Chapter 2

Useful Concepts from Information Theory

2.1 Quantifying Information

2.1.1 The entropy

It turns out that there is a way to quantify the intuitive notion that some messages contain more information than others. This was first realized by Claude Shannon, and presented in two now famous articles in 1948 [22]. There are two ways to approach the notion of the information contained in a message. The first is to view it as capturing how much the message reduces our uncertainty about something. This might seem like a natural approach, given that we are now familiar with the concept of a state of knowledge, and how it changes upon obtaining data. The other approach is to ask, how long must a message be in order to convey a given piece of information. In this case it is the length of the required message that tells us "how much" information we wish to convey. It turns out that these two approaches, although they seem rather different, are actually the same thing — although it is the second that picks out a unique measure for information in the most unambiguous way. We begin then, by considering this second approach.

To determine the length of a message that is required to tell someone something, we must first specify the alphabet we are using. Specifically, we need to fix the number of letters in our alphabet. It is simplest to take the smallest workable alphabet, that containing only two letters, and we will call these letters 0 and 1. Imagine now that there is a movie showing this coming weekend at the local theater, but you do not whether it is scheduled for Saturday or Sunday. So you phone up a friend who knows, and ask them. Since there are only two possibilities, your friend only needs to send you a message consisting of a single symbol that similarly has two possibilities: so long as we have agreed beforehand that 0 will represent Saturday, and 1 will represent Sunday, then the answer to your question is provided by sending just a single 0, or a single 1. Because the letter has two possibilities, it is referred to as

a *bit*, short for *binary digit*. Note that two people must *always* agree beforehand on the meaning of the symbols they will use to communicate - the only reason you can read this text is because we have a prior agreement about the meaning of each word.

But now think about the above scenario a little further. What happens if we need to ask about many movies, each of which will show on Saturday or Sunday? If there are N movies, then the message will consist of a sequence of N zeros and ones, which we will refer to as a string of N bits. If the movies are all equally likely to occur on either of the two days, then every possible sequence of zeros and ones that the message could contain is equally likely. But what happens when the movies occur preferentially on Saturday? Let us say that the probability that each movie occurs on Saturday is p. The possible messages that your friend may send to you are no longer equally likely. If $p = 0.99$ then on average only one in every hundred movies will occur on Sunday. This makes it mind-bogglingly unlikely that all the bits in the message will be equal to 1 (recall that 1 corresponds to Sunday). We now realize that if certain messages are sufficiently unlikely, then your friend will never have to send them. This means that the total number of different messages that she or he needs to able to send, to inform you of the days of the N movies, is less than 2^N. This would in turn mean that all these required messages could be represented by a bit-string with *less* than N bits, since the total number of messages that can be represented by N bits is 2^N.

If the probability of each bit being 0 is p, then the most likely strings of N bits (the most likely messages) are those in which pN of the bits are 0 (and thus the other $(1-p)N$ bits are 1). This is essentially just the law of large numbers (for a discussion of this law, see Appendix B). The most likely messages are not the only ones your friend will need to send - messages in which the number of zeros and ones are close to the mean number will also be quite frequent. The key to determining the number of bits required to send the message is to identify the fraction of messages whose probability goes to zero as $N \to \infty$. These are the messages we do not have to send, so long as N is sufficiently large, and this tells us how many bits we will require to convey the information.

Typical sets and typical states

In stating that there are messages that we will (almost) never need to send, what we really mean is that *there is a set of messages for which the total probability of any of its members occurring* is negligible for large N. The rest of the messages must therefore come from a set whose total probability is ≈ 1, and these are called the *typical messages*. This is an important concept not only for information theory, but also for statistical mechanics, and we will need it again Chapter 4. For this reason we pause to define the concept of typicality in a more general context.

Consider an object that has N possible states. All the states are mutually exclusive, so that the object must be in one and only one of them. The probabilities for the N states thus sum to unity. In addition, there is some specified way, one that is natural

in a given context, in which the number of states of the object can be increased, and this number can be made as large as one wishes. Given such an object, we say that a subset of the N states is a *typical set* if the total probability of all the states in the subset tends to unity as $N \to \infty$. An element of a typical set is called a *typical state*. The complement of such a subset (consisting of all the states not included in it) is called an *atypical* set. The important property of typical sets is that, for large N, if we pick a state at random, we will almost certainly obtain a typical state.

The Shannon entropy

In the scenario we discussed above, each question we wanted to know the answer to (the day when each movie was showing) had just two possibilities. In general, we may wish to communicate the state of something that has many possible states. Let us say that we want to communicate the values of N identical independent random variables, where each has M possible values. Let us label the states by $m = 1, \ldots M$, and the probability of each state as p_m. It turns out that, in the limit of large N, the number of typical messages that we might need to send is given by

$$N_{\text{typ}} = 2^{NH[\{P_m\}]}, \tag{2.1}$$

where

$$H[\{P_m\}] \equiv -\sum_{m=1}^{M} p_m \log_2(p_m), \tag{2.2}$$

and $\log_2$ denotes the logarithm to the base two. This means that the total number of bits required for the message is $NH[\{P_m\}]$, and thus the average number of bits required *per random variable* is

$$N_{\text{bits}} = H[\{P_m\}] = -\sum_{m=1}^{M} p_m \log_2(p_m). \tag{2.3}$$

Here $H[\{P_m\}]$ is called the *Shannon entropy* of the probability distribution $\{p_m\}$. The Shannon entropy is a *functional* of a probability distribution — that is, it takes a probability distribution as an input, and spits out a number.

