Chapter 5

Quantum Information Theory

Quantum information theory is a rich subject that could easily have occupied us all term. But because we are short of time (I'm anxious to move on to quantum computation), I won't be able to cover this subject in as much depth as I would have liked. We will settle for a brisk introduction to some of the main ideas and results. The lectures will perhaps be sketchier than in the first term, with more hand waving and more details to be filled in through homework exercises. Perhaps this chapter should have been called "quantum information theory for the impatient."

Quantum information theory deals with four main topics:

(1) Transmission of classical information over quantum channels (which we will discuss).

(2) The tradeoff between acquisition of information about a quantum state and disturbance of the state (briefly discussed in Chapter 4 in connection with quantum cryptography, but given short shrift here).

(3) Quantifying quantum entanglement (which we will touch on briefly).

(4) Transmission of quantum information over quantum channels. (We will discuss the case of a noiseless channel, but we will postpone discussion of the noisy channel until later, when we come to quantum error-correcting codes.)

These topics are united by a common recurring theme: the interpretation and applications of the Von Neumann entropy.

5.1 Shannon for Dummies

Before we can understand Von Neumann entropy and its relevance to quantum information, we must 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? (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? (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, with a minimum of ε's and δ's. That way, I can compress classical information theory to about 11 pages.

5.1.1 Shannon entropy and data compression

A message is a string of letters chosen from an alphabet of k letters

$$\{a_1, a_2, \dots , a_k\}. \tag{5.1}$$

Let us suppose that the letters in the message are statistically independent, and that each letter a_x occurs with an *a priori* probability $p(a_x)$, where $\sum_{x=1}^{k} p(a_x) = 1$. For example, the simplest case is a binary alphabet, where 0 occurs with probability $1 - p$ and 1 with probability p (where $0 \leq p \leq 1$).

Now consider long messages with n letters, $n \gg 1$. We ask: is it possible to compress the message to a shorter string of letters that conveys essentially the same information?

For n very large, the law of large numbers tells us that typical strings will contain (in the binary case) 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 $logn! = n\log n - n + 0(\log n)$ we obtain

$$\log\binom{n}{np} = \log\left(\frac{n!}{(np)![n(1-p)]!}\right) \cong$$

$$n\log n - n - [np\log np - np + n(1-p)\log n(1-p) - n(1-p)]$$

$$= nH(p), \tag{5.2}$$

where

$$H(p) = -p\log p - (1-p)\log(1-p) \tag{5.3}$$

is the *entropy* function. Hence, the number of typical strings is of order $2^{nH(p)}$. (Logs are understood to have base 2 unless otherwise specified.)

To convey essentially all the information carried by a string of n bits, it suffices to choose a block code that assigns a positive integer to each of the typical strings. This block code has about $2^{nH(p)}$ letters (all occurring with equal *a priori* probability), so we may specify any one of the letters using a binary string of length $nH(p)$. Since $0 \leq H(p) \leq 1$ for $0 \leq p \leq 1$, and $H(p) = 1$ only for $p = \frac{1}{2}$, the block code shortens the message for any $p \neq \frac{1}{2}$ (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 $n \to \infty$.

This reasoning generalizes easily to the case of k letters, where letter x occurs with probability $p(x)$.[1] In a string of n letters, x typically occurs about $np(x)$ times, and the number of typical strings is of order

$$\frac{n!}{\prod_x (np(x))!} \simeq 2^{-nH(X)}, \tag{5.4}$$

where we have again invoked the Stirling approximation and

$$H(X) = \sum_x -p(x)\log p(x). \tag{5.5}$$

[1]The ensemble in which each of n letters is drawn from the distribution X will be denoted X^n.

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 $H(X)$ 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 in a slightly different language. A particular n-letter message

$$x_1 x_2 \cdots x_n, \tag{5.6}$$

occurs with *a priori* probability

$$P(x_1 \cdots x_n) = p(x_1)p(x_2)\cdots p(x_n) \tag{5.7}$$

$$\log P(x_1 \cdots x_n) = \sum_{i=1}^{n} \log p(x_i). \tag{5.8}$$

Applying the central limit theorem to this sum, we conclude that for "most sequences"

$$-\frac{1}{n}\log P(x_1, \cdots, x_n) \sim \langle -\log p(x) \rangle \equiv H(X), \tag{5.9}$$

where the brackets denote the mean value with respect to the probability distribution that governs the random variable x.

Of course, with ε's and δ's we can formulate these statements precisely. For any $\varepsilon, \delta > 0$ and for n sufficiently large, each "typical sequence" has a probability P satisfying

$$H(X) - \delta < -\frac{1}{n}\log P(x_1 \cdots x_n) < H(X) + \delta, \tag{5.10}$$

and the total probability of all typical sequences exceeds $1-\varepsilon$. Or, in other words, sequences of letters occurring with a total probability greater than $1-\varepsilon$ ("typical sequences") each have probability P such that

$$2^{-n(H-\delta)} \geq P \geq 2^{-n(H+\delta)}, \tag{5.11}$$

and from eq. (5.11) we may infer upper and lower bounds on the *number* $N(\varepsilon, \delta)$ of typical sequences (since the sum of the probabilities of all typical sequences must lie between $1-\varepsilon$ and 1):

$$2^{n(H+\delta)} \geq N(\varepsilon, \delta) \geq (1-\varepsilon)2^{n(H-\delta)}. \tag{5.12}$$

With a block code of length $n(H+\delta)$ bits we can encode all typical sequences. Then no matter how the atypical sequences are encoded, the probability of error will still be less than ε.

Conversely, if we attempt to compress the message to less than $H-\delta'$ bits per letter, we will be unable to achieve a small error rate as $n \to \infty$, because we will be unable to assign unique codewords to all typical sequences. The probability P_{success} of successfully decoding the message will be bounded by

$$P_{\text{success}} \leq 2^{n(H-\delta')}2^{-n(H-\delta)} + \varepsilon' = 2^{-n(\delta'-\delta)} + \varepsilon'; \tag{5.13}$$

we can correctly decode only $2^{n(H-\delta')}$ typical messages, each occurring with probability less than $2^{-n(H-\delta)}$ (the ε' is added to allow for the possibility that we manage to decode the atypical messages correctly). Since we may choose δ as small as we please, this success probability becomes small as $n \to \infty$.

We conclude that the optimal code compresses each letter to $H(X)$ bits asymptotically. This is Shannon's noiseless coding theorem.

5.1.2 Mutual information

The Shannon entropy $H(X)$ quantifies how much information is conveyed, on the average, by a letter drawn from the ensemble X, for it tells us how many bits are required (asymptotically as $n \to \infty$, where n is the number of letters drawn) to encode that information.

The mutual information $I(X;Y)$ quantifies how *correlated* two messages are. How much do we know about a message drawn from X^n when we have read a message drawn from Y^n?

For example, suppose we want to send a message from a transmitter to a receiver. But the communication channel is noisy, so that the message received (y) might differ from the message sent (x). The noisy channel can be characterized by the conditional probabilities $p(y|x)$ – the probability that y is received when x is sent. We suppose that the letter x is sent with *a priori* probability $p(x)$. We want to quantify how much we learn about x when we receive y; how much information do we gain?

As we have already seen, the entropy $H(X)$ quantifies my *a priori* ignorance per letter, before any message is received; that is, you would need to convey nH (noiseless) bits to me to completely specify (asymptotically) a particular message of n letters. But after I learn the value of y, I can use

Bayes' rule to update my probability distribution for x:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}. \tag{5.14}$$

(I know $p(y|x)$ if I am familiar with the properties of the channel, and $p(x)$ if I know the *a priori* probabilities of the letters; thus I can compute $p(y) = \sum_x p(y|x)p(x)$.) Because of the new knowledge I have acquired, I am now less ignorant about x than before. Given the y's I have received, using an optimal code, you can specify a particular string of n letters by sending me

$$H(X|Y) = \langle -\log p(x|y)\rangle, \tag{5.15}$$

bits per letter. $H(X|Y)$ is called the "conditional entropy." From $p(x|y) = p(x,y)/p(y)$, we see that

$$\begin{aligned} H(X|Y) &= \langle -\log p(x,y) + \log p(y)\rangle \\ &= H(X,Y) - H(Y), \end{aligned} \tag{5.16}$$

and similarly

$$\begin{aligned} H(Y|X) &\equiv \langle -\log p(y|x)\rangle \\ &= \langle -\log \left(\frac{p(x,y)}{p(x)}\right)\rangle = H(X,Y) - H(X). \end{aligned} \tag{5.17}$$

We may interpret $H(X|Y)$, then, as the number of *additional* bits per letter needed to specify *both* x and y once y is known. Obviously, then, this quantity cannot be negative.

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

$$\begin{aligned} I(X;Y) &\equiv H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(Y) - H(Y|X). \end{aligned} \tag{5.18}$$

$I(X;Y)$ is called the mutual information. It is obviously 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 can never *reduce* my knowledge of X, so $I(X;Y)$ is obviously nonnegative. (The inequalities $H(X) \geq H(X|Y) \geq 0$ are easily proved using the convexity of the log function; see for example *Elements of Information Theory* by T. Cover and J. Thomas.)

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

$$I(X;Y) \equiv \langle \log \frac{p(x,y)}{p(x)p(y)} \rangle = 0; \tag{5.19}$$

naturally, we can't find out about X by learning Y if there is no correlation!

5.1.3 The noisy channel coding theorem

If we want to communicate over a noisy channel, it is obvious that we can improve the reliability of transmission through redundancy. For example, I might send each bit many times, and the receiver could use majority voting to decode the bit.

But given a channel, is it always possible to find a code that can ensure arbitrarily good reliability (as $n \to \infty$)? And what can be said about the *rate* of such codes; *i.e.*, how many bits are required per letter of the message?

In fact, Shannon showed that any channel can be used for arbitrarily reliable communication at a finite (nonzero) rate, as long as there is *some* correlation between input and output. Furthermore, he found a useful expression for the optimal rate that can be attained. These results are the content of the "noisy channel coding theorem."

Suppose, to be concrete, that we are using a binary alphabet, 0 and 1 each occurring with *a priori* probability $\frac{1}{2}$. And suppose that the channel is the "binary symmetric channel" – it acts on each bit independently, flipping its value with probability p, and leaving it intact with probability $1-p$. That is, the conditional probabilities are

$$\begin{array}{ll} p(0|0) = 1-p, & p(0|1) = p, \\ p(1|0) = p, & p(1|1) = 1-p. \end{array} \tag{5.20}$$

We want to construct a family of codes of increasing block size n, such that the probability of a decoding error goes to zero as $n \to \infty$. If the number of bits encoded in the block is k, then the code consists of a choice of

2^k "codewords" among the 2^n possible strings of n bits. We define the rate R of the code (the number of data bits carried per bit transmitted) as

$$R = \frac{k}{n}. \tag{5.21}$$

We should design our code so that the code strings are as "far apart" as possible. That is for a given rate R, we want to maximize the number of bits that must be flipped to change one codeword to another (this number is called the "Hamming distance" between the two codewords).

For any input string of length n bits, errors will typically cause about np of the bits to flip – hence the input typically diffuses to one of about $2^{nH(p)}$ typical output strings (occupying a "sphere" of "Hamming radius" np about the input string). To decode reliably, we will want to choose our input codewords so that the error spheres of two different codewords are unlikely to overlap. Otherwise, two different inputs will sometimes yield the same output, and decoding errors will inevitably occur. If we are to avoid such decoding ambiguities, the total number of strings contained in all 2^{nR} error spheres must not exceed the total number 2^n of bits in the output message; we require

$$2^{nH(p)}2^{nR} \le 2^n \tag{5.22}$$

or

$$R \le 1 - H(p) \equiv C(p). \tag{5.23}$$

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 *attainable* (asymptotically)?

In fact transmission with R arbitrarily close to C and arbitrarily small error probability is possible. Perhaps the most ingenious of Shannon's ideas was to demonstrate that C can be attained by considering an average over "random codes." (Obviously, choosing a code at random is not the most clever possible procedure, but, perhaps surprisingly, it turns out that random coding achieves as high a rate (asymptotically for large n) as any other coding scheme.) Since C is the optimal rate for reliable transmission of data over the noisy channel it is called the *channel capacity.*

Suppose that 2^{nR} codewords are chosen at random by sampling the ensemble X^n. A message (one of the codewords) is sent. To decode the message, we draw a "Hamming sphere" around the message received that contains

$$2^{n(H(p)+\delta)}, \tag{5.24}$$

strings. The message is decoded as the codeword contained in this sphere, assuming such a codeword exists and is unique. If no such codeword exists, or the codeword is not unique, then we will assume that a decoding error occurs.

How likely is a decoding error? We have chosen the decoding sphere large enough so that failure of a valid codeword to appear in the sphere is atypical, so we need only worry about more than one valid codeword occupying the sphere. Since there are altogether 2^n possible strings, the Hamming sphere around the output contains a fraction

$$\frac{2^{n(H(p)+\delta)}}{2^n} = 2^{-n(C(p)-\delta)}, \tag{5.25}$$

of all strings. Thus, the probability that one of the 2^{nR} randomly chosen codewords occupies this sphere "by accident" is

$$2^{-n(C(p)-R-\delta)}, \tag{5.26}$$

Since we may choose δ as small as we please, R can be chosen as close to C as we please (but below C), and this error probability will still become exponentially small as $n \to \infty$.

So far we have 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. Thus there must exist one particular code with average probability of error (averaged over codewords) less than ε. But 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. We have demonstrated the existence of a code such that

$$\frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} P_i < \varepsilon. \tag{5.27}$$

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

$$\frac{1}{2^{nR}}(N_{2\varepsilon})2\varepsilon < \varepsilon \text{ or } N_{2\varepsilon} < 2^{nR-1}, \tag{5.28}$$

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

$$\text{Rate} = R - \frac{1}{n}, \tag{5.29}$$

which approaches R as $n \to \infty$

We have seen, then, that $C(p) = 1 - H(p)$ is the maximum rate that can be attained asymptotically with an arbitrarily small probability of error.

Consider now how these arguments generalize to more general alphabets and channels. We are given a channel specified by the $p(y|x)$'s, and let us specify a probability distribution $X = \{x, p(x)\}$ for the input letters. We will send strings of n letters, and we will assume that the channel acts on each letter independently. (A channel acting this way is said to be "memoryless.") Of course, once $p(y|x)$ and X are specified, $p(x|y)$ and $Y = \{y, p(y)\}$ are determined.

To establish an attainable rate, we again consider averaging over random codes, where codewords are chosen with *a priori* probability governed by X^n. Thus with high probability, these codewords will be chosen from a typical set of strings of letters, where there are about $2^{nH(X)}$ such typical strings.

For a typical received message in Y^n, there are about $2^{nH(X|Y)}$ messages that could have been sent. We may decode by associating with the received message a "sphere" containing $2^{n(H(X|Y)+\delta)}$ possible inputs. If there exists a unique codeword in this sphere, we decode the message as that codeword.

As before, it is unlikely that no codeword will be in the sphere, but we must exclude the possibility that there are more than one. Each decoding sphere contains a fraction

$$\begin{aligned} \frac{2^{n(H(X|Y)+\delta)}}{2^{nH(X)}} &= 2^{-n(H(X)-H(X|Y)-\delta)} \\ &= 2^{-n(I(X;Y)-\delta)}, \end{aligned} \tag{5.30}$$

of the typical inputs. If there are 2^{nR} codewords, the probability that any one falls in the decoding sphere by accident is

$$2^{nR}2^{-n(I(X;Y)-\delta)} = 2^{-n(I(X;Y)-R-\delta)}. \tag{5.31}$$

Since δ can be chosen arbitrarily small, we can choose R as close to I as we please (but less than I), and still have the probability of a decoding error become exponentially small as $n \to \infty$.

This argument shows that when we average over random codes and over codewords, the probability of an error becomes small for any rate $R < I$. The same reasoning as before then demonstrates the existence of a particular code with error probability $< \varepsilon$ for every codeword. This is a satisfying result, as it is consistent with our interpretation of I as the information that we gain about the input X when the signal Y is received – that is, I is the information per letter that we can send over the channel.

