otimes |0\rangle). \qquad (10.134)$$

If Alice's measurement of the third qubit yields $|1\rangle$, she has projected her input state onto the low-probability subspace $\Lambda^\perp$. In this event, the best thing she can do is send the state that Bob will decompress to the most likely state $|0'0'0'\rangle$ – that is, she sends the state $|\psi_{\text{comp}}\rangle$ such that

$$|\psi'\rangle = \boldsymbol{U}^{-1}(|\psi_{\text{comp}}\rangle \otimes |0\rangle) = |0'0'0'\rangle. \qquad (10.135)$$

Thus, if Alice encodes the three-qubit signal state $|\psi\rangle$, sends two qubits to Bob, and Bob decodes as just described, then Bob obtains the state $\boldsymbol{\rho}'$

$$|\psi\rangle\langle\psi| \to \boldsymbol{\rho}' = \boldsymbol{E}|\psi\rangle\langle\psi|\boldsymbol{E} + |0'0'0'\rangle\langle\psi|(\boldsymbol{I} - \boldsymbol{E})|\psi\rangle\langle 0'0'0'|, \qquad (10.136)$$

where $\boldsymbol{E}$ is the projection onto Λ. The fidelity achieved by this procedure is

$$\begin{aligned} F &= \langle\psi|\boldsymbol{\rho}'|\psi\rangle = (\langle\psi|\boldsymbol{E}|\psi\rangle)^2 + (\langle\psi|(\boldsymbol{I}-\boldsymbol{E})|\psi\rangle)(\langle\psi|0'0'0'\rangle)^2 \\ &= (.9419)^2 + (.0581)(.6219) = .9234. \end{aligned} \tag{10.137}$$

This is indeed better than the naive procedure of sending two of the three qubits each with perfect fidelity.

As we consider longer messages with more letters, the fidelity of the compression improves, as long as we don't try to compress too much. The Von-Neumann entropy of the one-qubit ensemble is

$$H(\boldsymbol{\rho}) = H\left(\cos^2\frac{\pi}{8}\right) = .60088\ldots \tag{10.138}$$

Therefore, according to Schumacher's theorem, we can shorten a long message by the factor, say, .6009, and still achieve very good fidelity.

10.3.2 Schumacher compression in general

The key to Shannon's noiseless coding theorem is that we can code the typical sequences and ignore the rest, without much loss of fidelity. To quantify the compressibility of quantum information, we promote the notion of a typical *sequence* to that of a typical *subspace*. The key to Schumacher's noiseless quantum coding theorem is that we can code the typical subspace and ignore its orthogonal complement, without much loss of fidelity.

We consider a message of n letters where each letter is a pure quantum state drawn from the ensemble $\{|\varphi(x)\rangle, p(x)\}$, so that the density operator of a single letter is

$$\boldsymbol{\rho} = \sum_x p(x)|\varphi(x)\rangle\langle\varphi(x)|. \tag{10.139}$$

Since the letters are drawn independently, the density operator of the entire message is

$$\boldsymbol{\rho}^{\otimes n} \equiv \boldsymbol{\rho}\otimes\cdots\otimes\boldsymbol{\rho}. \tag{10.140}$$

We claim that, for n large, this density matrix has nearly all of its support on a subspace of the full Hilbert space of the messages, where the dimension of this subspace asymptotically approaches $2^{nH(\boldsymbol{\rho})}$.

This claim follows directly from the corresponding classical statement, for we may consider $\boldsymbol{\rho}$ to be realized by an ensemble of orthonormal pure states, its eigenstates, where the probability assigned to each eigenstate is the corresponding eigenvalue. In this basis our source of quantum information is effectively classical, producing messages which are tensor

products of $\boldsymbol{\rho}$ eigenstates, each with a probability given by the product of the corresponding eigenvalues. For a specified n and δ, define the δ-*typical subspace* Λ as the space spanned by the eigenvectors of $\boldsymbol{\rho}^{\otimes n}$ with eigenvalues λ satisfying

$$2^{-n(H-\delta)} \geq \lambda \geq 2^{-n(H+\delta)}. \tag{10.141}$$

Borrowing directly from Shannon's argument, we infer that for any $\delta, \varepsilon > 0$ and n sufficiently large, the sum of the eigenvalues of $\boldsymbol{\rho}^{\otimes n}$ that obey this condition satisfies

$$\mathrm{tr}(\boldsymbol{\rho}^{\otimes n}\boldsymbol{E}) \geq 1 - \varepsilon, \tag{10.142}$$

where $\boldsymbol{E}$ denotes the projection onto the typical subspace Λ, and the number $\dim(\Lambda)$ of such eigenvalues satisfies

$$2^{n(H+\delta)} \geq \dim(\Lambda) \geq (1-\varepsilon)2^{n(H-\delta)}. \tag{10.143}$$

Our coding strategy is to send states in the typical subspace faithfully. We can make a measurement that projects the input message onto either Λ or $\Lambda^{\perp}$; the outcome will be Λ with probability $p_\Lambda = \mathrm{tr}(\boldsymbol{\rho}^{\otimes n}\boldsymbol{E}) \geq 1 - \varepsilon$. In that event, the projected state is coded and sent. Asymptotically, the probability of the other outcome becomes negligible, so it matters little what we do in that case.

The coding of the projected state merely packages it so it can be carried by a minimal number of qubits. For example, we apply a unitary change of basis $\boldsymbol{U}$ that takes each state $|\psi_{\mathrm{typ}}\rangle$ in Λ to a state of the form

$$\boldsymbol{U}|\psi_{\mathrm{typ}}\rangle = |\psi_{\mathrm{comp}}\rangle \otimes |0_{\mathrm{rest}}\rangle, \tag{10.144}$$

where $|\psi_{\mathrm{comp}}\rangle$ is a state of $n(H+\delta)$ qubits, and $|0_{\mathrm{rest}}\rangle$ denotes the state $|0\rangle \otimes \ldots \otimes |0\rangle$ of the remaining qubits. Alice sends $|\psi_{\mathrm{comp}}\rangle$ to Bob, who decodes by appending $|0_{\mathrm{rest}}\rangle$ and applying $\boldsymbol{U}^{-1}$.

Suppose that

$$|\varphi(\vec{x})\rangle = |\varphi(x_1)\rangle \otimes \ldots \otimes |\varphi(x_n)\rangle, \tag{10.145}$$

denotes any one of the n-letter pure state messages that might be sent. After coding, transmission, and decoding are carried out as just described, Bob has reconstructed a state

$$\begin{aligned} |\varphi(\vec{x})\rangle\langle\varphi(\vec{x})| \mapsto \boldsymbol{\rho}'(\vec{x}) = \boldsymbol{E}|\varphi(\vec{x})\rangle\langle\varphi(\vec{x})|\boldsymbol{E} \\ + \boldsymbol{\rho}_{\mathrm{Junk}}(\vec{x})\langle\varphi(\vec{x})|(\boldsymbol{I}-\boldsymbol{E})|\varphi(\vec{x})\rangle, \end{aligned} \tag{10.146}$$

where $\boldsymbol{\rho}_{\mathrm{Junk}}(\vec{x})$ is the state we choose to send if the measurement yields the outcome $\Lambda^{\perp}$. What can we say about the fidelity of this procedure?

The fidelity varies from message to message, so we consider the fidelity averaged over the ensemble of possible messages:

$$\begin{aligned}\bar{F} &= \sum_{\vec{x}} p(\vec{x})\langle\varphi(\vec{x})|\boldsymbol{\rho}'(\vec{x})|\varphi(\vec{x})\rangle \\ &= \sum_{\vec{x}} p(\vec{x})\langle\varphi(\vec{x})|\boldsymbol{E}|\varphi(\vec{x})\rangle\langle\varphi(\vec{x})|\boldsymbol{E}|\varphi(\vec{x})\rangle \\ &+ \sum_{\vec{x}} p(\vec{x})\langle\varphi(\vec{x})|\boldsymbol{\rho}_{\text{Junk}}(\vec{x})|\varphi(\vec{x})\rangle\langle\varphi(\vec{x})|\boldsymbol{I}-\boldsymbol{E}|\varphi(\vec{x})\rangle \\ &\geq \sum_{\vec{x}} p(\vec{x})\langle\varphi(\vec{x})|\boldsymbol{E}|\varphi(\vec{x})\rangle^2, \end{aligned} \tag{10.147}$$

where the last inequality holds because the "Junk" term is nonnegative. Since any real number z satisfies

$$(z-1)^2 \geq 0, \quad \text{or} \quad z^2 \geq 2z-1, \tag{10.148}$$

we have (setting $z = \langle\varphi(\vec{x})|\boldsymbol{E}|\varphi(\vec{x})\rangle$)

$$\langle\varphi(\vec{x})|\boldsymbol{E}|\varphi(\vec{x})\rangle^2 \geq 2\langle\varphi(\vec{x})|\boldsymbol{E}|\varphi(\vec{x})\rangle - 1, \tag{10.149}$$

and hence

$$\begin{aligned}\bar{F} &\geq \sum_{\vec{x}} p(\vec{x})(2\langle\varphi(\vec{x})|\boldsymbol{E}|\varphi(\vec{x})\rangle - 1) \\ &= 2\ \text{tr}(\boldsymbol{\rho}^{\otimes n}\boldsymbol{E}) - 1 \geq 2(1-\varepsilon) - 1 = 1 - 2\varepsilon. \end{aligned} \tag{10.150}$$

Since ε and δ can be as small as we please, we have shown that it is possible to compress the message to $n(H + o(1))$ qubits, while achieving an average fidelity that becomes arbitrarily good as n gets large.

Is further compression possible? Let us suppose that Bob will decode the message $\boldsymbol{\rho}_{\text{comp}}(\vec{x})$ that he receives by appending qubits and applying a unitary transformation $\boldsymbol{U}^{-1}$, obtaining

$$\boldsymbol{\rho}'(\vec{x}) = \boldsymbol{U}^{-1}(\boldsymbol{\rho}_{\text{comp}}(\vec{x}) \otimes |0\rangle\langle 0|)\boldsymbol{U} \tag{10.151}$$

("unitary decoding"), and suppose that $\boldsymbol{\rho}_{\text{comp}}(\vec{x})$ has been compressed to $n(H - \delta')$ qubits. Then, no matter how the input messages have been encoded, the decoded messages are all contained in a subspace Λ' of Bob's Hilbert space with $\dim(\Lambda') = 2^{n(H-\delta')}$.

If the input message is $|\varphi(\vec{x})\rangle$, then the density operator reconstructed by Bob can be diagonalized as

$$\boldsymbol{\rho}'(\vec{x}) = \sum_{a_{\vec{x}}} |a_{\vec{x}}\rangle\lambda_{a_{\vec{x}}}\langle a_{\vec{x}}|, \tag{10.152}$$

where the $|a_{\vec{x}}\rangle$'s are mutually orthogonal states in Λ'. The fidelity of the reconstructed message is

$$\begin{aligned} F(\vec{x}) &= \langle\varphi(\vec{x})|\boldsymbol{\rho}'(\vec{x})|\varphi(\vec{x})\rangle \\ &= \sum_{a_{\vec{x}}} \lambda_{a_{\vec{x}}} \langle\varphi(\vec{x})|a_{\vec{x}}\rangle\langle a_{\vec{x}}|\varphi(\vec{x})\rangle \\ &\leq \sum_{a_{\vec{x}}} \langle\varphi(\vec{x})|a_{\vec{x}}\rangle\langle a_{\vec{x}}|\varphi(\vec{x})\rangle \leq \langle\varphi(\vec{x})|\boldsymbol{E}'|\varphi(\vec{x})\rangle, \end{aligned} \tag{10.153}$$

where $\boldsymbol{E}'$ denotes the orthogonal projection onto the subspace Λ'. The average fidelity therefore obeys

$$\bar{F} = \sum_{\vec{x}} p(\vec{x}) F(\vec{x}) \leq \sum_{\vec{x}} p(\vec{x}) \langle\varphi(\vec{x})|\boldsymbol{E}'|\varphi(\vec{x})\rangle = \mathrm{tr}(\boldsymbol{\rho}^{\otimes n}\boldsymbol{E}'). \tag{10.154}$$

But, according to the Ky Fan dominance principle discussed in §10.2.2, since $\boldsymbol{E}'$ projects onto a space of dimension $2^{n(H-\delta')}$, $\mathrm{tr}(\boldsymbol{\rho}^{\otimes n}\boldsymbol{E}')$ can be no larger than the sum of the $2^{n(H-\delta')}$ largest eigenvalues of $\boldsymbol{\rho}^{\otimes n}$. The δ-typical eigenvalues of $\boldsymbol{\rho}^{\otimes n}$ are no smaller than $2^{-n(H-\delta)}$, so the sum of the $2^{n(H-\delta')}$ largest eigenvalues can be bounded above:

$$\mathrm{tr}(\boldsymbol{\rho}^{\otimes n}\boldsymbol{E}') \leq 2^{n(H-\delta')} 2^{-n(H-\delta)} + \varepsilon = 2^{-n(\delta'-\delta)} + \varepsilon, \tag{10.155}$$

where the $+\,\varepsilon$ accounts for the contribution from the atypical eigenvalues. Since we may choose ε and δ as small as we please for sufficiently large n, we conclude that the average fidelity $\bar{F}$ gets small as $n \to \infty$ if we compress to $H(\boldsymbol{\rho}) - \Omega(1)$ qubits per letter. We find, then, that $H(\boldsymbol{\rho})$ qubits per letter is the optimal compression of the quantum information that can be achieved if we are to obtain good fidelity as n goes to infinity. This is Schumacher's quantum source coding theorem.

The above argument applies to any conceivable encoding scheme, but only to a restricted class of decoding schemes, unitary decodings. The extension of the argument to general decoding schemes is sketched in §10.6.3. The conclusion is the same. The point is that $n(H-\delta)$ qubits are too few to faithfully encode the typical subspace.

There is another useful way to think about Schumacher's quantum compression protocol. Suppose that Alice's density operator $\boldsymbol{\rho}_A^{\otimes n}$ has a *purification* $|\psi\rangle_{RA}$ which Alice shares with Robert. Alice wants to convey her share of $|\psi\rangle_{RA}$ to Bob with high fidelity, sending as few qubits to Bob as possible. To accomplish this task, Alice can use the same procedure as described above, attempting to compress the state of A by projecting onto its typical subspace Λ. Alice's projection succeeds with probability

$$P(\boldsymbol{E}) = \langle\psi|\boldsymbol{I}\otimes\boldsymbol{E}|\psi\rangle = \mathrm{tr}\left(\boldsymbol{\rho}^{\otimes n}\boldsymbol{E}\right) \geq 1-\varepsilon, \tag{10.156}$$

where $\boldsymbol{E}$ projects onto Λ, and when successful prepares the state

$$\frac{(\boldsymbol{I}\otimes\boldsymbol{E})\,|\psi\rangle}{\sqrt{P(\boldsymbol{E})}}. \tag{10.157}$$

Therefore, after Bob decompresses, the state he shares with Robert has fidelity F_e with $|\psi\rangle$ satisfying

$$F_e \geq \langle\psi|\boldsymbol{I}\otimes\boldsymbol{E}|\psi\rangle\langle\psi|\boldsymbol{I}\otimes\boldsymbol{E}|\psi\rangle = \left(\mathrm{tr}\left(\boldsymbol{\rho}^{\otimes n}\boldsymbol{E}\right)\right)^2 \geq (1-\varepsilon)^2 \geq 1-2\varepsilon. \tag{10.158}$$

We conclude that Alice can transfer her share of the pure state $|\psi\rangle_{RA}$ to Bob by sending $nH(\boldsymbol{\rho})+o(n)$ qubits, achieving arbitrarily good *entanglement fidelity* F_e as $n\to\infty$.

To summarize, there is a close analogy between Shannon's classical source coding theorem and Schumacher's quantum source coding theorem. In the classical case, nearly all long messages are typical sequences, so we can code only these and still have a small probability of error. In the quantum case, nearly all long messages have nearly perfect overlap with the typical subspace, so we can code only the typical subspace and still achieve good fidelity.

Alternatively, Alice could send classical information to Bob, the string $x_1x_2\cdots x_n$, and Bob could follow these classical instructions to reconstruct Alice's state $|\varphi(x_1)\rangle\otimes\ldots\otimes|\varphi(x_n)\rangle$. By this means, they could achieve high-fidelity compression to $H(X)+o(1)$ bits — or qubits — per letter, where X is the classical ensemble $\{x,p(x)\}$. But if $\{|\varphi(x)\rangle,p(x)\}$ is an ensemble of *nonorthogonal* pure states, this classically achievable amount of compression is not optimal; some of the classical information about the preparation of the state is redundant, because the nonorthogonal states cannot be perfectly distinguished. Schumacher coding goes further, achieving optimal compression to $H(\boldsymbol{\rho})+o(1)$ qubits per letter. Quantum compression packages the message more efficiently than classical compression, but at a price — Bob receives the quantum state Alice intended to send, but Bob doesn't know what he has. In contrast to the classical case, Bob can't fully decipher Alice's quantum message accurately. An attempt to read the message will unavoidably disturb it.

10.4 Entanglement Concentration and Dilution

Any bipartite pure state that is not a product state is entangled. But *how* entangled? Can we compare two states and say that one is more entangled than the other?

For example, consider the two bipartite states

$$|\phi^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle),$$
$$|\psi\rangle = \frac{1}{\sqrt{2}}|00\rangle + \frac{1}{2}|11\rangle + \frac{1}{2}|22\rangle. \tag{10.159}$$

$|\phi^+\rangle$ is a maximally entangled state of two qubits, while $|\psi\rangle$ is a *partially* entangled state of two *qutrits*. Which is more entangled?

It is not immediately clear that the question has a meaningful answer. Why should it be possible to find an unambiguous way of ordering all bipartite pure states according to their degree of entanglement? Can we compare a pair of qutrits with a pair of qubits any more than we can compare apples and oranges?

A crucial feature of entanglement is that it cannot be created by local operations and classical communication (LOCC). In particular, if Alice and Bob share a bipartite pure state, its Schmidt number does not increase if Alice or Bob performs a unitary transformation on her/his share of the state, nor if Alice or Bob measures her/his share, even if Alice and Bob exchange classical messages about their actions and measurement outcomes. Therefore, any quantitative measure of entanglement should have the property that LOCC cannot increase it, and it should also vanish for an unentangled product state. An obvious candidate is the Schmidt number, but on reflection it does not seem very satisfactory. Consider

$$|\psi_\varepsilon\rangle = \sqrt{1 - 2|\varepsilon|^2}\ |00\rangle + \varepsilon|11\rangle + \varepsilon|22\rangle, \tag{10.160}$$

which has Schmidt number 3 for any $|\varepsilon| > 0$. Do we really want to say that $|\psi_\varepsilon\rangle$ is "more entangled" than $|\phi^+\rangle$? Entanglement, after all, can be regarded as a resource — we might plan to use it for teleportation, for example — and it seems clear that $|\psi_\varepsilon\rangle$ (for $|\varepsilon| \ll 1$) is a less valuable resource than $|\phi^+\rangle$.

It turns out, though, that there is a natural and useful way to quantify the entanglement of any bipartite pure state. To compare two states, we use LOCC to convert both states to a common currency that can be compared directly. The common currency is *maximal* entanglement, and the amount of shared entanglement can be expressed in units of Bell pairs (maximally entangled two-qubit states), also called *ebits* of entanglement.

To quantify the entanglement of a particular bipartite pure state, $|\psi\rangle_{AB}$, imagine preparing n identical copies of that state. Alice and Bob share a large supply of maximally entangled *Bell pairs*. Using LOCC, they are to convert k Bell pairs $(|\phi^+\rangle_{AB})^{\otimes k}$ to n high-fidelity copies of the desired state $(|\psi\rangle_{AB})^{\otimes n}$. What is the minimum number $k_{\min}$ of Bell pairs with which they can perform this task?

To obtain a precise answer, we consider the *asymptotic* setting, requiring arbitrarily high-fidelity conversion in the limit of large n. We say that a rate R of conversion from $|\phi^+\rangle$ to $|\psi\rangle$ is asymptotically achievable if for any $\varepsilon, \delta > 0$, there is an LOCC protocol with

$$\frac{k}{n} \leq R + \delta, \tag{10.161}$$

which prepares the target state $|\psi^+\rangle^{\otimes n}$ with fidelity $F \geq 1 - \varepsilon$. We define the *entanglement cost* E_C of $|\psi\rangle$ as the infimum of achievable conversion rates:

$$E_C(|\psi\rangle) := \inf \{\text{achievable rate for creating } |\psi\rangle \text{ from Bell pairs}\}. \tag{10.162}$$

Asymptotically, we can create many copies of $|\psi\rangle$ by consuming E_C Bell pairs per copy.

Now imagine that n copies of $|\psi\rangle_{AB}$ are already shared by Alice and Bob. Using LOCC, Alice and Bob are to convert $(|\psi\rangle_{AB})^{\otimes n}$ back to the standard currency: k' Bell pairs $|\phi^+\rangle_{AB}^{\otimes k'}$. What is the maximum number $k'_{\max}$ of Bell pairs they can extract from $|\psi\rangle_{AB}^{\otimes n}$? In this case we say that a rate R' of conversion from $|\psi\rangle$ to $|\phi^+\rangle$ is asymptotically achievable if for any $\varepsilon, \delta > 0$, there is an LOCC protocol with

$$\frac{k'}{n} \geq R' - \delta, \tag{10.163}$$

which prepares the target state $|\phi^+\rangle^{\otimes k'}$ with fidelity $F \geq 1 - \varepsilon$. We define the *distillable entanglement* E_D of $|\psi\rangle$ as the supremum of achievable conversion rates:

$$E_D(|\psi\rangle) := \sup \{\text{achievable rate for distilling Bell pairs from } |\psi\rangle\}. \tag{10.164}$$

Asymptotically, we can convert many copies of $|\psi\rangle$ to Bell pairs, obtaining E_D Bell pairs per copy of $|\psi\rangle$ consumed.

Since it is an in inviolable principle that LOCC cannot create entanglement, it is certain that

$$E_D(|\psi\rangle) \leq E_C(|\psi\rangle); \tag{10.165}$$

otherwise Alice and Bob could increase their number of shared Bell pairs by converting them to copies of $|\psi\rangle$ and then back to Bell pairs. In fact the entanglement cost and distillable entanglement are *equal* for bipartite pure states. (The story is more complicated for bipartite mixed states; see §10.5.) Therefore, for pure states at least we may drop the subscript, using $E(|\psi\rangle)$ to denote the *entanglement* of $|\psi\rangle$. We don't need to distinguish between entanglement cost and distillable entanglement because

conversion of entanglement from one form to another is an asymptotically *reversible* process. E quantifies both what we have to pay in Bell pairs to create $|\psi\rangle$, and value of $|\psi\rangle$ in Bell pairs for performing tasks like quantum teleportation which consume entanglement.

But what is the value of $E(|\psi\rangle_{AB})$? Perhaps you can guess — it is

$$E(|\psi\rangle_{AB}) = H(\boldsymbol{\rho}_A) = H(\boldsymbol{\rho}_B), \tag{10.166}$$

the Von Neumann entropy of Alice's density operator $\boldsymbol{\rho}_A$ (or equivalently Bob's density operator $\boldsymbol{\rho}_B$). This is clearly the right answer in the case where $|\psi\rangle_{AB}$ is a product of k Bell pairs. In that case $\boldsymbol{\rho}_A$ (or $\boldsymbol{\rho}_B$) is $\frac{1}{2}\boldsymbol{I}$ for each qubit in Alice's possession

$$\boldsymbol{\rho}_A = \left(\frac{1}{2}\boldsymbol{I}\right)^{\otimes k}, \tag{10.167}$$

and

$$H(\boldsymbol{\rho}_A) = k\ H\left(\frac{1}{2}\boldsymbol{I}\right) = k. \tag{10.168}$$

How do we see that $E = H(\boldsymbol{\rho}_A)$ is the right answer for any bipartite pure state?

Though it is perfectly fine to use Bell pairs as the common currency for comparing bipartite entangled states, in the asymptotic setting it is simpler and more natural to allow fractions of a Bell pair, which is what we'll do here. That is, we'll consider a maximally entangled state of two d-dimensional systems to be $\log_2 d$ Bell pairs, even if d is not a power of two. So our goal will be to show that Alice and Bob can use LOCC to convert shared maximal entanglement of systems with dimension $d = 2^{n(H(\boldsymbol{\rho}_A)+\delta)}$ into n copies of $|\psi\rangle$, for any positive δ and with arbitrarily good fidelity as $n \to \infty$, and conversely that Alice and Bob can use LOCC to convert n copies of $|\psi\rangle$ into a shared maximally entangled state of d-dimensional systems with arbitrarily good fidelity, where $d = 2^{n(H(\boldsymbol{\rho}_A)-\delta)}$. This suffices to demonstrate that $E_C(|\psi\rangle) = E_D(|\psi\rangle) = H(\boldsymbol{\rho}_A)$.

First let's see that if Alice and Bob share $k = n(H(\boldsymbol{\rho}_A) + \delta)$ Bell pairs, then they can prepare $|\psi\rangle_{AB}^{\otimes n}$ with high fidelity using LOCC. They perform this task, called *entanglement dilution*, by combining quantum teleportation with Schumacher compression. To get started, Alice locally creates n copies of $|\psi\rangle_{AC}$, where A and C are systems she controls in her laboratory. Next she wishes to teleport the C^n share of these copies to Bob, but to minimize the consumption of Bell pairs, she should compress C^n before teleporting it.

If A and C are d-dimensional, then the bipartite state $|\psi\rangle_{AC}$ can be expressed in terms of its Schmidt basis as

$$|\psi\rangle_{AC} = \sqrt{p_0}\ |00\rangle + \sqrt{p_1}\ |11\rangle + \ldots + \sqrt{p_{d-1}}\ |d{-}1, d{-}1\rangle, \tag{10.169}$$

and n copies of the state can be expressed as

$$|\psi\rangle_{AC}^{\otimes n} = \sum_{x_1,\ldots,x_n=0}^{d-1} \sqrt{p(x_1)\ldots p(x_n)}\ |x_1 x_2 \ldots x_n\rangle_{A^n} \otimes |x_1 x_2 \ldots x_n\rangle_{C^n}$$
$$= \sum_{\vec{x}} \sqrt{p(\vec{x})}\ |\vec{x}\rangle_{A^n} \otimes |\vec{x}\rangle_{C^n}, \tag{10.170}$$

where $\sum_{\vec{x}} p(\vec{x}) = 1$. If Alice attempts to project onto the δ-typical subspace of C^n, she succeeds with high probability

$$P = \sum_{\delta-\text{typical } \vec{x}} p(\vec{x}) \geq 1 - \varepsilon \tag{10.171}$$

and when successful prepares the post-measurement state

$$|\Psi\rangle_{A^n C^n} = P^{-1/2} \sum_{\delta-\text{typical } \vec{x}} \sqrt{p(\vec{x})}\ |\vec{x}\rangle_{A^n} \otimes |\vec{x}\rangle_{C^n}, \tag{10.172}$$

such that

$$\langle\Psi|\psi^{\otimes n}\rangle = P^{-1/2} \sum_{\delta-\text{typical } \vec{x}} p(\vec{x}) = \sqrt{P} \geq \sqrt{1-\varepsilon}. \tag{10.173}$$

Since the typical subspace has dimension at most $2^{n(H(\boldsymbol{\rho})+\delta)}$, Alice can teleport the C^n half of $|\Psi\rangle$ to Bob with perfect fidelity using no more than $n(H(\boldsymbol{\rho}) + \delta)$ Bell pairs shared by Alice and Bob. The teleportation uses LOCC: Alice's entangled measurement, classical communication from Alice to Bob to convey the measurement outcome, and Bob's unitary transformation conditioned on the outcome. Finally, after the teleportation, Bob decompresses, so that Alice and Bob share a state which has high fidelity with $|\psi\rangle_{AB}^{\otimes n}$. This protocol demonstrates that the entanglement cost E_C of $|\psi\rangle$ is not more than $H(\boldsymbol{\rho}_A)$.