The result given in Eq.(2.1) is the heart of “Shannon’s noiseless coding theorem". The term “coding" refers to the act of associating a sequence of letters (a code word) with each possible message we might need to send. (The term noiseless means that the messages are sent to the receiver without any errors being introduced *en route*.) Shannon’s theorem tells us how many bits we need in each code word. To further relate this result to real communication, consider sending english text across the internet. Each english letter (symbol) that is sent comes from an effectively random distribution that we can determine by examining lots of typical text. While each letter in english prose is not actually independent of the next (given the letter “t", for example, the letter “h" is more likely to follow than “s", even though the latter

appears more frequently overall) it is not that bad an approximation to assume that they are independent. In this case Shannon's theorem tells us that if we send N letters we must send on average H bits per letter, where H is the entropy of the distribution (relative frequency) of english letters. In a real transmission protocol N can be very large (e.g. we may be sending the manuscript of a large book). However, since N is never infinite, we sometimes have to send atypical messages. Real coding schemes handle this by having a code word for every possible message (not merely the typical messages), but they use short code words for typical messages, and long code words for atypical messages. This allows one to get close to the smallest number of bits per letter (the ultimate compression limit) set by the Shannon entropy.

As a simple example of a real coding scheme, consider our initial scenario concerning the N movies. If almost all movies are shown on a Saturday, then our typical messages will consist of long strings of 0's, separated by the occasional 1. In this case a good coding scheme is to use a relatively long sequence for 1 (say 0000), and for each string of zeros send a binary sequence giving the *number* of zeros in the string. (To actually implement the coding and decoding we must remember that the binary sequences used for sending the lengths of the strings of zeros are not allowed to contain the sequence for 1.)

Note that while we have used base-two for the logarithm in defining the Shannon entropy, any base is fine; the choice of base simply reflects the size of the alphabet we are using to encode the information. We can just as well use the natural logarithm, and even though the natural logarithm does not correspond to an alphabet with an integer number of symbols, it is customary to refer to the units of this measure as *nats*. From now on we will use the natural logarithm for the entropy, unless indicated otherwise. We defer the proof of Shannon's noiseless coding theorem to Appendix B, and turn now to the relationship between the Shannon entropy and our intuitive notion of uncertainty.

Entropy and uncertainty

While the Shannon entropy (from now on, just the *entropy*) was derived by considering the resources required for communication, it also captures the intuitive notion of uncertainty. Consider first a random variable with just two possibilities, 0 and 1, where the probability for 0 is p. If p is close to zero or close to unity, then we can be pretty confident about the value the variable will take, and our uncertainty is low. Using the same reasoning, our uncertainty should be maximum when $p = 1/2$. The entropy has this property - it is maximal when $p = 1/2$, and minimal (specifically, equal to zero) when $p = 0$ or 1. Similarly, if we have a random variable with N outcomes, the entropy is maximal, and equal to $\ln N$, when all the outcomes are equally likely.

If we have two independent random variables X and Y, then the entropy of their joint probability distribution is the sum of their individual entropies. This property, which is simple to show (we leave it as an exercise) also captures a reasonable no-

tion of uncertainty - if we are uncertain about two separate things, then our total uncertainty should be the sum of each.

A less obvious, but also reasonable property of uncertainty, which we will refer to as the "corse-graining" property, is the following. Consider an object X with three possible states, $m = 1, 2$, and 3, with probabilities $p_1 = 1/2$, $p_2 = 1/3$, and $p_3 = 1/6$. The total entropy of X is $H(X) = H[\{1/2, 1/3, 1/6\}]$. Now consider two ways we could provide someone (Alice) with the information about the state of X. We could tell Alice the state in one go, reducing her uncertainty completely. Alternatively, we could give her the information in two steps. We first tell Alice whether m is less than 2, or greater than or equal to 2. The entropy of this piece of information is $H[\{1/2, 1/2\}]$. If the second possibility occurs, which happens half the time, then Alice is still left with two possibilities, whose entropy is given by $H[\{2/3, 1/3\}]$. We can then remove this uncertainty by telling her whether $m = 2$ or 3. It turns out that the entropy of X is equal to the entropy of the first step, plus the entropy of the second step, weighted by the probability that we are left with this uncertainty after the first step. More generally, if we have a set of states and we "corse-grain" it by lumping the states into a number of mutually exclusive subsets, then the total entropy of the set is give by the weighted average of the entropies of all the subsets.