The mutual information $I(X;Y)$ depends not only on the channel conditional probabilities $p(y|x)$ but also on the priori probabilities $p(x)$ of the letters. The above random coding argument applies for any choice of the $p(x)$'s, so we have demonstrated that errorless transmission is possible for any rate R less than

$$C \equiv \underset{\{p(x)\}}{\mathrm{Max}} I(X;Y). \tag{5.32}$$

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

We have now shown that any rate $R < C$ is attainable, but is it possible for R to exceed C (with the error probability still approaching 0 for large n)? To show that C is an upper bound on the rate may seem more subtle in the general case than for the binary symmetric channel – the probability of error is different for different letters, and we are free to exploit this in the design of our code. However, we may reason as follows:

Suppose we have chosen 2^{nR} strings of n letters as our codewords. Consider a probability distribution (denoted $\tilde{X}^n$) in which each codeword occurs with equal probability ($= 2^{-nR}$). Evidently, then,

$$H(\tilde{X}^n) = nR. \tag{5.33}$$

Sending the codewords through the channel we obtain a probability distribution $\tilde{Y}^n$ of output states.

Because we assume that the channel acts on each letter independently, the conditional probability for a string of n letters factorizes:

$$p(y_1 y_2 \cdots y_n | x_1 x_2 \cdots x_n) = p(y_1|x_1) p(y_2|x_2) \cdots p(y_n|x_n), \tag{5.34}$$

and it follows that the conditional entropy satisfies

$$\begin{aligned} H(\tilde{Y}^n|\tilde{X}^n) = \langle -\log p(y^n|x^n)\rangle &= \sum_i \langle -\log p(y_i|x_i)\rangle \\ &= \sum_i H(\tilde{Y}_i|\tilde{X}_i), \end{aligned} \tag{5.35}$$

where $\tilde{X}_i$ and $\tilde{Y}_i$ are the marginal probability distributions for the ith letter determined by our distribution on the codewords. Recall that we also know that $H(X,Y) \leq H(X) + H(Y)$, or

$$H(\tilde{Y}^n) \leq \sum_i H(\tilde{Y}_i). \tag{5.36}$$

It follows that

$$\begin{aligned} I(\tilde{Y}^n; \tilde{X}^n) &= H(\tilde{Y}^n) - H(\tilde{Y}^n|\tilde{X}^n) \\ &\leq \sum_i (H(\tilde{Y}_i) - H(\tilde{Y}_i|\tilde{X}_i)) \\ &= \sum_i I(\tilde{Y}_i; \tilde{X}_i) \leq nC; \end{aligned} \tag{5.37}$$

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 $I(X;Y)$).

Recalling the symmetry of mutual information, we have

$$\begin{aligned} I(\tilde{X}^n; \tilde{Y}^n) &= H(\tilde{X}^n) - H(\tilde{X}^n|\tilde{Y}^n) \\ &= nR - H(\tilde{X}^n|\tilde{Y}^n) \leq nC. \end{aligned} \tag{5.38}$$

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

$$\frac{1}{n} H(\tilde{X}^n|\tilde{Y}^n) \to 0. \tag{5.39}$$

If errorless transmission is possible, then, eq. (5.38) becomes

$$R \leq C, \tag{5.40}$$

in the limit $n \to \infty$. The rate cannot exceed the capacity. (Remember that the conditional entropy, unlike the mutual information, is *not* symmetric. Indeed $(1/n)H(\tilde{Y}^n|\tilde{X}^n)$ does *not* become small, because the channel introduces uncertainty about what message will be received. But if we can decode accurately, there is no uncertainty about what codeword was sent, once the signal has been received.)

We have now shown that the capacity C is the highest rate of communication through the noisy channel that can be attained, 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.

Of course the method we have used to show that $R = C$ is asymptotically attainable (averaging over random codes) is not very constructive. Since a random code has no structure or pattern, encoding and decoding would be quite unwieldy (we require an exponentially large code book). Nevertheless, the theorem is important and useful, because it tells us what is in principle attainable, and furthermore, what is not attainable, even in principle. Also, 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.

5.2 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 information $H(X)$ is the number of incompressible bits of information carried per letter (asymptotically as $n \to \infty$).

We may also be interested in correlations between 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

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X), \tag{5.41}$$

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 than can be transmitted through the channel (given the *a priori* distribution for the X's).

We would like to generalize these considerations to *quantum* information. So let us 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 have already discussed at length, 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 matrix

$$\boldsymbol{\rho} = \sum_x p_x \boldsymbol{\rho}_x; \tag{5.42}$$

for the POVM $\{\mathbf{F}_a\}$, we have

$$\mathrm{Prob}(a) = tr(\mathbf{F}_a \boldsymbol{\rho}). \tag{5.43}$$

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

$$S(\boldsymbol{\rho}) = -tr(\boldsymbol{\rho} \log \boldsymbol{\rho}). \tag{5.44}$$

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

$$\boldsymbol{\rho} = \sum_a \lambda_a |a\rangle\langle a|, \tag{5.45}$$

then

$$S(\boldsymbol{\rho}) = H(A), \tag{5.46}$$

where $H(A)$ is the Shannon entropy of the ensemble $A = \{a, \lambda_a\}$.

In the case where the signal alphabet consists of mutually orthogonal pure states, the quantum source reduces to a classical one; all of the signal states can be perfectly distinguished, and $S(\boldsymbol{\rho}) = H(X)$. The quantum source is more interesting when the signal states $\boldsymbol{\rho}$ 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 a dual role. It quantifies not only the *quantum* information content per letter of the 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 a third way: 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 conceptional context is really quite different. The central issue in quantum information theory is that nonorthogonal pure quantum states cannot be perfectly distinguished, a feature with no classical analog.

5.2.1 Mathematical properties of $S(\boldsymbol{\rho})$

There are a handful of properties of $S(\boldsymbol{\rho})$ that are frequently useful (many of which are closely analogous to properties of $H(X)$). I list some of these properties below. Most of the proofs are not difficult (a notable exception is the proof of strong subadditivity), and are included in the exercises at the end of the chapter. Some proofs can also be found in A. Wehrl, "General Properties of Entropy," Rev. Mod. Phys. **50** (1978) 221, or in Chapter 9 of A. Peres, *Quantum Theory: Concepts and Methods.*

(1) Purity. A pure state $\boldsymbol{\rho} = |\varphi\rangle\langle\varphi|$ has $S(\boldsymbol{\rho}) = 0$.

(2) Invariance. The entropy is unchanged by a unitary change of basis:

$$S(\mathbf{U}\boldsymbol{\rho}\mathbf{U}^{-1}) = S(\boldsymbol{\rho}). \tag{5.47}$$

This is obvious, since $S(\boldsymbol{\rho})$ depends only on the eigenvalues of $\boldsymbol{\rho}$.

(3) Maximum. If $\boldsymbol{\rho}$ has D nonvanishing eigenvalues, then

$$S(\boldsymbol{\rho}) \leq \log D, \tag{5.48}$$

with equality when all the nonzero eigenvalues are equal. (The entropy is maximized when the quantum state is chosen *randomly.*)

(4) Concavity. For $\lambda_1, \lambda_2, \cdots, \lambda_n \geq 0$ and $\lambda_1 + \lambda_2 + \cdots + \lambda_n = 1$

$$S(\lambda_1\boldsymbol{\rho}_1 + \cdots + \lambda_n\boldsymbol{\rho}_n) \geq \lambda_1 S(\boldsymbol{\rho}_1) + \cdots + \lambda_n S(\boldsymbol{\rho}_n). \tag{5.49}$$

That is, 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) Entropy of measurement. Suppose that, in a state $\boldsymbol{\rho}$, we measure the observable

$$\boldsymbol{A} = \sum_y |a_y\rangle a_y \langle a_y|, \tag{5.50}$$

so that the outcome a_y occurs with probability

$$p(a_y) = \langle a_y|\boldsymbol{\rho}|a_y\rangle. \tag{5.51}$$

Then the Shannon entropy of the ensemble of measurement outcomes $Y = \{a_y, p(a_y)\}$ satisfies

$$H(Y) \geq S(\boldsymbol{\rho}), \tag{5.52}$$

with equality when $\boldsymbol{A}$ and $\boldsymbol{\rho}$ commute. Mathematically, this is the statement that $S(\boldsymbol{\rho})$ increases if we replace all off-diagonal matrix elements of $\boldsymbol{\rho}$ by zero, in any basis. Physically, it says that the randomness of the measurement outcome is minimized if we choose to measure an observable that commutes with the density matrix. But if we measure a "bad" observable, the result will be less predictable.

(6) Entropy of preparation. If a pure state is drawn randomly from the ensemble $\{|\varphi_x\rangle, p_x\}$, so that the density matrix is

$$\boldsymbol{\rho} = \sum_x p_x |\varphi_x\rangle\langle\varphi_x|, \tag{5.53}$$

then

$$H(X) \geq S(\boldsymbol{\rho}), \tag{5.54}$$

with equality if the signal states $|\varphi_x\rangle$ are mutually orthogonal. This statement indicates that *distinguishability is lost* when we mix nonorthogonal pure states. (We can't fully recover the information about which state was prepared, because, as we'll discuss later on, the information gain attained by performing a measurement cannot exceed $S(\boldsymbol{\rho})$.)

(7) Subadditivity. Consider a bipartite system AB in the state $\boldsymbol{\rho}_{AB}$. Then

$$S(\boldsymbol{\rho}_{AB}) \leq S(\boldsymbol{\rho}_A) + S(\boldsymbol{\rho}_B), \tag{5.55}$$

(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 analogous to the property

$$H(X,Y) \leq H(X) + H(Y), \tag{5.56}$$

(or $I(X;Y) \geq 0$) of Shannon entropy; it holds because some of the information in XY (or AB) is encoded in the correlations between X and Y (A and B).

(8) Strong subadditivity. For any state $\boldsymbol{\rho}_{ABC}$ of a tripartite system,

$$S(\boldsymbol{\rho}_{ABC}) + S(\boldsymbol{\rho}_B) \leq S(\boldsymbol{\rho}_{AB}) + S(\boldsymbol{\rho}_{BC}). \tag{5.57}$$

This property is called "strong" subadditivity in that it reduces to subadditivity in the event that B is one-dimensional. The proof of the corresponding property of Shannon entropy is quite simple, but the proof for Von Neumann entropy turns out to be surprisingly difficult (it is sketched in Wehrl). You may find the strong subadditivity property easier to remember by thinking about it this way: AB and BC can be regarded as two *overlapping* subsystems. The entropy of their union (ABC) plus the entropy of their intersection (B) does not exceed the sum of the entropies of the subsystems (AB and BC). We will see that strong subadditivity has deep and important consequences.

(9) Triangle inequality (Araki-Lieb inequality): For a bipartite system,

$$S(\boldsymbol{\rho}_{AB}) \geq |S(\boldsymbol{\rho}_A) - S(\boldsymbol{\rho}_B)|. \tag{5.58}$$

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

$$H(X,Y) \geq H(X), H(Y), \tag{5.59}$$

or

$$H(X|Y), H(Y|X) \geq 0. \tag{5.60}$$

The Shannon entropy of a classical bipartite system exceeds the Shannon entropy of either part – there is more information in the whole system than in part of it! Not so for the Von Neumann entropy. In the extreme case of a bipartite pure quantum state, we have $S(\boldsymbol{\rho}_A) = S(\boldsymbol{\rho}_B)$ (and nonzero if the state is entangled) while $S(\boldsymbol{\rho}_{AB}) = 0$. The bipartite state has a definite preparation, but if we measure observables of the subsystems, the measurement outcomes are inevitably random and unpredictable. We cannot discern how the state was prepared by observing the two subsystems separately, rather, information is encoded in the nonlocal quantum correlations. The juxtaposition of the positivity of conditional Shannon entropy (in the classical case) with the triangle inequality (in the quantum case) nicely characterizes a key distinction between quantum and classical information.

5.2.2 Entropy and thermodynamics

Of course, the concept of entropy first entered science through the study of thermodynamics. I will digress briefly here on some thermodynamic implications of the mathematic properties of $S(\boldsymbol{\rho})$.

There are two distinct (but related) possible approaches to the foundations of quantum statistical physics. In the first, we consider the evolution of an isolated (closed) quantum system, but we perform some *coarse graining* to define our thermodynamic variables. In the second approach, which is perhaps better motivated physically, we consider an *open* system, a quantum system in contact with its environment, and we track the evolution of the open system without monitoring the environment.

For an open system, the crucial mathematical property of the Von Neumann entropy is *subadditivity.* If the system (A) and environment (E) are initially uncorrelated with one another

$$\boldsymbol{\rho}_{AE} = \boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_E, \tag{5.61}$$

then entropy is additive:

$$S(\boldsymbol{\rho}_{AE}) = S(\boldsymbol{\rho}_A) + S(\boldsymbol{\rho}_E). \tag{5.62}$$

Now suppose that the open system evolves for a while. The evolution is described by a unitary operator $\mathbf{U}_{AE}$ that acts on the combined system A plus E:

$$\boldsymbol{\rho}_{AE} \to \boldsymbol{\rho}'_{AE} = \mathbf{U}_{AE}\boldsymbol{\rho}_{AE}\mathbf{U}_{AE}^{-1}, \tag{5.63}$$

and since unitary evolution preserves S, we have

$$S(\boldsymbol{\rho}'_{AE}) = S(\boldsymbol{\rho}_{AE}). \tag{5.64}$$

Finally, we apply subadditivity to the state $\boldsymbol{\rho}'_{AE}$ to infer that

$$S(\boldsymbol{\rho}_A) + S(\boldsymbol{\rho}_E) = S(\boldsymbol{\rho}'_{AE}) \leq S(\boldsymbol{\rho}'_A) + S(\boldsymbol{\rho}'_E), \tag{5.65}$$

(with equality in the event that A and E remain uncorrelated). If we define the "total" entropy of the world as the sum of the entropy of the system and the entropy of the environment, we conclude that *the entropy of the world cannot decrease.* This is one form of the second law of thermodynamics. But note that we assumed that system and environment were initially uncorrelated to derive this "law."

Typically, the interaction of system and environment *will* induce correlations so that (assuming no initial correlations) the entropy will actually *increase*. From our discussion of the master equation, in §3.5 you'll recall that the environment typically "forgets" quickly, so that if our time resolution is coarse enough, we can regard the system and environment as "initially" uncorrelated (in effect) at each instant of time (the Markovian approximation). Under this assumption, the "total" entropy will increase monotonically, asymptotically approaching its theoretical maximum, the largest value it can attain consistent with all relevant conservation laws (energy, charge, baryon number, etc.)

Indeed, the usual assumption underlying quantum statistical physics is that system and environment are in the "most probable configuration," that which maximizes $S(\boldsymbol{\rho}_A)+S(\boldsymbol{\rho}_E)$. In this configuration, all "accessible" states are equally likely.

From a microscopic point of view, information initially encoded in the system (our ability to distinguish one initial state from another, initially orthogonal, state) is lost; it winds up encoded in quantum entanglement between system and environment. In principle that information could be recovered, but in practice it is totally inaccessible to localized observers. Hence thermodynamic irreversibility.

Of course, we can adapt this reasoning to apply to a large closed system (the whole universe?). We may divide the system into a small part of the whole and the rest (the environment of the small part). Then the sum of the entropies of the parts will be nondecreasing. This is a particular type of coarse graining. That part of a closed system behaves like an open system is why the microcanonical and canonical ensembles of statistical mechanics yield the same predictions for large systems.

5.3 Quantum Data Compression

What is the quantum analog of the noiseless coding theorem?

We consider a long message consisting of n letters, where each letter is chosen at random from the ensemble of pure states

$$\{|\varphi_x\rangle, p_x\}, \tag{5.66}$$

and the $|\varphi_x\rangle$'s are not necessarily mutually orthogonal. (For example, each $|\varphi_x\rangle$ might be the polarization state of a single photon.) Thus, each letter is described by the density matrix

$$\boldsymbol{\rho} = \sum_x p_x |\varphi_x\rangle\langle\varphi_x|, \tag{5.67}$$

and the entire message has the density matrix

$$\boldsymbol{\rho}^n = \boldsymbol{\rho} \otimes \cdots \otimes \boldsymbol{\rho}. \tag{5.68}$$

Now we ask, how *redundant* is this quantum information? We would like to devise a *quantum code* that enables us to compress the message to a smaller Hilbert space, but without compromising the fidelity of the message. For example, perhaps we have a quantum memory device (the hard disk of a quantum computer?), and we know the *statistical* properties of the recorded data (*i.e.*, we know $\boldsymbol{\rho}$). We want to conserve space on the device by compressing the data.