Now consider the distillable entanglement E_D. Suppose Alice and Bob share the state $|\psi\rangle_{AB}^{\otimes n}$. Since $|\psi\rangle_{AB}$ is, in general, a *partially* entangled state, the entanglement that Alice and Bob share is in a diluted form. They wish to *concentrate* their shared entanglement, squeezing it down to the smallest possible Hilbert space; that is, they want to convert it to maximally-entangled pairs. We will show that Alice and Bob can "distill" at least

$$k' = n(H(\boldsymbol{\rho}_A) - \delta) \tag{10.174}$$

Bell pairs from $|\psi\rangle_{AB}^{\otimes n}$, with high likelihood of success.

To illustrate the concentration of entanglement, imagine that Alice and Bob have n copies of the two-qubit state $|\psi\rangle$, which is

$$|\psi(p)\rangle = \sqrt{1-p}\ |00\rangle + \sqrt{p}\ |11\rangle, \tag{10.175}$$

where $0 \leq p \leq 1$, when expressed in its Schmidt basis. That is, Alice and Bob share the state

$$|\psi(p)\rangle^{\otimes n} = (\sqrt{1-p}\,|00\rangle + \sqrt{p}\,|11\rangle)^{\otimes n}. \tag{10.176}$$

When we expand this state in the $\{|0\rangle, |1\rangle\}$ basis, we find 2^n terms, in each of which Alice and Bob hold exactly the same binary string of length n.

Now suppose Alice (or Bob) performs a local measurement on her (his) n qubits, measuring the *total* spin along the z-axis

$$\boldsymbol{\sigma}_3^{(\text{total})} = \sum_{i=1}^{n} \boldsymbol{\sigma}_3^{(i)}. \tag{10.177}$$

Equivalently, the measurement determines the *Hamming weight* of Alice's n qubits, the number of $|1\rangle$'s in Alice's n-bit string; that is, the number of spins pointing up.

In the expansion of $|\psi(p)\rangle^{\otimes n}$ there are $\binom{n}{m}$ terms in which Alice's string has Hamming weight m, each occurring with the same amplitude: $(1-p)^{(n-m)/2}\, p^{m/2}$. Hence the probability that Alice's measurement finds Hamming weight m is

$$p(m) = \binom{n}{m}(1-p)^{n-m}p^m. \tag{10.178}$$

Furthermore, because Alice is careful not to acquire any additional information besides the Hamming weight when she conducts the measurement, by measuring the Hamming weight m she prepares a *uniform* superposition of all $\binom{n}{m}$ strings with m up spins. Because Alice and Bob have perfectly correlated strings, if Bob were to measure the Hamming weight of his qubits he would find the same outcome as Alice. Alternatively, Alice could report her outcome to Bob in a classical message, saving Bob the trouble of doing the measurement himself. Thus, Alice and Bob share a maximally entangled state

$$\sum_{i=1}^{D} |i\rangle_A \otimes |i\rangle_B, \tag{10.179}$$

where the sum runs over the $D = \binom{n}{m}$ strings with Hamming weight m.

For n large the binomial distribution $\{p(m)\}$ approaches a sharply peaked function of m with mean $\mu = np$ and variance $\sigma^2 = np(1-p)$. Hence the probability of a large deviation from the mean,

$$|m - np| = \Omega(n), \tag{10.180}$$

is $p = \exp(-\Omega(n))$. Using Stirling's approximation, it then follows that

$$2^{n(H(p)-o(1))} \leq D \leq 2^{n(H(p)+o(1))}. \tag{10.181}$$

with probability approaching one as $n \to \infty$, where $H(p) = -p\log_2 p - (1-p)\log_2(1-p)$ is the entropy function. Thus with high probability Alice and Bob share a maximally entangled state of Hilbert spaces $\mathcal{H}_A$ and $\mathcal{H}_B$ with $\dim(\mathcal{H}_A) = \dim(\mathcal{H}_B) = D$ and $\log_2 D \geq n(H(p) - \delta)$. In this sense Alice and Bob can distill $H(p) - \delta$ Bell pairs per copy of $|\psi\rangle_{AB}$.

Though the number m of up spins that Alice (or Bob) finds in her (his) measurement is typically close to np, it can fluctuate about this value. Sometimes Alice and Bob will be lucky, and then will manage to distill more than $H(p)$ Bell pairs per copy of $|\psi(p)\rangle_{AB}$. But the probability of doing substantially better becomes negligible as $n \to \infty$.

The same idea applies to bipartite pure states in larger Hilbert spaces. If A and B are d-dimensional systems, then $|\psi\rangle_{AB}$ has the Schmidt decomposition

$$|\psi(X)\rangle_{AB} = \sum_{i=0}^{d-1} \sqrt{p(x)}\ |x\rangle_A \otimes |x\rangle_B, \tag{10.182}$$

where X is the classical ensemble $\{x, p(x)\}$, and $H(\boldsymbol{\rho}_A) = H(\boldsymbol{\rho}_B) = H(X)$. The Schmidt decomposition of n copies of $\psi\rangle$ is

$$\sum_{x_1,x_2,\ldots,x_n=0}^{d-1} \sqrt{p(x_1)p(x_2)\ldots p(x_n)}\ |x_1x_2\ldots x_n\rangle_{A^n} \otimes |x_1x_2\ldots x_n\rangle_{B^n}. \tag{10.183}$$

Now Alice (or Bob) can measure the total number of $|0\rangle$'s, the total number of $|1\rangle$'s, etc. in her (his) possession. If she finds $m_0|0\rangle$'s, $m_1|1\rangle$'s, etc., then her measurement prepares a maximally entangled state with Schmidt number

$$D(m_0, m_1, \ldots, m_{d-1}) = \frac{n!}{m_0!m_1!\ldots m_{d-1}!} \tag{10.184}$$

and this outcome occurs with probability

$$p(m) = D(m_0, m_1, \ldots, m_{d-1})p(0)^{m_0}p(1)^{m_1}\ldots p(d-1)^{m_{d-1}}. \tag{10.185}$$

For n large, Alice will typically find $m_x \approx np(x)$, and again the probability of a large deviation is small, so that, from Stirling's approximation

$$2^{n(H(X)-o(1))} \leq D \leq 2^{n(H(X)+o(1))} \tag{10.186}$$

with high probability. Thus, asymptotically for $n \to \infty$, $n(H(\boldsymbol{\rho}_A) - o(1))$ high-fidelity Bell pairs can be distilled from n copies of $|\psi\rangle$, establishing that $E_D(|\psi\rangle) \geq H(\boldsymbol{\rho}_A)$, and therefore $E_D(|\psi\rangle) = E_C(|\psi\rangle) = E(|\psi\rangle)$.

This entanglement concentration protocol uses local operations but does not require any classical communication. When Alice and Bob do the same measurement they always get the same outcome, so there is no need for them to communicate. Classical communication really is necessary, though, to perform entanglement dilution. The protocol we described here, based on teleportation, requires two bits of classical one-way communication per Bell pair consumed; in a more clever protocol this can be reduced to $O(\sqrt{n})$ bits, but no further. Since the classical communication cost is sublinear in n, the number of bits of classical communication needed per copy of $|\psi\rangle$ becomes negligible in the limit $n \to \infty$.

10.5 Quantifying Mixed-State Entanglement

10.5.1 Asymptotic irreversibility under LOCC

The entanglement cost E_C and the distillable entanglement E_D are natural and operationally meaningful ways to quantify entanglement. It's quite satisfying to find that, because entanglement dilution and concentration are asymptotically reversible for pure states, these two measures of pure-state bipartite entanglement agree, and provide another operational role for the Von Neumann entropy of a marginal quantum state.

We can define E_C and E_D for bipartite mixed states just as we did for pure states, but the story is more complicated — when we prepare many copies of a mixed state shared by Alice and Bob, the dilution of Bell pairs is not in general reversible, even asymptotically, and the distillable entanglement can be strictly less than the entanglement cost, though it can never be larger. There are even bipartite mixed states with nonzero entanglement cost and zero distillable entanglement, a phenomenon called *bound entanglement*. This irreversibility is not shocking; any bipartite operation which maps many copies of the pure state $|\phi^+\rangle_{AB}$ to many copies of the mixed state $\boldsymbol{\rho}_{AB}$ necessarily discards some information to the environment, and we don't normally expect a process that forgets information to be reversible.

This separation between E_C and E_D raises the question, what is the preferred way to quantify the amount of entanglement when two parties share a mixed quantum state? The answer is, it depends. Many different measures of bipartite mixed-state entanglement have been proposed, each with its own distinctive advantages and disadvantages. Even though they do not always agree, both E_C and E_D are certainly valid measures. A further distinction can be made between the rate E_{D1} at which entanglement can be distilled with one-way communication between the parties, and the rate E_D with two-way communication. There are bipartite mixed states for which $E_D > E_{D1}$, and even states for which E_D is nonzero while

E_{D1} is zero. In contrast to the pure-state case, we don't have nice formulas for the values of the various entanglement measures, though there are useful upper and lower bounds. We will derive a lower bound on E_{D1} in §10.8.2 (the *hashing inequality*).

There are certain properties that any reasonable measure of bipartite quantum entanglement should have. The most important is that it must not increase under local operations and classical communication, because quantum entanglement cannot be created by LOCC alone. A function on bipartite states that is nonincreasing under LOCC is called an *entanglement monotone*. Note that an entanglement monotone will also be *invariant* under local unitary operations $U_{AB} = U_A \otimes U_B$, for if U_{AB} can reduce the entanglement for any state, its inverse can increase entanglement.

A second important property is that a bipartite entanglement measure must *vanish for separable states*. Recall from Chapter 4 that a bipartite mixed state is separable if it can be expressed as a convex combination of product states,

$$\boldsymbol{\rho}_{AB} = \sum_x p(x)\ |\alpha(x)\rangle\langle\alpha(x)|_A \ \otimes\ |\beta(x)\rangle\langle\beta(x)|_B. \tag{10.187}$$

A separable state is not entangled, as it can be created using LOCC. Via classical communication, Alice and Bob can establish a shared source of randomness, the distribution $X = \{x, p(x)\}$. Then they may jointly sample from X; if the outcome is x, Alice prepares $|\alpha(x)\rangle$ while Bob prepares $|\beta(x)\rangle$.

A third desirable property for a bipartite entanglement measure is that it should agree with $E = E_C = E_D$ for bipartite pure states. Both the entanglement cost and the distillable entanglement respect all three of these properties.

We remark in passing that, despite the irreversibility of entanglement dilution under LOCC, there is a mathematically viable way to formulate a reversible theory of bipartite entanglement which applies even to mixed states. In this formulation, we allow Alice and Bob to perform arbitrary bipartite operations that are incapable of creating entanglement; these include LOCC as well as additional operations which cannot be realized using LOCC. In this framework, dilution and concentration of entanglement become asymptotically reversible even for mixed states, and a unique measure of entanglement can be formulated characterizing the optimal rate of conversion between copies of $\boldsymbol{\rho}_{AB}$ and Bell pairs using these non-entangling operations.

Irreversible bipartite entanglement theory under LOCC, and also the reversible theory under non-entangling bipartite operations, are both examples of *resource theories*. In the resource theory framework, one or

more parties are able to perform some restricted class of operations, and they are capable of preparing a certain restricted class of states using these operations. In addition, the parties may also have access to *resource states*, which are outside the class they can prepare on their own. Using their restricted operations, they can transform resource states from one form to another, or consume resource states to perform operations beyond what they could achieve with their restricted operations alone. The name "resource state" conveys that such states are valuable because they may be consumed to do useful things.

In a two-party setting, where LOCC is allowed or more general non-entangling operations are allowed, bipartite entangled states may be regarded as a valuable resource. Resource theory also applies if the allowed operations are required to obey certain symmetries; then states breaking this symmetry become a resource. In thermodynamics, states deviating from thermal equilibrium are a resource. Entanglement theory, as a particularly well developed resource theory, provides guidance and tools which are broadly applicable to many different interesting situations.

10.5.2 Squashed entanglement

As an example of an alternative bipartite entanglement measure, consider the *squashed entanglement* $E_{\rm sq}$, defined by

$$E_{\rm sq}(\boldsymbol{\rho}_{AB}) = \inf\left\{\frac{1}{2}I(A;B|C) : \boldsymbol{\rho}_{AB} = \mathrm{tr}_C\left(\boldsymbol{\rho}_{ABC}\right)\right\} \tag{10.188}$$

The squashed entanglement of $\boldsymbol{\rho}_{AB}$ is the greatest lower bound on the quantum conditional mutual information of all possible extensions of $\boldsymbol{\rho}_{AB}$ to a tripartite state $\boldsymbol{\rho}_{ABC}$; it can be shown to be an entanglement monotone. The locution "squashed" conveys that choosing an optimal conditioning system C squashes out the non-quantum correlations between A and B.

For pure states the extension is superfluous, so that

$$E_{\rm sq}(|\psi\rangle_{AB}) = \frac{1}{2}I(A;B) = H(A) = H(B) = E(|\psi\rangle_{AB}). \tag{10.189}$$

For a separable state, we may choose the extension

$$\boldsymbol{\rho}_{ABC} = \sum_x p(x)\ |\alpha(x)\rangle\langle\alpha(x)|_A \ \otimes\ |\beta(x)\rangle\langle\beta(x)|_B \ \otimes\ |x\rangle\langle x|_C. \tag{10.190}$$

where $\{|x\rangle_C\}$ is an orthonormal set; the state $\boldsymbol{\rho}_{ABC}$ has the block-diagonal form eq.(10.82) and hence $I(A;B|C) = 0$. Conversely, if $\boldsymbol{\rho}_{AB}$ has any extension $\boldsymbol{\rho}_{ABC}$ with $I(A;B|C) = 0$, then $\boldsymbol{\rho}_{ABC}$ has the form eq.(10.82) and therefore $\boldsymbol{\rho}_{AB}$ is separable.

$E_{\rm sq}$ is difficult to compute, because the infimum is to be evaluated over all possible extensions, where the system C may have arbitrarily high dimension. This property also raises the logical possibility that there are nonseparable states for which the infimum vanishes; conceivably, though a nonseparable $\boldsymbol{\rho}_{AB}$ can have no finite-dimensional extension for which $I(A;B|C) = 0$, perhaps $I(A;B|C)$ can approach zero as the dimension of C increases. Fortunately, though this is not easy to show, it turns out that $E_{\rm sq}$ is strictly positive for any nonseparable state. In this sense, then, it is a faithful entanglement measure, strictly positive if and only if the state is nonseparable.

One desirable property of $E_{\rm sq}$, not shared by E_C and E_D, is its additivity on tensor products (Exercise 10.6),

$$E_{\rm sq}(\boldsymbol{\rho}_{AB} \otimes \boldsymbol{\rho}_{A'B'}) = E_{\rm sq}(\boldsymbol{\rho}_{AB}) + E_{\rm sq}(\boldsymbol{\rho}_{A'B'}). \tag{10.191}$$

Though, unlike E_C and E_D, squashed entanglement does not have an obvious operational meaning, any additive entanglement monotone which matches E for bipartite pure states is bounded above and below by E_C and E_D respectively,

$$E_C \geq E_{\rm sq} \geq E_D. \tag{10.192}$$

10.5.3 Entanglement monogamy

Classical correlations are *polyamorous*; they can be shared among many parties. If Alice and Bob read the same newspaper, then they have information in common and become correlated. Nothing prevents Claire from reading the same newspaper; then Claire is just as strongly correlated with Alice and with Bob as Alice and Bob are with one another. Furthermore, David, Edith, and all their friends can read the newspaper and join the party as well.

Quantum correlations are not like that; they are harder to share. If Bob's state is pure, then the tripartite quantum state is a product $\boldsymbol{\rho}_B \otimes \boldsymbol{\rho}_{AC}$, and Bob is completely uncorrelated with Alice and Claire. If Bob's state is mixed, then he can be entangled with other parties. But if Bob is fully entangled with Alice (shares a pure state with Alice), then the state is a product $\boldsymbol{\rho}_{AB} \otimes \boldsymbol{\rho}_C$; Bob has used up all his ability to entangle by sharing with Alice, and Bob cannot be correlated with Claire at all. Conversely, if Bob shares a pure state with Claire, the state is $\boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_{BC}$, and Bob is uncorrelated with Alice. Thus we say that quantum entanglement is *monogamous*.

Entanglement measures obey monogamy inequalities which reflect this tradeoff between Bob's entanglement with Alice and with Claire in a three-party state. Squashed entanglement, in particular, obeys a

monogamy relation following easily from its definition, which was our primary motivation for introducing this quantity; we have

$$E_{\text{sq}}(A;B) + E_{\text{sq}}(A;C) \leq E_{\text{sq}}(A;BC). \tag{10.193}$$

In particular, in the case of a pure tripartite state, $E_{\text{sq}} = H(A)$ is the (pure-state) entanglement shared between A and BC. The inequality is saturated if Alice's system is divided into subsystems A_1 and A_2 such that the tripartite pure state is

$$|\psi\rangle_{ABC} = |\psi_1\rangle_{A_1B} \otimes |\psi_2\rangle_{A_2C}. \tag{10.194}$$

In general, combining eq.(10.192) with eq.(10.193) yields

$$E_D(A;B) + E_D(A;C) \leq E_C(A;BC); \tag{10.195}$$

loosely speaking, the entanglement cost $E_C(A;BC)$ imposes a ceiling on Alice's ability to entangle with Bob and Claire individually, requiring her to trade in some distillable entanglement with Bob to increase her distillable entanglement with Claire.

To prove the monogamy relation eq.(10.193), we note that mutual information obeys a *chain rule* which is really just a restatement of the definition of conditional mutual information:

$$I(A;BC) = I(A;C) + I(A;B|C). \tag{10.196}$$

A similar equation follows directly from the definition if we condition on a fourth system D,

$$I(A;BC|D) = I(A;C|D) + I(A;B|CD). \tag{10.197}$$

Now, $E_{\text{sq}}(A;BC)$ is the infimum of $I(A;BC|D)$ over all possible extensions of $\boldsymbol{\rho}_{ABC}$ to $\boldsymbol{\rho}_{ABCD}$. But since $\boldsymbol{\rho}_{ABCD}$ is also an extension of $\boldsymbol{\rho}_{AB}$ and $\boldsymbol{\rho}_{AC}$, we have

$$I(A;BC|D) \geq E_{\text{sq}}(A;C) + E_{\text{sq}}(A;B) \tag{10.198}$$

for any such extension. Taking the infimum over all $\boldsymbol{\rho}_{ABCD}$ yields eq.(10.193).

A further aspect of monogamy arises when we consider extending a quantum state to more parties. We say that the bipartite state $\boldsymbol{\rho}_{AB}$ of systems A and B is k-*extendable* if there is a $(k+1)$-part state $\boldsymbol{\rho}_{AB_1\ldots B_k}$ whose marginal state on AB_j matches $\boldsymbol{\rho}_{AB}$ for each $j = 1, 2, \ldots k$, and such that $\boldsymbol{\rho}_{AB_1\ldots B_k}$ is invariant under permutations of the k systems $B_1, B_2 \ldots B_k$. Separable states are k-extendable for every k, and entangled pure states are not even 2-extendable. Every entangled mixed state fails to be k-extendable for some finite k, and we may regard the maximal value $k_{\max}$ for which such a symmetric extension exists as a rough measure of how entangled the state is — bipartite entangled states with larger and larger $k_{\max}$ are closer and closer to being separable.

10.6 Accessible Information

10.6.1 How much can we learn from a measurement?

Consider a game played by Alice and Bob. Alice prepares a quantum state drawn from the ensemble $\mathcal{E} = \{\boldsymbol{\rho}(x), p(x)\}$ and sends the state to Bob. Bob knows this ensemble, but not the particular state that Alice chose to send. After receiving the state, Bob performs a POVM with elements $\{\boldsymbol{E}(y)\} \equiv \boldsymbol{E}$, hoping to find out as much as he can about what Alice sent. The conditional probability that Bob obtains outcome y if Alice sent $\boldsymbol{\rho}(x)$ is $p(y|x) = \mathrm{tr}\,(\boldsymbol{E}(y)\boldsymbol{\rho}(x))$, and the joint distribution governing Alice's preparation and Bob's measurement is $p(x, y) = p(y|x)p(x)$.

Before he measures, Bob's ignorance about Alice's state is quantified by $H(X)$, the number of "bits per letter" needed to specify x; after he measures his ignorance is reduced to $H(X|Y) = H(XY) - H(Y)$. The improvement in Bob's knowledge achieved by the measurement is Bob's *information gain*, the mutual information

$$I(X;Y) = H(X) - H(X|Y). \tag{10.199}$$

Bob's best strategy (his *optimal measurement*) maximizes this information gain. The best information gain Bob can achieve,

$$\mathrm{Acc}(\mathcal{E}) = \max_{\boldsymbol{E}} I(X;Y), \tag{10.200}$$

is a property of the ensemble $\mathcal{E}$ called the *accessible information* of $\mathcal{E}$.

If the states $\{\boldsymbol{\rho}(x)\}$ are mutually orthogonal they are perfectly distinguishable. Bob can identify Alice's state with certainty by choosing $\boldsymbol{E}(x)$ to be the projector onto the support of $\boldsymbol{\rho}(x)$; Then $p(y|x) = \delta_{x,y} = p(x|y)$, hence $H(X|Y) = \langle -\log p(x|y)\rangle = 0$ and $\mathrm{Acc}(\mathcal{E}) = H(X)$. Bob's task is more challenging if Alice's states are not orthogonal. Then no measurement will identify the state perfectly, so $H(X|Y)$ is necessarily positive and $\mathrm{Acc}(\mathcal{E}) < H(X)$.

Though there is no simple general formula for the accessible information of an ensemble, we can derive a useful upper bound, called the *Holevo bound*. For the special case of an ensemble of pure states $\mathcal{E} = \{|\varphi(x)\rangle, p(x)\}$, the Holevo bound becomes

$$\mathrm{Acc}(\mathcal{E}) \le H(\boldsymbol{\rho}), \quad \text{where} \quad \boldsymbol{\rho} = \sum_x p(x)|\varphi(x)\rangle\langle\varphi(x)|, \tag{10.201}$$

and a sharper statement is possible for an ensemble of mixed states, as we will see. Since the entropy for a quantum system with dimension d can be no larger than $\log d$, the Holevo bound asserts that Alice, by sending n qubits to Bob ($d = 2^n$) can convey no more than n bits of information.

This is true even if Bob performs a sophisticated collective measurement on all the qubits at once, rather than measuring them one at a time.

Therefore, if Alice wants to convey classical information to Bob by sending qubits, she can do no better than treating the qubits as though they were classical, sending each qubit in one of the two orthogonal states $\{|0\rangle, |1\rangle\}$ to transmit one bit. This statement is not so obvious. Alice might try to stuff more classical information into a single qubit by sending a state chosen from a large alphabet of pure single-qubit signal states, distributed uniformly on the Bloch sphere. But the enlarged alphabet is to no avail, because as the number of possible signals increases the signals also become less distinguishable, and Bob is not able to extract the extra information Alice hoped to deposit in the qubit.

If we can send information more efficiently by using an alphabet of mutually orthogonal states, why should we be interested in the accessible information for an ensemble of non-orthogonal states? There are many possible reasons. Perhaps Alice finds it easier to send signals, like coherent states, which are imperfectly distinguishable rather than mutually orthogonal. Or perhaps Alice sends signals to Bob through a noisy channel, so that signals which are orthogonal when they enter the channel are imperfectly distinguishable by the time they reach Bob.

The accessible information game also arises when an experimental physicist tries to measure an unknown classical force using a quantum system as a probe. For example, to measure the z-component of a magnetic field, we may prepare a spin-$\frac{1}{2}$ particle pointing in the x-direction; the spin precesses for time t in the unknown field, producing an ensemble of possible final states (which will be an ensemble of mixed states if the initial preparation is imperfect, or if decoherence occurs during the experiment). The more information we can gain about the final state of the spin, the more accurately we can determine the value of the magnetic field.

10.6.2 Holevo bound

Recall that quantum mutual information obeys monotonicity — if a quantum channel maps B to B', then $I(A;B) \geq I(A;B')$. We derive the Holevo bound by applying monotonicity of mutual information to the accessible information game. We will suppose that Alice records her chosen state in a classical register X and Bob likewise records his measurement outcome in another register Y, so that Bob's information gain is the mutual information $I(X;Y)$ of the two registers. After Alice's preparation

of her system A, the joint state of XA is

$$\boldsymbol{\rho}_{XA} = \sum_x p(x)|x\rangle\langle x| \otimes \boldsymbol{\rho}(x). \tag{10.202}$$

Bob's measurement is a quantum channel mapping A to AY according to

$$\boldsymbol{\rho}(x) \mapsto \sum_y \boldsymbol{M}(y)\boldsymbol{\rho}(x)\boldsymbol{M}(y)^\dagger \otimes |y\rangle\langle y|, \tag{10.203}$$

where $\boldsymbol{M}(y)^\dagger \boldsymbol{M}(y) = \boldsymbol{E}(y)$, yielding the state for XAY

$$\boldsymbol{\rho}'_{XAY} = \sum_x p(x)|x\rangle\langle x| \otimes \boldsymbol{M}(y)\boldsymbol{\rho}(x)\boldsymbol{M}(y)^\dagger \otimes |y\rangle\langle y|. \tag{10.204}$$

Now we have

$$I(X;Y)_{\boldsymbol{\rho}'} \leq I(X;AY)_{\boldsymbol{\rho}'} \leq I(X;A)_{\boldsymbol{\rho}}, \tag{10.205}$$

where the subscript indicates the state in which the mutual information is evaluated; the first inequality uses strong subadditivity in the state $\boldsymbol{\rho}'$, and the second uses monotonicity under the channel mapping $\boldsymbol{\rho}$ to $\boldsymbol{\rho}'$.

The quantity $I(X;A)$ is an intrinsic property of the ensemble $\mathcal{E}$; it is denoted $\chi(\mathcal{E})$ and called the *Holevo chi* of the ensemble. We have shown that however Bob chooses his measurement his information gain is bounded above by the Holevo chi; therefore,

$$\mathrm{Acc}(\mathcal{E}) \leq \chi(\mathcal{E}) := I(X;A)_{\boldsymbol{\rho}}. \tag{10.206}$$

This is the Holevo bound.

Now let's calculate $I(X;A)_{\boldsymbol{\rho}}$ explicitly. We note that

$$\begin{aligned} H(XA) &= -\mathrm{tr}\left(\sum_x p(x)|x\rangle\langle x| \otimes \boldsymbol{\rho}(x) \log\left(\sum_{x'} p(x')|x'\rangle\langle x'| \otimes \boldsymbol{\rho}(x')\right)\right) \\ &= -\sum_x \mathrm{tr}\; p(x)\boldsymbol{\rho}(x)\left(\log p(x) + \log \boldsymbol{\rho}(x)\right) \\ &= H(X) + \sum_x p(x) H(\boldsymbol{\rho}(x)), \end{aligned} \tag{10.207}$$

and therefore

$$H(A|X) = H(XA) - H(X) = \sum_x p(x) H(\boldsymbol{\rho}(x)). \tag{10.208}$$

Using $I(X;A) = H(A) - H(A|X)$, we then find

$$\chi(\mathcal{E}) = I(X;A) = H(\boldsymbol{\rho}_A) - \sum_x p(x) H(\boldsymbol{\rho}_A(x)) \equiv H(A)_{\mathcal{E}} - \langle H(A)\rangle_{\mathcal{E}} \tag{10.209}$$

For an ensemble of pure states, χ is just the entropy of the density operator arising from the ensemble, but for an ensemble $\mathcal{E}$ of mixed states it is a strictly smaller quantity – the difference between the entropy $H(\boldsymbol{\rho}_{\mathcal{E}})$ of the convex sum of signal states and the convex sum $\langle H\rangle_{\mathcal{E}}$ of the signal state entropies; this difference is always nonnegative because of the concavity of the entropy function (or because mutual information is nonnegative).