It turns out that the following three properties are enough to specify the entropy as the unique measure of uncertainty for a probability distribution $\{p_m\}$, $m = 1, \ldots, M$ (The proof is given in [23]):

1. The entropy is continuous in each of the p_m.
2. If all the M possibilities are equally likely, then the entropy increases with M.
3. The entropy has the corse-graining property.

To summarize, the amount of communication resources required to inform us about something is also a measure of our initial uncertainty about that thing. In information theory, the entropy of the thing we are trying to send information about is called the "entropy of the source" of information. We can therefore think of the entropy as measuring both the uncertainty about something, and the amount of information "contained" by that thing.

2.1.2 The mutual information

Now we have a quantitative measure of information, we can quantify the information obtained by a classical measurement. If we denote the unknown quantity as x, and the measurement result as y, then as we saw in section 1.1, the measurement is described by the conditional probability $P(y|x)$. We will define x as having a discrete set of values labeled by n, with $n = 1, \ldots, N$, and the measurement as having a discrete set of outcomes labeled by j, with $j = 1, \ldots, M$. If our state-of-knowledge

regarding x prior to the measurement is $P(n)$, then upon obtaining the result j, our state-of-knowledge becomes

$$P(n|j) = \left[\frac{P(j|n)}{P(j)}\right] P(n). \tag{2.4}$$

Our uncertainty before the measurement is the entropy of $P(n)$, $H[x] = H[\{P(n)\}]$, and our uncertainty after the measurement is $H[\{P(n|j)\}]$. The reduction in our uncertainty, given outcome j, is therefore $\Delta H_j = H[\{P(n)\}] - H[\{P(n|j)\}]$. A reasonable measure of the amount of information the measurement provides about x is the reduction in the entropy of our state-of-knowlegde about x, *averaged* over all the possible measurement results j. This is

$$\langle \Delta H \rangle = H[\{P(n)\}] - \sum_j P(j) H[\{P(n|j)\}]. \tag{2.5}$$

Like the entropy, this new quantity is important in information theory, and we explain why below. Before we do, however, it is worth examining the expression for $\langle \Delta H \rangle$ a little further. First note that the quantity that is subtracted in Eq.(2.5) is the entropy of x after the measurement, averaged over the measurement results. This is called, reasonably enough, the *conditional entropy* of x given y, and written as $H[x|y]$:

$$H[x|y] \equiv \sum_j P(j) H[\{P(n|j)\}]. \tag{2.6}$$

Next, the joint entropy of x and y, being the total uncertainty when we know neither of them, is the entropy of their joint probability density, $P(n, j)$. We write this as $H[x, y] \equiv H[\{P(n, j)\}]$. It turns out that by rearranging the expression for $\langle \Delta H \rangle$, we can alternatively write it as

$$\langle \Delta H \rangle = H[x] + H[y] - H[x, y]. \tag{2.7}$$

One way to interpret this expression is the following. The random variables x and y both contain uncertainty. However, if they are correlated, then some of this uncertainty is shared between them. So when we add the entropies $H[x]$ and $H[y]$, we are counting the entropy that they have in common *twice*. This in turn means that the total uncertainty of x and y together, $H[x, y]$ is in general *less* than $H[x] + H[y]$. The difference $(H[x] + H[y]) - H[x, y]$ is therefore a measure of how correlated y is with x (how much y knows about x), and it turns out that this difference corresponds precisely to what we learn about x when we know y. Since the expression in Eq.(2.7) is a measure of the uncertainty that is common to x and y, it is called the *mutual information* between x and y, and written as $H[x : y]$. Thus

$$\langle \Delta H \rangle = H[x : y] = H[x] + H[y] - H[x, y] = H[x] - H[x|y]. \tag{2.8}$$

The expression for $\langle \Delta H \rangle$ in Eq.(2.7) is also useful because it shows us that $\langle \Delta H \rangle$ is symmetric in x and y. That is, the amount of information we obtain about x upon learning the value of y, is exactly the same as the amount of information we would obtain about y if we learned the value of x.

The mutual information and communication channels

In the context of measurements, we think of two correlated variables x and y as describing, respectively, the unknown quantity and the measurement result. In this context the conditional probability $P(y|x)$ characterizes the measurement. To instead relate the mutual information to communication, one considers x to be the input to a communication channel, and y to be the output. Thus the sender encodes a message into some alphabet, and the values of x are the symbols of this alphabet. The output of the channel, however, may not be a faithful copy of the input. In general the channel may be noisy, so that the output of the channel, just like the result of a measurement, is related probabilistically to the input. In the context of information theory, the conditional probability $P(y|x)$ characterizes a *channel*.