The optimal compression that can be attained was found by Ben Schumacher. Can you guess the answer? The best possible compression compatible with arbitrarily good fidelity as $n \to \infty$ is compression to a Hilbert space $\mathcal{H}$ with

$$\log(\dim \mathcal{H}) = nS(\boldsymbol{\rho}). \tag{5.69}$$

In this sense, the Von Neumann entropy is the number of *qubits* of quantum information carried per letter of the message. For example, if the message consists of n photon polarization states, we can compress the message to

$m = nS(\boldsymbol{\rho})$ photons – compression is always possible unless $\boldsymbol{\rho} = \frac{1}{2}\mathbf{1}$. (We can't compress random qubits just as we can't compress random bits.)

Once Shannon's results are known and understood, the proof of Schumacher's theorem is not difficult. Schumacher's important contribution was to ask the right question, and so to establish for the first time a precise (quantum) information theoretic interpretation of Von Neumann entropy.[2]

5.3.1 Quantum data compression: an example

Before discussing Schumacher's quantum data compression protocol in full generality, it is helpful to consider a simple example. So suppose that our letters are single qubits drawn from the ensemble

$$\begin{aligned} |\uparrow_z\rangle &= \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad p = \tfrac{1}{2}, \\ |\uparrow_x\rangle &= \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \quad p = \tfrac{1}{2}, \end{aligned} \tag{5.70}$$

so that the density matrix of each letter is

$$\begin{aligned} \boldsymbol{\rho} &= \frac{1}{2}|\uparrow_z\rangle\langle\uparrow_z| + \frac{1}{2}|\uparrow_x\rangle\langle\uparrow_x| \\ &= \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix}. \end{aligned} \tag{5.71}$$

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

$$\begin{aligned} |0'\rangle \equiv |\uparrow_{\hat{n}}\rangle &= \begin{pmatrix} \cos\frac{\pi}{8} \\ \sin\frac{\pi}{8} \end{pmatrix}, \\ |1'\rangle \equiv |\downarrow_{\hat{n}}\rangle &= \begin{pmatrix} \sin\frac{\pi}{8} \\ -\cos\frac{\pi}{8} \end{pmatrix}; \end{aligned} \tag{5.72}$$

the eigenvalues are

$$\begin{aligned} \lambda(0') &= \frac{1}{2} + \frac{1}{2\sqrt{2}} = \cos^2\frac{\pi}{8}, \\ \lambda(1') &= \frac{1}{2} - \frac{1}{2\sqrt{2}} = \sin^2\frac{\pi}{8}; \end{aligned} \tag{5.73}$$

[2] An interpretation of $S(\boldsymbol{\rho})$ in terms of *classical* information encoded in quantum states was actually known earlier, as we'll soon discuss.

(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

$$|\langle 0'|\uparrow_z\rangle|^2 = |\langle 0'|\uparrow_x\rangle|^2 = \cos^2\frac{\pi}{8} = .8535, \tag{5.74}$$

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

$$|\langle 1'|\uparrow_z\rangle|^2 = |\langle 1'|\uparrow_x\rangle|^2 = \sin^2\frac{\pi}{8} = .1465. \tag{5.75}$$

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*

$$F = \frac{1}{2}|\langle\uparrow_z|\psi\rangle|^2 + \frac{1}{2}|\langle\uparrow_x|\psi\rangle|^2, \tag{5.76}$$

among all possible 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 (quantum channels are very expensive!). 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 the two letters with $F = 1$, and he has $F = .8535$ for the third; hence $F = .8535$ overall. But is there a more clever procedure that achieves higher fidelity?

There *is* a better procedure. 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 $|\psi\rangle = |\psi_1\rangle|\psi_2\rangle|\psi_3\rangle$ is any signal state (with each of three qubits in either the $|\uparrow_z\rangle$ or $|\uparrow_x\rangle$ state), we have

$$\begin{aligned}
&|\langle 0'0'0'|\psi\rangle|^2 = \cos^6\left(\frac{\pi}{8}\right) = .6219,\\
&|\langle 0'0'1'|\psi\rangle|^2 = |\langle 0'1'0'|\psi\rangle|^2 = |\langle 1'0'0'|\psi\rangle|^2 = \cos^4\left(\frac{\pi}{8}\right)\sin^2\left(\frac{\pi}{8}\right) = .1067,\\
&|\langle 0'1'1'|\psi\rangle|^2 = |\langle 1'0'1'|\psi\rangle|^2 = |\langle 1'1'0'|\psi\rangle|^2 = \cos^2\left(\frac{\pi}{8}\right)\sin^4\left(\frac{\pi}{8}\right) = .0183,\\
&|\langle 1'1'1'|\psi\rangle|^2 = \sin^6\left(\frac{\pi}{8}\right) = .0031.
\end{aligned} \tag{5.77}$$

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 a ("fuzzy") measurement that projects a signal state onto Λ or $\Lambda^\perp$, the probability of projecting onto the likely subspace is

$$P_{likely} = .6219 + 3(.1067) = .9419, \tag{5.78}$$

while the probability of projecting onto the unlikely subspace is

$$P_{unlikely} = 3(.0183) + .0031 = .0581. \tag{5.79}$$

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

$$|\cdot\rangle|\cdot\rangle|0\rangle, \tag{5.80}$$

and the four low-probability basis states to

$$|\cdot\rangle|\cdot\rangle|1\rangle; \tag{5.81}$$

then Alice measures the third qubit to complete the fuzzy measurement. If the outcome is $|0\rangle$, then Alice's input state has been projected (in effect) 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 $\mathbf{U}^{-1}$, obtaining

$$|\psi'\rangle = \mathbf{U}^{-1}(|\psi_{\text{comp}}\rangle|0\rangle). \tag{5.82}$$

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

$$|\psi'\rangle = \mathbf{U}^{-1}(|\psi_{\text{comp}}\rangle|0\rangle) = |0'0'0'\rangle. \tag{5.83}$$

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 $\boldsymbol{\rho}'$

$$|\psi\rangle\langle\psi| \to \boldsymbol{\rho}' = \mathbf{E}|\psi\rangle\langle\psi|\mathbf{E} + |0'0'0'\rangle\langle\psi|(1-\mathbf{E})|\psi\rangle\langle 0'0'0'|, \tag{5.84}$$

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

$$\begin{aligned} F = \langle\psi|\boldsymbol{\rho}'|\psi\rangle &= (\langle\psi|\mathbf{E}|\psi\rangle)^2 + (\langle\psi|(1-\mathbf{E})|\psi\rangle)(\langle\psi|0'0'0'\rangle)^2 \\ &= (.9419)^2 + (.0581)(.6219) = .9234. \end{aligned} \tag{5.85}$$

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. The Von-Neumann entropy of the one-qubit ensemble is

$$S(\boldsymbol{\rho}) = H\left(\cos^2\frac{\pi}{8}\right) = .60088\ldots \tag{5.86}$$

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

5.3.2 Schumacher encoding 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 matrix of a single letter is

$$\boldsymbol{\rho} = \sum_x p_x |\varphi_x\rangle\langle\varphi_x|. \tag{5.87}$$

Furthermore, the letters are drawn independently, so that the density matrix of the entire message is

$$\boldsymbol{\rho}^n \equiv \boldsymbol{\rho} \otimes \cdots \otimes \boldsymbol{\rho}. \tag{5.88}$$

We wish to argue 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 $2^{nS(\boldsymbol{\rho})}$.

This conclusion follows directly from the corresponding classical statement, if we consider the orthonormal basis in which $\boldsymbol{\rho}$ is diagonal. Working in this basis, we may regard our quantum information source as an effectively classical source, producing messages that are strings of $\boldsymbol{\rho}$ eigenstates, each with a probability given by the product of the corresponding eigenvalues.

For a specified n and δ, define the typical subspace Λ as the space spanned by the eigenvectors of $\boldsymbol{\rho}^n$ with eigenvalues λ satisfying

$$2^{-n(S-\delta)} \geq \lambda \geq e^{-n(S+\delta)}. \tag{5.89}$$

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

$$\mathrm{tr}(\boldsymbol{\rho}^n \mathbf{E}) > 1 - \varepsilon, \tag{5.90}$$

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

$$2^{n(S+\delta)} \geq \dim(\Lambda) \geq (1-\varepsilon) 2^{n(S-\delta)}. \tag{5.91}$$

Our coding strategy is to send states in the typical subspace faithfully. For example, we can make a fuzzy measurement that projects the input message onto either Λ or $\Lambda^\perp$; the outcome will be Λ with probability $P_\Lambda = \mathrm{tr}(\boldsymbol{\rho}^n \mathbf{E}) > 1 - \varepsilon$. 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 $\mathbf{U}$ that takes each state $|\psi_{\rm typ}\rangle$ in Λ to a state of the form

$$\mathbf{U}|\psi_{\rm typ}\rangle = |\psi_{\rm comp}\rangle |0_{\rm rest}\rangle, \tag{5.92}$$

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

Suppose that

$$|\varphi_i\rangle = |\varphi_{x_1(i)}\rangle \ldots |\varphi_{x_n(i)}\rangle, \tag{5.93}$$

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

$$\begin{aligned} |\varphi_i\rangle\langle\varphi_i| \to \boldsymbol{\rho}_i' = {} & \mathbf{E}|\varphi_i\rangle\langle\varphi_i|\mathbf{E} \\ & + \boldsymbol{\rho}_{i,\rm Junk}\langle\varphi_i|(\mathbf{1}-\mathbf{E})|\varphi_i\rangle, \end{aligned} \tag{5.94}$$

where $\boldsymbol{\rho}_{i,\text{Junk}}$ is the state we choose to send if the fuzzy measurement yields the outcome $\Lambda^\perp$. What can we say about the fidelity of this procedure?

The fidelity varies from message to message (in contrast to the example discussed above), so we consider the fidelity averaged over the ensemble of possible messages:

$$\begin{aligned} F &= \sum_i p_i \langle \varphi_i | \boldsymbol{\rho}'_i | \varphi_i \rangle \\ &= \sum_i p_i \langle \varphi_i | \mathbf{E} | \varphi_i \rangle \langle \varphi_i | \mathbf{E} | \varphi_i \rangle + \sum_i p_i \langle \varphi_i | \boldsymbol{\rho}_{i,\text{Junk}} | \varphi_i \rangle \langle \varphi_i | \mathbf{1} - \mathbf{E} | \varphi_i \rangle \\ &\geq \sum_i p_i \parallel \mathbf{E} | \varphi_i \rangle \parallel^4, \end{aligned} \tag{5.95}$$

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

$$(x-1)^2 \geq 0, \text{ or } x^2 \geq 2x - 1, \tag{5.96}$$

we have (setting $x = \parallel \mathbf{E}|\varphi_i\rangle \parallel^2$)

$$\parallel \mathbf{E}|\varphi_i\rangle \parallel^4 \geq 2 \parallel \mathbf{E}|\varphi_i\rangle \parallel^2 - 1 = 2\langle \varphi_i | \mathbf{E} | \varphi_i \rangle - 1, \tag{5.97}$$

and hence

$$\begin{aligned} F &\geq \sum_i p_i (2\langle \varphi_i | \mathbf{E} | \varphi_i \rangle - 1) \\ &= 2 \operatorname{tr}(\boldsymbol{\rho}^n \mathbf{E}) - 1 > 2(1-\varepsilon) - 1 = 1 - 2\varepsilon. \end{aligned} \tag{5.98}$$

We have shown, then, that it is possible to compress the message to fewer than $n(S+\delta)$ qubits, while achieving an average fidelity that becomes arbitrarily good a n gets large.

So we have established that the message may be compressed, with insignificant loss of fidelity, to $S+\delta$ qubits per letter. Is further compression possible?

Let us suppose that Bob will decode the message $\boldsymbol{\rho}_{\text{comp},i}$ that he receives by appending qubits and applying a unitary transformation $\mathbf{U}^{-1}$, obtaining

$$\boldsymbol{\rho}'_i = \mathbf{U}^{-1}(\boldsymbol{\rho}_{\text{comp},i} \otimes |0\rangle\langle 0|)\mathbf{U} \tag{5.99}$$

("unitary decoding"). Suppose that $\boldsymbol{\rho}_{\text{comp}}$ has been compressed to $n(S-\delta)$ qubits. Then, *no matter how the input message have been encoded*, the decoded messages are all contained in a subspace Λ' of Bob's Hilbert space of dimension $2^{n(S-\delta)}$. (We are *not* assuming now that Λ' has anything to do with the typical subspace.)

If the input message is $|\varphi_i\rangle$, then the message reconstructed by Bob is $\boldsymbol{\rho}'_i$ which can be diagonalized as

$$\boldsymbol{\rho}'_i = \sum_{a_i} |a_i\rangle\lambda_{a_i}\langle a_i|, \tag{5.100}$$

where the $|a_i\rangle$'s are mutually orthogonal states in Λ'. The fidelity of the reconstructed message is

$$\begin{aligned} F_i &= \langle\varphi_i|\boldsymbol{\rho}'_i|\varphi_i\rangle \\ &= \sum_{a_i} \lambda_{a_i}\langle\varphi_i|a_i\rangle\langle a_i|\varphi_i\rangle \\ &\leq \sum_{a_i}\langle\varphi_i|a_i\rangle\langle a_i|\varphi_i\rangle \leq \langle\varphi_i|\mathbf{E}'|\varphi_i\rangle, \end{aligned} \tag{5.101}$$

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

$$F = \sum_i p_i F_i \leq \sum_i p_i\langle\varphi_i|\mathbf{E}'|\varphi_i\rangle = \mathrm{tr}(\boldsymbol{\rho}^n\mathbf{E}'). \tag{5.102}$$

But since $\mathbf{E}'$ projects onto a space of dimension $2^{n(S-\delta)}$, $\mathrm{tr}(\boldsymbol{\rho}^n\mathbf{E}')$ can be no larger than the sum of the $2^{n(S-\delta)}$ largest eigenvalues of $\boldsymbol{\rho}^n$. It follows from the properties of typical subspaces that this sum becomes as small as we please; for n large enough

$$F \leq \mathrm{tr}(\boldsymbol{\rho}^n\mathbf{E}') < \varepsilon. \tag{5.103}$$

Thus we have shown that, if we attempt to compress to $S-\delta$ qubits per letter, then the fidelity inevitably becomes poor for n sufficiently large. We conclude then, that $S(\boldsymbol{\rho})$ qubits per letter is the optimal compression of the quantum information that can be attained if we are to obtain good fidelity as n goes to infinity. This is Schumacher's noiseless quantum coding theorem.

The above argument applies to any conceivable encoding scheme, but only to a restricted class of decoding schemes (unitary decodings). A more general decoding scheme can certainly be contemplated, described by a *superoperator*. More technology is then required to prove that better compression than S

qubits per letter is not possible. But the conclusion is the same. The point is that $n(S-\delta)$ qubits are not sufficient to distinguish all of the typical states.

To summarize, there is a close analogy between Shannon's noiseless coding theorem and Schumacher's noiseless quantum 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 unit overlap with the typical subspace, so we can code only the typical subspace and still achieve good fidelity.

In fact, Alice could send effectively classical information to Bob—the string $x_1x_2\cdots x_n$ encoded in mutually orthogonal quantum states—and Bob could then follow these classical instructions to reconstruct Alice's state. By this means, they could achieve high-fidelity compression to $H(X)$ bits—or qubits—per letter. But if the letters are drawn from an ensemble of *nonorthogonal* pure states, this amount of compression is not optimal; some of the classical information about the preparation of the state has become redundant, because the nonorthogonal states cannot be perfectly distinguished. Thus Schumacher coding can go further, achieving optimal compression to $S(\boldsymbol{\rho})$ qubits per letter. The information has been packaged more efficiently, but at a price—Bob has received what Alice intended, but Bob can't know what he has. In contrast to the classical case, Bob can't make any measurement that is certain to decipher Alice's message correctly. An attempt to read the message will unavoidably disturb it.