10.6.3 Monotonicity of Holevo χ

Since Holevo χ is the mutual information $I(X;A)$ of the classical register X and the quantum system A, the monotonicity of mutual information also implies the monotonicity of χ. If $\mathcal{N}: A \to A'$ is a quantum channel, then $I(X;A') \leq I(X;A)$ and therefore

$$\chi(\mathcal{E}') \leq \chi(\mathcal{E}), \tag{10.210}$$

where

$$\mathcal{E} = \{\boldsymbol{\rho}(x)), p(x)\} \quad \text{and} \quad \mathcal{E}' = \{\boldsymbol{\rho}'(x) = \mathcal{N}(\boldsymbol{\rho}(x)), p(x)\}. \tag{10.211}$$

A channel cannot increase the Holevo χ of an ensemble.

Its monotonicity provides a further indication that $\chi(\mathcal{E})$ is a useful measure of the information encoded in an ensemble of quantum states; the decoherence described by a quantum channel can reduce this quantity, but never increases it. In contrast, the Von Neumann entropy may either increase or decrease under the action of a channel. Mapping pure states to mixed states can increase H, but a channel might instead map the mixed states in an ensemble to a fixed pure state $|0\rangle\langle 0|$, decreasing H and improving the purity of each signal state, but without improving the distinguishability of the states.

We discussed the asymptotic limit $H(\boldsymbol{\rho})$ on quantum compression per letter in §10.3.2. There we considered unitary decoding; invoking the monotonicity of Holevo χ clarifies why more general decoders cannot do better. Suppose we compress and decompress the ensemble $\mathcal{E}^{\otimes n}$ using an encoder $\mathcal{N}_e$ and a decoder $\mathcal{N}_d$, where both maps are quantum channels:

$$\mathcal{E}^{\otimes n} \xrightarrow{\mathcal{N}_e} \tilde{\mathcal{E}}^{(n)} \xrightarrow{\mathcal{N}_d} \tilde{\mathcal{E}}'^{(n)} \approx \mathcal{E}^{\otimes n} \tag{10.212}$$

The Holevo χ of the input pure-state product ensemble is additive, $\chi(\mathcal{E}^{\otimes n}) = H(\boldsymbol{\rho}^{\otimes n}) = nH(\boldsymbol{\rho})$, and χ of a d-dimensional system is no larger than $\log_2 d$; therefore if the ensemble $\tilde{\mathcal{E}}^{(n)}$ is compressed to q qubits per letter, then because of the monotonicity of χ the decompressed ensemble $\tilde{\mathcal{E}}'^{(n)}$ has Holevo chi per letter $\frac{1}{n}\chi(\tilde{\mathcal{E}}'^{(n)}) \leq q$. If the decompressed output

ensemble has high fidelity with the input ensemble, its χ per letter should nearly match the χ per letter of the input ensemble, hence

$$q \geq \frac{1}{n}\chi(\tilde{\mathcal{E}}'^{(n)}) \geq H(\boldsymbol{\rho}) - \delta \tag{10.213}$$

for any positive δ and sufficiently large n. We conclude that high-fidelity compression to fewer than $H(\boldsymbol{\rho})$ qubits per letter is impossible asymptotically, even when the compression and decompression maps are arbitrary channels.

10.6.4 Improved distinguishability through coding: an example

To better acquaint ourselves with the concept of accessible information, let's consider a single-qubit example. Alice prepares one of the three possible pure states

$$\begin{aligned} |\varphi_1\rangle &= |\uparrow_{\hat{n}_1}\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \\ |\varphi_2\rangle &= |\uparrow_{\hat{n}_2}\rangle = \begin{pmatrix} -\frac{1}{2} \\ \frac{\sqrt{3}}{2} \end{pmatrix}, \\ |\varphi_3\rangle &= |\uparrow_{\hat{n}_3}\rangle = \begin{pmatrix} -\frac{1}{2} \\ -\frac{\sqrt{3}}{2} \end{pmatrix}; \end{aligned} \tag{10.214}$$

a spin-$\frac{1}{2}$ object points in one of three directions that are symmetrically distributed in the xz-plane. Each state has *a priori* probability $\frac{1}{3}$. Evidently, Alice's signal states are nonorthogonal:

$$\langle\varphi_1|\varphi_2\rangle = \langle\varphi_1|\varphi_3\rangle = \langle\varphi_2|\varphi_3\rangle = -\frac{1}{2}. \tag{10.215}$$

Bob's task is to find out as much as he can about what Alice prepared by making a suitable measurement. The density matrix of Alice's ensemble is

$$\boldsymbol{\rho} = \frac{1}{3}(|\varphi_1\rangle\langle\varphi_1| + |\varphi_2\rangle\langle\varphi_3| + |\varphi_3\rangle\langle\varphi_3|) = \frac{1}{2}\boldsymbol{I}, \tag{10.216}$$

which has $H(\boldsymbol{\rho}) = 1$. Therefore, the Holevo bound tells us that the mutual information of Alice's preparation and Bob's measurement outcome cannot exceed 1 bit.

In fact, though, the accessible information is considerably less than the one bit allowed by the Holevo bound. In this case, Alice's ensemble has enough symmetry that it is not hard to guess the optimal measurement. Bob may choose a POVM with three outcomes, where

$$\boldsymbol{E}_a = \frac{2}{3}(\boldsymbol{I} - |\varphi_a\rangle\langle\varphi_a|), \quad a = 1, 2, 3; \tag{10.217}$$

we see that

$$p(a|b) = \langle\varphi_b|\boldsymbol{E}_a|\varphi_b\rangle = \begin{cases} 0 & a = b, \\ \frac{1}{2} & a \neq b. \end{cases} \tag{10.218}$$

The measurement outcome a *excludes* the possibility that Alice prepared a, but leaves equal *a posteriori* probabilities $\left(p = \frac{1}{2}\right)$ for the other two states. Bob's information gain is

$$I = H(X) - H(X|Y) = \log_2 3 - 1 = .58496. \tag{10.219}$$

To show that this measurement is really optimal, we may appeal to a variation on a theorem of Davies, which assures us that an optimal POVM can be chosen with three $\boldsymbol{E}_a$'s that share the same three-fold symmetry as the three states in the input ensemble. This result restricts the possible POVM's enough so that we can check that eq. (10.217) is optimal with an explicit calculation. Hence we have found that the ensemble $\mathcal{E} = \{|\varphi_a\rangle, p_a = \frac{1}{3}\}$ has accessible information.

$$\mathrm{Acc}(\mathcal{E}) = \log_2\left(\frac{3}{2}\right) = .58496... \tag{10.220}$$

The Holevo bound is not saturated.

Now suppose that Alice has enough cash so that she can afford to send two qubits to Bob, where again each qubit is drawn from the ensemble $\mathcal{E}$. The obvious thing for Alice to do is prepare one of the *nine* states

$$|\varphi_a\rangle \otimes |\varphi_b\rangle, \quad a, b = 1, 2, 3, \tag{10.221}$$

each with $p_{ab} = 1/9$. Then Bob's best strategy is to perform the POVM eq. (10.217) on each of the two qubits, achieving a mutual information of .58496 bits per qubit, as before.

But, determined to do better, Alice and Bob decide on a different strategy. Alice will prepare one of *three* two-qubit states

$$|\Phi_a\rangle = |\varphi_a\rangle \otimes |\varphi_a\rangle, \quad a = 1, 2, 3, \tag{10.222}$$

each occurring with *a priori* probability $p_a = 1/3$. Considered one-qubit at a time, Alice's choice is governed by the ensemble $\mathcal{E}$, but now her two qubits have (classical) correlations – both are prepared the same way.

The three $|\Phi_a\rangle$'s are linearly independent, and so span a three-dimensional subspace of the four-dimensional two-qubit Hilbert space. In Exercise 10.4, you will show that the density operator

$$\boldsymbol{\rho} = \frac{1}{3}\left(\sum_{a=1}^{3} |\Phi_a\rangle\langle\Phi_a|\right), \tag{10.223}$$

has the nonzero eigenvalues $1/2, 1/4, 1/4$, so that

$$H(\boldsymbol{\rho}) = -\frac{1}{2}\log_2\frac{1}{2} - 2\left(\frac{1}{4}\log_2\frac{1}{4}\right) = \frac{3}{2}. \tag{10.224}$$

The Holevo bound requires that the accessible information *per qubit* is no more than 3/4 bit, which is at least consistent with the possibility that we can exceed the .58496 bits per qubit attained by the nine-state method.

Naively, it may seem that Alice won't be able to convey as much classical information to Bob, if she chooses to send one of only three possible states instead of nine. But on further reflection, this conclusion is not obvious. True, Alice has fewer signals to choose from, but the signals are *more distinguishable;* we have

$$\langle\Phi_a|\Phi_b\rangle = \frac{1}{4}, \quad a \neq b, \tag{10.225}$$

instead of eq. (10.215). It is up to Bob to exploit this improved distinguishability in his choice of measurement. In particular, Bob will find it advantageous to perform *collective* measurements on the two qubits instead of measuring them one at a time.

It is no longer obvious what Bob's optimal measurement will be. But Bob can invoke a general procedure that, while not guaranteed optimal, is usually at least pretty good. We'll call the POVM constructed by this procedure a "pretty good measurement" (or PGM).

Consider some collection of vectors $|\tilde{\Phi}_a\rangle$ that are not assumed to be orthogonal or normalized. We want to devise a POVM that can distinguish these vectors reasonably well. Let us first construct

$$\boldsymbol{G} = \sum_a |\tilde{\Phi}_a\rangle\langle\tilde{\Phi}_a|; \tag{10.226}$$

This is a positive operator on the space spanned by the $|\tilde{\Phi}_a\rangle$'s. Therefore, on that subspace, $\boldsymbol{G}$ has an inverse, $\boldsymbol{G}^{-1}$ and that inverse has a positive square root $\boldsymbol{G}^{-1/2}$. Now we define

$$\boldsymbol{E}_a = \boldsymbol{G}^{-1/2}|\tilde{\Phi}_a\rangle\langle\tilde{\Phi}_a|\boldsymbol{G}^{-1/2}, \tag{10.227}$$

and we see that

$$\begin{aligned}\sum_a \boldsymbol{E}_a &= \boldsymbol{G}^{-1/2}\left(\sum_a |\tilde{\Phi}_a\rangle\langle\tilde{\Phi}_a|\right)\boldsymbol{G}^{-1/2} \\ &= \boldsymbol{G}^{-1/2}\boldsymbol{G}\boldsymbol{G}^{-1/2} = \boldsymbol{I},\end{aligned} \tag{10.228}$$

on the span of the $|\tilde{\Phi}_a\rangle$'s. If necessary, we can augment these $\boldsymbol{E}_a$'s with one more positive operator, the projection $\boldsymbol{E}_0$ onto the orthogonal complement of the span of the $|\tilde{\Phi}_a\rangle$'s, and so construct a POVM. This POVM is the PGM associated with the vectors $|\tilde{\Phi}_a\rangle$.

In the special case where the $|\tilde{\Phi}_a\rangle$'s are orthogonal,

$$|\tilde{\Phi}_a\rangle = \sqrt{\lambda_a}|\phi_a\rangle, \tag{10.229}$$

(where the $|\Phi_a\rangle$'s are orthonormal), we have

$$\begin{aligned} \boldsymbol{E}_a &= \sum_{a,b,c} (|\phi_b\rangle\lambda_b^{-1/2}\langle\phi_b|)(|\phi_a\rangle\lambda_a\langle\phi_a|)(|\phi_c\rangle\lambda_c^{-1/2}\langle\phi_c|) \\ &= |\phi_a\rangle\langle\phi_a|; \end{aligned} \tag{10.230}$$

this is the orthogonal measurement that perfectly distinguishes the $|\Phi_a\rangle$'s and so clearly is optimal. If the $|\tilde{\Phi}_a\rangle$'s are linearly independent but not orthogonal, then the PGM is again an orthogonal measurement (because n one-dimensional operators in an n-dimensional space can constitute a POVM only if mutually orthogonal — see Exercise 3.11), but in that case the measurement may not be optimal.

In Exercise 10.4, you'll construct the PGM for the vectors $|\Phi_a\rangle$ in eq. (10.222), and you'll show that

$$\begin{aligned} p(a|a) &= \langle\Phi_a|\boldsymbol{E}_a|\Phi_a\rangle = \frac{1}{3}\left(1+\frac{1}{\sqrt{2}}\right)^2 = .971405 \\ p(b|a) &= \langle\Phi_a|\boldsymbol{E}_b|\Phi_a\rangle = \frac{1}{6}\left(1-\frac{1}{\sqrt{2}}\right)^2 = .0142977, \end{aligned} \tag{10.231}$$

(for $b \neq a$). It follows that the conditional entropy of the input is

$$H(X|Y) = .215893, \tag{10.232}$$

and since $H(X) = \log_2 3 = 1.58496$, the information gain is

$$I(X;Y) = H(X) - H(X|Y) = 1.36907, \tag{10.233}$$

a mutual information of .684535 bits per qubit. Thus, the improved distinguishability of Alice's signals has indeed paid off – we have exceeded the .58496 bits that can be extracted from a single qubit. We still didn't saturate the Holevo bound ($I \leq 1.5$ in this case), but we came a lot closer than before.

This example, first described by Peres and Wootters, teaches some useful lessons. First, Alice is able to convey more information to Bob by "pruning" her set of codewords. She is better off choosing among fewer signals that are more distinguishable than more signals that are less distinguishable. An alphabet of three letters encodes more than an alphabet of nine letters.

Second, Bob is able to read more of the information if he performs a collective measurement instead of measuring each qubit separately. His optimal orthogonal measurement projects Alice's signal onto a basis of *entangled* states.

10.6.5 Classical capacity of a quantum channel

This example illustrates how coding and collective measurement can enhance accessible information, but while using the code narrowed the gap between the accessible information and the Holevo chi of the ensemble, it did not close the gap completely. As is often the case in information theory, we can characterize the accessible information more precisely by considering an asymptotic i.i.d. setting. To be specific, we'll consider the task of sending classical information reliably through a noisy quantum channel $\mathcal{N}^{A\to B}$.

An ensemble of input signal states $\mathcal{E} = \{\boldsymbol{\rho}(x), p(x)\}$ prepared by Alice is mapped by the channel to an ensemble of output signals $\mathcal{E}' = \{\mathcal{N}(\boldsymbol{\rho}(x)), p(x)\}$. If Bob measures the output his information gain

$$\mathrm{Acc}(\mathcal{E}') \leq I(X;B) = \chi(\mathcal{E}'). \tag{10.234}$$

is bounded above by the Holevo chi of the output ensemble $\mathcal{E}'$. To convey as much information through the channel as possible, Alice and Bob may choose the input ensemble $\mathcal{E}$ that maximizes the Holevo chi of the output ensemble $\mathcal{E}'$. The maximum value

$$\chi(\mathcal{N}) := \max_{\mathcal{E}} \chi(\mathcal{E}') = \max_{\mathcal{E}} I(X;B), \tag{10.235}$$

of $\chi(\mathcal{E}')$ is a property of the channel, which we will call the Holevo chi of $\mathcal{N}$.

As we've seen, Bob's actual optimal information gain in this *single-shot* setting may fall short of $\chi(\mathcal{E}')$ in general. But instead of using the channel just once, suppose that Alice and Bob use the channel $n \gg 1$ times, where Alice sends signal states chosen from a code, and Bob performs an optimal measurement to decode the signals he receives. Then an information gain of $\chi(\mathcal{N})$ bits per letter really can be achieved asymptotically as $n \to \infty$.

Let's denote Alice's ensemble of encoded n-letter signal states by $\tilde{\mathcal{E}}^{(n)}$, denote the ensemble of classical labels carried by the signals by $\tilde{X}^n$, and denote Bob's ensemble of measurement outcomes by $\tilde{Y}^n$. Let's say that the code has rate R if Alice may choose from among 2^{nR} possible signals to send. If classical information can be sent through the channel with rate $R - o(1)$ such that Bob can decode the signal with negligible error probability as $n \to \infty$, then we say the rate R is *achievable*. The classical capacity $C(\mathcal{N})$ of the quantum channel $\mathcal{N}^{\mathcal{A}\to\mathcal{B}}$ is the supremum of all achievable rates.

Just as in our discussion of the capacity of a classical channel in §10.1.4, the conditional entropy per letter $\frac{1}{n}H(\tilde{X}^n|\tilde{Y}^n))$ approaches zero as $n \to \infty$

if the error probability is asymptotically negligible; therefore

$$\begin{aligned} R &\leq \frac{1}{n}\left(I(\tilde{X}^n;\tilde{Y}^n)+o(1)\right) \\ &\leq \frac{1}{n}\left(\max_{\mathcal{E}^{(n)}} I(X^n;B^n)+o(1)\right) = \frac{1}{n}\left(\chi(\mathcal{N}^{\otimes n})+o(1)\right), \end{aligned} \tag{10.236}$$

where we obtain the first inequality as in eq.(10.47) and the second inequality by invoking the Holevo bound, optimized over all possible n-letter input ensembles. We therefore infer that

$$C(\mathcal{N}) \leq \lim_{n\to\infty} \frac{1}{n}\chi\left(\mathcal{N}^{\otimes n}\right); \tag{10.237}$$

the classical capacity is bounded above by the asymptotic Holevo χ per letter of the product channel $\mathcal{N}^{\otimes n}$.

In fact this upper bound is actually an achievable rate, and hence equal to the classical capacity $C(\mathcal{N})$. However, this formula for the classical capacity is not very useful as it stands, because it requires that we optimize the Holevo χ over message ensembles of arbitrary length; we say that the formula for capacity is *regularized* if, as in this case, it involves taking a limit in which the number of channel tends to infinity. It would be far preferable to reduce our expression for $C(\mathcal{N})$ to a *single-letter formula* involving just one use of the channel. In the case of a classical channel, the reduction of the regularized expression to a single-letter formula was possible, because the conditional entropy for n uses of the channel is additive as in eq.(10.44).

For quantum channels the situation is more complicated, as channels are known to exist such that the Holevo χ is strictly superadditive:

$$\chi\left(\mathcal{N}_1 \otimes \mathcal{N}_2\right) > \chi\left(\mathcal{N}_1\right) + \chi\left(\mathcal{N}_2\right). \tag{10.238}$$

Therefore, at least for some channels, we are stuck with the not-very-useful regularized formula for the classical capacity. But we can obtain a single-letter formula for the optimal achievable communication rate if we put a restriction on the code used by Alice and Bob. In general, Alice is entitled to choose input codewords which are entangled across the many uses of the channel, and when such entangled codes are permitted the computation of the classical channel capacity may be difficult. But suppose we demand that all of Alice's codewords are product states. With that proviso the Holevo chi becomes subadditive, and we may express the optimal rate as

$$C_1\left(\mathcal{N}\right) = \chi(\mathcal{N}). \tag{10.239}$$

$C_1(\mathcal{N})$ is called the *product-state capacity* of the channel.

Let's verify the subadditivity of χ for product-state codes. The product channel $\mathcal{N}^{\otimes n}$ maps product states to product states; hence if Alice's input signals are product states then so are Bob's output signals, and we can express Bob's n-letter ensemble as

$$\mathcal{E}^{(n)} = \{\boldsymbol{\rho}(x_1) \otimes \boldsymbol{\rho}(x_2) \otimes \cdots \otimes \boldsymbol{\rho}(x_n),\ p(x_1 x_2 \dots x_n)\}, \tag{10.240}$$

which has Holevo χ

$$\chi(\mathcal{E}^{(n)}) = I(X^n; B^n) = H(B^n) - H(B^n|X^n). \tag{10.241}$$

While the Von Neumann entropy is subadditive,

$$H(B^n) = \sum_{i=1}^{n} H(B_i); \tag{10.242}$$

the (negated) conditional entropy

$$-H(B^n|X^n) = -\sum_{\vec{x}} p(\vec{x})\ H\left(\boldsymbol{\rho}(\vec{x})\right) \tag{10.243}$$

(see eq.(10.209)) is not subadditive in general. But for the product-state ensemble eq.(10.240), since the entropy of a product is additive, we have

$$\begin{aligned} H(B^n|X^n) &= \sum_{x_1, x_2, \dots, x_n} p(x_1 x_2, \dots x_n) \left(\sum_{i=1}^{n} H\left(\boldsymbol{\rho}(x_i)\right) \right) \\ &= \sum_{i=1}^{n} p_i(x_i) H(\boldsymbol{\rho}(x_i)) = \sum_{i=1}^{n} H(B_i|X_i) \end{aligned} \tag{10.244}$$

where $x_i = \{x_i, p_i(x_i)\}$ is the marginal probability distribution for the ith letter. Eq.(10.244) is a quantum analog of eq.(10.44), which holds for product-state ensembles but not in general for entangled ensembles. Combining eq.(10.241), (10.242), (10.244), we have

$$I(X^n; B^n) \leq \sum_{i=1}^{n} \left(H(B_i) - H(B_i|X_i)\right) = \sum_{i} I(X_i; B_i) \leq n\chi(\mathcal{N}). \tag{10.245}$$

Therefore the Holevo χ of a channel is subadditive when restricted to product-state codewords, as we wanted to show.

We won't give a careful argument here that $C_1(\mathcal{N})$ is an asymptotically achievable rate using product-state codewords; we'll just give a rough sketch of the idea. We demonstrate achievability with a random coding argument similar to Shannon's. Alice fixes an input ensemble

$\mathcal{E} = \{\boldsymbol{\rho}(x), p(x)\}$, and samples from the product ensemble $\mathcal{E}^{\otimes n}$ to generate a codeword; that is, the codeword

$$\boldsymbol{\rho}(\vec{x}) = \boldsymbol{\rho}(x_1) \otimes \boldsymbol{\rho}(x_2) \otimes \cdots \otimes \boldsymbol{\rho}(x_n) \tag{10.246}$$

is selected with probability $p(\vec{x}) = p(x_1)p(x_2)\ldots p(x_n)$. (In fact Alice should choose each $\boldsymbol{\rho}(\vec{x})$ to be pure to optimize the communication rate.) This codeword is sent via n uses of the channel $\mathcal{N}$, and Bob receives the product state

$$\mathcal{N}^{\otimes n}\left(\boldsymbol{\rho}(\vec{x})\right) = \mathcal{N}(\boldsymbol{\rho}(x_1)) \otimes \mathcal{N}(\boldsymbol{\rho}(x_2)) \otimes \cdots \otimes \mathcal{N}(\boldsymbol{\rho}(x_n)). \tag{10.247}$$

Averaged over codewords, the joint state of Alice's classical register X^n and Bob's system B^n is

$$\boldsymbol{\rho}_{X^n B^n} = \sum_{\vec{x}} p(\vec{x})\ |\vec{x}\rangle\langle\vec{x}| \otimes \mathcal{N}^{\otimes n}(\boldsymbol{\rho}(\vec{x})). \tag{10.248}$$

To decode, Bob performs a POVM designed to distinguish the codewords effectively; a variant of the pretty good measurement described in §10.6.4 does the job well enough. The state Bob receives is mostly supported on a typical subspace with dimension $2^{n(H(B)+o(1))}$, and for each typical codeword that Alice sends, what Bob receives is mostly supported on a much smaller typical subspace with dimension $2^{n(H(B|X)+o(1))}$. The key point is that ratio of these spaces is exponential in the mutual information of X and B:

$$\frac{2^{n(H(B|X)+o(1))}}{2^{n(H(B)-o(1))}} = 2^{-n(I(X;B)-o(1))} \tag{10.249}$$

Each of Bob's POVM elements has support on the typical subspace arising from a particular one of Alice's codewords. The probability that any codeword is mapped purely by accident to the decoding subspace of a different codeword is suppressed by the ratio eq.(10.249). Therefore, the probability of a decoding error remains small even when there are 2^{nR} codewords to distinguish, for $R = I(X;B) - o(1)$.

We complete the argument with standard Shannonisms. Since the probability of decoding error is small when we average over codes, it must also be small, averaged over codewords, for a particular sequence of codes. Then by pruning half of the codewords, reducing the rate by a negligible amount, we can ensure that the decoding errors are improbable for every codeword in the code. Therefore $I(X;B)$ is an achievable rate for classical communication. Optimizing over all product-state input ensembles, we obtain eq.(10.239).

To turn this into an honest argument, we would need to specify Bob's decoding measurement more explicitly and do a careful error analysis.

This gets a bit technical, so we'll skip the details. Somewhat surprisingly, though, it turns out to be easier to prove capacity theorems when quantum channels are used for other tasks besides sending classical information. We'll turn to that in §10.7.

10.6.6 Entanglement-breaking channels

Though Holevo chi is superadditive for some quantum channels, there are classes of channels for which chi is additive, and for any such channel $\mathcal{N}$ the classical capacity is $C = \chi(\mathcal{N})$ without any need for regularization. For example, consider *entanglement-breaking channels*. We say that $\mathcal{N}^{A\to B}$ is entanglement breaking if for any input state $\boldsymbol{\rho}_{RA}$, $I \otimes \mathcal{N}(\boldsymbol{\rho}_{RA})$ is a separable state on RA — the action of $\mathcal{N}$ on A always breaks its entanglement with R. We claim that if $\mathcal{N}_1$ is entanglement breaking, and $\mathcal{N}_2$ is an arbitrary channel, then

$$\chi\left(\mathcal{N}_1 \otimes \mathcal{N}_2\right) \leq \chi(\mathcal{N}_1) + \chi(\mathcal{N}_2). \tag{10.250}$$

To bound the chi of the product channel, consider an input ensemble

$$\boldsymbol{\rho}_{XA_1A_2} = \sum_x p(x)\ |x\rangle\langle x| \otimes \boldsymbol{\rho}(x)_{A_1A_2}. \tag{10.251}$$

Because $\mathcal{N}_1$ is entanglement breaking, $\boldsymbol{\rho}(x)_{A_1A_2}$ is mapped by the product channel to a separable state:

$$\mathcal{N}_1 \otimes \mathcal{N}_2 : \boldsymbol{\rho}(x)_{A_1A_2} \mapsto \sum_y p(y|x)\ \boldsymbol{\sigma}(x,y)_{B_1} \otimes \boldsymbol{\tau}(x,y)_{B_2}. \tag{10.252}$$

Now $\chi(\mathcal{N}_1 \otimes \mathcal{N}_2)$ is the maximum of $I(X;B_1B_2)_{\boldsymbol{\rho}'}$, evaluated in the state

$$\boldsymbol{\rho}'_{XB_1B_2} = \sum_{x,y} p(x)p(y|x)|x\rangle\langle x| \otimes \boldsymbol{\sigma}(x,y)_{B_1} \otimes \boldsymbol{\tau}(x,y)_{B_2} \tag{10.253}$$

which may be regarded as the marginal state (after tracing out Y) of

$$\tilde{\boldsymbol{\rho}}'_{XYB_1B_2} = \sum_{x,y} p(x,y)|x,y\rangle\langle x,y| \otimes \boldsymbol{\sigma}(x,y)_{B_1} \otimes \tilde{\boldsymbol{\tau}}(x,y)_{B_2} \tag{10.254}$$

Because $\tilde{\boldsymbol{\rho}}'$ becomes a product state when conditioned on (x,y), it satisfies

$$H(B_1B_2|XY) = H(B_1|XY) + H(B_2|XY), \tag{10.255}$$

and from the subadditivity and strong subadditivity of entropy we have

$$\begin{aligned} I(X;B_1B_2) &\leq I(XY;B_1B_2) = H(B_1B_2) - H(B_1B_2|XY) \\ &\leq H(B_1) + H(B_2) - H(B_1|XY) - H(B_2|XY) \\ &= I(XY;B_1) + I(XY;B_2). \end{aligned} \tag{10.256}$$

The right-hand side is bounded above by $\chi(\mathcal{N}_1)+\chi(\mathcal{N}_2)$, and maximizing the left-hand side yields eq.(10.250).