The reason that we can regard the average reduction in entropy as being the *information* provided by the measurement, is because of the following result. The mutual information between the input and output of a channel, is precisely the amount of information, per symbol, that can be reliably sent down the channel. Specifically, consider that a sender has a message of N symbols, and for which the entropy per symbol is H. If she encodes the message appropriately, and sends it down a channel with mutual information $H[x:y]$, then in the limit of large N, the sender will be able to decode the message without error so long as $H \leq H[x:y]$. This result is called Shannon's noisy coding theorem, and along with the noiseless coding theorem makes up the fundamental theorems of information theory. Note that while the conditional probability $p(x|y)$ depends only on the channel, the mutual information depends on the probability distribution of the input to the channel, x. So to obtain the maximal transmission rate, the sender must encode her message so that the input symbols have the distribution that maximizes the mutual information for the given channel. There is no analytic form for this maximum in general, and it must often be computed numerically. The maximum of the mutual information for a given channel is called the *capacity* of the channel. Proofs of the noisy coding theorem can be found in most textbooks on information theory.

2.2 Quantifying Uncertainty About a Quantum System

2.2.1 The von Neumann entropy

Up until now we have usually written a probability distribution over a finite set of states, n, as the set of probabilities $\{p_n\}$. It is often useful to use instead a vector notation for a discrete probability distribution. An important probability distribution associated with a quantum system is the set of eigenvalues of the density matrix. We will denote the vector whose elements are the eigenvalues of a density matrix ρ as $\boldsymbol{\lambda}(\rho)$. We will write the elements of the vector $\boldsymbol{\lambda}$ as λ_n, so that $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)$. From now on vectors will be synonymous with probability distributions.

A useful measure of our uncertainty about the state of a quantum system is pro-

vided by the Shannon entropy of the eigenvalues of the density matrix, being

$$H[\boldsymbol{\lambda}(\rho)] = -\sum_n \lambda_n \ln \lambda_n. \tag{2.9}$$

This quantity is called the *von Neumann entropy* of the state ρ. One of the main reasons that the von Neumann entropy is useful is because there is a sense in which it represents the *minimum* uncertainty that we have about the future behavior of a quantum system. Before we explain why, we note that this entropy can be written in a very compact form. To write this form we must be familiar with the notion of an arbitrary function of a matrix. Consider a matrix ρ that is diagonalized by the unitary operator U. That is

$$\rho = UDU^\dagger = U \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_N \end{pmatrix} U^\dagger. \tag{2.10}$$

Consider also a function $f(x)$ of a number x that has the taylor series $f(x) = \sum_n c_n x^n$. Then two natural and completely equivalent definitions of the function $f(\rho)$ are

$$\begin{aligned} f(\rho) &\equiv U \begin{pmatrix} f(\lambda_1) & 0 & \cdots & 0 \\ 0 & f(\lambda_2) & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & f(\lambda_N) \end{pmatrix} U^\dagger \\ &\equiv \sum_n c_n \rho^n, \end{aligned} \tag{2.11}$$

With this definition, the von Neumann entropy of ρ is given by

$$\begin{aligned} S(\rho) &\equiv -\text{Tr}[\rho \ln \rho] \\ &= -\text{Tr}[U\rho U^\dagger U \ln \rho U^\dagger] \\ &= -\text{Tr}[D \ln D] \\ &= -\sum_n \lambda_n \ln \lambda_n \;=\; H[\boldsymbol{\lambda}(\rho)]. \end{aligned} \tag{2.12}$$

There are two senses in which $S(\rho)$ represents the minimum uncertainty associated with the state-of-knowledge ρ. The first is the following. If we make a measurement that provides us with complete information about the system — that is, a measurement in which each measurement operator projects the system onto a pure state — then the von Neumann entropy is the minimum possible entropy of the measurement

outcomes. Specifically, if we make any complete measurement that has M outcomes labeled by $m = 1, \ldots, m$, then their respective probabilities, p_m, will always satisfy

$$H[\{p_m\}] = -\sum_m p_m \ln p_m \geq S(\rho). \tag{2.13}$$

Since the measurement is complete,

$$p_m = \alpha_m \langle\psi_m|\rho|\psi_m\rangle \quad \text{and} \quad \sum_m \alpha_m |\psi_m\rangle\langle\psi_m| = I, \tag{2.14}$$

for some set of states $|\psi_m\rangle$ and positive numbers α_m. The value of the lower bound is achieved by a von Neumann measurement in the eigenbasis of the density matrix. This lower bound is certainly not obvious, and is a special case of a deep result regarding measurements that we will describe later in section 2.3.1. Since ultimately questions about the future behavior of a quantum system are questions about the results of future measurements that will be performed on it, the von Neumann entropy measures our maximum predictability about future behavior.

The second sense in which $S(\rho)$ represents the minimum uncertainty inherent in ρ regards the pure-state ensembles that generate ρ: if $\rho = \sum_n p_n |\psi_n\rangle\langle\psi_n|$, then the ensemble probabilities p_n always satisfy

$$H[\{p_n\}] = -\sum_n p_n \ln p_n \geq S(\rho). \tag{2.15}$$

This result is also not obvious, and will be discussed in section 2.2.3.