5.3.3 Mixed-state coding: Holevo information

The Schumacher theorem characterizes the compressibility of an ensemble of pure states. But what if the letters are drawn from an ensemble of *mixed* states? The compressibility in that case is not firmly established, and is the subject of current research.[3]

It is easy to see that $S(\boldsymbol{\rho})$ won't be the answer for mixed states. To give a trivial example, suppose that a particular mixed state $\boldsymbol{\rho}_0$ with $S(\boldsymbol{\rho}_0)\neq 0$ is chosen with probability $p_0=1$. Then the message is always $\boldsymbol{\rho}_0\otimes\boldsymbol{\rho}_0\otimes\cdots\otimes\boldsymbol{\rho}_0$ and it carries no information; Bob can reconstruct the message perfectly without receiving *anything* from Alice. Therefore, the message can be compressed to zero qubits per letters, which is less than $S(\boldsymbol{\rho})>0$.

To construct a slightly less trivial example, recall that for an ensemble of

[3]See M. Horodecki, `quant-ph/9712035`.

mutually orthogonal pure states, the Shannon entropy of the ensemble equals the Von Neumann entropy

$$H(X) = S(\boldsymbol{\rho}), \tag{5.104}$$

so that the classical and quantum compressibility coincide. This makes sense, since the orthogonal states are perfectly distinguishable. In fact, if Alice wants to send the message

$$|\varphi_{x_1}\rangle\varphi_{x_2}\rangle \cdots |\varphi_{x_n}\rangle \tag{5.105}$$

to Bob, she can send the classical message $x_1 \ldots x_n$ to Bob, who can reconstruct the state with perfect fidelity.

But now suppose that the letters are drawn from an ensemble of mutually orthogonal *mixed* states $\{\boldsymbol{\rho}_x, p_x\}$,

$$tr \boldsymbol{\rho}_x \boldsymbol{\rho}_y = 0 \text{ for } x \neq y; \tag{5.106}$$

that is, $\boldsymbol{\rho}_x$ and $\boldsymbol{\rho}_y$ have support on mutually orthogonal subspaces of the Hilbert space. These mixed states are also perfectly distinguishable, so again the messages are essentially classical, and therefore can be compressed to $H(X)$ qubits per letter. For example, we can extend the Hilbert space $\mathcal{H}_A$ of our letters to the larger space $\mathcal{H}_A \otimes \mathcal{H}_B$, and choose a purification of each $\boldsymbol{\rho}_x$, a pure state $|\varphi_x\rangle_{AB} \in \mathcal{H}_A \otimes \mathcal{H}_B$ such that

$$tr_B(|\varphi_x\rangle_{AB}\ {}_{AB}\langle\varphi_x|) = (\boldsymbol{\rho}_x)_A. \tag{5.107}$$

These pure states are mutually orthogonal, and the ensemble $\{|\varphi_x\rangle_{AB}, p_x\}$ has Von Neumann entropy $H(X)$; hence we may Schumacher compress a message

$$|\varphi_{x_1}\rangle_{AB} \cdots |\varphi_{x_n}\rangle_{AB}, \tag{5.108}$$

to $H(X)$ qubits per letter (asymptotically). Upon decompressing this state, Bob can perform the partial trace by "throwing away" subsystem B, and so reconstruct Alice's message.

To make a reasonable guess about what expression characterizes the compressibility of a message constructed from a mixed state alphabet, we might seek a formula that reduces to $S(\boldsymbol{\rho})$ for an ensemble of pure states, and to

$H(X)$ for an ensemble of mutually orthogonal mixed states. Choosing a basis in which

$$\boldsymbol{\rho} = \sum_x p_x \boldsymbol{\rho}_x, \tag{5.109}$$

is block diagonalized, we see that

$$\begin{aligned} S(\boldsymbol{\rho}) &= -\mathrm{tr}\boldsymbol{\rho}\log\boldsymbol{\rho} = -\sum_x \mathrm{tr}(p_x\boldsymbol{\rho}_x)\log(p_x\boldsymbol{\rho}_x) \\ &= -\sum_x p_x \log p_x - \sum_x p_x \mathrm{tr}\boldsymbol{\rho}_x \log\boldsymbol{\rho}_x \\ &= H(X) + \sum_x p_x S(\boldsymbol{\rho}_x), \end{aligned} \tag{5.110}$$

(recalling that $\mathrm{tr}\boldsymbol{\rho}_x = 1$ for each x). Therefore we may write the Shannon entropy as

$$H(X) = S(\boldsymbol{\rho}) - \sum_x p_x S(\boldsymbol{\rho}_x) \equiv \chi(\mathcal{E}). \tag{5.111}$$

The quantity $\chi(\mathcal{E})$ is called the *Holevo information* of the ensemble $\mathcal{E} = \{\boldsymbol{\rho}_x, p_x\}$. Evidently, it depends not just on the density matrix $\boldsymbol{\rho}$, but also on the particular way that $\boldsymbol{\rho}$ is realized as an ensemble of mixed states. We have found that, for either an ensemble of pure states, or for an ensemble of *mutually orthogonal* mixed states, the Holevo information $\chi(\mathcal{E})$ is the optimal number of qubits per letter that can be attained if we are to compress the messages while retaining good fidelity for large n.

The Holevo information can be regarded as a generalization of Von Neumann entropy, reducing to $S(\boldsymbol{\rho})$ for an ensemble of pure states. It also bears a close resemblance to the mutual information of classical information theory:

$$I(Y;X) = H(Y) - H(Y|X) \tag{5.112}$$

tells us how much, on the average, the Shannon entropy of Y is reduced once we learn the value of X; similarly,

$$\chi(\mathcal{E}) = S(\boldsymbol{\rho}) - \sum_x p_x S(\boldsymbol{\rho}_x) \tag{5.113}$$

tells us how much, on the average, the Von Neumann entropy of an ensemble is reduced when we know which preparation was chosen. Like the classical mutual information, the Holevo information is always nonnegative, as follows from the concavity property of $S(\boldsymbol{\rho})$,

$$S(\sum p_x \boldsymbol{\rho}_x) \geq \sum_x p_x S(\boldsymbol{\rho}_x). \tag{5.114}$$

Now we wish to explore the connection between the Holevo information and the compressibility of messages constructed from an alphabet of *nonorthogonal* mixed states. In fact, it can be shown that, in general, high-fidelity compression to less than χ qubits per letter is not possible.

To establish this result we use a "monotonicity" property of χ that was proved by Lindblad and by Uhlmann: A superoperator cannot increase the Holevo information. That is, if \$ is any superoperator, let it act on an ensemble of mixed states according to

$$\$: \mathcal{E} = \{\boldsymbol{\rho}_x, p_x\} \to \mathcal{E}' = \{\$(\boldsymbol{\rho}_x), p_x\}; \tag{5.115}$$

then

$$\chi(\mathcal{E}') \leq \chi(\mathcal{E}). \tag{5.116}$$

Lindblad–Uhlmann monotonicity is closely related to the strong subadditivity of the Von Neumann entropy, as you will show in a homework exercise.

The monotonicity of χ provides a further indication that χ quantifies an amount of information encoded in a quantum system. The decoherence described by a superoperator can only retain or reduce this quantity of information – it can never increase it. Note that, in contrast, the Von Neumann entropy is not monotonic. A superoperator might take an initial pure state to a mixed state, increasing $S(\boldsymbol{\rho})$. But another superoperator takes every mixed state to the "ground state" $|0\rangle\langle 0|$, and so reduces the entropy of an initial mixed state to zero. It would be misleading to interpret this reduction of S as an "information gain," in that our ability to distinguish the different possible preparations has been completely destroyed. Correspondingly, decay to the ground state reduces the Holevo information to zero, reflecting that we have lost the ability to reconstruct the initial state.

We now consider messages of n letters, each drawn independently from the ensemble $\mathcal{E} = \{\boldsymbol{\rho}_x, p_x\}$; the ensemble of all such input messages is denoted $\mathcal{E}^{(n)}$. A code is constructed that compresses the messages so that they all occupy a Hilbert space $\tilde{\mathcal{H}}^{(n)}$; the ensemble of compressed messages is denoted $\tilde{\mathcal{E}}^{(n)}$. Then decompression is performed with a superoperator \$,

$$\$: \tilde{\mathcal{E}}^{(n)} \to \mathcal{E}'^{(n)}, \tag{5.117}$$

to obtain an ensemble $\mathcal{E}'^{(n)}$ of output messages.

Now suppose that this coding scheme has high fidelity. To minimize technicalities, let us not specify in detail how the fidelity of $\mathcal{E}'^{(n)}$ relative to $\mathcal{E}^{(n)}$ should be quantified. Let us just accept that if $\mathcal{E}'^{(n)}$ has high fidelity, then for any δ and n sufficiently large

$$\frac{1}{n}\chi(\mathcal{E}^{(n)}) - \delta \le \frac{1}{n}\chi(\mathcal{E}'^{(n)}) \le \frac{1}{n}\chi(\mathcal{E}^{(n)}) + \delta; \tag{5.118}$$

the Holevo information per letter of the output approaches that of the input. Since the input messages are product states, it follows from the additivity of $S(\boldsymbol{\rho})$ that

$$\chi(\mathcal{E}^{(n)}) = n\chi(\mathcal{E}), \tag{5.119}$$

and we also know from Lindblad–Uhlmann monotonicity that

$$\chi(\mathcal{E}'^{(n)}) \le \chi(\tilde{\mathcal{E}}^{(n)}). \tag{5.120}$$

By combining eqs. (5.118)-(5.120), we find that

$$\frac{1}{n}\chi(\tilde{\mathcal{E}}^{(n)}) \ge \chi(\mathcal{E}) - \delta. \tag{5.121}$$

Finally, $\chi(\tilde{\mathcal{E}}^{(n)})$ is bounded above by $S(\tilde{\boldsymbol{\rho}}^{(n)})$, which is in turn bounded above by $\log \dim \tilde{\mathcal{H}}^{(n)}$. Since δ may be as small as we please, we conclude that, asymptotically as $n \to \infty$,

$$\frac{1}{n}\log(\dim \tilde{\mathcal{H}}^{(n)}) \ge \chi(\mathcal{E}); \tag{5.122}$$

high-fidelity compression to fewer than $\chi(\mathcal{E})$ qubits per letter is not possible.

One is sorely tempted to conjecture that compression to $\chi(\mathcal{E})$ qubits per letter is asymptotically attainable. As of mid-January, 1998, this conjecture still awaits proof or refutation.

5.4 Accessible Information

The close analogy between the Holevo information $\chi(\mathcal{E})$ and the classical mutual information $I(X;Y)$, as well as the monotonicity of χ, suggest that χ is related to the amount of *classical* information that can be stored in and recovered from a quantum system. In this section, we will make this connection precise.

The previous section was devoted to quantifying the *quantum* information content – measured in *qubits* – of messages constructed from an alphabet of quantum states. But now we will turn to a quite different topic. We want to quantify the *classical* information content – measured in bits – that can be extracted from such messages, particularly in the case where the alphabet includes letters that are not mutually orthogonal.

Now, why would we be so foolish as to store classical information in nonorthogonal quantum states that cannot be perfectly distinguished? Storing information this way should surely be avoided as it will degrade the classical signal. But perhaps we can't help it. For example, maybe I am a communications engineer, and I am interested in the intrinsic physical limitations on the classical capacity of a high bandwidth optical fiber. Clearly, to achieve a higher throughout of classical information per unit power, we should choose to encode information in single photons, and to attain a high rate, we should increase the number of photons transmitted per second. But if we squeeze photon wavepackets together tightly, the wavepackets will overlap, and so will not be perfectly distinguishable. How do we maximize the classical information transmitted in that case? As another important example, maybe I am an experimental physicist, and I want to use a delicate quantum system to construct a very sensitive instrument that measures a classical force acting on the system. We can model the force as a free parameter x in the system's Hamiltonian $\mathbf{H}(x)$. Depending on the value of x, the state of the system will evolve to various possible final (nonorthogonal) states $\boldsymbol{\rho}_x$. How much information about x can our apparatus acquire?

While physically this is a much different issue than the compressibility of quantum information, mathematically the two questions are related. We will find that the Von Neumann entropy and its generalization the Holevo information will play a central role in the discussion.

Suppose, for example, that Alice prepares a pure quantum state drawn from the ensemble $\mathcal{E} = \{|\varphi_x\rangle, p_x\}$. Bob knows the ensemble, but not the particular state that Alice chose. He wants to acquire as much information as possible about x.

Bob collects his information by performing a generalized measurement, the POVM $\{\mathbf{F}_y\}$. If Alice chose preparation x, Bob will obtain the measurement outcome y with conditional probability

$$p(y|x) = \langle \varphi_x | \mathbf{F}_y | \varphi_x \rangle. \tag{5.123}$$

These conditional probabilities, together with the ensemble X, determine the amount of information that Bob gains on the average, the mutual information $I(X;Y)$ of preparation and measurement outcome.

Bob is free to perform the measurement of his choice. The "best" possible measurement, that which maximizes his information gain, is called the *optimal measurement* determined by the ensemble. The maximal information gain is

$$\mathrm{Acc}(\mathcal{E}) = \underset{\{\mathbf{F}_y\}}{\mathrm{Max}} \, I(X;Y), \tag{5.124}$$

where the Max is over all POVM's. This quantity is called the *accessible information* of the ensemble $\mathcal{E}$.

Of course, if the states $|\varphi_x\rangle$ are mutually orthogonal, then they are perfectly distinguishable. The orthogonal measurement

$$\mathbf{E}_y = |\varphi_y\rangle\langle\varphi_y|, \tag{5.125}$$

has conditional probability

$$p(y|x) = \delta_{y,x}, \tag{5.126}$$

so that $H(X|Y) = 0$ and $I(X;Y) = H(X)$. This measurement is clearly optimal – the preparation is completely determined – so that

$$\mathrm{Acc}(\mathcal{E}) = H(X), \tag{5.127}$$

for an ensemble of mutually orthogonal (pure *or* mixed) states.

But the problem is much more interesting when the signal states are nonorthogonal pure states. In this case, no useful general formula for $\mathrm{Acc}(\mathcal{E})$ is known, but there is an upper bound

$$\mathrm{Acc}(\mathcal{E}) \leq S(\boldsymbol{\rho}). \tag{5.128}$$

We have seen that this bound is saturated in the case of orthogonal signal states, where $S(\boldsymbol{\rho}) = H(X)$. In general, we know from classical information theory that $I(X;Y) \leq H(X)$; but for nonorthogonal states we have $S(\boldsymbol{\rho}) <$

$H(X)$, so that eq. (5.128) is a better bound. Even so, this bound is not tight; in many cases $\mathrm{Acc}(\mathcal{E})$ is strictly less than $S(\boldsymbol{\rho})$.

We obtain a sharper relation between $\mathrm{Acc}(\mathcal{E})$ and $S(\boldsymbol{\rho})$ if we consider the accessible information per letter in a message containing n letters. Now Bob has more flexibility – he can choose to perform a collective measurement on all n letters, and thereby collect more information than if he were restricted to measuring only one letter at a time. Furthermore, Alice can choose to prepare, rather than arbitrary messages with each letter drawn from the ensemble $\mathcal{E}$, an ensemble of special messages (a code) designed to be maximally distinguishable.

We will then see that Alice and Bob can find a code such that the marginal ensemble for each letter is $\mathcal{E}$, and the accessible information per letter asymptotically approaches $S(\boldsymbol{\rho})$ as $n \to \infty$. In this sense, $S(\boldsymbol{\rho})$ characterizes the accessible information of an ensemble of *pure* quantum states.