An example of an entanglement-breaking channel is a *classical-quantum channel*, also called a *c-q channel*, which acts according to

$$\mathcal{N}^{A\to B} : \boldsymbol{\rho}_A \mapsto \sum_x \langle x|\boldsymbol{\rho}_A|x\rangle \sigma(x)_B, \tag{10.257}$$

where $\{|x\rangle\}$ is an orthonormal basis. In effect, the channel performs a complete orthogonal measurement on the input state and then prepares an output state conditioned on the measurement outcome. The measurement breaks the entanglement between system A and any other system with which it was initially entangled. Therefore, c-q channels are entanglement breaking and have additive Holevo chi.

10.7 Quantum Channel Capacities and Decoupling

10.7.1 Coherent information and the quantum channel capacity

As we have already emphasized, it's marvelous that the capacity for a classical channel can be expressed in terms of the optimal correlation between input and output for a *single use* of the channel,

$$C := \max_X I(X;Y). \tag{10.258}$$

Another pleasing feature of this formula is its *robustness*. For example, the capacity does not increase if we allow the sender and receiver to share randomness, or if we allow feedback from receiver to sender. But for quantum channels the story is more complicated. We've seen already that no simple single-letter formula is known for the classical capacity of a quantum channel, if we allow entanglement among the channel inputs, and we'll soon see that the same is true for the quantum capacity. In addition, it turns out that entanglement shared between sender and receiver can boost the classical and quantum capacities of some channels, and so can "backward" communication from receiver to sender. There are a variety of different notions of capacity for quantum channels, all reasonably natural, and all with different achievable rates.

While Shannon's theory of classical communication over noisy classical channels is pristine and elegant, the same cannot be said for the theory of communication over noisy quantum channels, at least not in its current state. It's still a work in progress. Perhaps some day another genius like Shannon will construct a beautiful theory of quantum capacities. For now, at least there are a lot of interesting things we can say about achievable rates. Furthermore, the tools that have been developed

to address questions about quantum capacities have other applications beyond communication theory.

The most direct analog of the classical capacity of a classical channel is the quantum capacity of a quantum channel, unassisted by shared entanglement or feedback. The quantum channel $\mathcal{N}^{A\to B}$ is a TPCP map from $\mathcal{H}_A$ to $\mathcal{H}_B$, and Alice is to use the channel n times to convey a quantum state to Bob with high fidelity. She prepares her state $|\psi\rangle$ in a code subspace

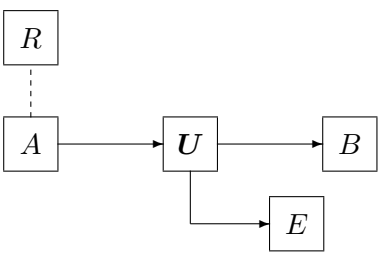
$$\mathcal{H}^{(n)} \subseteq \mathcal{H}_A^{\otimes n} \tag{10.259}$$

and sends it to Bob, who applies a decoding map, attempting to recover $|\psi\rangle$. The rate R of the code is the number of encoded qubits sent per channel use,

$$R = \log_2 \dim \left(\mathcal{H}^{(n)}\right), \tag{10.260}$$

We say that the rate R is *achievable* if there is a sequence of codes with increasing n such that for any $\varepsilon, \delta > 0$ and for sufficiently large n the rate is at least $R-\delta$ and Bob's recovered state $\boldsymbol{\rho}$ has fidelity $F = \langle\psi|\boldsymbol{\rho}|\psi\rangle \geq 1-\varepsilon$. The *quantum channel capacity* $Q(\mathcal{N})$ is the supremum of all achievable rates.

There is a regularized formula for $Q(\mathcal{N})$. To understand the formula we first need to recall that any channel $\mathcal{N}^{A\to B}$ has an isometric Stinespring dilation $\boldsymbol{U}^{A\to BE}$ where E is the channel's "environment." Furthermore, any input density operator $\boldsymbol{\rho}_A$ has a purification; if we introduce a *reference system* R, for any $\boldsymbol{\rho}_A$ there is a pure state ψ_{RA} such that $\boldsymbol{\rho}_A = \mathrm{tr}_R\left(|\psi\rangle\langle\psi|\right)$. (I will sometimes use ψ rather than the Dirac ket $|\psi\rangle$ to denote a pure state vector, when the context makes the meaning clear and the ket notation seems unnecessarily cumbersome.) Applying the channel's dilation to ψ_{RA}, we obtain an output pure state ϕ_{RBE}, which we represent graphically as:



We then define the *one-shot quantum capacity* of the channel $\mathcal{N}$ by

$$Q_1(\mathcal{N}) := \max_A \left(-H(R|B)_{\phi_{RBE}}\right). \tag{10.261}$$

Here the maximum is taken over all possible input density operators $\{\boldsymbol{\rho}_A\}$, and $H(R|B)$ is the quantum conditional entropy

$$H(R|B) = H(RB) - H(B) = H(E) - H(B), \tag{10.262}$$

where in the last equality we used $H(RB) = H(E)$ in a pure state of RBE. The quantity $-H(R|B)$ has such a pivotal role in quantum communication theory that it deserves to have its own special name. We call it the *coherent information* from R to B and denote it

$$I_c(R\rangle B)_\phi = -H(R|B)_\phi = H(B)_\phi - H(E)_\phi. \tag{10.263}$$

This quantity does not depend on how the purification ϕ of the density operator $\boldsymbol{\rho}_A$ is chosen; any one purification can be obtained from any other by a unitary transformation acting on R alone, which does not alter $H(B)$ or $H(E)$. Indeed, since the expression $H(B) - H(E)$ only depends on the marginal state of BE, for the purpose of computing this quantity we could just as well consider the input to the channel to be the mixed state $\boldsymbol{\rho}_A$ obtained from ψ_{RA} by tracing out the reference system R.

For a classical channel, $H(R|B)$ is always nonnegative and the coherent information is never positive. In the quantum setting, $I_c(R\rangle B)$ is positive if the reference system R is more strongly correlated with the channel output B than with the environment E. Indeed, an alternative way to express the coherent information is

$$I_c(R\rangle B) = \frac{1}{2}\left(I(R;B) - I(R;E)\right) = H(B) - H(E), \tag{10.264}$$

where we note that (because ϕ_{RBE} is pure)

$$\begin{aligned} I(R;B) &= H(R) + H(B) - H(RB) = H(R) + H(B) - H(E), \\ I(R;E) &= H(R) + H(E) - H(RE) = H(R) + H(E) - H(B). \end{aligned} \tag{10.265}$$

Now we can state the regularized formula for the quantum channel capacity — it is the optimal asymptotic coherent information per letter

$$Q(\mathcal{N}^{A\to B}) = \lim_{n\to\infty} \max_{A^n} \frac{1}{n} I_c(R^n\rangle B^n)_{\phi_{R^n B^n E^n}}, \tag{10.266}$$

where the input density operator $\boldsymbol{\rho}_{A^n}$ is allowed to be entangled across the n channel uses. If coherent information were subadditive, we could reduce this expression to a single-letter quantity, the one-shot capacity $Q_1(\mathcal{N})$. But, unfortunately, for some channels the coherent information can be superadditive, in which case the regularized formula is not very informative. At least we can say that $Q_1(\mathcal{N})$ is an achievable rate, and therefore a lower bound on the capacity.

10.7.2 The decoupling principle

Before we address achievability, let's understand why eq.(10.266) is an upper bound on the capacity. First we note that the monotonicity of mutual information implies a corresponding monotonicity property for the coherent information. Suppose that the channel $\mathcal{N}_1^{A\to B}$ is followed by a channel $\mathcal{N}_2^{B\to C}$. Because mutual information is monotonic we have

$$I(R;A) \geq I(R;B) \geq I(R;C), \tag{10.267}$$

which can also be expressed as

$$H(R) - H(R|A) \geq H(R) - H(R|B) \geq H(R) - H(R|C), \tag{10.268}$$

and hence

$$I_c(R\rangle A) \geq I_c(R\rangle B) \geq I_c(R\rangle C). \tag{10.269}$$

A quantum channel cannot increase the coherent information, which has been called the *quantum data-processing inequality*.

Suppose now that $\boldsymbol{\rho}_A$ is a quantum code state, and that the two channels acting in succession are a noisy channel $\mathcal{N}^{A\to B}$ and the decoding map $\mathcal{D}^{B\to\hat{B}}$ applied by Bob to the channel output in order to recover the channel input. Consider the action of the dilation $\boldsymbol{U}^{A\to BE}$ of $\mathcal{N}$ followed by the dilation $\boldsymbol{V}^{B\to\hat{B}B'}$ of $\mathcal{D}$ on the input purification ψ_{RA}, under the assumption that Bob is able to recover *perfectly*:

$$\psi_{RA} \xrightarrow{\boldsymbol{U}} \phi_{RBE} \xrightarrow{\boldsymbol{V}} \tilde{\psi}_{R\hat{B}B'E} = \psi_{R\hat{B}} \otimes \chi_{B'E}. \tag{10.270}$$

If the decoding is perfect, then after decoding Bob holds in system $\hat{B}$ the purification of the state of R, so that

$$H(R) = I_c(R\rangle A)_\psi = I_c(R\rangle \hat{B})_{\tilde{\psi}}. \tag{10.271}$$

Since the initial and final states have the same coherent information, the quantum data processing inequality implies that the same must be true for the intermediate state ϕ_{RBE}:

$$\begin{aligned} H(R) = I_c(R\rangle B) &= H(B) - H(E) \\ \implies H(B) &= H(RE) = H(R) + H(E). \end{aligned} \tag{10.272}$$

Thus the state of RE is a product state. We have found that if Bob is able to recover perfectly from the action of the channel dilation $\boldsymbol{U}^{A\to BE}$ on the pure state ψ_{RA}, then, in the resulting channel output pure state ϕ_{RBE}, the marginal state $\boldsymbol{\rho}_{RE}$ must be the product $\boldsymbol{\rho}_R \otimes \boldsymbol{\rho}_E$.

Conversely, suppose that ψ_{RA} is an entangled pure state, and Alice wishes to transfer the purification of R to Bob by sending it through the noisy channel $\boldsymbol{U}^{A\to BE}$. And suppose that in the resulting tripartite pure state ϕ_{RBE}, the marginal state of RE factorizes as $\boldsymbol{\rho}_{RE} = \boldsymbol{\rho}_R \otimes \boldsymbol{\rho}_E$. Then B decomposes into subsystems $B = \hat{B}_1 B_2$ such that

$$\phi_{RBE} = \tilde{\psi}_{RB_1} \otimes \chi_{B_2 E}. \tag{10.273}$$

Now Bob can construct an isometric decoder $\boldsymbol{V}^{B_1\to\hat{B}}$, which extracts the purification of R into Bob's preferred subsystem $\hat{B}$. Since all purifications of R differ by an isometry on Bob's side, Bob can choose his decoding map to output the state $\psi_{R\hat{B}}$; then the input state of RA is successfully transmitted to $R\hat{B}$ as desired. Furthermore, we may choose the initial state to be a maximally entangled state Φ_{RA} of the reference system with the code space of a quantum code; if the marginal state of RE factorizes in the resulting output pure state ϕ_{RBE}, then by the relative state method of Chapter 3 we conclude that *any* state in the code space can be sent through the channel and decoded with perfect fidelity by Bob.

We have found that purified quantum information transmitted through the noisy channel is exactly correctable if and only if the reference system is completely uncorrelated with the channel's environment, or as we sometimes say, *decoupled* from the environment. This is the *decoupling principle*, a powerful notion underlying many of the key results in the theory of quantum channels.

So far we have shown that exact correctability corresponds to exact decoupling. But we can likewise see that approximate correctability corresponds to approximate decoupling. Suppose for example that the state of RE is close to a product state in the L^1 norm:

$$\|\boldsymbol{\rho}_{RE} - \boldsymbol{\rho}_R \otimes \boldsymbol{\rho}_E\|_1 \leq \varepsilon. \tag{10.274}$$

As we learned in Chapter 2, if two density operators are close together in this norm, that means they also have fidelity close to one and hence purifications with a large overlap. Any purification of the product state $\boldsymbol{\rho}_R \otimes \boldsymbol{\rho}_E$ has the form

$$\tilde{\phi}_{RBE} = \tilde{\psi}_{RB_1} \otimes \chi_{B_2 E}, \tag{10.275}$$

and since all purifications of $\boldsymbol{\rho}_{RE}$ can be transformed to one another by an isometry acting on the purifying system B, there is a way to choose the decomposition $B = B_1 B_2$ such that

$$F(\boldsymbol{\rho}_{RE}, \boldsymbol{\rho}_R \otimes \boldsymbol{\rho}_E) = \left\| \langle \phi_{RBE} | \tilde{\phi}_{RBE} \rangle \right\|^2 \geq 1 - \|\boldsymbol{\rho}_{RE} - \boldsymbol{\rho}_R \otimes \boldsymbol{\rho}_E\|_1 \geq 1 - \varepsilon. \tag{10.276}$$

Furthermore, because fidelity is monotonic, both under tracing out E and under the action of Bob's decoding map, and because Bob can decode $\tilde{\phi}_{RBE}$ perfectly, we conclude that

$$F\left(\mathcal{D}^{B\to\hat{B}}\left(\boldsymbol{\rho}_{RB}\right),\psi_{R\hat{B}}\right)\geq 1-\varepsilon \tag{10.277}$$

if Bob chooses the proper decoding map $\mathcal{D}$. Thus approximate decoupling in the L^1 norm implies high-fidelity correctability. It is convenient to note that the argument still works the same way if $\boldsymbol{\rho}_{RE}$ is ε-close in the L^1 norm to $\tilde{\boldsymbol{\rho}}_R\otimes\tilde{\boldsymbol{\rho}}_E$, where $\tilde{\boldsymbol{\rho}}_R$ is not necessarily $\mathrm{tr}_E\left(\boldsymbol{\rho}_{RE}\right)$ and $\tilde{\boldsymbol{\rho}}_E$ is not necessarily $\mathrm{tr}_R\left(\boldsymbol{\rho}_{RE}\right)$. We'll use this form of the argument in what follows.

On the other hand, if (approximate) decoupling fails, the fidelity of Bob's decoded state will be seriously compromised. Suppose that in the state ϕ_{RBE} we have

$$H(R)+H(E)-H(RE)=\varepsilon>0. \tag{10.278}$$

Then the coherent information of ϕ is

$$I_c(R\rangle B)_\phi=H(B)_\phi-H(E)_\phi=H(RE)_\phi-H(E)_\phi=H(R)_\phi-\varepsilon. \tag{10.279}$$

By the quantum data processing inequality, we know that the coherent information of Bob's decoded state $\tilde{\psi}_{R\hat{B}}$ is no larger; hence

$$I_c(R\rangle\hat{B})_{\tilde{\psi}}=H(R)_\psi-H(R\hat{B})_{\tilde{\psi}}\leq H(R)_\psi-\varepsilon, \tag{10.280}$$

and therefore

$$H(R\hat{B})_{\tilde{\psi}}\geq\varepsilon \tag{10.281}$$

The deviation from perfect decoupling means that the decoded state of $R\hat{B}$ has some residual entanglement with the environment E, and is therefore impure.

Now we have the tools to derive an upper bound on the quantum channel capacity $Q(\mathcal{N})$. For n channel uses, let $\psi^{(n)}$ be a maximally entangled state of a reference system $\mathcal{H}_R^{(n)}\subseteq\mathcal{H}_R^{\otimes n}$ with a code space $\mathcal{H}_A^{(n)}\subseteq\mathcal{H}_A^{\otimes n}$, where $\dim\mathcal{H}_A^{(n)}=2^{nR}$, so that

$$I_c(R^n\rangle A^n)_{\psi^{(n)}}=H(R^n)_{\psi^{(n)}}=nR. \tag{10.282}$$

Now A^n is transmitted to B^n through $\left(\boldsymbol{U}^{A\to BE}\right)^{\otimes n}$, yielding the pure state $\phi^{(n)}$ of $R^nB^nE^n$. If Bob can decode with high fidelity, then his decoded state must have coherent information $H(R^n)_{\psi^{(n)}}-o(n)$, and the quantum data processing inequality then implies that

$$I_c(R^n\rangle B^n)_{\phi^{(n)}}=H(R^n)_{\psi^{(n)}}-o(n)=nR-o(n) \tag{10.283}$$

and hence

$$R = \frac{1}{n} I_c(R^n\rangle B^n)_{\phi^{(n)}} + o(1). \tag{10.284}$$

Taking the limit $n \to \infty$ we see that the expression for $Q(\mathcal{N})$ in eq.(10.266) is an upper bound on the quantum channel capacity. In Exercise 10.9, you will sharpen the statement eq.(10.283), showing that

$$H(R^n) - I_c(R^n\rangle B^n) \leq 2H_2(\varepsilon) + 4\varepsilon nR. \tag{10.285}$$

To show that $Q(\mathcal{N})$ is an achievable rate, rather than just an upper bound, we will need to formulate a quantum version of Shannon's random coding argument. Our strategy (see §10.9.3) will be to demonstrate the existence of codes that achieve approximate decoupling of E^n from R^n.

10.7.3 Degradable channels

Though coherent information can be superadditive in some cases, there are classes of channels for which the coherent information is additive, and therefore the quantum channel capacity matches the single-shot capacity, for which there is a single-letter formula. One such class is the class of *degradable channels*.

To understand what a degradable channel is, we first need the concept of a *complementary channel*. Any channel $\mathcal{N}^{A\to B}$ has a Stinespring dilation $\boldsymbol{U}^{A\to BE}$, from which we obtain $\mathcal{N}^{A\to B}$ by tracing out the environment E. Alternatively we obtain the channel $\mathcal{N}_c^{A\to E}$ complementary to $\mathcal{N}^{A\to B}$ by tracing out B instead. Since we have the freedom to compose $\boldsymbol{U}^{A\to BE}$ with an isometry $\boldsymbol{V}^{E\to E}$ without changing $\mathcal{N}^{A\to B}$, the complementary channel is defined only up to an isometry acting on E. This lack of uniqueness need not trouble us, because the properties of interest for the complementary channel are invariant under such isometries.

We say that the channel $\mathcal{N}^{A\to B}$ is degradable if we can obtain its complementary channel by composing $\mathcal{N}^{A\to B}$ with a channel mapping B to E:

$$\mathcal{N}_c^{A\to E} = \mathcal{T}^{B\to E} \circ \mathcal{N}^{A\to B}. \tag{10.286}$$

In this sense, when Alice sends a state through the channel, Bob, who holds system B, receives a less noisy copy than Eve, who holds system E.

Now suppose that $\boldsymbol{U}_1^{A_1\to B_1E_1}$ and $\boldsymbol{U}_2^{A_2\to B_2E_2}$ are dilations of the degradable channels $\mathcal{N}_1$ and $\mathcal{N}_2$. Alice introduces a reference system R and prepares an input pure state $\psi_{RA_1A_2}$, then sends the state to Bob via $\mathcal{N}_1 \otimes \mathcal{N}_2$, preparing the output pure state $\phi_{RB_1B_2E_1E_2}$. We would like to evaluate the coherent information $I_c(R\rangle B_1B_2)_\phi$ in this state.

The key point is that because both channels are degradable, there is a product channel $\mathcal{T}_1 \otimes \mathcal{T}_2$ mapping B_1B_2 to E_1E_2, and the monotonicity of mutual information therefore implies

$$I(B_1; B_2) \geq I(E_1; E_2). \tag{10.287}$$

Therefore, the coherent information satisfies

$$\begin{aligned} I_c(R\rangle B_1B_2) &= H(B_1B_2) - H(E_1E_2) \\ &= H(B_1) + H(B_2) - I(B_1; B_2) - H(E_1) - H(E_2) + I(E_1; E_2) \\ &\leq H(B_1) - H(E_1) + H(B_2) - H(E_2). \end{aligned} \tag{10.288}$$

These quantities are all evaluated in the state $\phi_{RB_1B_2E_1E_2}$. But notice that for the evaluation of $H(B_1) - H(E_1)$, the isometry $\boldsymbol{U}_2^{A_2 \to B_2E_2}$ is irrelevant. This quantity is really the same as the coherent information $I_c(RA_2\rangle B_1)$, where now we regard A_2 as part of the reference system for the input to channel $\mathcal{N}_1$. Similarly $H(B_2) - H(E_2) = I_c(RA_1\rangle B_2)$, and therefore,

$$I_c(R\rangle B_1B_2) \leq I_c(RA_2\rangle B_1) + I_c(RA_1\rangle B_2) \leq Q_1(\mathcal{N}_1) + Q_1(\mathcal{N}_2), \tag{10.289}$$

where in the last inequality we use the definition of the one-shot capacity as coherent information maximized over all inputs. Since $Q_1(\mathcal{N}_1 \otimes \mathcal{N}_2)$ is likewise defined by maximizing the coherent information $I_c(R\rangle B_1B_2)$, we find that

$$Q_1(\mathcal{N}_1 \otimes \mathcal{N}_2) \leq Q_1(\mathcal{N}_1) + Q_1(\mathcal{N}_2) \tag{10.290}$$

if $\mathcal{N}_1$ and $\mathcal{N}_2$ are degradable.

The regularized formula for the capacity of $\mathcal{N}$ is

$$Q(\mathcal{N}) = \lim_{n\to\infty} \frac{1}{n} Q_1(\mathcal{N}^{\otimes n}) \leq Q_1(\mathcal{N}), \tag{10.291}$$

where the last inequality follows from eq.(10.290) assuming that $\mathcal{N}$ is degradable. We'll see that $Q_1(\mathcal{N})$ is actually an achievable rate, and therefore a single-letter formula for the quantum capacity of a degradable channel.

As a concrete example of a degradable channel, consider the *generalized dephasing channel* with dilation

$$U^{A\to BE} : |x\rangle_A \mapsto |x\rangle_B \otimes |\alpha_x\rangle_E, \tag{10.292}$$

where $\{|x\rangle_A\}$, $\{|x\rangle_B\}$ are orthonormal bases for $\mathcal{H}_A$, $\mathcal{H}_B$ respectively, and the states $\{|\alpha_x\rangle_E\}$ of the environment are not necessarily orthogonal. The

corresponding channel is

$$\mathcal{N}^{A\to B} : \boldsymbol{\rho} \mapsto \sum_{x,x'} |x\rangle\langle x|\boldsymbol{\rho}|x'\rangle\langle\alpha_{x'}|\alpha_x\rangle\langle x'|, \tag{10.293}$$

which has the complementary channel

$$\mathcal{N}_c^{A\to E} : \boldsymbol{\rho} \mapsto \sum_{x} |\alpha_x\rangle\langle x|\boldsymbol{\rho}|x\rangle\langle\alpha_x|. \tag{10.294}$$

In the special case where the states $\{|\alpha_x\rangle_E = |x\rangle_E\}$ *are* orthonormal, we obtain the *completely dephasing channel*

$$\Delta^{A\to B} : \boldsymbol{\rho} \mapsto \sum_{x} |x\rangle\langle x|\boldsymbol{\rho}|x\rangle\langle x|, \tag{10.295}$$

whose complement $\Delta^{A\to E}$ has the same form as $\Delta^{A\to B}$. We can easily check that

$$\mathcal{N}_c^{A\to E} = \mathcal{N}_c^{C\to E} \circ \Delta^{B\to C} \circ \mathcal{N}^{A\to B}; \tag{10.296}$$

therefore $\mathcal{N}_c \circ \Delta$ degrades $\mathcal{N}$ to $\mathcal{N}_c$. Thus $\mathcal{N}$ is degradable and $Q(\mathcal{N}) = Q_1(\mathcal{N})$.

Further examples of degradable channels are discussed in Exercise 10.11.

10.8 Quantum Protocols

Using the decoupling principle in an i.i.d. setting, we can prove achievable rates for two fundamental quantum protocols. These are fondly known as the father and mother protocols, so named because each spawns a brood of interesting corollaries. We will formulate these protocols and discuss some of their "children" in this section, postponing the proofs until §10.9.

10.8.1 Father: Entanglement-assisted quantum communication

The father protocol is a scheme for entanglement-assisted quantum communication. Through many uses of a noisy quantum channel $\mathcal{N}^{A\to B}$, this protocol sends quantum information with high fidelity from Alice to Bob, while also consuming some previously prepared quantum entanglement shared by Alice and Bob. The task performed by the protocol is summarized by the *father resource inequality*

$$\left\langle \mathcal{N}^{A\to B} : \boldsymbol{\rho}_A \right\rangle + \frac{1}{2} I(R;E)[qq] \geq \frac{1}{2} I(R;B)[q \to q], \tag{10.297}$$

where the resources on the left-hand side can be used to achieve the result on the right-hand side, in an asymptotic i.i.d. setting. That is, for any positive ε, the quantum channel $\mathcal{N}$ may be used n times to transmit $\frac{n}{2}I(R;B) - o(n)$ qubits with fidelity $F \geq 1 - \varepsilon$, while consuming $\frac{n}{2}I(R;E) + o(n)$ ebits of entanglement shared between sender and receiver. These entropic quantities are evaluated in a tripartite pure state ϕ_{RBE}, obtained by applying the Stinespring dilation $\boldsymbol{U}^{A\to BE}$ of $\mathcal{N}^{A\to B}$ to the purification ψ_{RA} of the input density operator $\boldsymbol{\rho}_A$. Eq.(10.297) means that for any input density operator $\boldsymbol{\rho}_A$, there exists a coding procedure that achieves the quantum communication at the specified rate by consuming entanglement at the specified rate.

To remember the father resource inequality, it helps to keep in mind that $I(R;B)$ quantifies something *good*, the correlation with the reference system which survives transmission through the channel, while $I(R;E)$ quantifies something *bad*, the correlation between the reference system R and the channel's environment E, which causes the transmitted information to decohere. The larger the good quantity $I(R;B)$, the higher the rate of quantum communication. The larger the bad quantity $I(R;E)$, the more entanglement we need to consume to overcome the noise in the channel. To remember the factor of $\frac{1}{2}$ in front of $I(R;B)$, consider the case of a noiseless quantum channel, where ψ_{RA} is maximally entangled; in that case there is no environment,

$$\phi_{RB} = \frac{1}{\sqrt{d}} \sum_{i=0}^{d-1} |i\rangle_R \otimes |i\rangle_B, \tag{10.298}$$

and $\frac{1}{2}I(R;B) = H(R) = H(B) = \log_2 d$ is just the number of qubits in A. To remember the factor of $\frac{1}{2}$ in front of $I(R;E)$, consider the case of a noiseless *classical* channel, where the quantum information completely decoheres in a preferred basis; in that case

$$\phi_{RBE} = \frac{1}{\sqrt{d}} \sum_{i=0}^{d-1} |i\rangle_R \otimes |i\rangle_B \otimes |i\rangle_E, \tag{10.299}$$

and $I(R;B) = I(R;E) = H(R) = H(B) = \log_2 d$. Then the father inequality merely says that we can teleport $\frac{n}{2}$ qubits by consuming $\frac{n}{2}$ ebits and sending n classical bits.