There is a further reason why the von Neumann entropy is an important quantity for quantum states: because it is an entropy, it is intimately connected with various information theoretic questions one can ask about measurements on quantum systems. We will consider some of these questions in section 2.3.1. The von Neumann entropy also has a number of very useful properties. Some of these are

1. For a system of dimension N: $\ln N \geq S(\rho) \geq 0$.
2. For $p \in [0, 1]$: $S(p\rho_1 + (1-p)\rho_2) \geq pS(\rho_1) + (1-p)S(\rho_2)$.
3. For $\rho = \sum_n p_n \rho_n$: $S(\rho) \leq \sum_n p_n S(\rho_n) + H[\{p_n\}]$.
4. If ρ_{A} is the state of system A, ρ_{B} is the state of system B, and ρ is the joint state of the combined systems, then
   i) $S(\rho) \leq S(\rho_{\mathrm{A}}) + S(\rho_{\mathrm{B}})$ with equality if and only if $\rho = \rho_{\mathrm{A}} \otimes \rho_{\mathrm{B}}$.
   ii) If ρ is pure then $S(\rho_{\mathrm{A}}) = S(\rho_{\mathrm{B}})$.
5. Consider three systems A, B and C, where we denote the states of each by ρ_{X}, with $X = A, B$ or C, the joint states of any two systems X and Y by ρ_{XY}, and the joint state of all three by ρ_{ABC}. The von Neumann entropy satisfies
   $$S(\rho_{\mathrm{ABC}}) + S(\rho_{\mathrm{B}}) \leq S(\rho_{\mathrm{AB}}) + S(\rho_{\mathrm{BC}}).$$

The first property gives the range of possible values of the von Neumann entropy: it is zero if and only if the state is pure, and is equal to $\ln N$ if and only if the state is completely mixed. Property two is called *concavity*, property 4 i) is *subaditivity*, and property five is *strong subaditivity*. The last of these is extremely powerful. Many important results in quantum information theory follow from it, and we will encounter one of these in section 2.3.1. The proof's of all the above properties, except for the last, are quite short, and all can be found in [21]. The simplest proof of strong subadditivity to date is due to Ruskai, and can be found in [24].

2.2.2 Majorization and Density Matrices

While the Shannon entropy is a measure of uncertainty, it is also more than that — it is the *specific* measure of uncertainty that corresponds to the resources required to send information that will eliminate the uncertainty. There is, in fact, a more basic way to capture the notion of uncertainty alone, without the extra requirement that it provide a measure of information. This is the concept of *majorization* [25, 15], and it turns out to be very useful in quantum measurement and information theory.

Consider two vectors $\mathbf{p} = (p_1, \ldots, p_N)$ and $\mathbf{q} = (q_1, \ldots, q_N)$ that represent probability distributions. To define majorization we always label the elements of our probability vectors so as to put these elements in decreasing order. This means that $p_1 \geq p_2 \geq \cdots p_N$. With this ordering, the vector $\mathbf{p}$ is said to *majorize* $\mathbf{q}$ if (and only if)

$$\sum_{n=1}^{k} p_n \geq \sum_{n=1}^{k} q_n, \quad \text{for } k = 1, \ldots, N. \tag{2.16}$$

In words, this means that the largest element of $\mathbf{p}$ is greater than the largest element of $\mathbf{q}$, the largest two elements of $\mathbf{p}$ are greater than the largest two element of $\mathbf{q}$, *etc.* If you think about this, it means that the probability distribution given by the elements of $\mathbf{p}$ is more narrowly peaked than $\mathbf{q}$. This in turn means that the $\mathbf{p}$ probability distribution has *less uncertainty* than that of $\mathbf{q}$. The relation "$\mathbf{p}$ majorizes $\mathbf{q}$" is written as

$$\mathbf{p} \succ \mathbf{q} \quad \text{or} \quad \mathbf{q} \prec \mathbf{p}. \tag{2.17}$$

Majorization also has a connection to the intuitive notion of what it means to randomize, or mix. Consider a probability vector $\mathbf{p}$, and a some set of permutation matrices $\{T_j\}$, with $j = 1, \ldots, M$. (Multiplying a vector by a permutation matrix permutes the elements of the vector, but does nothing else.) If we choose to perform at random one of the permutations T_j on the elements of $\mathbf{p}$, and choose permutation T_j with probability λ_n, then our resulting state of knowledge is given by the probability vector $\mathbf{q} = \sum_n \lambda_n T_n \mathbf{p}$. It turns out that for any set of permutations

$$\mathbf{q} = \sum_n \lambda_n T_n \mathbf{p} \prec \mathbf{p}. \tag{2.18}$$

A proof of this result is given in Chapter 2 of [15].