Furthermore, these results generalize to ensembles of mixed quantum states, with the Holevo information replacing the Von Neumann entropy. The accessible information of an ensemble of mixed states $\{\boldsymbol{\rho}_x, p_x\}$ satisfies

$$\mathrm{Acc}(\mathcal{E}) \leq \chi(\mathcal{E}), \tag{5.129}$$

a result known as the *Holevo bound*. This bound is not tight in general (though it is saturated for ensembles of mutually orthogonal mixed states). However, if Alice and Bob choose an n-letter code, where the marginal ensemble for each letter is $\mathcal{E}$, and Bob performs an optimal POVM on all n letters collectively, then the best attainable accessible information per letter is $\chi(\mathcal{E})$ – *if* all code words are required to be *product* states. In this sense, $\chi(\mathcal{E})$ characterizes the accessible information of an ensemble of *mixed* quantum states.

One way that an alphabet of mixed quantum states might arise is that Alice might try to send pure quantum states to Bob through a noisy quantum channel. Due to decoherence in the channel, Bob receives mixed states that he must decode. In this case, then, $\chi(\mathcal{E})$ characterizes the maximal amount of classical information that can be transmitted to Bob through the noisy quantum channel.

For example, Alice might send to Bob n photons in certain polarization states. If we suppose that the noise acts on each photon independently, and that Alice sends unentangled states of the photons, then $\chi(\mathcal{E})$ is the maximal amount of information that Bob can acquire per photon. Since

$$\chi(\mathcal{E}) \leq S(\boldsymbol{\rho}) \leq 1, \tag{5.130}$$

it follows in particular that a single (unentangled) photon can carry at most one bit of classical information.

5.4.1 The Holevo Bound

The Holevo bound on the accessible information is not an easy theorem, but like many good things in quantum information theory, it follows easily once the strong subadditivity of Von Neumann entropy is established. Here we will assume strong subadditivity and show that the Holevo bound follows.

Recall the setting: Alice prepares a quantum state drawn from the ensemble $\mathcal{E} = \{\boldsymbol{\rho}_x, p_x\}$, and then Bob performs the POVM $\{\mathbf{F}_y\}$. The joint probability distribution governing Alice's preparation x and Bob's outcome y is

$$p(x,y) = p_x tr\{\mathbf{F}_y \boldsymbol{\rho}_x\}. \tag{5.131}$$

We want to show that

$$I(X;Y) \leq \chi(\mathcal{E}). \tag{5.132}$$

Since strong subadditivity is a property of three subsystems, we will need to identify three systems to apply it to. Our strategy will be to prepare an input system X that stores a classical record of what preparation was chosen and an output system Y whose classical correlations with x are governed by the joint probability distribution $p(x,y)$. Then applying strong subadditivity to X, Y, and our quantum system Q, we will be able to relate $I(X;Y)$ to $\chi(\mathcal{E})$.

Suppose that the initial state of the system XQY is

$$\boldsymbol{\rho}_{XQY} = \sum_x p_x |x\rangle\langle x| \otimes \boldsymbol{\rho}_x \otimes |0\rangle\langle 0|, \tag{5.133}$$

where the $|x\rangle$'s are mutually orthogonal pure states of the input system X, and $|0\rangle$ is a particular pure state of the output system Y. By performing partial traces, we see that

$$\begin{aligned}
\boldsymbol{\rho}_X &= \sum_x p_x |x\rangle\langle x| \rightarrow S(\boldsymbol{\rho}_X) = H(X) \\
\boldsymbol{\rho}_Q &= \sum_x p_x \boldsymbol{\rho}_x \equiv \boldsymbol{\rho} \rightarrow S(\boldsymbol{\rho}_{QY}) = S(\boldsymbol{\rho}_Q) = S(\boldsymbol{\rho}).
\end{aligned} \tag{5.134}$$

and since the $|x\rangle$'s are mutually orthogonal, we also have

$$\begin{aligned} S(\boldsymbol{\rho}_{XQY}) = S(\boldsymbol{\rho}_{XQ}) &= \sum_x -\mathrm{tr}(p_x \boldsymbol{\rho}_x \log p_x \boldsymbol{\rho}_x) \\ &= H(X) + \sum_x p_x S(\boldsymbol{\rho}_x). \end{aligned} \tag{5.135}$$

Now we will perform a unitary transformation that "imprints" Bob's measurement result in the output system Y. Let us suppose, for now, that Bob performs an orthogonal measurement $\{\mathbf{E}_y\}$, where

$$\mathbf{E}_y \mathbf{E}_{y'} = \delta_{y,y'} \mathbf{E}_y, \tag{5.136}$$

(we'll consider more general POVM's shortly). Our unitary transformation $\mathbf{U}_{QY}$ acts on QY according to

$$\mathbf{U}_{QY} : |\varphi\rangle_Q \otimes |0\rangle_Y = \sum_y \mathbf{E}_y |\varphi\rangle_Q \otimes |y\rangle_Y, \tag{5.137}$$

(where the $|y\rangle$'s are mutually orthogonal), and so transforms $\boldsymbol{\rho}_{XQY}$ as

$$\mathbf{U}_{QY} : \boldsymbol{\rho}_{XQY} \to \boldsymbol{\rho}'_{XQY} = \sum_{x,y,y'} p_x |x\rangle\langle x| \otimes \mathbf{E}_y \boldsymbol{\rho}_x \mathbf{E}_{y'} \otimes |y\rangle\langle y'|. \tag{5.138}$$

Since Von Neumann entropy is invariant under a unitary change of basis, we have

$$\begin{aligned} S(\boldsymbol{\rho}'_{XQY}) &= S(\boldsymbol{\rho}_{XQY}) = H(x) + \sum_x p_x S(\boldsymbol{\rho}_x), \\ S(\boldsymbol{\rho}'_{QY}) &= S(\boldsymbol{\rho}_{QY}) = S(\boldsymbol{\rho}), \end{aligned} \tag{5.139}$$

and taking a partial trace of eq. (5.138) we find

$$\begin{aligned} \boldsymbol{\rho}'_{XY} &= \sum_{x,y} p_x \mathrm{tr}(\mathbf{E}_y \boldsymbol{\rho}_x) |x\rangle\langle x| \otimes |y\rangle\langle y| \\ &= \sum_{x,y} p(x,y) |x,y\rangle\langle x,y| \to S(\boldsymbol{\rho}'_{XY}) = H(X,Y), \end{aligned} \tag{5.140}$$

(using eq. (5.136). Evidently it follows that

$$\boldsymbol{\rho}'_Y = \sum_y p(y)|y\rangle\langle y| \to S(\boldsymbol{\rho}'_Y) = H(Y). \tag{5.141}$$

Now we invoke strong subadditivity in the form

$$S(\boldsymbol{\rho}'_{XQY}) + S(\boldsymbol{\rho}'_Y) \leq S(\boldsymbol{\rho}'_{XY}) + S(\boldsymbol{\rho}'_{QY}), \tag{5.142}$$

which becomes

$$H(X) + \sum_x p_x S(\boldsymbol{\rho}_x) + H(Y) \leq H(X,Y) + S(\boldsymbol{\rho}), \tag{5.143}$$

or

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \leq S(\boldsymbol{\rho}) - \sum_x p_x S(\boldsymbol{\rho}_x) = \chi(\mathcal{E}). \tag{5.144}$$

This is the Holevo bound.

One way to treat more general POVM's is to enlarge the system by appending one more subsystem Z. We then construct a unitary $\mathbf{U}_{QYZ}$ acting as

$$\mathbf{U}_{QYZ} : |\varphi\rangle_Q \otimes |0\rangle_Y \otimes |0\rangle_Z = \sum_y \sqrt{\mathbf{F}_y}|\varphi\rangle_A \otimes |y\rangle_Y \otimes |y\rangle_Z, \tag{5.145}$$

so that

$$\boldsymbol{\rho}'_{XQYZ} = \sum_{x,y,y'} p_x |x\rangle\langle x| \otimes \sqrt{\mathbf{F}_y}\boldsymbol{\rho}_x\sqrt{\mathbf{F}_{y'}} \otimes |y\rangle\langle y'| \otimes |y\rangle\langle y'|. \tag{5.146}$$

Then the partial trace over Z yields

$$\boldsymbol{\rho}'_{XQY} = \sum_{x,y} p_x |x\rangle\langle x| \otimes \sqrt{\mathbf{F}_y}\boldsymbol{\rho}_x\sqrt{\mathbf{F}_y} \otimes |y\rangle\langle y|, \tag{5.147}$$

and

$$\begin{aligned}
\boldsymbol{\rho}'_{XY} &= \sum_{x,y} p_x tr(\mathbf{F}_y\boldsymbol{\rho}_x)|x\rangle\langle x| \otimes |y\rangle\langle y| \\
&= \sum_{x,y} p(x,y)|x,y\rangle\langle x,y| \\
&\to S(\boldsymbol{\rho}'_{XY}) = H(X,Y).
\end{aligned} \tag{5.148}$$

The rest of the argument then runs as before.

5.4.2 Improving distinguishability: the Peres–Wootters method

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

$$
\begin{aligned}
|\varphi_1\rangle &= |\uparrow_{\hat{n}_1}\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \\
|\varphi_2\rangle &= |\uparrow_{\hat{n}_2}\rangle = \begin{pmatrix} -\frac{1}{2} \\ \frac{\sqrt{3}}{2} \end{pmatrix}, \\
|\varphi_3\rangle &= |\uparrow_{\hat{n}_3}\rangle = \begin{pmatrix} -\frac{1}{2} \\ -\frac{\sqrt{3}}{2} \end{pmatrix};
\end{aligned} \tag{5.149}
$$

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:

$$
\langle\varphi_1|\varphi_2\rangle = \langle\varphi_1|\varphi_3\rangle = \langle\varphi_2|\varphi_3\rangle = -\frac{1}{2}. \tag{5.150}
$$

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

$$
\boldsymbol{\rho} = \frac{1}{3}(|\varphi_1\rangle\langle\varphi_1| + |\varphi_2\rangle\langle\varphi_3| + |\varphi_3\rangle\langle\varphi_3|) = \frac{1}{2}\mathbf{1}, \tag{5.151}
$$

which has $S(\boldsymbol{\rho}) = 1$. 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

$$
\mathbf{F}_{\bar{a}} = \frac{2}{3}(\mathbf{1} - |\varphi_a\rangle\langle\varphi_a|), \quad a = 1, 2, 3; \tag{5.152}
$$

we see that

$$
p(a|b) = \langle\varphi_b|\mathbf{F}_{\bar{a}}|\varphi_b\rangle = \begin{cases} 0 & a = b, \\ \frac{1}{2} & a \neq b. \end{cases} \tag{5.153}
$$

Therefore, 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

$$I = H(X) - H(X|Y) = \log_2 3 - 1 = .58496. \qquad (5.154)$$

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 $\mathbf{F}_a$'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. (5.152) 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.

$$\mathrm{Acc}(\mathcal{E}) = \log_2 \left(\frac{3}{2}\right) = .58496... \qquad (5.155)$$

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

$$|\varphi_a\rangle|\varphi_b\rangle, \quad a, b = 1, 2, 3, \qquad (5.156)$$

each with $p_{ab} = 1/9$. Then Bob's best strategy is to perform the POVM eq. (5.152) on each of the two qubits, achieving a mutual information of .58496 bits per qubit, as before.

But Alice and Bob are determined to do better. After discussing the problem with A. Peres and W. Wootters, they decide on a different strategy. Alice will prepare one of *three* two-qubit states

$$|\Phi_a\rangle = |\varphi_a\rangle|\varphi_a\rangle, \quad a = 1, 2, 3, \qquad (5.157)$$

each occurring with *a priori* probability $p_a = 1/2$. 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 a homework exercise, you will show that the density matrix

$$\boldsymbol{\rho} = \frac{1}{3}\left(\sum_{a=1}^{3} |\Phi_a\rangle\langle\Phi_a|\right), \qquad (5.158)$$

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

$$S(\boldsymbol{\rho}) = -\frac{1}{2}\log\frac{1}{2} - 2\left(\frac{1}{4}\log\frac{1}{4}\right) = \frac{3}{2}. \tag{5.159}$$

The Holevo bound requires that the accessible information *per qubit* is less than 3/4 bit. This would at least be 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

$$\langle\Phi_a|\Phi_b\rangle = \frac{1}{4}, \quad a \neq b, \tag{5.160}$$

instead of eq. (5.150). 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

$$\mathbf{G} = \sum_a |\tilde{\Phi}_a\rangle\langle\tilde{\Phi}_a|; \tag{5.161}$$

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

$$\mathbf{F}_a = \mathbf{G}^{-1/2}|\tilde{\Phi}_a\rangle\langle\tilde{\Phi}_a|\mathbf{G}^{-1/2}, \tag{5.162}$$

and we see that

$$\begin{aligned}\sum_a \mathbf{F}_a &= \mathbf{G}^{-1/2}\left(\sum_a |\tilde{\Phi}_a\rangle\langle\tilde{\Phi}_a|\right)\mathbf{G}^{-1/2} \\ &= \mathbf{G}^{-1/2}\mathbf{G}\mathbf{G}^{-1/2} = \mathbf{1}, \end{aligned} \tag{5.163}$$

on the span of the $|\tilde{\Phi}_a\rangle$'s. If necessary, we can augment these $\mathbf{F}_a$'s with one more positive operator, the projection $\mathbf{F}_0$ 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,

$$|\tilde{\Phi}_a\rangle = \sqrt{\lambda_a}|\Phi_a\rangle, \tag{5.164}$$

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

$$\begin{aligned}\mathbf{F}_a &= \sum_{a,b,c}(|\Phi_b\rangle\lambda_b^{-1/2}\langle\Phi_b|)(\lambda_a|\Phi_a\rangle\langle\Phi_a|)(|\Phi_c\rangle\lambda_c^{-1/2}\langle\Phi_c|)\\ &= |\Phi_a\rangle\langle\Phi_a|;\end{aligned} \tag{5.165}$$

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), but in that case the measurement may not be optimal.

In the homework, you'll construct the PGM for the vectors $|\Phi_a\rangle$ in eq. (5.157), and you'll show that

$$\begin{aligned}p(a|a) = \langle\Phi_a|\mathbf{F}_a|\Phi_a\rangle &= \frac{1}{3}\left(1+\frac{1}{\sqrt{2}}\right)^2 = .971405\\ p(b|a) = \langle\Phi_a|\mathbf{F}_b|\Phi_a\rangle &= \frac{1}{6}\left(1-\frac{1}{\sqrt{2}}\right)^2 = .0142977,\end{aligned} \tag{5.166}$$

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

$$H(X|Y) = .215893, \tag{5.167}$$

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

$$I = H(X) - H(X|Y) = 1.36907, \tag{5.168}$$

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 ($I < 1.5$ 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.

The PGM described here is "optimal" in the sense that it gives the best information gain of any *known* measurement. Most likely, this is really the highest I that can be achieved with *any* measurement, but I have not proved it.

5.4.3 Attaining Holevo: pure states

With these lessons in mind, we can proceed to show that, given an ensemble of pure states, we can construct n-letter codewords that asymptotically attain an accessible information of $S(\boldsymbol{\rho})$ per letter.

We must select a code, the ensemble of codewords that Alice can prepare, and a "decoding observable," the POVM that Bob will use to try to distinguish the codewords. Our task is to show that Alice can choose $2^{n(S-\delta)}$ codewords, such that Bob can determine which one was sent, with negligible probability of error as $n \to \infty$. We won't go through all the details of the argument, but will be content to understand why the result is highly plausible.

The main idea, of course, is to invoke random coding. Alice chooses product signal states

$$|\varphi_{x_1}\rangle|\varphi_{x_2}\rangle \ldots |\varphi_{x_n}\rangle, \tag{5.169}$$

by drawing each letter at random from the ensemble $\mathcal{E} = \{|\varphi_x\rangle, p_x\}$. As we have seen, for a typical code each typical codeword has a large overlap with a typical subspace $\Lambda^{(n)}$ that has dimension $\dim \Lambda^{(n)} > 2^{n(S(\boldsymbol{\rho})-\delta)}$. Furthermore, for a typical code, the marginal ensemble governing each letter is close to $\mathcal{E}$.