Before proving the father resource inequality, we will first discuss a few of its interesting consequences.

Entanglement-assisted classical communication. Suppose Alice wants to send classical information to Bob, rather than quantum information. Then we can use superdense coding to turn the quantum communication

achieved by the father protocol into classical communication, at the cost of consuming some additional entanglement. By invoking the superdense coding resource inequality

$$SD: \quad [q \to q] + [qq] \geq 2[c \to c] \tag{10.300}$$

$\frac{n}{2} I(R;B)$ times, and combining with the father resource inequality, we obtain $I(R;B)$ bits of classical communication per use of the channel while consuming a number of ebits

$$\frac{1}{2} I(R;E) + \frac{1}{2} I(R;B) = H(R) \tag{10.301}$$

per channel use. Thus we obtain an achievable rate for entanglement-assisted classical communication through the noisy quantum channel:

$$\left\langle \mathcal{N}^{A \to B} : \boldsymbol{\rho}_A \right\rangle + H(R)[qq] \geq I(R;B)[c \to c]. \tag{10.302}$$

We may define the *entanglement-assisted classical capacity* $C_E(\mathcal{N})$ as the supremum over achievable rates of classical communication per channel use, assuming that an unlimited amount of entanglement is available at no cost. Then the resource inequality eq.(10.302) implies

$$C_E(\mathcal{N}) \geq \max_A I(R;B). \tag{10.303}$$

In this case there is a matching upper bound, so $C_E(\mathcal{N})$ is really an equality, and hence a single-letter formula for the entanglement-assisted classical capacity. Furthermore, eq.(10.302) tells us a rate of entanglement consumption which suffices to achieve the capacity. If we disregard the cost of entanglement, the father protocol shows that a rate can be achieved for entanglement-assisted quantum communication which is half the entanglement-assisted classical capacity $C_E(\mathcal{N})$ of the noisy channel $\mathcal{N}$. That's clearly true, since by consuming entanglement we can use teleportation to convert n bits of classical communication into $n/2$ qubits of quantum communication.

Quantum channel capacity. It may be that Alice wants to send quantum information to Bob, but Alice and Bob are not so fortunate as to have pre-existing entanglement at their disposal. They can still make use of the father protocol, if we are willing to loan them some entanglement, which they are later required to repay. In this case we say that the entanglement *catalyzes* the quantum communication. Entanglement is needed to activate the process to begin with, but at the conclusion of the process no net entanglement has been consumed.

In this catalytic setting, Alice and Bob borrow $\frac{1}{2} I(R;E)$ ebits of entanglement per use of the channel to get started, execute the father protocol,

and then sacrifice some of the quantum communication they have generated to replace the borrowed entanglement via the resource inequality

$$[q \to q] \geq [qq]. \tag{10.304}$$

After repaying their debt, Alice and Bob retain a number of qubits of quantum communication per channel use

$$\frac{1}{2}I(R;B) - \frac{1}{2}I(R;E) = H(B) - H(E) = I_c(R\rangle B), \tag{10.305}$$

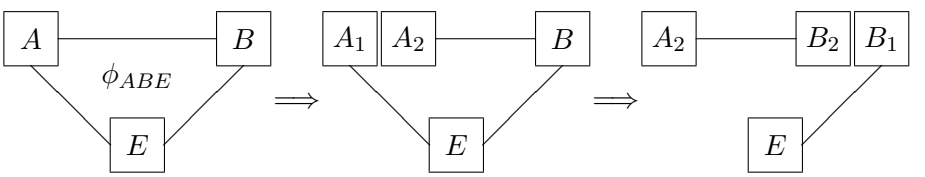
the channel's coherent information from R to B. We therefore obtain the achievable rate for quantum communication

$$\left\langle \mathcal{N}^{A\to B} : \boldsymbol{\rho}_A \right\rangle \geq I_c(R\rangle B)[q \to q], \tag{10.306}$$

albeit in the catalyzed setting. It can actually be shown that this same rate is achievable without invoking catalysis (see §10.9.4). As already discussed in §10.7.1, though, because of the superadditivity of coherent information this resource inequality does not yield a general single-letter formula for the quantum channel capacity $Q(\mathcal{N})$.

10.8.2 Mother: Quantum state transfer

In the mother protocol, Alice, Bob, and Eve initially share a tripartite pure state ϕ_{ABE}; thus Alice and Bob together hold the purification of Eve's system E. Alice wants to send her share of this purification to Bob, using as few qubits of noiseless quantum communication as possible. Therefore, Alice divides her system A into two subsystems A_1 and A_2, where A_1 is as small as possible and A_2 is uncorrelated with E. She keeps A_2 and sends A_1 to Bob. After receiving A_1, Bob divides A_1B into two subsystems B_1 and B_2, where B_1 purifies E and B_2 purifies A_2. Thus, at the conclusion of the protocol, Bob holds the purification of E in B_1, and in addition Alice and Bob share a bipartite pure state in A_2B_2. The protocol is portrayed in the following diagram:



In the i.i.d. version of the mother protocol, the initial state is $\phi_{ABE}^{\otimes n}$, and the task achieved by the protocol is summarized by the mother resource inequality

$$\langle \phi_{ABE} \rangle + \frac{1}{2}I(A;E)[q \to q] \geq \frac{1}{2}I(A;B)[qq] + \langle \phi'_{\tilde{B}E} \rangle, \tag{10.307}$$

where the resources on the left-hand side can be used to achieve the result on the right-hand side, in an asymptotic i.i.d. setting, and the entropic quantities are evaluated in the state ϕ_{ABE}. That is, if $A_1^{(n)}$ denotes the state Alice sends and $A_2^{(n)}$ denotes the state she keeps, then for any positive ε, the state of $A_2^{(n)}E^n$ is ε-close in the L^1 norm to a product state, where $\log \left|A_1^{(n)}\right| = \frac{n}{2}I(A;E)+o(n)$, while $A_2^{(n)}B_2^{(n)}$ contains $\frac{n}{2}I(A;B)-o(n)$ shared ebits of entanglement. Eq.(10.307) means that for any input pure state ϕ_{ABE} there is a way to choose the subsystem $A_2^{(n)}$ of the specified dimension such that $A_2^{(n)}$ and E^n are nearly uncorrelated and the specified amount of entanglement is harvested in $A_2^{(n)}B_2^{(n)}$.

The mother protocol is in a sense *dual* to the father protocol. While the father protocol consumes entanglement to achieve quantum communication, the mother protocol consumes quantum communication and harvests entanglement. For the mother, $I(A;B)$ quantifies the correlation between Alice and Bob at the beginning of the protocol (something *good*), and $I(A;E)$ quantifies the noise in the initial shared entanglement (something *bad*). The mother protocol can also be viewed as a quantum generalization of the Slepian-Wolf distributed compression protocol discussed in §10.1.3. The mother protocol merges Alice's and Bob's shares of the purification of E by sending Alice's share to Bob, much as distributed source coding merges the classical correlations shared by Alice and Bob by sending Alice's classical information to Bob. For this reason the mother protocol has been called the *fully quantum Slepian-Wolf protocol*; the modifier "fully" will be clarified in §10.8.2, when we discuss a variant on quantum state transfer in which classical communication is assumed to be freely available.

We may also view the mother protocol as a generalization of the entanglement concentration protocol discussed in §10.4, extending that discussion in three ways:

1. The initial entangled state shared by Alice and Bob may be mixed rather than pure.

2. The communication from Alice to Bob is quantum rather than classical.

3. The amount of communication that suffices to execute the protocol is quantified by the resource inequality.

Also note that if the state of AE is pure (uncorrelated with B), then the mother protocol reduces to Schumacher compression. In that case $\frac{1}{2}I(A;E) = H(A)$, and the mother resource inequality states that the

purification of A^n can be transferred to Bob with high fidelity using $nH(A) + o(n)$ qubits of quantum communication.

Before proving the mother resource inequality, we will first discuss a few of its interesting consequences.

Hashing inequality. Suppose Alice and Bob wish to distill entanglement from many copies of the state ϕ_{ABE}, using only local operations and classical communication (LOCC). In the catalytic setting, they can borrow some quantum communication, use the mother protocol to distill some shared entanglement, and then use classical communication and their harvested entanglement to repay their debt via quantum teleportation. Using the teleportation resource inequality

$$TP: \quad [qq] + 2[c \to c] \geq [q \to q] \tag{10.308}$$

$\frac{n}{2} I(A;E)$ times, and combining with the mother resource inequality, we obtain

$$\langle \phi_{ABE} \rangle + I(A;E)[c \to c] \geq I_c(A\rangle B)[qq] + \langle \phi'_{\tilde{B}E} \rangle, \tag{10.309}$$

since the net amount of distilled entanglement is $\frac{1}{2} I(A;B)$ per copy of ϕ achieved by the mother minus the $\frac{1}{2} I(A;E)$ per copy consumed by teleportation, and

$$\frac{1}{2} I(A;B) - \frac{1}{2} I(A;E) = H(B) - H(E) = I_c(A\rangle B). \tag{10.310}$$

Eq.(10.309) is the *hashing inequality*, which quantifies an achievable rate for distilling ebits of entanglement shared by Alice and Bob from many copies of a mixed state $\boldsymbol{\rho}_{AB}$, using one-way classical communication, assuming that $I_c(A\rangle B) = -H(A|B)$ is positive. Furthermore, the hashing inequality tells us how much classical communication suffices for this purpose.

In the case where the state $\boldsymbol{\rho}_{AB}$ is pure, $I_c(A\rangle B) = H(A) - H(AB) = H(A)$ and there is no environment E; thus we recover our earlier conclusion about concentration of pure-state bipartite entanglement — that $H(A)$ Bell pairs can be extracted per copy, with a negligible classical communication cost.

State merging. Suppose Alice and Bob share the purification of Eve's state, and Alice wants to transfer her share of the purification to Bob, where now unlimited *classical* communication from Alice to Bob is available at no cost. In contrast to the mother protocol, Alice wants to achieve the transfer with as little one-way quantum communication as possible, even if she needs to send more bits in order to send fewer qubits.

In the catalytic setting, Alice and Bob can borrow some quantum communication, perform the mother protocol, then use teleportation and the entanglement extracted by the mother protocol to repay some of the borrowed quantum communication. Combining teleportation of $\frac{n}{2}I(A;B)$ qubits with the mother resource inequality, we obtain

$$\langle \phi_{ABE} \rangle + H(A|B)[q \to q] + I(A;B)[c \to c] \geq \langle \phi'_{\tilde{B}E} \rangle, \tag{10.311}$$

using

$$\frac{1}{2}I(A;E) - \frac{1}{2}I(A;B) = H(E) - H(B) = H(AB) - H(B) = H(A|B). \tag{10.312}$$

Eq.(10.311) is the *state-merging inequality*, expressing how much quantum and classical communication suffices to achieve the state transfer in an i.i.d. setting, assuming that $H(A|B)$ is nonnegative.

Like the mother protocol, this state merging protocol can be viewed as a (partially) quantum version of the Slepian-Wolf protocol for merging classical correlations. In the classical setting, $H(X|Y)$ quantifies Bob's remaining ignorance about Alice's information X when Bob knows only Y; correspondingly, Alice can reveal X to Bob by sending $H(X|Y)$ bits per letter of X. Similarly, state merging provides an operational meaning to the quantum conditional information $H(A|B)$, as the number of qubits per copy of ϕ that Alice sends to Bob to convey her share of the purification of E, assuming classical communication is free. In this sense we may regard $H(A|B)$ as a measure of Bob's remaining "ignorance" about the shared purification of E when he holds only B.

Classically, $H(X|Y)$ is nonnegative, and zero if and only if Bob is already certain about XY, but quantumly $H(A|B)$ can be negative. How can Bob have "negative uncertainty" about the quantum state of AB? If $H(A|B) < 0$, or equivalently if $I(A;E) < I(A;B)$, then the mother protocol yields more quantum entanglement than the amount of quantum communication it consumes. Therefore, when $H(A|B)$ is negative (*i.e.* $I_c(A\rangle B)$ is positive), the mother resource inequality implies the Hashing inequality, asserting that classical communication from Alice to Bob not only achieves state transfer, but also distills $-H(A|B)$ ebits of entanglement per copy of ϕ. These distilled ebits can be deposited in the entanglement bank, to be withdrawn as needed in future rounds of state merging, thus reducing the quantum communication cost of those future rounds. Bob's "negative uncertainty" today reduces the quantum communication cost of tasks to be performed tomorrow.

10.8.3 Operational meaning of strong subadditivity

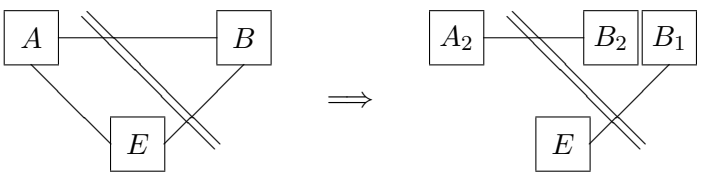
The observation that $H(A|B)$ is the quantum communication cost of state merging allows us to formulate a simple operational proof of the strong subadditivity of Von Neumann entropy, expressed in the form

$$H(A|BC) \leq H(A|B), \quad \text{or} \quad -H(A|B) \leq -H(A|BC). \tag{10.313}$$

When $H(A|B)$ is positive, eq.(10.313) is the obvious statement that it is no harder to merge Alice's system with Bob's if Bob holds C as well as B. When $H(A|B)$ is negative, eq.(10.313) is the obvious statement that Alice and Bob can distill no less entanglement using one-way classical communication if Bob holds C as well as B.

To complete this argument, we need to know that $H(A|B)$ is not only *achievable* but also that it is the *optimal* quantum communication cost of state merging, and that $-H(A|B)$ ebits is the optimal yield of hashing. The optimality follows from the principle that, for a bipartite pure state, k qubits of quantum communication cannot increase the shared entanglement of AB by more than k ebits.

If $H(A|B)$ is negative, consider cutting the system ABE into the two parts AE and B, as in the following figure:



In the hashing protocol, applied to n copies of ϕ_{ABE}, the entanglement across this cut at the beginning of the protocol is $nH(B)$. By the end of the protocol E^n has decoupled from $A_2^{(n)}$ and has entanglement $nH(E)$ with $B_1^{(n)}$, ignoring $o(n)$ corrections. If k ebits shared by Alice and Bob are distilled, the final entanglement across the AE-B cut is

$$nH(E) + k \leq nH(B) \implies \frac{k}{n} \leq H(B) - H(E) = -H(A|B). \tag{10.314}$$

This inequality holds because LOCC cannot increase the entanglement across the cut, and implies that no more than $-H(A|B)$ ebits of entanglement per copy of ϕ_{ABE} can be distilled in the hashing protocol, asymptotically.

On the other hand, if $H(A|B)$ is positive, at the conclusion of state merging $B_1^{(n)}$ is entangled with E^n, and the entanglement across the AE-B cut is at least $nH(E)$. To achieve this increase in entanglement, the

number of qubits sent from Alice to Bob must be at least

$$k \geq nH(E) - nH(B) \implies \frac{k}{n} \geq H(E) - H(B) = H(A|B) \qquad (10.315)$$

This inequality holds because the entanglement across the cut cannot increase by more than the quantum communication across the cut, and implies that at least $H(A|B)$ qubits must be sent per copy of ϕ_{ABE} to achieve state merging.

To summarize, we have proven strong subadditivity, not by the traditional route of sophisticated matrix analysis, but via a less direct method. This proof is built on two cornerstones of quantum information theory — the decoupling principle and the theory of typical subspaces — which are essential ingredients in the proof of the mother resource inequality.

10.8.4 Negative conditional entropy in thermodynamics

As a further application of the decoupling mother resource inequality, we now revisit *Landauer's Principle*, developing another perspective on the implications of negative quantum conditional entropy. Recall that *erasure* of a bit is a process which maps the bit to 0 irrespective of its initial value. This process is *irreversible* — knowing only the final state 0 after erasure, we cannot determine whether the initial state before erasure was 0 or 1. Irreversibility implies that erasure incurs an unavoidable thermodynamic cost. According to Landauer's Principle, erasing a bit at temperature T requires work no less than $W = kT \ln 2$.

A specific erasure procedure is analyzed in Exercise 10.14. Suppose a two-level quantum system has energy eigenstates $|0\rangle$, $|1\rangle$ with corresponding eigenvalues E_0 and E_1, where $E = E_1 - E_0 \geq 0$. Initially the qubit is in an unknown mixture of these two states, and the energy splitting is $E = 0$. We erase the bit in three steps. In the first step, we bring the bit into contact with a heat bath at temperature $T > 0$, and wait for the bit to come to thermal equilibrium with the bath. In this step the bit "forgets" its initial value, but the bit is not yet erased because it has not been reset. In the second step, with the bit still in contact with the bath, we turn on a control field which slowly increases E_1 to a value much larger than kT while maintaining thermal equilibrium all the while, thus resetting the bit to $|0\rangle$. In the third step, we isolate the bit from the bath and turn off the control field, so the two states of the bit become degenerate again. As shown in Exercise 10.14, work $W = kT \ln 2$ is required to execute step 2, with the energy dissipated as heat flowing from bit to bath.

We can also run the last two steps backward, increasing E_1 while the bit is isolated from the bath, then decreasing E_1 with the bit in contact

with the bath. This procedure maps the state $|0\rangle$ to the maximally mixed state of the bit, extracting work $W = kT \ln 2$ from the bath in the process.

Erasure is irreversible because the agent performing the erasure does not know the information being erased. (If a copy of the information were stored in her memory, survival of that copy would mean that the erasure had not succeeded). From an information-theoretic perspective, the reduction in the thermodynamic entropy of the erased bit, and hence the work required to perform the erasure, arises because erasure reduces the agent's *ignorance* about the state of the bit, ignorance which is quantified by the Shannon entropy. But to be more precise, it is the *conditional* entropy of the system, given the state of the agent's memory, which captures the agent's ignorance before erasure and therefore also the thermodynamic cost of erasing. Thus the minimal work needed to erase system A should be expressed as

$$W(A|O) = H(A|O)kT \ln 2, \tag{10.316}$$

where O is the memory of the *observer* who performs the erasure, and $H(A|O)$ quantifies that observer's ignorance about the state of A.

But what if A and O are quantum systems? We know that if A and O are entangled, then the conditional entropy $H(A|O)$ can be negative. Does that mean we can erase A while *extracting* work rather than doing work?

Yes, we can! Suppose for example that A and O are qubits and their initial state is maximally entangled. By controlling the contact between AO and the heat bath, the observer can extract work $W = 2kT \log 2$ while transforming AO to a maximally mixed state, using the same work extraction protocol as described above. Then she can do work $W = kT \log 2$ to return A to the state $|0\rangle$. The net effect is to erase A while extracting work $W = kT \log 2$, satisfying the equality eq.(10.316).

To appreciate why this trick works, we should consider the joint state of AO rather than the state of A alone. Although the marginal state of A is mixed at the beginning of the protocol and pure at the end, the state of AO is pure at the beginning and mixed at the end. Positive work is extracted by sacrificing the purity of AO.

To generalize this idea, let's consider $n \gg 1$ copies of the state $\boldsymbol{\rho}_{AO}$ of system A and memory O. Our goal is to map the n copies of A to the erased state $|000\ldots0\rangle$ while using or extracting the optimal amount of work. In fact, the optimal work per copy is given by eq.(10.316) in the $n \to \infty$ limit.

To achieve this asymptotic work per copy, the observer first projects A^n onto its typical subspace, succeeding with probability $1 - o(1)$. A unitary transformation then rotates the typical subspace to a subsystem

$\bar{A}$ containing $n(H(A) + o(1))$ qubits, while erasing the complementary qubits as in eq.(10.144). Now it only remains to erase $\bar{A}$.

The mother resource inequality ensures that we may decompose $\bar{A}$ into subsystems A_1A_2 such that A_2 contains $\frac{n}{2}\left(I(A;O) - o(1)\right)$ qubits and is nearly maximally entangled with a subsystem of O^n. What is important for the erasure protocol is that we may identify a subsystem of $\bar{A}O^n$ containing $n\left(I(A;O) - o(1)\right)$ qubits which is only distance $o(1)$ away from a pure state. By controlling the contact between this subsystem and the heat bath, we may extract work $W = n(I(A;O) - o(1))kT\log 2$ while transforming the subsystem to a maximally mixed state. We then proceed to erase $\bar{A}$, expending work $kT\log|\bar{A}| = n(H(A)+o(1))kT\log 2$. The net work cost of the erasure, per copy of $\boldsymbol{\rho}_{AO}$, is therefore

$$W = \left(H(A) - I(A;O) + o(1)\right)kT\log 2 = \left(H(A|O) + o(1)\right)kT\log 2, \tag{10.317}$$

and the erasure succeeds with probability $1 - o(1)$. A notable feature of the protocol is that only the subsystem of O^n which is entangled with A_2 is affected. Any correlation of the memory O with other systems remains intact, and can be exploited in the future to reduce the cost of erasure of those other systems.

As does the state merging protocol, this erasure protocol provides an operational interpretation of strong subadditivity. For positive $H(A|O)$, $H(A|O) \geq H(A|OO')$ means that it is no harder to erase A if the observer has access to both O and O' than if she has access to O alone. For negative $H(A|O)$, $-H(A|OO') \geq -H(A|O)$ means that we can extract at least as much work from AOO' as from its subsystem AO.

To carry out this protocol and extract the optimal amount of work while erasing A, we need to know which subsystem of O^n provides the purification of A_2. The decoupling argument ensures that this subsystem exists, but does not provide a constructive method for finding it, and therefore no concrete protocol for erasing at optimal cost. This quandary is characteristic of Shannon theory; for example, Shannon's noisy channel coding theorem ensures the existence of a code that achieves the channel capacity, but does not provide any explicit code construction.

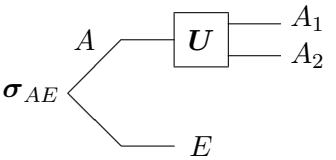
10.9 The Decoupling Inequality

Achievable rates for quantum protocols are derived by using random codes, much as in classical Shannon theory. But this similarity between classical and quantum Shannon theory is superficial — at a deeper conceptual level, quantum protocols differ substantially from classical ones. Indeed, the decoupling principle underlies many of the key findings of

quantum Shannon theory, providing a unifying theme that ties together many different results. In particular, the mother and father resource inequalities, and hence all their descendants enumerated above, follow from an inequality that specifies a sufficient condition for decoupling.

This *decoupling inequality* addresses the following question: Suppose that Alice and Eve share a quantum state $\boldsymbol{\sigma}_{AE}$, where A is an n-qubit system. This state may be mixed, but in general A and E are correlated; that is, $I(A;E) > 0$. Now Alice starts discarding qubits one at a time, where each qubit is a randomly selected two-dimensional subsystem of what Alice holds. Each time Alice discards a qubit, her correlation with E grows weaker. How many qubits should she discard so that the subsystem she retains has a negligible correlation with Eve's system E?

To make the question precise, we need to formalize what it means to discard a random qubit. More generally, suppose that A has dimension $|A|$, and Alice decomposes A into subsystems A_1 and A_2, then discards A_1 and retains A_2. We would like to consider many possible ways of choosing the discarded system with specified dimension $|A_1|$. Equivalently, we may consider a fixed decomposition $A = A_1 A_2$, where we apply a unitary transformation $\boldsymbol{U}$ to A before discarding A_1. Then discarding a random subsystem with dimension $|A_1|$ is the same thing as applying a random unitary $\boldsymbol{U}$ before discarding the fixed subsystem A_1:



To analyze the consequences of discarding a random subsystem, then, we will need to be able to compute the expectation value of a function $f(\boldsymbol{U})$ when we average $\boldsymbol{U}$ uniformly over the group of unitary $|A| \times |A|$ matrices. We denote this expectation value as $\mathbb{E}_{\boldsymbol{U}}[f(\boldsymbol{U})]$; to perform computations we will only need to know that $\mathbb{E}_{\boldsymbol{U}}$ is suitably normalized, and is invariant under left or right multiplication by any constant unitary matrix $\boldsymbol{V}$:

$$\mathbb{E}_{\boldsymbol{U}}[\boldsymbol{I}] = 1, \quad \mathbb{E}_{\boldsymbol{U}}[f(\boldsymbol{U})] = \mathbb{E}_{\boldsymbol{U}}[f(\boldsymbol{V}\boldsymbol{U})] = \mathbb{E}_{\boldsymbol{U}}[f(\boldsymbol{U}\boldsymbol{V})]. \tag{10.318}$$

These conditions uniquely define $\mathbb{E}_{\boldsymbol{U}}[f(\boldsymbol{U})]$, which is sometimes described as the integral over the unitary group using the *invariant measure* or *Haar measure* on the group.

If we apply the unitary transformation $\boldsymbol{U}$ to A, and then discard A_1, the marginal state of A_2E is

$$\boldsymbol{\sigma}_{A_2E}(\boldsymbol{U}) := \mathrm{tr}_{A_1}\left((\boldsymbol{U}_A \otimes \boldsymbol{I}_E)\,\boldsymbol{\sigma}_{AE}\left(\boldsymbol{U}_A^\dagger \otimes \boldsymbol{I}_E\right)\right). \tag{10.319}$$

The decoupling inequality expresses how close (in the L^1 norm) $\boldsymbol{\sigma}_{A_2E}$ is to a product state when we average over $\boldsymbol{U}$:

$$\left(\mathbb{E}_{\boldsymbol{U}}\left[\ \|\boldsymbol{\sigma}_{A_2E}(\boldsymbol{U}) - \boldsymbol{\sigma}_{A_2}^{\max}\otimes\boldsymbol{\sigma}_E\|_1\ \right]\right)^2 \leq \frac{|A_2|\cdot|E|}{|A_1|}\ \mathrm{tr}\left(\boldsymbol{\sigma}_{AE}^2\right), \quad (10.320)$$

where

$$\boldsymbol{\sigma}_{A_2}^{\max} := \frac{1}{|A_2|}\ \boldsymbol{I} \quad (10.321)$$

denotes the maximally mixed state on A_2, and $\boldsymbol{\sigma}_E$ is the marginal state $\mathrm{tr}_A\boldsymbol{\sigma}_{AE}$.

This inequality has interesting consequences even in the case where there is no system E at all and $\boldsymbol{\sigma}_A$ is pure, where it becomes

$$\left(\mathbb{E}_{\boldsymbol{U}}\left[\ \|\boldsymbol{\sigma}_{A_2}(\boldsymbol{U}) - \boldsymbol{\sigma}_{A_2}^{\max}\|_1\ \right]\right)^2 \leq \frac{|A_2|}{|A_1|}\ \mathrm{tr}\left(\boldsymbol{\sigma}_A^2\right) = \frac{|A_2|}{|A_1|}. \quad (10.322)$$

Eq.(10.322) implies that, for a randomly chosen pure state of the bipartite system $A = A_1A_2$, where $|A_2|/|A_1| \ll 1$, the density operator on A_2 is very nearly maximally mixed with high probability. One can likewise show that the expectation value of the entanglement entropy of A_1A_2 is very close to the maximal value: $\mathbb{E}\left[H(A_2)\right] \geq \log_2|A_2| - |A_2|/\left(2|A_1|\ln 2\right)$. Thus, if for example A_2 is 50 qubits and A_1 is 100 qubits, the typical entropy deviates from maximal by only about $2^{-50} \approx 10^{-15}$.