It turns out that it is very useful to extend the concept of Majorization so as to compare probability vectors that have different lengths. To do this, one merely adds new elements, all of which are zero, to the shorter of the two vectors so that both vectors are the same length. We then apply the usual definition of majorization to these two vectors of equal length. For example, to determine the majorization relation between the vectors $\mathbf{p} = (2/3, 1/3)$ and $\mathbf{q} = (1/2, 1/4, 1/4)$, we first add a zero to the end of $\mathbf{p}$, giving the vector $\tilde{\mathbf{p}} = (2/3, 1/3, 0)$. Now applying the majorization criteria to $\tilde{\mathbf{p}}$ and $\mathbf{q}$ gives us

$$\mathbf{p} = \begin{pmatrix} 2/3 \\ 1/3 \end{pmatrix} \succ \begin{pmatrix} 1/2 \\ 1/4 \\ 1/4 \end{pmatrix} = \mathbf{q}. \tag{2.19}$$

The concept of majorization imposes a *partial* order on probability vectors. The order is partial because there are many pairs of vectors for which neither vector majorizes the other. This property reflects the fact that under this basic notion of uncertainty, some vectors cannot be considered to be strictly more or less uncertain than others.

One of the reasons that majorization is an important concept is that it does have a direct connection to the entropy. This is a result of a theorem connecting a certain class of concave functions to majorization. Note that the entropy of a vector $\mathbf{p}$, $H[\mathbf{p}]$, is a function of N inputs (being the N elements of $\mathbf{p}$), and it is concave in all these inputs. An way to state this property is that, for all vectors $\mathbf{p}$ and $\mathbf{q}$,

$$H[\alpha\mathbf{p} + (1-\alpha)\mathbf{q}] \geq \alpha H[\mathbf{p}] + (1-\alpha)H[\mathbf{q}], \qquad \alpha \in [0,1]. \tag{2.20}$$

The entropy is also *symmetric* in its inputs. This means that its value does not change when its inputs are permuted in any way. Functions that are concave in all their inputs, as well as being symmetric, are called *Schur-concave* functions.

The strong relationship between majorization and Schur-concave functions is that

$$\mathbf{p} \succ \mathbf{q} \quad \Rightarrow \quad f[\mathbf{p}] \leq f[\mathbf{q}] \qquad \text{for } f \text{ Schur-concave.} \tag{2.21}$$

If we apply this to the entropy H, it shows us that if $\mathbf{p}$ majorizes $\mathbf{q}$, then $\mathbf{p}$ has less entropy than $\mathbf{q}$, and is thus less uncertain that $\mathbf{q}$. The relation given by Eq.(2.21) is a direct result of the inequality in Eq.(2.18).

We can apply the concept of majorization to density matrices via their eigenvalues. If the vector of eigenvalues of ρ is $\boldsymbol{\lambda}(\rho)$, and that of σ is $\boldsymbol{\mu}(\sigma)$, then we write

$$\rho \succ \sigma \qquad \text{meaning that} \qquad \boldsymbol{\lambda}(\rho) \succ \boldsymbol{\mu}(\sigma). \tag{2.22}$$

The reason that majorization is especially useful in quantum mechanics is a relation, called *Horn's lemma*, that connects majorization to unitary transformations. To present this relation we need to define something called a *stochastic matrix*. A stochastic matrix is a matrix whose elements are all greater than or equal to zero, and

for which the elements in each row sum to unity (that is, each row is a probability vector). If $\mathbf{p}$ is a probability vector, and S is a stochastic matrix, then the relation

$$\mathbf{q} = S\mathbf{p} \tag{2.23}$$

means that the elements of $\mathbf{q}$ are averages over the elements of S. (The term *stochastic matrix* comes from the fact that these matrices represent transition probabilities between states in a stochastic jump process, but we are not concerned with this here. Further details regarding jump processes can be found in [1, 26].)

The relation given by Eq.(2.23) does not imply any particular majorization relationship between $\mathbf{q}$ and $\mathbf{p}$: while it implies that every element of $\mathbf{q}$ is an average of the elements of $\mathbf{p}$, not all the elements of $\mathbf{p}$ (and thus, not all the probability contained in $\mathbf{p}$) need be distributed among the elements of $\mathbf{q}$. As a result the elements of $\mathbf{q}$ need not sum to unity, so $\mathbf{q}$ need not be a probability vector. A *doubly-stochastic* matrix, on the other hand, is one in which both the rows and columns all sum to unity. In this case, not only is $\mathbf{q}$ guaranteed to be a probability vector, but it must also have the same number of elements as $\mathbf{p}$. A doubly-stochastic matrix never decreases uncertainty. In fact, it is true that

$$\mathbf{q} \prec \mathbf{p} \quad \textit{if and only if} \quad \mathbf{q} = D\mathbf{p} \tag{2.24}$$

for some doubly-stochastic matrix D. The proof of this can found in, e.g. [27, 14, 15]. The condition that the rows and columns of a matrix, D, sum to unity is equivalent to the following two conditions: 1. As mentioned above, D is "probability preserving", so that if $\mathbf{p}$ is correctly normalized, then so is $D\mathbf{p}$, and 2. That D is "unital", meaning that it leaves the uniform distribution unchanged.