Because the typical subspace is very large for n large, Alice can choose many codewords, yet be assured that the typical overlap of two typical codewords is very small. Heuristically, the typical codewords are randomly distributed in the typical subspace, and on average, two random unit vectors in a space of dimension D have overlap $1/D$. Therefore if $|u\rangle$ and $|w\rangle$ are two codewords

$$\langle |\langle u|w\rangle|^2\rangle_\Lambda < 2^{-n(S-\delta)}. \tag{5.170}$$

Here $< \cdot >_\Lambda$ denotes an average over random typical codewords.

You can convince yourself that the typical codewords really are uniformly distributed in the typical subspace as follows: Averaged over the ensemble, the overlap of random codewords $|\varphi_{x_1}\rangle \ldots |\varphi_{x_n}\rangle$ and $|\varphi_{y_1}\rangle \ldots |\varphi_{y_n}\rangle$ is

$$\begin{aligned} &= \sum p_{x_1} \ldots p_{x_n} p_{y_1} \ldots p_{y_n} (|\langle \varphi_{x_1}|\varphi_{y_1}\rangle|^2 \ldots |\langle \varphi_{x_n}|\varphi_{y_n}\rangle|^2) \\ &= \mathrm{tr}(\boldsymbol{\rho} \otimes \ldots \otimes \boldsymbol{\rho})^2. \end{aligned} \tag{5.171}$$

Now suppose we restrict the trace to the typical subspace $\Lambda^{(n)}$; this space has $\dim \Lambda^{(n)} < 2^{n(S+\delta)}$ and the eigenvalues of $\boldsymbol{\rho}^{(n)} = \boldsymbol{\rho} \otimes \ldots \otimes \boldsymbol{\rho}$ restricted to $\Lambda^{(n)}$ satisfy $\lambda < 2^{-n(S-\delta)}$. Therefore

$$\langle |\langle u|w\rangle|^2\rangle_\Lambda = \mathrm{tr}_\Lambda[\boldsymbol{\rho}^{(n)}]^2 < 2^{n(S+\delta)}[2^{-n(S-\delta)}]^2 = 2^{-n(S-3\delta)}, \tag{5.172}$$

where tr_Λ denotes the trace in the typical subspace.

Now suppose that $2^{n(S-\delta)}$ random codewords $\{|u_i\rangle\}$ are selected. Then if $|u_j\rangle$ is any fixed codeword

$$\sum_{i\neq j} \langle |\langle u_i|u_j\rangle|^2\rangle < 2^{n(S-\delta)} 2^{-n(S-\delta')} + \varepsilon = 2^{-n(\delta-\delta')} + \varepsilon; \tag{5.173}$$

here the sum is over all codewords, and the average is no longer restricted to the typical codewords – the ε on the right-hand side arises from the atypical case. Now for any fixed δ, we can choose δ' and ε as small as we please for n sufficiently large; we conclude that when we average over both codes and codewords within a code, the codewords become highly distinguishable as $n \to \infty$.

Now we invoke some standard Shannonisms: Since eq. (5.173) holds when we average over codes, it also holds for a particular code. (Furthermore, since nearly all codes have the property that the marginal ensemble for each letter is close to $\mathcal{E}$, there is a code with this property satisfying eq. (5.173).) Now eq. (5.173) holds when we average over the particular codeword $|u_j\rangle$. But by throwing away at most half of the codewords, we can ensure that each and every codeword is highly distinguishable from all the others.

We see that Alice can choose $2^{n(S-\delta)}$ highly distinguishable codewords, which become mutually orthogonal as $n \to \infty$. Bob can perform a PGM at finite n that approaches an optimal orthogonal measurement as $n \to \infty$. Therefore the accessible information per letter

$$\frac{1}{n}\mathrm{Acc}(\tilde{\mathcal{E}}^{(n)}) = S(\boldsymbol{\rho}) - \delta, \tag{5.174}$$

is attainable, where $\tilde{\mathcal{E}}^{(n)}$ denotes Alice's ensemble of n-letter codewords.

Of course, for any finite n, Bob's POVM will be a complicated collective measurement performed on all n letters. To give an honest proof of attainability, we should analyze the POVM carefully, and bound its probability of error. This has been done by Hausladen, *et al.*[4] The handwaving argument here at least indicates why their conclusion is not surprising.

It also follows from the Holevo bound and the subadditivity of the entropy that the accessible information per letter cannot exceed $S(\boldsymbol{\rho})$ asymptotically. The Holevo bound tells us that

$$\mathrm{Acc}(\tilde{\mathcal{E}}^{(n)}) \leq S(\tilde{\boldsymbol{\rho}}^{(n)}), \tag{5.175}$$

where $\tilde{\boldsymbol{\rho}}^{(n)}$ denotes the density matrix of the codewords, and subadditivity implies that

$$S(\tilde{\boldsymbol{\rho}}^{(n)}) \leq \sum_{i=1}^{n} S(\tilde{\boldsymbol{\rho}}_i), \tag{5.176}$$

where $\tilde{\boldsymbol{\rho}}_i$ is the reduced density matrix of the ith letter. Since each $\tilde{\boldsymbol{\rho}}_i$ approaches $\boldsymbol{\rho}$ asymptotically, we have

$$\lim_{n\to\infty} \frac{1}{n}\mathrm{Acc}(\tilde{\mathcal{E}}^{(n)}) \leq \lim_{n\to\infty} \frac{1}{n} S(\tilde{\boldsymbol{\rho}}^{(n)}) \leq S(\boldsymbol{\rho}). \tag{5.177}$$

To derive this bound, we did not assume anything about the code, except that the marginal ensemble for each letter asymptotically approaches $\mathcal{E}$. In

[4] P. Hausladen, R. Jozsa, B. Schumacher, M. Westmoreland, and W. K. Wootters, "Classical information capacity of a quantum channel," *Phys. Rev.* A **54** (1996) 1869-1876.

particular the bound applies even if the codewords are entangled states rather than product states. Therefore we have shown that $S(\boldsymbol{\rho})$ is the optimal accessible information per letter.

We can define a kind of channel capacity associated with a specified alphabet of pure quantum states, the "fixed-alphabet capacity." We suppose that Alice is equipped with a source of quantum states. She can produce any one of the states $|\varphi_x\rangle$, but it is up to her to choose the *a priori* probabilities of these states. The fixed-alphabet capacity C_{f_a} is the maximum accessible information per letter she can achieve with the best possible distribution $\{p_x\}$. We have found that

$$C_{f_a} = \underset{\{p_x\}}{\text{Max}}\, S(\boldsymbol{\rho}). \tag{5.178}$$

C_{f_a} is the optimal number of classical bits we can encode per letter (asymptotically), given the specified quantum-state alphabet of the source.

5.4.4 Attaining Holevo: mixed states

Now we would like to extend the above reasoning to a more general context. We will consider n-letter messages, where the marginal ensemble for each letter is the ensemble of *mixed* quantum states

$$\mathcal{E} = \{\boldsymbol{\rho}_x, p_x\}. \tag{5.179}$$

We want to argue that it is possible (asymptotically as $n \to \infty$) to convey $\chi(\mathcal{E})$ bits of classical information per letter. Again, our task is to: (1) specify a code that Alice and Bob can use, where the ensemble of codewords yields the ensemble $\mathcal{E}$ letter by letter (at least asymptotically). (2) Specify Bob's decoding observable, the POVM he will use to attempt to distinguish the codewords. (3) Show that Bob's probability of error approaches zero as $n \to \infty$. As in our discussion of the pure-state case, I will not exhibit the complete proof (see Holevo[5] and Schumacher and Westmoreland[6]). Instead, I'll offer an argument (with even more handwaving than before, if that's possible) indicating that the conclusion is reasonable.

[5]A.S. Holevo, "The Capacity of the Quantum Channel with General Signal States," quant-ph/9611023

[6]B. Schumacher and M.D. Westmoreland, "Sending Classical Information Via Noisy Quantum Channels," *Phys. Rev.* **A 56** (1997) 131-138.

As always, we will demonstrate attainability by a random coding argument. Alice will select mixed-state codewords, with each letter drawn from the ensemble $\mathcal{E}$. That is, the codeword

$$\boldsymbol{\rho}_{x_1} \otimes \boldsymbol{\rho}_{x_2} \otimes \ldots \otimes \boldsymbol{\rho}_{x_n}, \tag{5.180}$$

is chosen with probability $p_{x_1} p_{x_2} \ldots p_{x_n}$. The idea is that *each* typical codeword can be regarded as an ensemble of pure states, with nearly all of its support on a certain typical subspace. If the typical subspaces of the various codewords have little overlap, then Bob will be able to perform a POVM that identifies the typical subspace characteristic of Alice's message, with small probability of error.

What is the dimension of the typical subspace of a typical codeword? If we *average* over the codewords, the mean entropy of a codeword is

$$\langle S^{(n)} \rangle = \sum_{x_1 \ldots x_n} p_{x_1} p_{x_2} \ldots p_{x_n} S(\boldsymbol{\rho}_{x_1} \otimes \boldsymbol{\rho}_{x_2} \otimes \ldots \otimes \boldsymbol{\rho}_{x_n}). \tag{5.181}$$

Using additivity of the entropy of a product state, and $\Sigma_x\, p_x = 1$, we obtain

$$\langle S^{(n)} \rangle = n \sum_x p_x S(\boldsymbol{\rho}_x) \equiv n \langle S \rangle. \tag{5.182}$$

For n large, the entropy of a codeword is, with high probability, close to this mean, and furthermore, the high probability eigenvalues of $\boldsymbol{\rho}_{x_1} \otimes \ldots \otimes \boldsymbol{\rho}_{x_2}$ are close to $2^{-n\langle S \rangle}$. In other words a typical $\boldsymbol{\rho}_{x_1} \otimes \ldots \otimes \boldsymbol{\rho}_{x_n}$ has its support on a typical subspace of dimension $2^{n\langle S \rangle}$.

This statement is closely analogous to the observation (crucial to the proof of Shannon's noisy channel coding theorem) that the number of typical messages received when a typical message is sent through a noisy classical channel is $2^{nH(Y|X)}$.

Now the argument follows a familiar road. For each typical message $x_1 x_2 \ldots x_n$, Bob can construct a "decoding subspace" of dimension $2^{n(\langle S \rangle + \delta)}$, with assurance that Alice's message is highly likely to have nearly all its support on this subspace. His POVM will be designed to determine in which decoding subspace Alice's message lies. Decoding errors will be unlikely if typical decoding subspaces have little overlap.

Although Bob is really interested only in the value of the decoding subspace (and hence $x_1 x_2 \ldots x_n$), let us suppose that he performs the complete PGM determined by all the vectors that span all the typical subspaces of

Alice's codewords. (And this PGM will approach an orthogonal measurement for large n, as long as the number of codewords is not too large.) He obtains a particular result which is likely to be in the typical subspace of dimension $2^{nS(\boldsymbol{\rho})}$ determined by the source $\boldsymbol{\rho} \otimes \boldsymbol{\rho} \otimes \ldots \otimes \boldsymbol{\rho}$, and furthermore, is likely to be in the decoding subspace of the message that Alice actually sent. Since Bob's measurement results are uniformly distributed in a space on dimension 2^{nS}, and the pure-state ensemble determined by a particular decoding subspace has dimension $2^{n(\langle S \rangle + \delta)}$, the average overlap of the vector determined by Bob's result with a typical decoding subspace is

$$\frac{2^{n(\langle S \rangle + \delta)}}{2^{nS}} = 2^{-n(S - \langle S \rangle - \delta)} = 2^{-n(\chi - \delta)}. \tag{5.183}$$

If Alice chooses 2^{nR} codewords, the average probability of a decoding error will be

$$2^{nR} 2^{-n(\chi - \delta)} = 2^{-n(\chi - R - \delta)}. \tag{5.184}$$

We can choose any R less than χ, and this error probability will get very small as $n \to \infty$.

This argument shows that the probability of error is small, averaged over both random codes and codewords. As usual, we can choose a particular code, and throw away some codewords to achieve a small probability of error for every codeword. Furthermore, the particular code may be chosen to be typical, so that the marginal ensemble for each codeword approaches $\mathcal{E}$ as $n \to \infty$. We conclude that an accessible information of χ per letter is asymptotically attainable.

The structure of the argument closely follows that for the corresponding classical coding theorem. In particular, the quantity χ arose much as I does in Shannon's theorem. While 2^{-nI} is the probability that a particular typical sequence lies in a specified decoding sphere, $2^{-n\chi}$ is the overlap of a particular typical state with a specified decoding subspace.

5.4.5 Channel capacity

Combining the Holevo bound with the conclusion that χ bits per letter is attainable, we obtain an expression for the *classical* capacity of a quantum channel (But with a caveat: we are assured that this "capacity" cannot be exceeded only if we disallow entangled codewords.)

Alice will prepare n-letter messages and send them through a noisy quantum channel to Bob. The channel is described by a superoperator, and we will assume that the same superoperator \$ acts on each letter independently (*memoryless* quantum channel). Bob performs the POVM that optimizes his information going about what Alice prepared.

It will turn out, in fact, that Alice is best off preparing pure-state messages (this follows from the subadditivity of the entropy). If a particular letter is prepared as the pure state $|\varphi_x\rangle$, Bob will receive

$$|\varphi_x\rangle\langle\varphi_x| \to \$(|\varphi_x\rangle\langle\varphi_x|) \equiv \boldsymbol{\rho}_x. \tag{5.185}$$

And if Alice sends the pure state $|\varphi_{x_1}\rangle \ldots |\varphi_{x_n}\rangle$, Bob receives the mixed state $\boldsymbol{\rho}_{x_1} \otimes \ldots \otimes \boldsymbol{\rho}_{x_n}$. Thus, the ensemble of Alice's codewords determines as ensemble $\tilde{\mathcal{E}}^{(n)}$ of mixed states received by Bob. Hence Bob's optimal information gain is by definition $\mathrm{Acc}(\tilde{\mathcal{E}}^{(n)})$, which satisfies the Holevo bound

$$\mathrm{Acc}(\tilde{\mathcal{E}}^{(n)}) \le \chi(\tilde{\mathcal{E}}^{(n)}). \tag{5.186}$$

Now Bob's ensemble is

$$\{\boldsymbol{\rho}_{x_1} \otimes \ldots \otimes \boldsymbol{\rho}_{x_n},\ p(x_1, x_2, \ldots, x_n)\}, \tag{5.187}$$

where $p(x_1, x_2 \ldots, x_n)$ is a completely arbitrary probability distribution on Alice's codewords. Let us calculate χ for this ensemble. We note that

$$\begin{aligned}
&\sum_{x_1 \ldots x_n} p(x_1, x_2, \ldots, x_n) S(\boldsymbol{\rho}_{x_1} \otimes \ldots \otimes \boldsymbol{\rho}_{x_n}) \\
&= \sum_{x_1 \ldots x_n} p(x_1, x_2, \ldots, x_n) \left[S(\boldsymbol{\rho}_{x_1}) + S(\boldsymbol{\rho}_{x_2}) + \ldots + S(\boldsymbol{\rho}_{x_n}) \right] \\
&= \sum_{x_1} p_1(x_1) S(\boldsymbol{\rho}_{x_1}) + \sum_{x_2} p_2(x_2) S(\boldsymbol{\rho}_{x_2}) + \ldots + \sum_{x_n} p_n(x_n) S(\boldsymbol{\rho}_{x_n}),
\end{aligned} \tag{5.188}$$

where, e.g., $p_1(x_1) = \sum_{x_2 \ldots x_n} p(x_1, x_2, \ldots, x_n)$ is the marginal probability distribution for the first letter. Furthermore, from subadditivity we have

$$S(\tilde{\boldsymbol{\rho}}^{(n)}) \le S(\tilde{\boldsymbol{\rho}}_1) + S(\tilde{\boldsymbol{\rho}}_2) + \ldots + S(\tilde{\boldsymbol{\rho}}_n), \tag{5.189}$$

where $\tilde{\boldsymbol{\rho}}_i$ is the reduced density matrix for the ith letter. Combining eq. (5.188) and eq. (5.189) we find that

$$\chi(\tilde{\mathcal{E}}^{(n)}) \le \chi(\tilde{\mathcal{E}}_1) + \ldots + \chi(\tilde{\mathcal{E}}_n), \tag{5.190}$$

where $\tilde{\mathcal{E}}_i$ is the marginal ensemble governing the ith letter that Bob receives. Eq. (5.190) applies to any ensemble of product states.