10.9.1 Proof of the decoupling inequality

To prove the decoupling inequality, we will first bound the distance between $\boldsymbol{\sigma}_{A_2E}$ and a product state in the L^2 norm, and then use the Cauchy-Schwarz inequality to obtain a bound on the L^1 distance. Eq.(10.320) follows from

$$\mathbb{E}_{\boldsymbol{U}}\left[\ \|\boldsymbol{\sigma}_{A_2E}(\boldsymbol{U}) - \boldsymbol{\sigma}_{A_2}^{\max}\otimes\boldsymbol{\sigma}_E\|_2^2\ \right] \leq \frac{1}{|A_1|}\ \mathrm{tr}\left(\boldsymbol{\sigma}_{AE}^2\right), \quad (10.323)$$

combined with

$$\left(\mathbb{E}\left[f\right]\right)^2 \leq \mathbb{E}\left[f\right]^2 \quad \text{and} \quad \|M\|_1^2 \leq d\|M\|_2^2 \quad (10.324)$$

(for nonnegative f), which implies

$$\left(\mathbb{E}\left[\|\cdot\|_1\right]\right)^2 \leq \mathbb{E}\left[\|\cdot\|_1^2\right] \leq |A_2|\cdot|E|\cdot\mathbb{E}\left[\|\cdot\|_2^2\right]. \quad (10.325)$$

We also note that

$$\begin{aligned}\|\boldsymbol{\sigma}_{A_2E} - \boldsymbol{\sigma}_{A_2}^{\max}\otimes\boldsymbol{\sigma}_E\|_2^2 &= \mathrm{tr}\left(\boldsymbol{\sigma}_{A_2E} - \boldsymbol{\sigma}_{A_2}^{\max}\otimes\boldsymbol{\sigma}_E\right)^2 \\ &= \mathrm{tr}\left(\boldsymbol{\sigma}_{A_2E}^2\right) - \frac{1}{|A_2|}\mathrm{tr}\left(\boldsymbol{\sigma}_E^2\right), \end{aligned} \quad (10.326)$$

because

$$\mathrm{tr}\left(\boldsymbol{\sigma}_{A_2}^{\max}\right)^2 = \frac{1}{|A_2|}; \tag{10.327}$$

therefore, to prove eq.(10.323) it suffices to show

$$\mathbb{E}_{\boldsymbol{U}}\left[\mathrm{tr}\left(\boldsymbol{\sigma}_{A_2E}^2(\boldsymbol{U})\right)\right] \leq \frac{1}{|A_2|}\mathrm{tr}\left(\boldsymbol{\sigma}_E^2\right) + \frac{1}{|A_1|}\mathrm{tr}\left(\boldsymbol{\sigma}_{AE}^2\right). \tag{10.328}$$

We can facilitate the computation of $\mathbb{E}_{\boldsymbol{U}}\left[\mathrm{tr}\left(\boldsymbol{\sigma}_{A_2E}^2(\boldsymbol{U})\right)\right]$ using a clever trick. For any bipartite system BC, imagine introducing a second copy $B'C'$ of the system. Then (Exercise 10.7)

$$\mathrm{tr}_C\left(\boldsymbol{\sigma}_C^2\right) = \mathrm{tr}_{BCB'C'}\left(\boldsymbol{I}_{BB'} \otimes \boldsymbol{S}_{CC'}\right)\left(\boldsymbol{\sigma}_{BC} \otimes \boldsymbol{\sigma}_{B'C'}\right), \tag{10.329}$$

where $\boldsymbol{S}_{CC'}$ denotes the swap operator, which acts as

$$\boldsymbol{S}_{CC'} : |i\rangle_C \otimes |j\rangle_{C'} \mapsto |j\rangle_C \otimes |i\rangle_{C'}. \tag{10.330}$$

In particular, then,

$$\begin{aligned}
&\mathrm{tr}_{A_2E}\left(\boldsymbol{\sigma}_{A_2E}^2(\boldsymbol{U})\right) \\
&= \mathrm{tr}_{AEA'E'}\left(\boldsymbol{I}_{A_1A_1'} \otimes \boldsymbol{S}_{A_2A_2'} \otimes \boldsymbol{S}_{EE'}\right)\left(\boldsymbol{\sigma}_{AE}(\boldsymbol{U}) \otimes \boldsymbol{\sigma}_{A'E'}(\boldsymbol{U})\right) \\
&= \mathrm{tr}_{AEA'E'}\left(\boldsymbol{M}_{AA'}(\boldsymbol{U}) \otimes \boldsymbol{S}_{EE'}\right)\left(\boldsymbol{\sigma}_{AE} \otimes \boldsymbol{\sigma}_{A'E'}\right),
\end{aligned} \tag{10.331}$$

where

$$\boldsymbol{M}_{AA'}(\boldsymbol{U}) = \left(\boldsymbol{U}_A^\dagger \otimes \boldsymbol{U}_{A'}^\dagger\right)\left(\boldsymbol{I}_{A_1A_1'} \otimes \boldsymbol{S}_{A_2A_2'}\right)\left(\boldsymbol{U}_A \otimes \boldsymbol{U}_{A'}\right). \tag{10.332}$$

The expectation value of $\boldsymbol{M}_{AA'}(\boldsymbol{U})$ is evaluated in Exercise 10.7; there we find

$$\mathbb{E}_{\boldsymbol{U}}\left[\boldsymbol{M}_{AA'}(\boldsymbol{U})\right] = c_{\boldsymbol{I}}\boldsymbol{I}_{AA'} + c_{\boldsymbol{S}}\boldsymbol{S}_{AA'} \tag{10.333}$$

where

$$\begin{aligned}
c_{\boldsymbol{I}} &= \frac{1}{|A_2|}\left(\frac{1-1/|A_1|}{1-1/|A|}\right) \leq \frac{1}{|A_2|}, \\
c_{\boldsymbol{S}} &= \frac{1}{|A_1|}\left(\frac{1-1/|A_2|}{1-1/|A|}\right) \leq \frac{1}{|A_1|}.
\end{aligned} \tag{10.334}$$

Plugging into eq.(10.331), we then obtain

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{U}}\left[\mathrm{tr}_{A_2E}\left(\boldsymbol{\sigma}_{A_2E}^2(\boldsymbol{U})\right)\right] \\
&\leq \mathrm{tr}_{AEA'E'}\left(\left(\frac{1}{|A_2|}\boldsymbol{I}_{AA'} + \frac{1}{|A_1|}\boldsymbol{S}_{AA'}\right) \otimes \boldsymbol{S}_{EE'}\right)\left(\boldsymbol{\sigma}_{AE} \otimes \boldsymbol{\sigma}_{A'E'}\right) \\
&= \frac{1}{|A_2|}\mathrm{tr}\left(\boldsymbol{\sigma}_E^2\right) + \frac{1}{|A_1|}\left(\boldsymbol{\sigma}_{AE}^2\right),
\end{aligned} \tag{10.335}$$

thus proving eq.(10.328) as desired.

10.9.2 Proof of the mother inequality

The mother inequality eq.(10.307) follows from the decoupling inequality eq.(10.320) in an i.i.d. setting. Suppose Alice, Bob, and Eve share the pure state $\phi_{ABE}^{\otimes n}$. Then there are jointly typical subspaces of A^n, B^n, and E^n, which we denote by $\bar{A}$, $\bar{B}$, $\bar{E}$, such that

$$\left|\bar{A}\right| = 2^{nH(A)+o(n)}, \quad \left|\bar{B}\right| = 2^{nH(B)+o(n)}, \quad \left|\bar{E}\right| = 2^{nH(E)+o(n)}. \quad (10.336)$$

Furthermore, the normalized pure state $\phi'_{\bar{A}\bar{B}\bar{E}}$ obtained by projecting $\phi_{ABE}^{\otimes n}$ onto $\bar{A} \otimes \bar{B} \otimes \bar{E}$ deviates from $\phi_{ABE}^{\otimes n}$ by distance $o(1)$ in the L^1 norm.

In order to transfer the purification of E^n to Bob, Alice first projects A^n onto its typical subspace, succeeding with probability $1 - o(1)$, and compresses the result. She then divides her compressed system $\bar{A}$ into two parts $\bar{A}_1\bar{A}_2$, and applies a random unitary to $\bar{A}$ before sending $\bar{A}_1$ to Bob. Quantum state transfer is achieved if $\bar{A}_2$ decouples from $\bar{E}$.

Because $\phi'_{\bar{A}\bar{B}\bar{E}}$ is close to $\phi_{ABE}^{\otimes n}$, we can analyze whether the protocol is successful by supposing the initial state is $\phi'_{\bar{A}\bar{B}\bar{E}}$ rather than $\phi_{ABE}^{\otimes n}$. According to the decoupling inequality

$$\begin{aligned}\left(\mathbb{E}_{\boldsymbol{U}}\left[\ \|\boldsymbol{\sigma}_{\bar{A}_2\bar{E}}(\boldsymbol{U}) - \boldsymbol{\sigma}_{\bar{A}_2}^{\max} \otimes \boldsymbol{\sigma}_{\bar{E}}\|_1\ \right]\right)^2 &\leq \frac{|\bar{A}| \cdot |\bar{E}|}{|\bar{A}_1|^2}\ \mathrm{tr}\left(\boldsymbol{\sigma}_{\bar{A}\bar{E}}^2\right) \\ = \frac{1}{|\bar{A}_1|^2} 2^{n(H(A)+H(E)+o(1))}\ \mathrm{tr}\left(\boldsymbol{\sigma}_{\bar{A}\bar{E}}^2\right) &= \frac{1}{|\bar{A}_1|^2} 2^{n(H(A)+H(E)-H(B)+o(1))};\end{aligned} \quad (10.337)$$

here we have used properties of typical subspaces in the second line, as well as the property that $\boldsymbol{\sigma}_{\bar{A}\bar{E}}$ and $\boldsymbol{\sigma}_{\bar{B}}$ have the same nonzero eigenvalues, because $\phi'_{\bar{A}\bar{B}\bar{E}}$ is pure.

Eq.(10.337) bounds the L^1 distance of $\boldsymbol{\sigma}_{\bar{A}_2\bar{E}}(\boldsymbol{U})$ from a product state when averaged over all unitaries, and therefore suffices to ensure the existence of at least one unitary transformation $\boldsymbol{U}$ such that the L^1 distance is bounded above by the right-hand side. Therefore, by choosing this $\boldsymbol{U}$, Alice can decouple $\bar{A}_2$ from E^n to $o(1)$ accuracy in the L^1 norm by sending to Bob

$$\log_2 |\bar{A}_1| = \frac{n}{2}\left(H(A) + H(E) - H(B) + o(1)\right) = \frac{n}{2}\left(I(A;E) + o(1)\right) \quad (10.338)$$

qubits, randomly chosen from the (compressed) typical subspace of A^n. Alice retains $nH(A) - \frac{n}{2}I(A;E) - o(n)$ qubits of her compressed system, which are nearly maximally mixed and uncorrelated with E^n; hence at the end of the protocol she shares with Bob this many qubit pairs, which

have high fidelity with a maximally entangled state. Since ϕ_{ABE} is pure, and therefore $H(A) = \frac{1}{2}\left(I(A;E) - I(A;B)\right)$, we conclude that Alice and Bob distill $\frac{n}{2}I(A;B) - o(n)$ ebits of entanglement, thus proving the mother resource inequality.

We can check that this conclusion is plausible using a crude counting argument. Disregarding the $o(n)$ corrections in the exponent, the state $\phi_{ABE}^{\otimes n}$ is nearly maximally mixed on a typical subspace of $A^n E^n$ with dimension $2^{nH(AE)}$, *i.e.* the marginal state on $\bar{A}\bar{E}$ can be realized as a nearly uniform ensemble of this many mutually orthogonal states. If $\bar{A}_1$ is randomly chosen and sufficiently small, we expect that, for each state in this ensemble, $\bar{A}_1$ is nearly maximally entangled with a subsystem of the much larger system $\bar{A}_2\bar{E}$, and that the marginal states on $\bar{A}_2\bar{E}$ arising from different states in the $\bar{A}\bar{E}$ ensemble have a small overlap. Therefore, we anticipate that tracing out $\bar{A}_1$ yields a state on $\bar{A}_2\bar{E}$ which is nearly maximally mixed on a subspace with dimension $|\bar{A}_1|2^{nH(AE)}$. Approximate decoupling occurs when this state attains full rank on $\bar{A}_2\bar{E}$, since in that case it is close to maximally mixed on $\bar{A}_2\bar{E}$ and therefore close to a product state on its support. The state transfer succeeds, therefore, provided

$$\begin{aligned} |\bar{A}_1|2^{nH(AE)} \approx |\bar{A}_2| \cdot |\bar{E}| = \frac{|\bar{A}| \cdot |\bar{E}|}{|\bar{A}_1|} \approx \frac{2^{n(H(A)+H(E))}}{|\bar{A}_1|} \\ \implies |\bar{A}_1|^2 \approx 2^{nI(A;E)}, \end{aligned} \tag{10.339}$$

as in eq.(10.338).

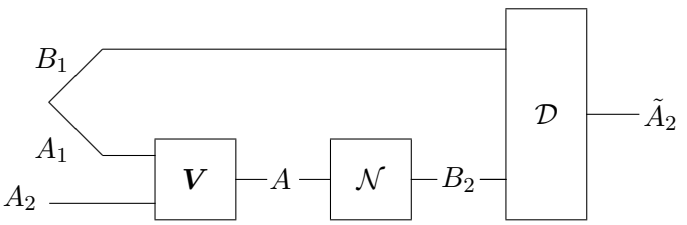
Our derivation of the mother resource inequality, based on random coding, does not exhibit any concrete protocol that achieves the claimed rate, nor does it guarantee the existence of any protocol in which the required quantum processing can be executed efficiently. Concerning the latter point, it is notable that our derivation of the decoupling inequality applies not just to the expectation value averaged uniformly over the unitary group, but also to any average over unitary transformations which satisfies eq.(10.333). In fact, this identity is satisfied by a uniform average over the Clifford group, which means that there is some Clifford transformation on $\bar{A}$ which achieves the rates specified in the mother resource inequality. Any Clifford transformation on n qubits can be reached by a circuit with $O(n^2)$ gates. Since it is also known that Schumacher compression can be achieved by a polynomial-time quantum computation, Alice's encoding operation can be carried out efficiently.

In fact, after compressing, Alice encodes the quantum information she sends to Bob using a stabilizer code (with Clifford encoder $\boldsymbol{U}$), and Bob's task, after receiving $\bar{A}_1$ is to correct the erasure of $\bar{A}_2$. Bob can replace each erased qubit by the standard state $|0\rangle$, and then measure the code's

check operators. With high probability, there is a unique Pauli operator acting on the erased qubits that restores Bob's state to the code space, and the recovery operation can be efficiently computed using linear algebra. Hence, Bob's part of the mother protocol, like Alice's, can be executed efficiently.

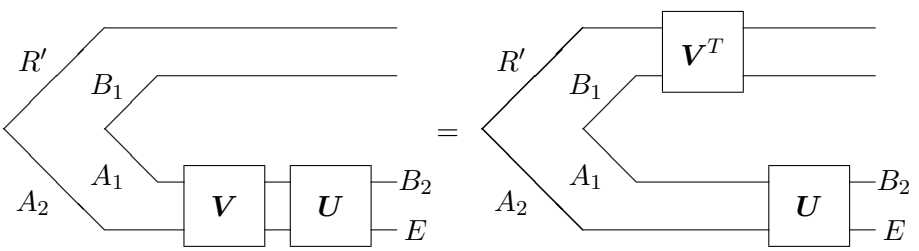
10.9.3 Proof of the father inequality

One-shot version. In the one-shot version of the father protocol, Alice and Bob share a pair of maximally entangled systems A_1B_1, and in addition Alice holds input state $\boldsymbol{\rho}_{A_2}$ of system A_2 which she wants to convey to Bob. Alice encodes $\boldsymbol{\rho}_{A_2}$ by applying a unitary transformation $\boldsymbol{V}$ to $A = A_1A_2$, then sends A to Bob via the noisy quantum channel $\mathcal{N}^{A\to B_2}$. Bob applies a decoding map $\mathcal{D}^{B_1B_2\to\tilde{A}_2}$ jointly to the channel output and his half of the entangled state he shares with Alice, hoping to recover Alice's input state with high fidelity:



We would like to know how much shared entanglement suffices for Alice and Bob to succeed.

This question can be answered using the decoupling inequality. First we introduce a reference system R' which is maximally entangled with A_2; then Bob succeeds if his decoder can extract the purification of R'. Because the systems $R'B_1$ and A_1A_1 are maximally entangled, the encoding unitary $\boldsymbol{V}$ acting on A_1A_2 can be replaced by its transpose $\boldsymbol{V}^T$ acting on $R'B_1$. We may also replace $\mathcal{N}$ by its Stinespring dilation $\boldsymbol{U}^{A_1A_2\to B_2E}$, so that the extended output state ϕ of $R'B_1B_2E$ is pure:



Finally we invoke the decoupling principle — if R' and E decouple, then R' is purified by a subsystem of B_1B_2, which means that Bob can recover $\boldsymbol{\rho}_{A_2}$ with a suitable decoding map.

If we consider $\boldsymbol{V}$, and hence also $\boldsymbol{V}^T$, to be a random unitary, then we may describe the situation this way: We have a tripartite pure state ϕ_{RB_2E}, where $R = R'B_1$, and we would like to know whether the marginal state of $R'E$ is close to a product state when the random subsystem B_1 is discarded from R. This is exactly the question addressed by the decoupling inequality, which in this case may be expressed as

$$\left(\mathbb{E}_{\boldsymbol{V}}\left[\ \|\boldsymbol{\sigma}_{R'E}(\boldsymbol{V}) - \boldsymbol{\sigma}_{R'}^{\max} \otimes \boldsymbol{\sigma}_E\|_1\ \right]\right)^2 \leq \frac{|R| \cdot |E|}{|B_1|^2}\ \mathrm{tr}\left(\boldsymbol{\sigma}_{RE}^2\right), \qquad (10.340)$$

Eq.(10.340) asserts that the L^1 distance from a product state is bounded above when averaged uniformly over all unitary $\boldsymbol{V}$'s; therefore there must be some particular encoding unitary $\boldsymbol{V}$ that satisfies the same bound. We conclude that near-perfect decoupling of $R'E$, and therefore high-fidelity decoding of B_2, is achievable provided that

$$|A_1| = |B_1| \gg |R'| \cdot |E|\ \mathrm{tr}\left(\boldsymbol{\sigma}_{RE}^2\right) = |A_2| \cdot |E|\ \mathrm{tr}\left(\boldsymbol{\sigma}_{B_2}^2\right), \qquad (10.341)$$

where to obtain the second equality we use the purity of ϕ_{RB_2E} and recall that the reference system R' is maximally entangled with A_2.

i.i.d. version. In the i.i.d. version of the father protocol, Alice and Bob achieve high fidelity entanglement-assisted quantum communication through n uses of the quantum channel $\mathcal{N}^{A\to B}$. The code they use for this purpose can be described in the following way: Consider an input density operator $\boldsymbol{\rho}_A$ of system A, which is purified by a reference system R. Sending the purified input state ψ_{RA} through $\boldsymbol{U}^{A\to BE}$, the isometric dilation of $\mathcal{N}^{A\to B}$, generates the tripartite pure state ϕ_{RBE}. Evidently applying $\left(\boldsymbol{U}^{A\to BE}\right)^{\otimes n}$ to $\psi_{RA}^{\otimes n}$ produces $\phi_{RBE}^{\otimes n}$.

But now suppose that before transmitting the state to Bob, Alice projects A^n onto its typical subspace $\bar{A}$, succeeding with probability $1 - o(1)$ in preparing a state of $\bar{A}\bar{R}$ that is nearly maximally entangled, where $\bar{R}$ is the typical subspace of R^n. Imagine dividing $\bar{R}$ into a randomly chosen subsystem B_1 and its complementary subsystem R'; then there is a corresponding decomposition of $A = A_1A_2$ such that A_1 is very nearly maximally entangled with B_1 and A_2 is very nearly maximally entangled with R'.

If we interpret B_1 as Bob's half of an entangled state of A_1B_1 shared with Alice, this becomes the setting where the one-shot father protocol applies, if we ignore the small deviation from maximal entanglement in A_1B_1 and $R'A_2$. As for our analysis of the i.i.d. mother protocol, we apply

the one-shot father inequality not to $\phi_{RBE}^{\otimes n}$, but rather to the nearby state $\phi'_{\bar{R}\bar{B}\bar{E}}$, where $\bar{B}$ and $\bar{E}$ are the typical subspaces of B^n and E^n respectively. Applying eq.(10.340), and using properties of typical subspaces, we can bound the square of the L^1 deviation of $R'E$ from a product state, averaged over the choice of B_1, by

$$\frac{|\bar{R}| \cdot |\bar{E}|}{|B_1|^2} \operatorname{tr}\left(\boldsymbol{\sigma}_{\bar{B}}^2\right) = \frac{2^{n(H(R)+H(E)-H(B)+o(1))}}{|B_1|^2} = \frac{2^{n(I(R;E)+o(1))}}{|B_1|^2}; \quad (10.342)$$

hence the bound also applies for some particular way of choosing B_1. This choice defines the code used by Alice and Bob in a protocol which consumes

$$\log_2 |B_1| = \frac{n}{2} I(R;E) + o(n) \quad (10.343)$$

ebits of entanglement, and conveys from Alice to Bob

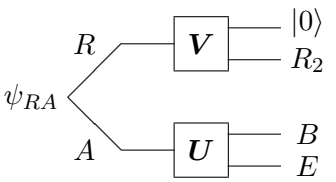
$$nH(B) - \frac{n}{2} I(R;E) - o(n) = \frac{n}{2} I(R;B) - o(n) \quad (10.344)$$

high-fidelity qubits. This proves the father resource inequality.

10.9.4 Quantum channel capacity revisited

In §10.8.1 we showed that the coherent information is an achievable rate for quantum communication over a noisy quantum channel. That derivation, a corollary of the father resource inequality, applied to a catalytic setting, in which shared entanglement between sender and receiver can be borrowed and later repaid. It is useful to see that the same rate is achievable without catalysis, a result we can derive from an alternative version of the decoupling inequality.

This version applies to the setting depicted here:



A density operator $\boldsymbol{\rho}_A$ for system A, with purification ψ_{RA}, is transmitted through a channel $\mathcal{N}^{A\to B}$ which has the isometric dilation $\boldsymbol{U}^{A\to BE}$. The reference system R has a decomposition into subsystems R_1R_2. We apply a random unitary transformation $\boldsymbol{V}$ to R, then project R_1 onto a fixed vector $|0\rangle_{R_1}$, and renormalize the resulting state. In effect, then we are projecting R onto a subspace with dimension $|R_2|$, which purifies a

corresponding code subspace of A. This procedure prepares a normalized pure state ϕ_{R_2BE}, and a corresponding normalized marginal state σ_{R_2E} of R_2E.

If R_2 decouples from E, then R_2 is purified by a subsystem of B, which means that the code subspace of A can be recovered by a decoder applied to B. A sufficient condition for approximate decoupling can be derived from the inequality

$$\left(\mathbb{E}_{\boldsymbol{V}}\left[\ \|\boldsymbol{\sigma}_{R_2E}(\boldsymbol{V}) - \boldsymbol{\sigma}_{R_2}^{\max} \otimes \boldsymbol{\sigma}_E\|_1\ \right]\right)^2 \leq |R_2| \cdot |E|\ \mathrm{tr}\left(\boldsymbol{\sigma}_{RE}^2\right). \quad (10.345)$$

Eq.(10.345) resembles eq.(10.320) and can be derived by a similar method. Note that the right-hand side of eq.(10.345) is enhanced by a factor of $|R_1|$ relative to the right-hand side of eq.(10.320). This factor arises because after projecting R_1 onto the fixed state $|0\rangle$ we need to renormalize the state by multiplying by $|R_1|$, while on the other hand the projection suppresses the expected distance squared from a product state by a factor $|R_1|$.

In the i.i.d. setting where the noisy channel is used n times, we consider $\phi_{RBE}^{\otimes n}$, and project onto the jointly typical subspaces $\bar{R}$, $\bar{B}$, $\bar{E}$ of R^n, B^n, E^n respectively, succeeding with high probability. We choose a code by projecting $\bar{R}$ onto a random subspace with dimension $|R_2|$. Then, the right-hand side of eq.(10.345) becomes

$$|R_2| \cdot 2^{n(H(E)-H(B)+o(1))}, \quad (10.346)$$

and since the inequality holds when we average uniformly over $\boldsymbol{V}$, it surely holds for some particular $\boldsymbol{V}$. That unitary defines a code which achieves decoupling and has the rate

$$\frac{1}{n}\log_2|R_2| = H(E) - H(B) - o(1) = I_c(R\rangle B) - o(1). \quad (10.347)$$

Hence the coherent information is an achievable rate for high-fidelity quantum communication over the noisy channel.

10.9.5 Black holes as mirrors

As our final application of the decoupling inequality, we consider a highly idealized model of black hole dynamics. Suppose that Alice holds a k-qubit system A which she wants to conceal from Bob. To be safe, she discards her qubits by tossing them into a large black hole, where she knows Bob will not dare to follow. The black hole B is an $(n-k)$-qubit system, which grows to n qubits after merging with A, where n is much larger than k.

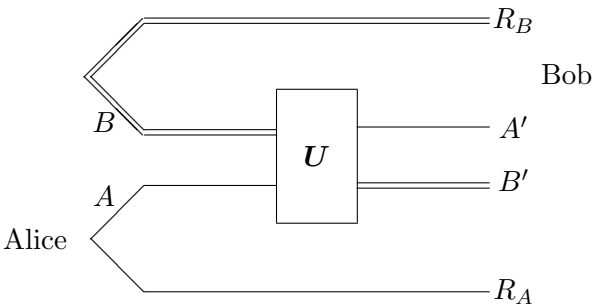
Black holes are not really completely black — they emit Hawking radiation. But qubits leak out of an evaporating black hole very slowly, at a rate per unit time which scales like $n^{-1/2}$. Correspondingly, it takes time $\Theta(n^{3/2})$ for the black hole to radiate away a significant fraction of its qubits. Because the black hole Hilbert space is so enormous, this is a very long time, about 10^{67} years for a solar mass black hole, for which $n \approx 10^{78}$. Though Alice's qubits might not remain secret forever, she is content knowing that they will be safe from Bob for 10^{67} years.

But in her haste, Alice fails to notice that her black hole is very, very old. It has been evaporating for so long that it has already radiated away more than half of its qubits. Let's assume that the joint state of the black hole and its emitted radiation is pure, and furthermore that the radiation is a Haar-random subsystem of the full system.

Because the black hole B is so old, $|B|$ is much smaller than the dimension of the radiation subsystem; therefore, as in eq.(10.322), we expect the state of B to be very nearly maximally mixed with high probability. We denote by R_B the subsystem of the emitted radiation which purifies B; thus the state of BR_B is very nearly maximally entangled. We assume that R_B has been collected by Bob and is under his control.

To keep track of what happens to Alice's k qubits, we suppose that her k-qubit system A is maximally entangled with a reference system R_A. After A enters the black hole, Bob waits for a while, until the k'-qubit system A' is emitted in the black hole's Hawking radiation. After retrieving A', Bob hopes to recover the purification of R_A by applying a suitable decoding map to $A'R_B$. Can he succeed?