Now consider a unitary transformation U, whose elements are u_{ij}. If we define a matrix D whose elements are $d_{ij} \equiv |u_{ij}|^2$, then D is doubly stochastic. When the elements of a doubly-stochastic matrix can be written as the square moduli of those of a unitary matrix, then D is called *unitary-stochastic*, or unistochastic. Not all doubly-stochastic matrices are unitary-stochastic. Nevertheless, it turns out that if $\mathbf{q} \prec \mathbf{p}$, then one can always find a unitary-stochastic matrix, D_u, for which $\mathbf{q} = D_u\mathbf{p}$. In addition, since D_u is doubly-stochastic, the relation $\mathbf{q} = D_u\mathbf{p}$ also implies that $\mathbf{q} \prec \mathbf{p}$. This is Horn's lemma:

Lemma 1. [Horn [28]] If u_{ij} are the elements of a unitary matrix, and the elements of the matrix D_u are given by $(D_u)_{ij} = |u_{ij}|^2$, then

$$\mathbf{q} \prec \mathbf{p} \quad \textit{if and only if} \quad \mathbf{q} = D_u\mathbf{p}. \tag{2.25}$$

A simple proof of Horn's lemma is given by Nielsen in [29]. The following useful fact about density matrices is an immediate consequence of this lemma.

Theorem 3. The diagonal elements of a density matrix ρ in any basis are majorized by its vector of eigenvalues, $\boldsymbol{\lambda}(\rho)$. This means that the probability distribution for finding the system in any set of basis vectors is always broader (more uncertain) than that for the eigenbasis of the density matrix.

Proof. Denote the vector of the diagonal elements of a matrix A as $\mathbf{diag}[A]$, the density operator when written in its eigenbasis as the matrix ρ, and the vector of eigenvalues of ρ as $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)$. Since ρ is diagonal we have $\boldsymbol{\lambda} = \mathbf{diag}[\rho]$. Now the diagonal elements of the density matrix written in any other basis are $\mathbf{d} = \mathbf{diag}[U\rho U^\dagger]$. Denoting the elements of U as u_{ij}, and those of $\mathbf{d}$ as d_i, this becomes

$$d_i = \sum_j |u_{ij}|^2 \lambda_j, \tag{2.26}$$

so that Horn's lemma gives us immediately $\mathbf{d} \prec \boldsymbol{\lambda}$. □

Concave functions, "impurity", and the "effective dimension"

While in an information theoretic context the von Neumann entropy is the most appropriate measure of the uncertainty of the state of a quantum system, it is not the most analytically tractable. The concept of Majorization shows us that any Schur-concave function of a probability vector provides a measure of uncertainty. Similarly, every Schur-*convex* function provides a measure of certainty. (A Schur-convex function is a symmetric function that is convex in all its arguments, rather than concave).

A particularly simple Schur-convex function of the eigenvalues of a density matrix ρ, and one that is often used as a measure of certainty, is called the *purity*, defined by

$$\mathcal{P}(\rho) \equiv \mathrm{Tr}[\rho^2]. \tag{2.27}$$

This useful because it has a much simpler algebraic form that the von Neumann entropy.

One can obtain a positive measure of uncertainty directly from the purity by subtracting it from unity. This quantity, sometimes called the "linear entropy", is thus

$$\mathcal{I}(\rho) \equiv 1 - \mathrm{Tr}[\rho^2]. \tag{2.28}$$

We will call $\mathcal{I}(\rho)$ the *impurity*. For an N-dimensional system, the impurity ranges from 0 (for a pure state) to $(N-1)/N$ (for a completely mixed state).

A second measure of uncertainty, which is also very simple, and has in addition a more physical meaning, is

$$\mathcal{D}(\rho) \equiv \frac{1}{\mathrm{Tr}[\rho^2]}. \tag{2.29}$$

This ranges between 1 (pure) and N (completely mixed), and provides a measure of the number of basis states that the system *could* be in, given our state-of-knowledge. This measure was introduced by Linden *et al.* [30], who call it the *effective dimension* of ρ. This measure is important in statistical mechanics, and will be used in Chapter 4.

2.2.3 Ensembles corresponding to a density matrix

It turns out that there is a very simple way to characterize all the possible pure-state ensembles that correspond to a given density matrix, ρ. This result is known as the *classification theorem for ensembles*.

Theorem 4. Consider an N-dimensional density matrix ρ, that has K non-zero eigenvalues $\{\lambda_j\}$, with the associated eigenvectors $\{|j\rangle\}$. Consider also an ensemble of M pure states $\{|\psi_m\rangle\}$, with associated probabilities $\{p_m\}$. It is true that

$$\rho = \sum_m p_m |\psi_m\rangle\langle\psi_m| \tag{2.30}$$

if and only if

$$\sqrt{p_m}|\psi_m\rangle = \sum_{j=1}^{N} u_{mj}\sqrt{\lambda_j}|j\rangle, \qquad m = 1, \ldots, M, \tag{2.31}$$

where the u_{mj} are the elements of an M dimensional unitary matrix. We also note that the ensemble must contain at least K linearly independent states. Thus M can be less than N only if $K < N$. There is no upper bound on M.