Now, for the channel described by the superoperator \$, we define the product-state *channel capacity*

$$C(\$) = \max_{\mathcal{E}} \chi(\$(\mathcal{E})). \tag{5.191}$$

Therefore, $\chi(\tilde{\mathcal{E}}_i) \le C$ for each term in eq. (5.190) and we obtain

$$\chi(\tilde{\mathcal{E}}^{(n)}) \le nC, \tag{5.192}$$

where $\tilde{\mathcal{E}}^{(n)}$ is any ensemble of product states. In particular, we infer from the Holevo bound that Bob's information gain is bounded above by nC. But we have seen that $\chi(\$(\mathcal{E}))$ bits per letter can be attained asymptotically for any $\mathcal{E}$, with the right choice of code and decoding observable. Therefore, C is the optimal number of bits per letter that can be sent through the noisy channel with negligible error probability, *if* the messages that Alice prepares are required to be product states.

We have left open the possibility that the product-state capacity $C(\$)$ might be exceeded if Alice is permitted to prepare *entangled* states of her n letters. It is not known (in January, 1998) whether there are quantum channels for which a higher rate can be attained by using entangled messages. This is one of the many interesting open questions in quantum information theory.

5.5 Entanglement Concentration

Before leaving our survey of quantum information theory, we will visit one more topic where Von Neumann entropy plays a central role: quantifying entanglement.

Consider two bipartite pure states. One is a *maximally* entangled state of two qubits

$$|\phi^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle). \tag{5.193}$$

The other is a *partially* entangled state of two *qutrits*

$$|\Psi\rangle = \frac{1}{\sqrt{2}}|00\rangle + \frac{1}{2}|11\rangle + \frac{1}{2}|22\rangle. \tag{5.194}$$

which state 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 placing all bipartite states on a continuum, of ordering them according to their degree of entanglement? Can we compare a pair of qutrits with a pair of qubits any more than we can compare an apple and an orange?

A crucial feature of entanglement is that it cannot be created by local operations. In particular, if Alice and Bob share a bipartite pure state, they cannot increase its Schmidt number by any local operations – any unitary transformation or POVM performed by Alice or Bob, even if Alice and Bob exchange classical messages about their actions and measurement outcomes. So a number used to quantify entanglement ought to have the property that local operations do not increase it. An obvious candidate is the Schmidt number, but on reflection it does not seem very satisfactory. Consider

$$|\Psi_\varepsilon\rangle = \sqrt{1-2|\varepsilon|^2}|00\rangle + \varepsilon|11\rangle + \varepsilon|22\rangle, \tag{5.195}$$

which has Schmidt number 3 for any $|\varepsilon| > 0$. Should we really 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. It seems clear that $|\Psi_\varepsilon\rangle$ (for $|\varepsilon| \ll 1$) is a less valuable resource than $|\varphi^+\rangle$.

It turns out, though, that there is a natural and sensible way to quantify the entanglement of any bipartite pure state. To compare two states, we perform local operations to change their entanglement to a common currency that can be compared directly. The common currency is a maximally entangled state.

A precise statement about interchangeability (via local operations) of various forms of entanglement will necessarily be an *asymptotic* statement. That is, to precisely quantify the entanglement of a particular bipartite pure state, $|\psi\rangle_{AB}$, let us imagine that we wish to prepare n identical copies of that state. We have available a large supply of maximally entangled *Bell pairs* shared by Alice and Bob. Alice and Bob are to use k of the Bell pairs $\left((|\phi^+\rangle_{AB})^k\right)$, and with local operations and classical communication, to prepare n copies of the desired state $((|\psi\rangle_{AB})^n)$. What is the minimum number $k_{\rm min}$ of Bell pairs with which they can perform this task?

And now suppose that n copies of $|\psi\rangle_{AB}$ have already been prepared. Alice and Bob are to perform local operations that will transform the entanglement of $(|\psi\rangle_{AB})^n$ back to the standard form; that is, they are to extract

k' Bell pairs $\left((|\phi^+\rangle_{AB})^{k'}\right)$. What is the maximum number $k'_{\rm max}$ of Bell pairs that can be extracted (locally) from $(|\psi\rangle_{AB})^n$?

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

$$k'_{\rm max} \le k_{\rm min}. \tag{5.196}$$

But we can show that

$$\lim_{n\to\infty} \frac{k_{\rm min}}{n} = \lim_{n\to\infty} \frac{k'_{\rm max}}{n} \equiv E(|\psi\rangle_{AB}). \tag{5.197}$$

In this sense, then, locally transforming n copies of the bipartite pure state $|\psi\rangle_{AB}$ into k' maximally entangled pairs is an asymptotically *reversible* process. Since n copies of $|\psi\rangle_{AB}$ can be exchanged for k Bell pairs and vice versa, we see that $\frac{k}{n}$ Bell pairs unambiguously characterizes the amount of entanglement carried by the state $|\psi\rangle_{AB}$. We will call the ratio k/n (in the $n\to\infty$ limit) the *entanglement* E of $|\psi\rangle_{AB}$. The quantity E measures both what we need to pay (in Bell pairs) to create $|\psi\rangle_{AB}$, and the value of $|\psi\rangle_{AB}$ as a resource (e.g., the number of qubits that can be faithfully teleported using $|\psi\rangle_{AB}$).

Now, given a particular pure state $|\psi\rangle_{AB}$, what is the value of E? Can you guess the answer? It is

$$E = S(\boldsymbol{\rho}_A) = S(\boldsymbol{\rho}_B); \tag{5.198}$$

the entanglement is the Von Neumann entropy of Alice's density matrix $\boldsymbol{\rho}_A$ (or Bob's density matrix $\boldsymbol{\rho}_B$). 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 $\frac{1}{2}\mathbf{1}$ for each qubit in Alice's possession

$$\boldsymbol{\rho}_A = \frac{1}{2}\mathbf{1} \otimes \frac{1}{2}\mathbf{1} \otimes \ldots \otimes \frac{1}{2}\mathbf{1}, \tag{5.199}$$

and

$$S(\boldsymbol{\rho}_A) = kS\left(\frac{1}{2}\mathbf{1}\right) = k. \tag{5.200}$$

We must now see why $E = S(\boldsymbol{\rho}_A)$ is the right answer for any bipartite pure state.

First we want to show that if Alice and Bob share $k = n(S(\boldsymbol{\rho}_A) + \delta)$ Bell pairs, than they can (by local operations) prepare $(|\psi\rangle_{AB})^n$ with high fidelity. They may perform this task by combining quantum teleportation with Schumacher compression. First, by locally manipulating a bipartite system AC that is under her control, Alice constructs (n copies of) the state $|\psi\rangle_{AC}$. Thus, we may regard the state of system C as a pure state drawn from an ensemble described by $\boldsymbol{\rho}_C$, where $S(\boldsymbol{\rho}_C) = S(\boldsymbol{\rho}_A)$. Next Alice performs Schumacher compression on her n copies of C, retaining good fidelity while squeezing the typical states in $(\mathcal{H}_C)^n$ down to a space $\tilde{\mathcal{H}}_C^{(n)}$) with

$$\dim \tilde{\mathcal{H}}_C^{(n)} = 2^{n(S(\boldsymbol{\rho}_A)+\delta)}. \tag{5.201}$$

Now Alice and Bob can use the $n(S(\boldsymbol{\rho}_A)+\delta)$ Bell pairs they share to teleport the compressed state from Alice's $\tilde{\mathcal{H}}_C^{(n)}$ to Bob's $\tilde{\mathcal{H}}_B^{(n)}$. The teleportation, which in principle has perfect fidelity, requires only local operations and classical communication, if Alice and Bob share the required number of Bell pairs. Finally, Bob Schumacher decompresses the state he receives; then Alice and Bob share $(|\psi\rangle_{AB})^n$ (with arbitrarily good fidelity as $n \to \infty$).

Let us now suppose that Alice and Bob have prepared the state $(|\psi\rangle_{AB})^n$. 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 by 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

$$k' = n(S(\boldsymbol{\rho}_A) - \delta) \tag{5.202}$$

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

Since we know that Alice and Bob are not able to create entanglement locally, they can't turn k Bell pairs into $k' > k$ pairs through local operations, at least not with high fidelity and success probability. It follows then that $nS(\boldsymbol{\rho}_A)$ is the minimum number of Bell pairs needed to create n copies of $|\psi\rangle_{AB}$, and that $nS(\boldsymbol{\rho}_A)$ is the maximal number of Bell pairs that can be distilled from n copies of $|\psi\rangle_{AB}$. If we could create $|\psi\rangle_{AB}$ from Bell pairs more efficiently, or we could distill Bell pairs from $|\psi\rangle_{AB}$ more efficiently, then we would have a way for Alice and Bob to increase their supply of Bell pairs with local operations, a known impossibility. Therefore, if we can find a way to distill $k' = n(S(\boldsymbol{\rho}_A) - \delta)$ Bell pairs from n copies of $|\psi\rangle_{AB}$, we know that $E = S(\boldsymbol{\rho}_A)$.

To illustrate the concentration of entanglement, imagine that Alice and Bob have n copies of the partially entangled pure state of two qubits

$$|\psi(\theta)\rangle_{AB} = \cos\theta|00\rangle + \sin\theta|11\rangle. \tag{5.203}$$

(Any bipartite pure state of two qubits can be written this way, if we adopt the Schmidt basis and a suitable phase convention.) That is, Alice and Bob share the state

$$(|\psi(\theta)\rangle)^n = (\cos\theta|00\rangle + \sin\theta|11\rangle)^n. \tag{5.204}$$

Now let Alice (or Bob) perform a local measurement on her (his) n qubits. Alice measures the *total* spin of her n qubits along the z-axis

$$\boldsymbol{\sigma}_{3,A}^{(\text{total})} = \sum_{i=1}^{n} \boldsymbol{\sigma}_{3,A}^{(i)}. \tag{5.205}$$

A crucial feature of this measurement is its "*fuzziness.*" The observable $\boldsymbol{\sigma}_{3,A}^{(\text{total})}$ is highly *degenerate*; Alice projects the state of her n spins onto one of the large eigenspaces of this observable. She does not measure the spin of any single qubit; in fact, she is very careful not to acquire any information other than the value of $\boldsymbol{\sigma}_{3,A}^{(\text{total})}$, or equivalently, the number of up spins.

If we expand eq. (5.204), we find altogether 2^n terms. Of these, there are $\binom{n}{m}$ terms in which exactly m of the qubits that Alice holds have the value 1. And each of these terms has a coefficient $(\cos\theta)^{n-m}(\sin\theta)^m$. Thus, the probability that Alice's measurement reveals that m spins are "up" is

$$P(m) = \binom{n}{m}(\cos^2\theta)^{n-m}(\sin^2\theta)^m. \tag{5.206}$$

Furthermore, if she obtains this outcome, then her measurement has prepared an *equally weighted* superposition of all $\binom{n}{m}$ states that have m up spins. (Of course, since Alice's and Bob's spins are perfectly correlated, if Bob were to measure $\sigma_{3,B}^{(\text{total})}$, he would find exactly the same result as Alice. Alternatively, Alice could report her result to Bob in a classical message, and so save Bob the trouble of doing the measurement himself.) No matter what the measurement result, Alice and Bob now share a new state $|\psi'\rangle_{AB}$ such that all the nonzero eigenvalues of $\boldsymbol{\rho}'_A$ (and $\boldsymbol{\rho}'_B$) are equal.

For n large, the probability distribution $P(m)$ in eq. (5.206) peaks sharply – the probability is close to 1 that m/n is close to $\sin^2\theta$ and that

$$\binom{n}{m} \sim \binom{n}{n\sin^2\theta} \sim 2^{nH(\sin^2\theta)}, \tag{5.207}$$

where $H(p) = -p\log p - (1-p)\log(1-p)$ is the entropy function. That is, with probability greater than $1-\varepsilon$, the entangled state now shared by Alice and Bob has a Schmidt number $\binom{n}{m}$ with

$$2^{n(H(\sin^2\theta)-\delta)} < \binom{n}{m} < 2^{n(H(\sin^2\theta)+\delta)}. \tag{5.208}$$

Now Alice and Bob want to convert their shared entanglement to standard ($|\phi^+\rangle$) Bell pairs. If the Schmidt number of their shared maximally entangled state happened to be a power of 2, this would be easy. Both Alice and Bob could perform a unitary transformation that would rotate the 2^k-dimensional support of her/his density matrix to the Hilbert space of k-qubits, and then they could discard the rest of their qubits. The k pairs that they retain would then be maximally entangled.

Of course $\binom{n}{m}$ need not be close to a power of 2. But if Alice and Bob share many batches of n copies of the partially entangled state, they can concentrate the entanglement in each batch. After operating on ℓ batches, they will have obtained a maximally entangled state with Schmidt number

$$N_{\text{Schm}} = \binom{n}{m_1}\binom{n}{m_2}\binom{n}{m_3}\cdots\binom{n}{m_\ell}, \tag{5.209}$$

where each m_i is typically close to $n\sin^2\theta$. For any $\varepsilon > 0$, this Schmidt number will eventually, for some ℓ, be close to a power of 2,

$$2^{k_\ell} \le N_{\text{Schm}} < 2^{k_\ell}(1+\varepsilon). \tag{5.210}$$

At that point, either Alice or Bob can perform a measurement that attempts to project the support of dimension $2^{k_\ell}(1+\varepsilon)$ of her/his density matrix to a subspace of dimension 2^{k_ℓ}, succeeding with probability $1-\varepsilon$. *Then* they rotate the support to the Hilbert space of k_ℓ qubits, and discard the rest of their qubits. Typically, k_ℓ is close to $n\ell H(\sin^2\theta)$, so that they distill about $H(\sin^2\theta)$ maximally entangled pairs from each partially entangled state, with a success probability close to 1.

Of course, though the number m of up spins that Alice (or Bob) finds in her (his) measurement is typically close to $n\sin^2\theta$, it can fluctuate about this value. Sometimes Alice and Bob will be lucky, and then will manage to distill more than $H(\sin^2\theta)$ Bell pairs per copy of $|\psi(\theta)\rangle_{AB}$. But the probability of doing substantially better becomes negligible as $n\to\infty$.

These considerations easily generalize to bipartite pure states in larger Hilbert spaces. A bipartite pure state with Schmidt number s can be expressed, in the Schmidt basis, as

$$|\psi(a_1, a_2, \ldots, a_s)\rangle_{AB} = a_1|11\rangle + a_2|22\rangle + \ldots + a_s|ss\rangle. \tag{5.211}$$

Then in the state $(|\psi\rangle_{AB})^n$, Alice (or Bob) can measure the total number of $|1\rangle$'s, the total number of $|2\rangle$'s, etc. in her (his) possession. If she finds $m_1|1\rangle$'s, $m_2|2\rangle$'s, etc., then her measurement prepares a maximally entangled state with Schmidt number

$$N_{\text{Schm}} = \frac{n!}{(m_1)!(m_2)!\cdots(m_s)!}. \tag{5.212}$$

For m large, Alice will typically find

$$m_i \sim |a_i|^2 n, \tag{5.213}$$

and therefore

$$N_{\text{Sch}} \sim 2^{nH}, \tag{5.214}$$

where

$$H = \sum_i -|a_i|^2 \log |a_i|^2 = S(\boldsymbol{\rho}_A). \tag{5.215}$$

Thus, asymptotically for $n \to \infty$, close to $nS(\boldsymbol{\rho}_A)$ Bell pairs can be distilled from n copies of $|\psi\rangle_{AB}$.

5.5.1 Mixed-state entanglement

We have found a well-motivated and unambiguous way to quantify the entanglement of a bipartite pure state $|\psi\rangle_{AB}$: $E = S(\boldsymbol{\rho}_A)$, where

$$\boldsymbol{\rho}_A = \text{tr}_B(|\psi\rangle_{AB}\ {}_{AB}\langle\psi|). \tag{5.216}$$

It is also of considerable interest to quantify the entanglement of bipartite mixed states. Unfortunately, mixed-state entanglement is not nearly as well understood as pure-state entanglement, and is the topic of much current research.