We've learned that Bob can succeed with high fidelity if the remaining black hole system B' decouples from Alice's reference system R_A. Let's suppose that the qubits emitted in the Hawking radiation are chosen randomly; that is, A' is a Haar-random k'-qubit subsystem of the n-qubit system AB, as depicted here:



The double lines indicate the very large systems B and B', and single lines the smaller systems A and A'. Because the radiated qubits are

random, we can determine whether $R_A B'$ decouples using the decoupling inequality, which for this case becomes

$$\mathbb{E}_{\boldsymbol{U}}\left[\ \|\boldsymbol{\sigma}_{B'R_A}(\boldsymbol{U}) - \boldsymbol{\sigma}_{B'}^{\max} \otimes \boldsymbol{\sigma}_{R_A}\|_1\ \right] \leq \sqrt{\frac{|ABR_A|}{|A'|^2}\ \mathrm{tr}\left(\boldsymbol{\sigma}_{ABR_A}^2\right)}. \tag{10.348}$$

Because the state of AR_A is pure, and B is maximally entangled with R_B, we have $\mathrm{tr}\left(\boldsymbol{\sigma}_{ABR_A}^2\right) = 1/|B|$, and therefore the Haar-averaged L^1 distance of $\boldsymbol{\sigma}_{B'R_A}$ from a product state is bounded above by

$$\sqrt{\frac{|AR_A|}{|A'|^2}} = \frac{|A|}{|A'|}. \tag{10.349}$$

Thus, if Bob waits for only $k' = k + c$ qubits of Hawking radiation to be emitted after Alice tosses in her k qubits, Bob can decode her qubits with excellent fidelity $F \geq 1 - 2^{-c}$.

Alice made a serious mistake. Rather than waiting for $\Omega(n)$ qubits to emerge from the black hole, Bob can already decode Alice's secret quite well when he has collected just a few more than k qubits. And Bob is an excellent physicist, who knows enough about black hole dynamics to infer the encoding unitary transformation $\boldsymbol{U}$, information he uses to find the right decoding map.

We could describe the conclusion, more prosaically, by saying that the random unitary $\boldsymbol{U}$ applied to AB encodes a good quantum error-correcting code, which achieves high-fidelity entanglement-assisted transmission of quantum information though an erasure channel with a high erasure probability. Of the n input qubits, only k' randomly selected qubits are received by Bob; the rest remain inside the black hole and hence are inaccessible. The input qubits, then, are erased with probability $p = (n - k')/n$, while nearly error-free qubits are recovered from the input qubits at a rate

$$R = \frac{k}{n} = 1 - p - \frac{k' - k}{n}; \tag{10.350}$$

in the limit $n \to \infty$ with $c = k' - k$ fixed, this rate approaches $1 - p$, the entanglement-assisted quantum capacity of the erasure channel.

So far, we've assumed that the emitted system A' is a randomly selected subsystem of AB. That won't be true for a real black hole. However, it is believed that the internal dynamics of actually black holes mixes quantum information quite rapidly (the *fast scrambling conjecture*). For a black hole with temperature T, it takes time of order $\hbar/kT$ for each qubit to be

emitted in the Hawking radiation, and a time longer by only a factor of $\log n$ for the dynamics to mix the black hole degrees of freedom sufficiently for our decoupling estimate to hold with reasonable accuracy. For a solar mass black hole, Alice's qubits are revealed just a few milliseconds after she deposits them, much faster than the 10^{67} years she had hoped for! Because Bob holds the system R_B which purifies B, and because he knows the right decoding map to apply to $A'R_B$, the black hole behaves like an information mirror — Alice's qubits bounce right back!

If Alice is more careful, she will dump her qubits into a young black hole instead. If we assume that the initial black hole B is in a pure state, then σ_{ABR_A} is also pure, and the Haar-averaged L^1 distance of $\boldsymbol{\sigma}_{B'R_A}$ from a product state is bounded above by

$$\sqrt{\frac{|ABR_A|}{|A'|^2}} = \frac{2^{n+k}}{2^{2k'}} = \frac{1}{2^c} \tag{10.351}$$

after

$$k' = \frac{1}{2}(n + k + c) \tag{10.352}$$

qubits are emitted. In this case, Bob needs to wait a long time, until more than half of the qubits in AB are radiated away. Once Bob has acquired $k + c$ more qubits than the number still residing in the black hole, he is empowered to decode Alice's k qubits with fidelity $F \geq 1 - 2^{-c}$. In fact, there is nothing special about Alice's subsystem A; by adjusting his decoding map appropriately, Bob can decode any k qubits he chooses from among the n qubits in the initial black hole AB.

There is far more to learn about quantum information processing by black holes, an active topic of current research (as of this writing in 2016), but we will not delve further into this fascinating topic here. We can be confident, though, that the tools and concepts of quantum information theory discussed in this book will be helpful for addressing the many unresolved mysteries of quantum gravity.

10.10 Summary

Shannon entropy and classical data compression. The *Shannon entropy* of an ensemble $X = \{x, p(x)\}$ is $H(X) \equiv \langle -\log p(x) \rangle$; it quantifies the compressibility of classical information. A message n letters long, where each letter is drawn independently from X, can be compressed to $H(X)$ bits per letter (and no further), yet can still be decoded with arbitrarily good accuracy as $n \to \infty$.

Conditional entropy and information merging. The *conditional entropy* $H(X|Y) = H(XY) - H(Y)$ quantifies how much the information

source X can be compressed when Y is known. If n letters are drawn from XY, where Alice holds X and Bob holds Y, Alice can convey X to Bob by sending $H(X|Y)$ bits per letter, asymptotically as $n \to \infty$.

Mutual information and classical channel capacity. The *mutual information* $I(X;Y) = H(X) + H(Y) - H(XY)$ quantifies how information sources X and Y are correlated; when we learn the value of y we acquire (on the average) $I(X;Y)$ bits of information about x, and vice versa. The capacity of a memoryless noisy classical communication channel is $C = \max_X I(X;Y)$. This is the highest number of bits per letter that can be transmitted through n uses of the channel, using the best possible code, with negligible error probability as $n \to \infty$.

Von Neumann entropy and quantum data compression. The *Von Neumann entropy* of a density operator $\boldsymbol{\rho}$ is

$$H(\boldsymbol{\rho}) = -\mathrm{tr}\boldsymbol{\rho}\log\boldsymbol{\rho}; \tag{10.353}$$

it quantifies the compressibility of an ensemble of pure quantum states. A message n letters long, where each letter is drawn independently from the ensemble $\{|\varphi(x)\rangle, p(x)\}$, can be compressed to $H(\boldsymbol{\rho})$ qubits per letter (and no further) where $\boldsymbol{\rho} = \sum_X p(x)|\varphi(x)\rangle\langle\varphi(x)|$, yet can still be decoded with arbitrarily good fidelity as $n \to \infty$.

Entanglement concentration and dilution. The *entanglement* E of a bipartite pure state $|\psi\rangle_{AB}$ is $E = H(\boldsymbol{\rho}_A)$ where $\boldsymbol{\rho}_A = \mathrm{tr}_B(|\psi\rangle\langle\psi|)$. With local operations and classical communication, we can prepare n copies of $|\psi\rangle_{AB}$ from nE Bell pairs (but not from fewer), and we can distill nE Bell pairs (but not more) from n copies of $|\psi\rangle_{AB}$, asymptotically as $n \to \infty$.

Accessible information. The *Holevo chi* of an ensemble $\mathcal{E} = \{\boldsymbol{\rho}(x), p(x)\}$ of quantum states is

$$\chi(\mathcal{E}) = H\left(\sum_x p(x)\boldsymbol{\rho}(x)\right) - \sum_x p(x)H(\boldsymbol{\rho}(x)). \tag{10.354}$$

The *accessible information* of an ensemble $\mathcal{E}$ of quantum states is the maximal number of bits of information that can be acquired about the preparation of the state (on the average) with the best possible measurement. The accessible information cannot exceed the Holevo chi of the ensemble. The product-state capacity of a quantum channel $\mathcal{N}$ is

$$C_1(\mathcal{N}) = \max_{\mathcal{E}} \chi(\mathcal{N}(\mathcal{E})). \tag{10.355}$$

This is the highest number of classical bits per letter that can be transmitted through n uses of the quantum channel, with negligible error probability as $n \to \infty$, assuming that each codeword is a product state.

Decoupling and quantum communication. In a tripartite pure state ϕ_{RBE}, we say that systems R and E *decouple* if the marginal density operator of RE is a product state, in which case R is purified by a subsystem of B. A quantum state transmitted through a noisy quantum channel $\mathcal{N}^{A\to B}$ (with isometric dilation $\boldsymbol{U}^{A\to BE}$) can be accurately decoded if a reference system R which purifies channel's input A nearly *decouples* from the channel's environment E.

Father and mother protocols. The *father and mother resource inequalities* specify achievable rates for entanglement-assisted quantum communication and quantum state transfer, respectively. Both follow from the *decoupling inequality*, which establishes a sufficient condition for approximate decoupling in a tripartite mixed state. By combining the father and mother protocols with superdense coding and teleportation, we can derive achievable rates for other protocols, including entanglement-assisted classical communication, quantum communication, entanglement distillation, and quantum state merging.

Homage to Ben Schumacher:

Ben.
He rocks.
I remember
When
He showed me how to fit
A qubit
In a small box.

I wonder how it feels
To be compressed.
And then to pass
A fidelity test.

Or does it feel
At all, and if it does
Would I squeal
Or be just as I was?

If not undone
I'd become as I'd begun
And write a memorandum
On being random.
Had it felt like a belt
Of rum?

And might it be predicted
That I'd become addicted,

Longing for my session
Of compression?

I'd crawl
To Ben again.
And call,
Put down your pen!
Don't stall!
Make me small!

10.11 Bibliographical Notes

Cover and Thomas [2] is an excellent textbook on classical information theory. Shannon's original paper [3] is still very much worth reading.

Nielsen and Chuang [4] provide a clear introduction to some aspects of quantum Shannon theory. Wilde [1] is a more up-to-date and very thorough account.

Properties of entropy are reviewed in [5]. Strong subadditivity of Von Neumann entropy was proven by Lieb and Ruskai [6], and the condition for equality was derived by Hayden *et al.* [7]. The connection between separability and majorization was pointed out by Nielsen and Kempe [8].

Bekenstein's entropy bound was formulated in [9] and derived by Casini [10]. Entropic uncertainty relations are reviewed in [11], and I follow their derivation. The original derivation, by Maassen and Uffink [12] uses different methods.

Schumacher compression was first discussed in [13, 14], and Bennett *et al.* [15] devised protocols for entanglement concentration and dilution. Measures of mixed-state entanglement are reviewed in [16]. The reversible theory of mixed-state entanglement was formulated by Brandão and Plenio [17]. Squashed entanglement was introduced by Christandl and Winter [18], and its monogamy discussed by Koashi and Winter [19]. Brandão, Christandl, and Yard [20] showed that squashed entanglement is positive for any nonseparable bipartite state. Doherty, Parrilo, and Spedalieri [21] showed that every nonseparable bipartite state fails to be k-extendable for some finite k.

The Holevo bound was derived in [22]. Peres-Wootters coding was discussed in [23]. The product-state capacity formula was derived by Holevo [24] and by Schumacher and Westmoreland [25]. Hastings [26] showed that Holevo chi can be superadditive. Horodecki, Shor, and Ruskai [27] introduced entanglement-breaking channels, and additivity of Holevo chi for these channels was shown by Shor [28].

Necessary and sufficient conditions for quantum error correction were formulated in terms of the decoupling principle by Schumacher and

Nielsen [29]; that (regularized) coherent information is an upper bound on quantum capacity was shown by Schumacher [30], Schumacher and Nielsen [29], and Barnum *et al.* [31]. That coherent information is an achievable rate for quantum communication was conjectured by Lloyd [32] and by Schumacher [30], then proven by Shor [33] and by Devetak [34]. Devetak and Winter [35] showed it is also an achievable rate for entanglement distillation. The quantum Fano inequality was derived by Schumacher [30].

Approximate decoupling was analyzed by Schumacher and Westmoreland [36], and used to prove capacity theorems by Devetak [34], by Horodecki *et al.* [37], by Hayden *et al.* [38], and by Abeyesinghe *et al.* [39]. The entropy of Haar-random subsystems had been discussed earlier, by Lubkin [40], Lloyd and Pagels [41], and Page [42]. Devetak, Harrow, and Winter [43, 44] introduced the mother and father protocols and their descendants. Devatak and Shor [45] introduced degradable quantum channels and proved that coherent information is additive for these channels. Bennett *et al.* [46, 47] found the single-letter formula for entanglement-assisted classical capacity. Superadditivity of coherent information was discovered by Shor and Smolin [48] and by DiVincenzo *et al.* [49]. Smith and Yard [50] found extreme examples of superadditivity, in which two zero-capacity channels have nonzero capacity when used jointly. The achievable rate for state merging was derived by Horodecki *et al.* [37], and used by them to prove strong subadditivity of Von Neumann entropy.

Decoupling was applied to Landuaer's principle by Renner *et al.* [51], and to black holes by Hayden and Preskill [52]. The fast scrambling conjecture was proposed by Sekino and Susskind [53].

10.12 Exercises

10.1 Positivity of quantum relative entropy

a) Show that $\ln x \leq x - 1$ for all positive real x, with equality iff $x = 1$.

b) The (classical) relative entropy of a probability distribution $\{p(x)\}$ relative to $\{q(x)\}$ is defined as

$$D(p \parallel q) \equiv \sum_x p(x) \left(\log p(x) - \log q(x)\right) \ . \tag{10.356}$$

Show that

$$D(p \parallel q) \geq 0 \ , \tag{10.357}$$

with equality iff the probability distributions are identical. **Hint**: Apply the inequality from (*a*) to $\ln\left(q(x)/p(x)\right)$.

c) The quantum relative entropy of the density operator $\boldsymbol{\rho}$ with respect to $\boldsymbol{\sigma}$ is defined as

$$D(\boldsymbol{\rho} \parallel \boldsymbol{\sigma}) = \mathrm{tr}\ \boldsymbol{\rho}\,(\log \boldsymbol{\rho} - \log \boldsymbol{\sigma}) \ . \tag{10.358}$$

Let $\{p_i\}$ denote the eigenvalues of $\boldsymbol{\rho}$ and $\{q_a\}$ denote the eigenvalues of $\boldsymbol{\sigma}$. Show that

$$D(\boldsymbol{\rho} \parallel \boldsymbol{\sigma}) = \sum_i p_i \left(\log p_i - \sum_a D_{ia} \log q_a \right) \ , \tag{10.359}$$

where D_{ia} is a doubly stochastic matrix. Express D_{ia} in terms of the eigenstates of $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$. (A matrix is doubly stochastic if its entries are nonnegative real numbers, where each row and each column sums to one.)

d) Show that if D_{ia} is doubly stochastic, then (for each i)

$$\log \left(\sum_a D_{ia} q_a \right) \geq \sum_a D_{ia} \log q_a \ , \tag{10.360}$$

with equality only if $D_{ia} = 1$ for some a.

e) Show that

$$D(\boldsymbol{\rho} \parallel \boldsymbol{\sigma}) \geq D(p \parallel r) \ , \tag{10.361}$$

where $r_i = \sum_a D_{ia} q_a$.

f) Show that $D(\boldsymbol{\rho} \parallel \boldsymbol{\sigma}) \geq 0$, with equality iff $\boldsymbol{\rho} = \boldsymbol{\sigma}$.

10.2 Properties of Von Neumann entropy

a) Use nonnegativity of quantum relative entropy to prove the *subadditivity* of Von Neumann entropy

$$H(\boldsymbol{\rho}_{AB}) \leq H(\boldsymbol{\rho}_A) + H(\boldsymbol{\rho}_B), \tag{10.362}$$

with equality iff $\boldsymbol{\rho}_{AB} = \boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B$. **Hint**: Consider the relative entropy of $\boldsymbol{\rho}_{AB}$ and $\boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B$.

b) Use subadditivity to prove the concavity of the Von Neumann entropy:

$$H(\sum_x p_x \boldsymbol{\rho}_x) \geq \sum_x p_x H(\boldsymbol{\rho}_x) \ . \tag{10.363}$$

Hint: Consider

$$\boldsymbol{\rho}_{AB} = \sum_x p_x \left(\boldsymbol{\rho}_x\right)_A \otimes \left(|x\rangle\langle x|\right)_B \ , \tag{10.364}$$

where the states $\{|x\rangle_B\}$ are mutually orthogonal.

c) Use the condition

$$H(\boldsymbol{\rho}_{AB}) = H(\boldsymbol{\rho}_A) + H(\boldsymbol{\rho}_B) \quad \text{iff} \quad \boldsymbol{\rho}_{AB} = \boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B \tag{10.365}$$

to show that, if all p_x's are nonzero,

$$H\left(\sum_x p_x \boldsymbol{\rho}_x\right) = \sum_x p_x H(\boldsymbol{\rho}_x) \tag{10.366}$$

iff all the $\boldsymbol{\rho}_x$'s are identical.

10.3 Monotonicity of quantum relative entropy

Quantum relative entropy has a property called *monotonicity*:

$$D(\boldsymbol{\rho}_A \| \boldsymbol{\sigma}_A) \leq D(\boldsymbol{\rho}_{AB} \| \boldsymbol{\sigma}_{AB}); \tag{10.367}$$

The relative entropy of two density operators on a system AB cannot be less than the induced relative entropy on the subsystem A.

a) Use monotonicity of quantum relative entropy to prove the strong subadditivity property of Von Neumann entropy. **Hint**: On a tripartite system ABC, consider the relative entropy of $\boldsymbol{\rho}_{ABC}$ and $\boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_{BC}$.

b) Use monotonicity of quantum relative entropy to show that the action of a quantum channel $\mathcal{N}$ cannot increase relative entropy:

$$D(\mathcal{N}(\boldsymbol{\rho}) \| \mathcal{N}(\boldsymbol{\sigma}) \leq D(\boldsymbol{\rho} \| \boldsymbol{\sigma}), \tag{10.368}$$

Hint: Recall that any quantum channel has an isometric dilation.

10.4 The Peres–Wootters POVM.

Consider the Peres–Wootters information source described in §10.6.4 of the lecture notes. It prepares one of the three states

$$|\Phi_a\rangle = |\varphi_a\rangle \otimes |\varphi_a\rangle, \quad a = 1, 2, 3, \tag{10.369}$$

each occurring with *a priori* probability $\frac{1}{3}$, where the $|\varphi_a\rangle$'s are defined in eq.(10.214).

a) Express the density matrix

$$\boldsymbol{\rho} = \frac{1}{3}\left(\sum_a |\Phi_a\rangle\langle\Phi_a|\right), \tag{10.370}$$

in terms of the Bell basis of maximally entangled states $\{|\phi^{\pm}\rangle, |\psi^{\pm}\rangle\}$, and compute $H(\boldsymbol{\rho})$.

b) For the three vectors $|\Phi_a\rangle, a = 1, 2, 3$, construct the "pretty good measurement" defined in eq.(10.227). (Again, expand the $|\Phi_a\rangle$'s in the Bell basis.) In this case, the PGM is an orthogonal measurement. Express the elements of the PGM basis in terms of the Bell basis.

c) Compute the mutual information of the PGM outcome and the preparation.

10.5 Separability and majorization

The hallmark of entanglement is that in an entangled state the whole is less random than its parts. But in a separable state the correlations are essentially classical and so are expected to adhere to the classical principle that the parts are less disordered than the whole. The objective of this problem is to make this expectation precise by showing that if the bipartite (mixed) state $\boldsymbol{\rho}_{AB}$ is separable, then

$$\lambda(\boldsymbol{\rho}_{AB}) \prec \lambda(\boldsymbol{\rho}_A) \ , \quad \lambda(\boldsymbol{\rho}_{AB}) \prec \lambda(\boldsymbol{\rho}_B) \ . \tag{10.371}$$

Here $\lambda(\boldsymbol{\rho})$ denotes the vector of eigenvalues of $\boldsymbol{\rho}$, and $\prec$ denotes majorization.

A separable state can be realized as an ensemble of pure product states, so that if $\boldsymbol{\rho}_{AB}$ is separable, it may be expressed as

$$\boldsymbol{\rho}_{AB} = \sum_a p_a \ |\psi_a\rangle\langle\psi_a| \otimes |\varphi_a\rangle\langle\varphi_a| \ . \tag{10.372}$$

We can also diagonalize $\boldsymbol{\rho}_{AB}$, expressing it as

$$\boldsymbol{\rho}_{AB} = \sum_j r_j |e_j\rangle\langle e_j| \ , \tag{10.373}$$

where $\{|e_j\rangle\}$ denotes an orthonormal basis for AB; then by the HJW theorem, there is a unitary matrix V such that

$$\sqrt{r_j}|e_j\rangle = \sum_a V_{ja}\sqrt{p_a}|\psi_a\rangle \otimes |\varphi_a\rangle \ . \tag{10.374}$$

Also note that $\boldsymbol{\rho}_A$ can be diagonalized, so that

$$\boldsymbol{\rho}_A = \sum_a p_a |\psi_a\rangle\langle\psi_a| = \sum_\mu s_\mu |f_\mu\rangle\langle f_\mu| \ ; \tag{10.375}$$

here $\{|f_\mu\rangle\}$ denotes an orthonormal basis for A, and by the HJW theorem, there is a unitary matrix U such that

$$\sqrt{p_a}|\psi_a\rangle = \sum_\mu U_{a\mu}\sqrt{s_\mu}|f_\mu\rangle \ . \tag{10.376}$$

Now show that there is a doubly stochastic matrix D such that

$$r_j = \sum_\mu D_{j\mu} s_\mu \ . \tag{10.377}$$

That is, you must check that the entries of $D_{j\mu}$ are real and nonnegative, and that $\sum_j D_{j\mu} = 1 = \sum_\mu D_{j\mu}$. Thus we conclude that $\lambda(\boldsymbol{\rho}_{AB}) \prec \lambda(\boldsymbol{\rho}_A)$. Just by interchanging A and B, the same argument also shows that $\lambda(\boldsymbol{\rho}_{AB}) \prec \lambda(\boldsymbol{\rho}_B)$.

Remark: Note that it follows from the Schur concavity of Shannon entropy that, if $\boldsymbol{\rho}_{AB}$ is separable, then the von Neumann entropy has the properties $H(AB) \geq H(A)$ and $H(AB) \geq H(B)$. Thus, for separable states, conditional entropy is nonnegative: $H(A|B) = H(AB) - H(B) \geq 0$ and $H(B|A) = H(AB) - H(A) \geq 0$. In contrast, if $H(A|B)$ is negative, then according to the hashing inequality the state of AB has positive distillable entanglement $-H(A|B)$, and therefore is surely not separable.

10.6 Additivity of squashed entanglement

Suppose that Alice holds systems A, A' and Bob holds systems B, B'. How is the entanglement of AA' with BB' related to the entanglement of A with B and A' with B'? In this problem we will show that the squashed entanglement is superadditive,

$$E_{\rm sq}(\boldsymbol{\rho}_{ABA'B'}) \geq E_{\rm sq}(\boldsymbol{\rho}_{AB}) + E_{\rm sq}(\boldsymbol{\rho}_{A'B'}) \tag{10.378}$$

and is strictly additive for a tensor product,

$$E_{\rm sq}(\boldsymbol{\rho}_{AB} \otimes \boldsymbol{\rho}_{A'B'}) = E_{\rm sq}(\boldsymbol{\rho}_{AB}) + E_{\rm sq}(\boldsymbol{\rho}_{A'B'}). \tag{10.379}$$

a) Use the chain rule for mutual information eq.(10.196) and eq.(10.197) and the nonnegativity of quantum conditional mutual information to show that

$$I(AA'; BB'|C) \geq I(A; B|C) + I(A'; B'|AC), \tag{10.380}$$

and show that eq.(10.378) follows.

b) Show that for any extension $\boldsymbol{\rho}_{ABC} \otimes \boldsymbol{\rho}_{A'B'C'}$ of the product state $\boldsymbol{\rho}_{AB} \otimes \boldsymbol{\rho}_{A'B'}$, we have

$$I(AA'; BB'|CC') \leq I(A; B|C) + I(A'; B'|C'). \tag{10.381}$$

Conclude that

$$E_{\rm sq}(\boldsymbol{\rho}_{AB} \otimes \boldsymbol{\rho}_{A'B'}) \leq E_{\rm sq}(\boldsymbol{\rho}_{AB}) + E_{\rm sq}(\boldsymbol{\rho}_{A'B'}), \tag{10.382}$$

which, when combined with eq.(10.378), implies eq.(10.379).

10.7 Proof of the decoupling inequality

In this problem we complete the derivation of the decoupling inequality sketched in §10.9.1.

a) Verify eq.(10.329).

To derive the expression for $\mathbb{E}_{\boldsymbol{U}}\left[\boldsymbol{M}_{AA'}(\boldsymbol{U})\right]$ in eq.(10.333), we first note that the invariance property eq.(10.318) implies that $\mathbb{E}_{\boldsymbol{U}}\left[\boldsymbol{M}_{AA'}(\boldsymbol{U})\right]$ commutes with $\boldsymbol{V}\otimes\boldsymbol{V}$ for any unitary $\boldsymbol{V}$. Therefore, by Schur's lemma, $\mathbb{E}_{\boldsymbol{U}}\left[\boldsymbol{M}_{AA'}(\boldsymbol{U})\right]$ is a weighted sum of projections onto irreducible representations of the unitary group. The tensor product of two fundamental representations of $\boldsymbol{U}(d)$ contains two irreducible representations — the symmetric and antisymmetric tensor representations. Therefore we may write

$$\mathbb{E}_{\boldsymbol{U}}\left[\boldsymbol{M}_{AA'}(\boldsymbol{U})\right] = c_{\mathrm{sym}}\,\boldsymbol{\Pi}^{(\mathrm{sym})}_{AA'} + c_{\mathrm{anti}}\,\boldsymbol{\Pi}^{(\mathrm{anti})}_{AA'}; \tag{10.383}$$

here $\boldsymbol{\Pi}^{(\mathrm{sym})}_{AA'}$ is the orthogonal projector onto the subspace of AA' symmetric under the interchange of A and A', $\boldsymbol{\Pi}^{(\mathrm{anti})}_{AA'}$ is the projector onto the antisymmetric subspace, and c_{sym}, c_{anti} are suitable constants. Note that

$$\begin{aligned}\boldsymbol{\Pi}^{(\mathrm{sym})}_{AA'} &= \frac{1}{2}\left(\boldsymbol{I}_{AA'} + \boldsymbol{S}_{AA'}\right),\\ \boldsymbol{\Pi}^{(\mathrm{anti})}_{AA'} &= \frac{1}{2}\left(\boldsymbol{I}_{AA'} - \boldsymbol{S}_{AA'}\right),\end{aligned} \tag{10.384}$$

where $\boldsymbol{S}_{AA'}$ is the swap operator, and that the symmetric and antisymmetric subspaces have dimension $\frac{1}{2}|A|\left(|A|+1\right)$ and dimension $\frac{1}{2}|A|\left(|A|-1\right)$ respectively.

Even if you are not familiar with group representation theory, you might regard eq.(10.383) as obvious. We may write $\boldsymbol{M}_{AA'}(\boldsymbol{U})$ as a sum of two terms, one symmetric and the other antisymmetric under the interchange of A and A'. The expectation of the symmetric part must be symmetric, and the expectation value of the antisymmetric part must be antisymmetric. Furthermore, averaging over the unitary group ensures that no symmetric state is preferred over any other.

b) To evaluate the constant c_{sym}, multiply both sides of eq.(10.383) by $\boldsymbol{\Pi}^{(\mathrm{sym})}_{AA'}$ and take the trace of both sides, thus finding

$$c_{\mathrm{sym}} = \frac{|A_1| + |A_2|}{|A|+1}. \tag{10.385}$$

c) To evaluate the constant c_{anti}, multiply both sides of eq.(10.383)) by $\mathbf{\Pi}^{(\text{anti})}_{AA'}$ and take the trace of both sides, thus finding

$$c_{\text{anti}} = \frac{|A_1| - |A_2|}{|A| - 1}. \tag{10.386}$$

d) Using

$$c_{\boldsymbol{I}} = \frac{1}{2}\left(c_{\text{sym}} + c_{\text{anti}}\right), \quad c_{\boldsymbol{S}} = \frac{1}{2}\left(c_{\text{sym}} - c_{\text{anti}}\right) \tag{10.387}$$

prove eq.(10.334).

10.8 Fano's inequality

Suppose $X = \{x, p(x)\}$ is a probability distribution for a letter x drawn from an alphabet of d possible letters, and that XY is the joint distribution for x and another random variable y which is correlated with x. Upon receiving y we estimate the value of x by evaluating a function $\hat{x}(y)$. We may anticipate that if our estimate is usually correct, then the conditional entropy $H(X|Y)$ must be small. In this problem we will confirm that expectation.

Let $e \in \{0, 1\}$ denote a binary random variable which takes the value $e = 0$ if $x = \hat{x}(y)$ and takes the value $e = 1$ if $x \neq \hat{x}(y)$, and let XYE denote the joint distribution for x, y, e. The *error probability* P_e is the probability that $e = 1$, averaged over this distribution. Our goal is to derive an upper bound on $H(X|Y)$ depending on P_e.

a) Show that

$$H(X|Y) = H(X|YE) + H(E|Y) - H(E|XY). \tag{10.388}$$

Note that $H(E|XY) = 0$ because e is determined when x and y are know, and that $H(E|Y) \leq H(E)$ because mutual information is nonnegative. Therefore,

$$H(X|Y) \leq H(X|YE) + H(E). \tag{10.389}$$

b) Noting that

$$H(X|YE) = p(e=0)H(X|Y, e=0) + p(e=1)H(X|Y, e=1), \tag{10.390}$$

and that $H(X|Y, e = 0) = 0$ (because $x = \hat{x}(y)$ is determined by y when there is no error), show that

$$H(X|YE) \leq P_e \log_2(d-1). \tag{10.391}$$

c) Finally, show that

$$H(X|Y) \leq H_2(P_e) + P_e \log_2(d-1), \tag{10.392}$$

which is *Fano's inequality.*

d) Use Fano's inequality to derive eq.(10.50), hence completing the proof that the classical channel capacity C is an upper bound on achievable rates for communication over a noisy channel with negligible error probability.

10.9 A quantum version of Fano's inequality

a) In a d-dimensional system, suppose a density operator $\boldsymbol{\rho}$ approximates the pure state $|\psi\rangle$ with fidelity

$$F = \langle\psi|\boldsymbol{\rho}|\psi\rangle = 1 - \varepsilon. \tag{10.393}$$

Show that

$$H(\boldsymbol{\rho}) \leq H_2(\varepsilon) + \varepsilon \log_2(d-1). \tag{10.394}$$

Hint: Recall that if a complete orthogonal measurement performed on the state $\boldsymbol{\rho}$ has distribution of outcomes X, then $H(\boldsymbol{\rho}) \leq H(X)$, where $H(X)$ is the Shannon entropy of X.

b) As in §10.7.2, suppose that the noisy channel $\mathcal{N}^{A\to B}$ acts on the pure state ψ_{RA}, and is followed by the decoding map $\mathcal{D}^{B\to C}$. Show that

$$H(R)_{\boldsymbol{\rho}} - I_c(R\rangle B)_{\boldsymbol{\rho}} \leq 2H(RC)_{\boldsymbol{\sigma}}, \tag{10.395}$$

where

$$\boldsymbol{\rho}_{RB} = \mathcal{N}(\psi_{RA}), \quad \boldsymbol{\sigma}_{RC} = \mathcal{D}\circ\mathcal{N}(\psi_{RA}). \tag{10.396}$$

Therefore, if the decoder's output (the state of RC) is almost pure, then the coherent information of the channel $\mathcal{N}$ comes close to matching its input entropy. **Hint**: Use the data processing inequality $I_c(R\rangle C)_{\boldsymbol{\sigma}} \leq I_c(R\rangle B)_{\boldsymbol{\rho}}$ and the subadditivity of von Neumann entropy. It is convenient to consider the joint pure state of the reference system, the output, and environments of the dilations of $\mathcal{N}$ and $\mathcal{D}$.

c) Suppose that the decoding map recovers the channel input with high fidelity,

$$F(\mathcal{D}\circ\mathcal{N}(\psi_{RA}), \psi_{RC}) = 1 - \varepsilon. \tag{10.397}$$

Show that

$$H(R)_{\boldsymbol{\rho}} - I_c(R\rangle B)_{\boldsymbol{\rho}} \leq 2H_2(\varepsilon) + 2\varepsilon \log_2(d^2 - 1), \quad (10.398)$$

assuming that R and C are d-dimensional. This is a quantum version of Fano's inequality, which we may use to derive an upper bound on the quantum channel capacity of $\mathcal{N}$.

10.10 Mother protocol for the GHZ state

The mother resource inequality expresses an asymptotic resource conversion that can be achieved if Alice, Bob, and Eve share n copies of the pure state ϕ_{ABE}: by sending $\frac{n}{2}I(A;E)$ qubits to Bob, Alice can destroy the correlations of her state with Eve's state, so that Bob alone holds the purification of Eve's state, and furthermore Alice and Bob share $\frac{n}{2}I(A;B)$ ebits of entanglement at the end of the protocol; here $I(A;E)$ and $I(A;B)$ denote quantum mutual informations evaluated in the state ϕ_{ABE}.

Normally, the resource conversion can be realized with arbitrarily good fidelity only in the limit $n \to \infty$. But in this problem we will see that the conversion can be perfect if Alice, Bob and Eve share only $n = 2$ copies of the three-qubit GHZ state

$$|\phi\rangle_{ABE} = \frac{1}{\sqrt{2}} \left(|000\rangle + |111\rangle \right) . \quad (10.399)$$

The protocol achieving this perfect conversion uses the notion of *coherent classical communication* defined in Chapter 4.

a) Show that in the GHZ state $|\phi\rangle_{ABE}$, $I(A;E) = I(A;B) = 1$. Thus, for this state, the mother inequality becomes

$$2\langle \phi_{ABE} \rangle + [q \to q]_{AB} \geq [qq]_{AB} + 2\langle \phi'_{\tilde{B}E} \rangle . \quad (10.400)$$

b) Suppose that in the GHZ state Alice measures the Pauli operator X, gets the outcome $+1$ and broadcasts her outcome to Bob and Eve. What state do Bob and Eve then share? What if Alice gets the outcome -1 instead?

c) Suppose that Alice, Bob, and Eve share just one copy of the GHZ state ϕ^{ABE}. Find a protocol such that, after one unit of *coherent classical communication* from Alice to Bob, the shared state becomes $|\phi^+\rangle_{AB} \otimes |\phi^+\rangle_{BE}$, where $|\phi^+\rangle = \left(|00\rangle + |11\rangle \right) / \sqrt{2}$ is a maximally entangled Bell pair.

d) Now suppose that Alice, Bob, and Eve start out with two copies of the GHZ state, and suppose that Alice and Bob can borrow

an ebit of entanglement, which will be repaid later, to catalyze the resource conversion. Use coherent superdense coding to construct a protocol that achieves the (catalytic) conversion eq. (10.400) *perfectly*.

10.11 Degradability of amplitude damping and erasure

The qubit amplitude damping channel $\mathcal{N}_{\text{a.d.}}^{A\to B}(p)$ discussed in §3.4.3 has the dilation $\boldsymbol{U}^{A\to BE}$ such that

$$\begin{aligned}\boldsymbol{U}:&|0\rangle_A \mapsto |0\rangle_B \otimes |0\rangle_E,\\ &|1\rangle_A \mapsto \sqrt{1-p}\,|1\rangle_B \otimes |0\rangle_E + \sqrt{p}\,|0\rangle_B \otimes |1\rangle_E;\end{aligned}$$

a qubit in its "ground state" $|0\rangle_A$ is unaffected by the channel, while a qubit in the "excited state" $|1\rangle_A$ decays to the ground state with probability p, and the decay process excites the environment. Note that $\boldsymbol{U}$ is invariant under interchange of systems B and E accompanied by transformation $p \leftrightarrow (1-p)$. Thus the channel complementary to $\mathcal{N}_{\text{a.d.}}^{A\to B}(p)$ is $\mathcal{N}_{\text{a.d.}}^{A\to E}(1-p)$.

a) Show that $\mathcal{N}_{\text{a.d.}}^{A\to B}(p)$ is degradable for $p \le 1/2$. Therefore, the quantum capacity of the amplitude damping channel is its optimized one-shot coherent information. **Hint**: It suffices to show that

$$\mathcal{N}_{\text{a.d.}}^{A\to E}(1-p) = \mathcal{N}_{\text{a.d.}}^{B\to E}(q) \circ \mathcal{N}_{\text{a.d.}}^{A\to B}(p), \tag{10.401}$$

where $0 \le q \le 1$.

The *erasure channel* $\mathcal{N}_{\text{erase}}^{A\to B}(p)$ has the dilation $\boldsymbol{U}^{A\to BE}$ such that

$$\boldsymbol{U}: |\psi\rangle_A \mapsto \sqrt{1-p}\,|\psi\rangle_B \otimes |e\rangle_E + \sqrt{p}\,|e\rangle_B \otimes |\psi\rangle_E; \tag{10.402}$$

Alice's system passes either to Bob (with probability $1-p$) or to Eve (with probability p), while the other party receives the "erasure symbol" $|e\rangle$, which is orthogonal to Alice's Hilbert space. Because $\boldsymbol{U}$ is invariant under interchange of systems B and E accompanied by transformation $p \leftrightarrow (1-p)$, the channel complementary to $\mathcal{N}_{\text{erase}}^{A\to B}(p)$ is $\mathcal{N}_{\text{erase}}^{A\to E}(1-p)$.

b) Show that $\mathcal{N}_{\text{erase}}^{A\to B}(p)$ is degradable for $p \le 1/2$. Therefore, the quantum capacity of the amplitude damping channel is its optimized one-shot coherent information. **Hint**: It suffices to show that

$$\mathcal{N}_{\text{erase}}^{A\to E}(1-p) = \mathcal{N}_{\text{erase}}^{B\to E}(q) \circ \mathcal{N}_{\text{erase}}^{A\to B}(p), \tag{10.403}$$

where $0 \le q \le 1$.

c) Show that for $p \leq 1/2$ the quantum capacity of the erasure channel is

$$Q(\mathcal{N}_{\text{erase}}^{A\to B}(p)) = (1-2p)\log_2 d, \tag{10.404}$$

where A is d-dimensional, and that the capacity vanishes for $1/2 \leq p \leq 1$.

10.12 Capacities of the depolarizing channel

Consider the depolarizing channel $\mathcal{N}_{\text{depol.}}(p)$, which acts on a pure state $|\psi\rangle$ of a single qubit according to

$$\mathcal{N}_{\text{depol.}}(p) : |\psi\rangle\langle\psi| \mapsto \left(1 - \frac{4}{3}p\right)|\psi\rangle\langle\psi| + \frac{4}{3}p \cdot \frac{1}{2}\boldsymbol{I}. \tag{10.405}$$

For this channel, compute the product-state classical capacity $C_1(p)$, the entanglement-assisted classical capacity $C_E(p)$, and the one-shot quantum capacity $Q_1(p)$. Plot the results as a function of p. For what value of p does Q_1 hit zero?

The depolarizing channel is not degradable, and in fact the quantum capacity $Q(p)$ is larger than $Q_1(p)$ when the channel is sufficiently noisy. The function $Q(p)$ is still unknown.

10.13 Noisy superdense coding and teleportation.

a) By converting the entanglement achieved by the mother protocol into classical communication, prove the noisy superdense coding resource inequality:

$$Noisy\ SD: \quad \langle\phi_{ABE}\rangle + H(A)[q\to q] \geq I(A;B)[c\to c]. \tag{10.406}$$

Verify that this matches the standard noiseless superdense coding resource inequality when ϕ is a maximally entangled state of AB.

b) By converting the entanglement achieved by the mother protocol into quantum communication, prove the noisy teleportation resource inequality:

$$Noisy\ TP: \quad \langle\phi_{ABE}\rangle + I(A;B)[c\to c] \geq I_c(A\rangle B)[q\to q]. \tag{10.407}$$

Verify that this matches the standard noiseless teleportation resource inequality when ϕ is a maximally entangled state of AB.

10.14 The cost of erasure

Erasure of a bit is a process in which the state of the bit is reset to 0. Erasure is *irreversible* — knowing only the final state 0 after erasure, we cannot determine whether the initial state before erasure was 0 or 1. This irreversibility implies that erasure incurs an unavoidable thermodynamic cost. According to *Landauer's Principle*, erasing a bit at temperature T requires work $W \geq kT \log 2$. In this problem you will verify that a particular procedure for achieving erasure adheres to Landauer's Principle.

Suppose that the two states of the bit both have zero energy. We erase the bit in two steps. In the first step, we bring the bit into contact with a reservoir at temperature $T > 0$, and wait for the bit to come to thermal equilibrium with the reservoir. In this step the bit "forgets" its initial value, but the bit is not yet erased because it has not been reset.

We reset the bit in the second step, by slowly turning on a control field λ which splits the degeneracy of the two states. For $\lambda \geq 0$, the state 0 has energy $E_0 = 0$ and the state 1 has energy $E_1 = \lambda$. After the bit thermalizes in step one, the value of λ increases gradually from the initial value $\lambda = 0$ to the final value $\lambda = \infty$; the increase in λ is slow enough that the qubit remains in thermal equilibrium with the reservoir at all times. As λ increases, the probability $P(0)$ that the qubit is in the state 0 approaches unity — *i.e.*, the bit is reset to the state 0, which has zero energy.

(*a*) For $\lambda \neq 0$, find the probability $P(0)$ that the qubit is in the state 0 and the probability $P(1)$ that the qubit is in the state 1.

(*b*) How much work is required to increase the control field from λ to $\lambda + d\lambda$?

(*c*) How much work is expended as λ increases slowly from $\lambda = 0$ to $\lambda = \infty$? (You will have to evaluate an integral, which can be done analytically.)

10.15 The first law of Von Neumann entropy

Writing the density operator in terms of its *modular Hamiltonian* $\boldsymbol{K}$ as in §10.2.6,

$$\boldsymbol{\rho} = \frac{e^{-\boldsymbol{K}}}{\mathrm{tr}\,(e^{-\boldsymbol{K}})}, \tag{10.408}$$

consider how the entropy $S(\boldsymbol{\rho}) = -\mathrm{tr}\,(\boldsymbol{\rho}\ln\boldsymbol{\rho})$ changes when the density operator is perturbed slightly:

$$\boldsymbol{\rho} \to \boldsymbol{\rho}' = \boldsymbol{\rho} + \delta\boldsymbol{\rho}. \tag{10.409}$$

Since $\boldsymbol{\rho}$ and $\boldsymbol{\rho}'$ are both normalized density operators, we have $\mathrm{tr}\,(\delta\boldsymbol{\rho}) = 0$. Show that

$$S(\boldsymbol{\rho}') - S(\boldsymbol{\rho}) = \mathrm{tr}\left(\boldsymbol{\rho}'\boldsymbol{K}\right) - \mathrm{tr}\,(\boldsymbol{\rho}\boldsymbol{K}) + O\left((\delta\boldsymbol{\rho})^2\right); \tag{10.410}$$

that is,

$$\delta S = \delta\langle\boldsymbol{K}\rangle \tag{10.411}$$

to first order in the small change in $\boldsymbol{\rho}$. This statement generalizes the first law of thermodynamics; for the case of a thermal density operator with $\boldsymbol{K} = T^{-1}\boldsymbol{H}$ (where $\boldsymbol{H}$ is the Hamiltonian and T is the temperature), it becomes the more familiar statement

$$\delta E = \delta\langle\boldsymbol{H}\rangle = T\delta S. \tag{10.412}$$

10.16 Information gain for a quantum state drawn from the uniform ensemble

Suppose Alice prepares a quantum state drawn from the ensemble $\{\boldsymbol{\rho}(x), p(x)\}$ and Bob performs a measurement $\{\boldsymbol{E}(y)\}$ yielding outcome y with probability $p(y|x) = \mathrm{tr}\,(\boldsymbol{E}(y)\boldsymbol{\rho}(x))$. As noted in §10.6.1, Bob's information gain about Alice's preparation is the mutual information $I(X;Y) = H(X) - H(X|Y)$. If x is a continuous variable, while y is discrete, it is more convenient to use the symmetry of mutual information to write $I(X;Y) = H(Y) - H(Y|X)$, where

$$H(Y|X) = \sum_y \int dx \cdot p(x) \cdot p(y|x) \cdot \log p(y|x); \tag{10.413}$$

here $p(x)$ is a probability *density* (that is, $p(x)dx$ is the probability for x to lie in the interval $[x, x+dx]$).

For example, suppose that Alice prepares an arbitrary pure state $|\varphi\rangle$ chosen from the uniform ensemble in a d-dimensional Hilbert space, and Bob performs an orthogonal measurement projecting onto the basis $\{|e_y\rangle\}$, hoping to learn something about what Alice prepared. Then Bob obtains outcome y with probability

$$p(y|\theta) = |\langle e_y|\varphi\rangle|^2 \equiv \cos^2\theta \tag{10.414}$$

where θ is the angle between $|\varphi\rangle$ and $|e_y\rangle$. Because Alice's ensemble is uniform, Bob's outcomes are also uniformly distributed; hence $H(Y) = \log d$. Furthermore, the measurement outcome y reveals only information about θ; Bob learns nothing else about $|\varphi\rangle$. Therefore, eq.(10.413) implies that the information gain may be expressed as

$$I(X;Y) = \log d - d \int d\theta \cdot p(\theta) \cdot \cos^2\theta \cdot \log \cos^2\theta. \tag{10.415}$$

Here $p(\theta)d\theta$ is the probability density for the vector $|\varphi\rangle$ to point in a direction making angle θ with the axis $|e_y\rangle$, where $0 \le \theta \le \pi/2$.

a) Show that

$$p(\theta) \cdot d\theta = -(d-1)\left[1-\cos^2\theta\right]^{d-2} \cdot d\cos^2\theta. \tag{10.416}$$

Hint: Choose a basis in which the fixed axis $|e_y\rangle$ is

$$|e_y\rangle = (1, \vec{0}) \tag{10.417}$$

and write

$$|\varphi\rangle = (e^{i\phi}\cos\theta, \psi^{\perp}), \tag{10.418}$$

where $\theta \in [0, \pi/2]$, and $|\psi^{\perp}\rangle$ denotes a complex $(d-1)$-component vector with length $\sin\theta$. Now note that the phase ϕ resides on a circle of radius $\cos\theta$ (and hence circumference $2\pi\cos\theta$), while $|\psi^{\perp}\rangle$ lies on a sphere of radius $\sin\theta$ (thus the volume of the sphere, up to a multiplicative numerical constant, is $\sin^{2d-3}\theta$).

b) Now evaluate the integral eq. (10.415) to show that the information gain from the measurement, in *nats*, is

$$I(X;Y) = \ln d - \left(\frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{d}\right) . \tag{10.419}$$

(Information is expressed in nats if logarithms are natural logarithms; I in nats is related to I in bits by $I_{\text{bits}} = I_{\text{nats}}/\ln 2$.)
Hint: To evaluate the integral

$$\int_0^1 dx(1-x)^p x \ln x \ , \tag{10.420}$$

observe that

$$x \ln x = \frac{d}{ds} x^s \Big|_{s=1} \ , \tag{10.421}$$

and then calculate $\int_0^1 dx(1-x)^p x^s$ by integrating by parts repeatedly.

c) Show that in the limit of large d, the information gain, in *bits*, approaches

$$I_{d=\infty} = \frac{1-\gamma}{\ln 2} = .60995\ldots, \tag{10.422}$$

where $\gamma = .57721\ldots$ is Euler's constant.

Our computed value of $H(Y|X)$ may be interpreted in another way: Suppose we fix an orthogonal measurement, choose a typical state, and perform the measurement repeatedly on that chosen state. Then the measurement outcomes will not be uniformly distributed. Instead the entropy of the outcomes will fall short of maximal by .60995 bits, in the limit of large Hilbert space dimension.

References

[1] M. M. Wilde, *Quantum Information Theory* (Cambridge, 2013).

[2] T. M. Cover and J. A. Thomas, *Information Theory* (Wiley, 1991).

[3] C. E Shannon and W. Weaver, *The Mathematical Theory of Communication* (Illinois, 1949).

[4] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge, 2000).

[5] A. Wehrl, General properties of entropy, Rev. Mod. Phys. 50, 221 (1978).

[6] E. H. Lieb and M. B. Ruskai, A fundamental property of quantum-mechanical entropy, Phys. Rev. Lett. 30, 434 (1973).

[7] P. Hayden, R. Jozsa, D. Petz, and A. Winter, Structure of states which satisfy strong subadditivity with equality, Comm. Math. Phys. 246, 359-374 (2003).

[8] M. A. Nielsen and J. Kempe, Separable states are more disordered globally than locally, Phys. Rev. Lett. 86, 5184 (2001).

[9] J. Bekenstein, Universal upper bound on the entropy-to-energy ration of bounded systems, Phys. Rev. D 23, 287 (1981).

[10] H. Casini, Relative entropy and the Bekenstein bound, Class. Quant. Grav. 25, 205021 (2008).

[11] P. J. Coles, M. Berta, M. Tomamichel, S. Wehner, Entropic uncertainty relations and their applications, arXiv:1511.04857 (2015).

[12] H. Maassen and J. Uffink, Phys. Rev. Lett. 60, 1103 (1988).

[13] B. Schumacher, Quantum coding, Phys. Rev. A 51, 2738 (1995).

[14] R. Jozsa and B. Schumacher, A new proof of the quantum noiseless coding theory, J. Mod. Optics 41, 2343-2349 (1994).

[15] C. H. Bennett, H. J. Bernstein, S. Popescu, and B. Schumacher, Concentrating partial entanglement by local operations, Phys. Rev. A 53, 2046 (1996).

[16] R. Horodecki, P. Horodecki, M. Horodecki, and K. Horodecki, Quantum entanglement, Rev. Mod. Phys. 81, 865 (2009).

[17] F. G. S. L. Brandão and M. B. Plenio, A reversible theory of entanglement and its relation to the second law, Comm. Math. Phys. 295, 829-851 (2010).

[18] M. Christandl and A. Winter, "Squashed entanglement": an additive entanglement measure, J. Math. Phys. 45, 829 (2004).

[19] M. Koashi and A. Winter, Monogamy of quantum entanglement and other correlations, Phys. Rev. A 69, 022309 (2004).

[20] F. G. S. L. Brandão, M. Cristandl, and J. Yard, Faithful squashed entanglement, Comm. Math. Phys. 306, 805-830 (2011).

[21] A. C. Doherty, P. A. Parrilo, and F. M. Spedalieri, Complete family of separability criteria, Phys. Rev. A 69, 022308 (2004).

[22] A. S. Holevo, Bounds for the quantity of information transmitted by a quantum communication channel, Probl. Peredachi Inf. 9, 3-11 (1973).

[23] A. Peres and W. K. Wootters, Optimal detection of quantum information, Phys. Rev. Lett 66, 1119 (1991).

[24] A. S. Holevo, The capacity of the quantum channel with general signal states, arXiv: quant-ph/9611023.

[25] B. Schumacher and M. D. Westmoreland, Sending classical information via noisy quantum channels, Phys. Rev. A 56, 131-138 (1997).

[26] M. B. Hastings, Superadditivity of communication capacity using entangled inputs, Nature Physics 5, 255-257 (2009).

[27] M. Horodecki, P. W. Shor, and M. B. Ruskai, Entanglement breaking channels, Rev. Math. Phys. 15, 629-641 (2003).

[28] P. W. Shor, Additivity of the classical capacity for entanglement-breaking quantum channels, J. Math. Phys. 43, 4334 (2002).

[29] B. Schumacher and M. A. Nielsen, Quantum data processing and error correction, Phys. Rev. A 54, 2629 (1996).

[30] B. Schumacher, Sending entanglement through noisy quantum channels, Phys. Rev. A 54, 2614 (1996).

[31] H. Barnum, E. Knill, and M. A. Nielsen, On quantum fidelities and channel capacities, IEEE Trans. Inf. Theory 46, 1317-1329 (2000).

[32] S. Lloyd, Capacity of the noisy quantum channel, Phys. Rev. A 55, 1613 (1997).

[33] P. W. Shor, unpublished (2002).

[34] I. Devetak, The private classical capacity and quantum capacity of a quantum channel, IEEE Trans. Inf. Theory 51, 44-55 (2005).

[35] I. Devetak and A. Winter, Distillation of secret key and entanglement from quantum states, Proc. Roy. Soc. A 461, 207-235 (2005).

[36] B. Schumacher and M. D. Westmoreland, Approximate quantum error correction, Quant. Inf. Proc. 1, 5-12 (2002).

[37] M. Horodecki, J. Oppenheim, and A. Winter, Quantum state merging and negative information, Comm. Math. Phys. 269, 107-136 (2007).

[38] P. Hayden, M. Horodecki, A. Winter, and J. Yard, Open Syst. Inf. Dyn. 15, 7-19 (2008).

[39] A. Abeyesinge, I. Devetak, P. Hayden, and A. Winter, Proc. Roy. Soc. A, 2537-2563 (2009).

[40] E. Lubkin, Entropy of an n-system from its correlation with a k-reservoir, J. Math. Phys. 19, 1028 (1978).

[41] S. Lloyd and H. Pagels, Complexity as thermodynamic depth, Ann. Phys. 188, 186-213 (1988)

[42] D. N. Page, Average entropy of a subsystem, Phys. Rev. Lett. 71, 1291 (1993).

[43] I. Devetak, A. W. Harrow, and A. Winter, A family of quantum protocols, Phys. Rev. Lett. 93, 230504 (2004).

[44] I. Devetak, A. W. Harrow, and A. Winter, A resource framework for quantum Shannon theory, IEEE Trans. Inf. Theory 54, 4587-4618 (2008).

[45] I. Devetak and P. W. Shor, The capacity of a quantum channel for simultaneous transmission of classical and quantum information, Comm. Math. Phys. 256, 287-303 (2005).

[46] C. H. Bennett, P. W. Shor, J. A. Smolin, and A. V. Thapliyal, Entanglement-assisted classical capacity of noisy quantum channels, Phys. Rev. Lett. 83, 3081 (1999).

[47] C. H. Bennett, P. W. Shor, J. A. Smolin, and A. V. Thapliyal, Entanglement-assisted classical capacity of a quantum channel and the reverse Shannon theorem, IEEE Trans. Inf. Theory 48, 2637-2655 (2002).

[48] P. W. Shor and J. A. Smolin, Quantum error-correcting codes need not completely reveal the error syndrome, arXiv:quant-ph/9604006.

[49] D. P. DiVincenzo, P. W. Shor, and J. A. Smolin, Quantum channel capacity of very noisy channels, Phys. Rev. A 57, 830 (1998).

[50] G. Smith and J. Yard, Quantum communication with zero-capacity channels, Science 321, 1812-1815 (2008).

[51] L. del Rio, J. Aberg, R. Renner, O. Dahlsten, and V. Vedral, The thermodynamic meaning of negative entropy, Nature 474, 61-63 (2011).

[52] P. Hayden and J. Preskill, Black holes as mirrors: quantum information in random subsystems, JHEP 09, 120 (2007).

[53] Y. Sekino and L. Susskind, Fast scramblers, JHEP 10, 065 (2008).