Suppose that $\boldsymbol{\rho}_{AB}$ is a mixed state shared by Alice and Bob, and that they have n identical copies of this state. And suppose that, asymptotically as $n \to \infty$, Alice and Bob can prepare $(\boldsymbol{\rho}_{AB})^n$, with good fidelity and high success probability, from k Bell pairs using local operations and classical communication. We define the *entanglement of formation* F of $\boldsymbol{\rho}_{AB}$ as

$$F(\boldsymbol{\rho}_{AB}) = \lim_{n\to\infty} \frac{k_{\min}}{n}. \tag{5.217}$$

Further, suppose that Alice and Bob can use local operations and classical communication to distill k' Bell pairs from n copies of $\boldsymbol{\rho}_{AB}$. We define the *entanglement of distillation* D of $\boldsymbol{\rho}_{AB}$ as

$$D(\boldsymbol{\rho}_{AB}) = \lim_{n\to\infty} \frac{k'_{\max}}{n}. \tag{5.218}$$

For pure states, we found $D = E = F$. But for mixed states, no explicit general formulas for D or F are known. Since entanglement cannot be created locally, we know that $D \leq F$, but it is not known (in January, 1998) whether $D = F$. However, one strongly suspects that, for mixed states, $D < F$. To prepare the mixed state $(\boldsymbol{\rho}_{AB})^n$ from the pure state $(|\phi^+\rangle_{AB\ AB}\langle\phi^+|)^k$, we must discard some quantum information. It would be quite surprising if this process turned out to be (asymptotically) reversible.

It is useful to distinguish two different types of entanglement of distillation. D_1 denotes the number of Bell pairs that can be distilled if only one-way classical communication is allowed (e.g., Alice can *send* messages to Bob but she cannot *receive* messages from Bob). $D_2 = D$ denotes the entanglement of distillation if the classical communication is unrestricted. It is known that $D_1 < D_2$, and hence that $D_1 < F$ for some mixed states (while $D_1 = D_2 = F$ for pure states).

One reason for the interest in mixed-state entanglement (and in D_1 in particular) is a connection with the transmission of quantum information through noisy quantum channels. If a quantum channel described by a superoperator \$ is not *too* noisy, then we can construct an n-letter block code such that quantum information can be encoded, sent through the channel $(\$)^n$, decoded, and recovered with arbitrarily good fidelity as $n \to \infty$. The optimal number of encoded qubits per letter that can be transmitted through the channel is called the quantum channel capacity $C(\$)$. It turns out that $C(\$)$ can be related to D_1 of a particular mixed state associated with the channel — but we will postpone further discussion of the quantum channel capacity until later.

5.6 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 $n \to \infty$.

Mutual information and classical channel capacity. The *mutual information* $I(X;Y) = H(X) + H(Y) - H(X,Y)$ quantifies how ensembles 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. The capacity of a memoryless noisy classical communication channel is $C = \max_{\{p(x)\}} I(X;Y)$. This is the highest number of bits per letter that can be transmitted through the channel (using the best possible code) with negligible error probability as $n \to \infty$.

Von Neumann entropy, Holevo information, and quantum data compression. The *Von Neumann entropy* of a density matrix $\boldsymbol{\rho}$ is

$$S(\boldsymbol{\rho}) = -\mathrm{tr}\boldsymbol{\rho} \log \boldsymbol{\rho}, \tag{5.219}$$

and the *Holevo information* of an ensemble $\mathcal{E} = \{\boldsymbol{\rho}_x, p_x\}$ of quantum states is

$$\chi(\mathcal{E}) = S(\sum_x p_x \boldsymbol{\rho}_x) - \sum_x p_x S(\boldsymbol{\rho}_x). \tag{5.220}$$

The Von Neumann entropy 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 $S(\boldsymbol{\rho})$ qubits per letter (and no further), yet can still be decoded with arbitrarily good fidelity as $n \to \infty$. If the letters are drawn from the ensemble $\mathcal{E}$ of mixed quantum states, then high-fidelity compression to fewer than $\chi(\mathcal{E})$ qubits per letter is not possible.

Accessible information. 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 information of the ensemble. An n-letter code can be constructed such that the marginal ensemble for each letter is close to $\mathcal{E}$, and the accessible information per letter is close to $\chi(\mathcal{E})$. The product-state capacity of a quantum channel \$ is

$$C(\$) = \max_{\mathcal{E}} \chi(\$(\mathcal{E})). \tag{5.221}$$

This is the highest number of classical bits per letter than can be transmitted through the quantum channel, with negligible error probability as $n \to \infty$, assuming that each codeword is a tensor product of letter states.

Entanglement concentration. The *entanglement* E of a bipartite pure state $|\psi\rangle_{AB}$ is $E = S(\boldsymbol{\rho}_A)$ where $\boldsymbol{\rho}_A = \mathrm{tr}_B(|\psi\rangle_{AB}\ {}_{AB}\langle\psi|)$. 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 Bells pairs (but not more) from n copies of $|\psi\rangle_{AB}$ (asymptotically as $n \to \infty$).

5.7 Exercises

5.1 Distinguishing nonorthogonal states.

Alice has prepared a single qubit in one of the two (nonorthogonal) states

$$|u\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \ |v\rangle = \begin{pmatrix} \cos\frac{\theta}{2} \\ \sin\frac{\theta}{2} \end{pmatrix}, \tag{5.222}$$

where $0 < \theta < \pi$. Bob knows the value of θ, but he has no idea whether Alice prepared $|u\rangle$ or $|v\rangle$, and he is to perform a measurement to learn what he can about Alice's preparation.

Bob considers three possible measurements:

a) An orthogonal measurement with

$$\mathbf{E}_1 = |u\rangle\langle u|, \quad \mathbf{E}_2 = \mathbf{1} - |u\rangle\langle u|. \tag{5.223}$$

(In this case, if Bob obtains outcome 2, he knows that Alice must have prepared $|v\rangle$.)

b) A three-outcome POVM with

$$\mathbf{F}_1 = A(\mathbf{1} - |u\rangle\langle u|), \quad \mathbf{F}_2 = A(\mathbf{1} - |v\rangle\langle v|)$$

$$\mathbf{F}_3 = (1 - 2A)\mathbf{1} + A(|u\rangle\langle u| + |v\rangle\langle v|), \tag{5.224}$$

where A has the largest value consistent with positivity of $\mathbf{F}_3$. (In this case, Bob determines the preparation unambiguously if he obtains outcomes 1 or 2, but learns nothing from outcome 3.)

c) An orthogonal measurement with

$$\mathbf{E}_1 = |w\rangle\langle w|, \quad \mathbf{E}_2 = \mathbf{1} - |w\rangle\langle w|, \tag{5.225}$$

where

$$|w\rangle = \begin{pmatrix} \cos\left[\frac{1}{2}\left(\frac{\theta}{2} + \frac{\pi}{2}\right)\right] \\ \sin\left[\frac{1}{2}\left(\frac{\theta}{2} + \frac{\pi}{2}\right)\right] \end{pmatrix}. \tag{5.226}$$

(In this case $\mathbf{E}_1$ and $\mathbf{E}_2$ are projections onto the spin states that are oriented in the $x - z$ plane normal to the axis that bisects the orientations of $|u\rangle$ and $|v\rangle$.)

Find Bob's average information gain $I(\theta)$ (the mutual information of the preparation and the measurement outcome) in all three cases, and plot all three as a function of θ. Which measurement should Bob choose?

5.2 Relative entropy.

The *relative entropy* $S(\boldsymbol{\rho}|\boldsymbol{\sigma})$ of two density matrices $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$ is defined by

$$S(\boldsymbol{\rho}|\boldsymbol{\sigma}) = \mathrm{tr}\boldsymbol{\rho}(\log\boldsymbol{\rho} - \log\boldsymbol{\sigma}). \tag{5.227}$$

You will show that $S(\boldsymbol{\rho}|\boldsymbol{\sigma})$ is nonnegative, and derive some consequences of this property.

a) A differentiable real-valued function of a real variable is *concave* if

$$f(y) - f(x) \le (y - x)f'(x), \tag{5.228}$$

for all x and y. Show that if $\boldsymbol{a}$ and $\boldsymbol{b}$ are observables, and f is concave, then

$$\mathrm{tr}(f(\boldsymbol{b}) - f(\boldsymbol{a})) \le \mathrm{tr}[(\boldsymbol{b} - \boldsymbol{a})f'(\boldsymbol{a})]. \tag{5.229}$$

b) Show that $f(x) = -x\log x$ is concave for $x > 0$.

c) Use (a) and (b) to show $S(\boldsymbol{\rho}|\boldsymbol{\sigma}) \geq 0$ for any two density matrices $\boldsymbol{\rho}$ and $\boldsymbol{\sigma}$.

d) Use nonnegativity of $S(\boldsymbol{\rho}|\boldsymbol{\sigma})$ to show that if $\boldsymbol{\rho}$ has its support on a space of dimension D, then

$$S(\boldsymbol{\rho}) \leq \log D. \tag{5.230}$$

e) Use nonnegativity of relative entropy to prove the *subadditivity* of entropy

$$S(\boldsymbol{\rho}_{AB}) \leq S(\boldsymbol{\rho}_A) + S(\boldsymbol{\rho}_B). \tag{5.231}$$

[Hint: Consider the relative entropy of $\boldsymbol{\rho}_A \otimes \boldsymbol{\rho}_B$ and $\boldsymbol{\rho}_{AB}$.]

f) Use subadditivity to prove the *concavity* of the entropy:

$$S(\sum_i \lambda_i \boldsymbol{\rho}_i) \geq \sum_i \lambda_i S(\boldsymbol{\rho}_i), \tag{5.232}$$

where the λ_i's are positive real numbers summing to one. [Hint: Apply subadditivity to

$$\boldsymbol{\rho}_{AB} = \sum_i \lambda_i \, (\boldsymbol{\rho}_i)_A \otimes (|e_i\rangle\langle e_i|)_B \, . \,] \tag{5.233}$$

g) Use subadditivity to prove the *triangle inequality* (also called the Araki-Lieb inequality):

$$S(\boldsymbol{\rho}_{AB}) \geq |S(\boldsymbol{\rho}_A) - S(\boldsymbol{\rho}_B)|. \tag{5.234}$$

[Hint: Consider a purification of $\boldsymbol{\rho}_{AB}$; that is, construct a pure state $|\psi\rangle$ such that $\boldsymbol{\rho}_{AB} = \mathrm{tr}_C|\psi\rangle\langle\psi|$. Then apply subadditivity to $\boldsymbol{\rho}_{BC}$.]

5.3 Lindblad–Uhlmann monotonicity.

According to a theorem proved by Lindblad and by Uhlmann, relative entropy on $\mathcal{H}_A \otimes \mathcal{H}_B$ has a property called *monotonicity*:

$$S(\boldsymbol{\rho}_A|\boldsymbol{\sigma}_A) \leq S(\boldsymbol{\rho}_{AB}|\boldsymbol{\sigma}_{AB}); \tag{5.235}$$

The relative entropy of two density matrices on a system AB cannot be less than the induced relative entropy on the subsystem A.

a) Use Lindblad-Uhlmann monotonicity to prove the strong subadditivity property of the 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 Lindblad–Uhlmann monotonicity to show that the action of a superoperator cannot increase relative entropy, that is,

$$S(\$\boldsymbol{\rho}|\$\boldsymbol{\sigma}) \leq S(\boldsymbol{\rho}|\boldsymbol{\sigma}), \tag{5.236}$$

Where \$ is any superoperator (completely positive map). [Hint: Recall that any superoperator has a unitary representation.]

c) Show that it follows from (b) that a superoperator cannot increase the Holevo information of an ensemble $\mathcal{E} = \{\boldsymbol{\rho}_x, p_x\}$ of mixed states:

$$\chi(\$(\mathcal{E})) \leq \chi(\mathcal{E}), \tag{5.237}$$

where

$$\chi(\mathcal{E}) = S\left(\sum_x p_x \boldsymbol{\rho}_x\right) - \sum_x p_x S(\boldsymbol{\rho}_x). \tag{5.238}$$

5.4 The Peres–Wootters POVM.

Consider the Peres–Wootters information source described in §5.4.2 of the lecture notes. It prepares one of the three states

$$|\Phi_a\rangle = |\varphi_a\rangle|\varphi_a\rangle, \quad a = 1, 2, 3, \tag{5.239}$$

each occurring with *a priori* probability $\frac{1}{3}$, where the $|\varphi_a\rangle$'s are defined in eq. (5.149).

a) Express the density matrix

$$\boldsymbol{\rho} = \frac{1}{3}\left(\sum_a |\Phi_a\rangle\langle\Phi_a|\right), \tag{5.240}$$

in terms of the Bell basis of maximally entangled states $\{|\phi^\pm\rangle, |\psi^\pm\rangle\}$, and compute $S(\boldsymbol{\rho})$.

b) For the three vectors $|\Phi_a\rangle, a = 1, 2, 3$, construct the "pretty good measurement" defined in eq. (5.162). (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.

5.5 Teleportation with mixed states.

An operational way to define entanglement is that an entangled state can be used to teleport an unknown quantum state with better fidelity than could be achieved with local operations and classical communication only. In this exercise, you will show that there are mixed states that are entangled in this sense, yet do not violate any Bell inequality. Hence, for mixed states (in contrast to pure states) "entangled" and "Bell-inequality-violating" are not equivalent.

Consider a "noisy" entangled pair with density matrix.

$$\boldsymbol{\rho}(\lambda) = (1-\lambda)|\psi^-\rangle\langle\psi^-| + \lambda\frac{1}{4}\mathbf{1}. \tag{5.241}$$

a) Find the fidelity F that can be attained if the state $\boldsymbol{\rho}(\lambda)$ is used to teleport a qubit from Alice to Bob. [Hint: Recall that you showed in an earlier exercise that a "random guess" has fidelity $F = \frac{1}{2}$.]

b) For what values of λ is the fidelity found in (a) better than what can be achieved if Alice measures her qubit and sends a classical message to Bob? [Hint: Earlier, you showed that $F = 2/3$ can be achieved if Alice measures her qubit. In fact this is the best possible F attainable with classical communication.]

c) Compute

$$\mathrm{Prob}(\uparrow_{\hat{n}}\uparrow_{\hat{m}}) \equiv \mathrm{tr}\left(\mathbf{E}_A(\hat{n})\mathbf{E}_B(\hat{m})\boldsymbol{\rho}(\lambda)\right), \tag{5.242}$$

where $\mathbf{E}_A(\hat{n})$ is the projection of Alice's qubit onto $|\uparrow_{\hat{n}}\rangle$ and $\mathbf{E}_B(\hat{m})$ is the projection of Bob's qubit onto $|\uparrow_{\hat{m}}\rangle$.

d) Consider the case $\lambda = 1/2$. Show that in this case the state $\boldsymbol{\rho}(\lambda)$ violates no Bell inequalities. Hint: It suffices to construct a local hidden variable model that correctly reproduces the spin correlations found in (c), for $\lambda = 1/2$. Suppose that the hidden variable $\hat{\alpha}$ is uniformly distributed on the unit sphere, and that there are functions f_A and f_B such that

$$\mathrm{Prob}_A(\uparrow_{\hat{n}}) = f_A(\hat{\alpha}\cdot\hat{n}), \quad \mathrm{Prob}_B(\uparrow_{\hat{m}}) = f_B(\hat{\alpha}\cdot\hat{m}). \tag{5.243}$$

The problem is to find f_A and f_B (where $0 \leq f_{A,B} \leq 1$) with the properties

$$\int_{\hat{\alpha}} f_A(\hat{\alpha}\cdot\hat{n}) = 1/2, \quad \int_{\hat{\alpha}} f_B(\hat{\alpha}\cdot\hat{m}) = 1/2,$$
$$\int_{\hat{\alpha}} f_A(\hat{\alpha}\cdot\hat{n}) f_B(\hat{\alpha}\cdot\hat{m}) = \mathrm{Prob}(\uparrow_{\hat{n}}\uparrow_{\hat{m}}). \tag{5.244}$$