Proof. The proof is surprisingly simple (in hindsight). Let ρ be an N-dimensional density matrix. Let us assume that the ensemble $\{|\phi_n\rangle, p_n\}$, with $n = 1, \ldots, M$, satisfies $\rho = \sum_n p_n |\phi_n\rangle\langle\phi_n|$. We first show that each of the states $|\phi_n\rangle$ are a linear combination of the eigenvectors of ρ with non-zero eigenvalues. (The space spanned by these eigenvectors is called the *support* of ρ.) Consider a state $|\psi\rangle$ that is outside the support of ρ. This means that $\langle\psi|\rho|\psi\rangle = 0$. Now writing ρ in terms of the ensemble $\{|\phi_n\rangle, p_n\}$ we have

$$0 = \langle\psi|\rho|\psi\rangle = \sum_n |\langle\phi_n|\psi\rangle|^2. \tag{2.32}$$

This means that every state, $|\psi\rangle$, that is outside the support of ρ is also outside the space spanned by the states $\{\phi_n\}$, and so every state $|\phi_n\rangle$ must be a linear combination of the eigenstates with non-zero eigenvalues. Further, an identical argument shows us that if $|\psi\rangle$ is any state in the space of the eigenvectors, then it must also be in the space spanned by the ensemble. Hence the ensemble must have at least K members ($M \geq K$). We can now write $\sqrt{p_n}|\phi_n\rangle = \sum_j c_{nj}\sqrt{\lambda_j}|j\rangle$, and this means that

$$\sum_j \lambda_j |j\rangle\langle j| = \rho = \sum_n p_n |\phi_n\rangle\langle\phi_n| = \sum_{jj'} \left(\sum_n c_{nj} c^*_{nj'} \sqrt{\lambda_j \lambda_{j'}} \right) |j\rangle\langle j'|. \tag{2.33}$$

Since the states $\{|j\rangle\}$ are all orthonormal, Eq.(2.33) can only be true if $\sum_n c_{nj} c^*_{nj'} = \delta_{jj'}$. If we take j to be the column index of the matrix c_{nj}, then this relation means

that all the columns of the matrix c_{nj} are orthonormal. So if $M = K$ the matrix c_{nj} is unitary. If $M >$,K then the relation $\sum_n c_{nj}c^*_{nj'} = \delta_{jj'}$ means that we can always extend the $M \times K$ matrix c_{nj} to a unitary $M \times M$ matrix, simply by adding more orthonormal columns.

We have now shown the "only if" part of the theorem. To complete the theorem, we have to show that if an ensemble is of the form given by Eq.(2.31), then it generates the density matrix $\rho = \sum_j \lambda_j |j\rangle\langle j|$. This is quite straightforward, and we leave it as an exercise. □

The classification theorem for ensembles gives us another very simple way to specify all the possible probability distributions $\{p_m\}$ for pure-state ensembles that generate a given density matrix:

Theorem 5. Define the probability vector $\mathbf{p} = (p_1, \ldots, p_M)$, and the vector of eigenvalues of ρ as $\boldsymbol{\lambda}(\rho)$. The relation

$$\rho = \sum_m p_m |\psi_m\rangle\langle\psi_m| \tag{2.34}$$

is true if and only if

$$\boldsymbol{\lambda}(\rho) \succ \mathbf{p}. \tag{2.35}$$

Proof. The proof follows from the classification theorem and Horn's lemma. To show that the "only if" part is true, we just have to realize that we can obtain p_m by multiplying $\sqrt{p_m}|\psi_m\rangle$ by its Hermitian conjugate. Using Eq.(2.31) this gives

$$\begin{aligned} p_m &= (\langle\psi_m|\sqrt{p_m})\,(\sqrt{p_m}|\psi_m\rangle) = \sum_{j=1}^{K}\sum_{k=1}^{K} u^*_{mk}u_{mj}\sqrt{\lambda_k\lambda_j}\langle k|j\rangle \\ &= \sum_{j=1}^{K} |u_{mj}|^2\lambda_j. \end{aligned} \tag{2.36}$$

Now if $M > K$, we can always extend the upper limit of the sum over j in the last line from K to M simply by defining, for $j = K+1, \ldots, M$, the "extra eigenvalues" $\lambda_j = 0$. We can then apply Horn's lemma, which tells us that $\boldsymbol{\lambda}(\rho) \succ \mathbf{p}$.

We now have the "only if" part. It remains to show that there exists an ensemble for *every* $\mathbf{p}$ that is majorized by $\boldsymbol{\lambda}(\rho)$. Horn's lemma tells us that if $\boldsymbol{\lambda}(\rho) \succ \mathbf{p}$ then there exists a unitary transformation u_{mj} such that Eq.(2.36) is true. We next define the set of ensemble states $\{|\psi_m\rangle\}$ using Eq.(2.31), which we can do since the u_{mj}, λ_j and $|j\rangle$ are all specified. We just have to check that the $|\psi_m\rangle$ are all correctly normalized. We have

$$\begin{aligned} p_m\langle\psi_m|\psi_m\rangle &= \sum_{j=1}^{K}\sum_{k=1}^{K} u^*_{mk}u_{mj}\sqrt{\lambda_k\lambda_j}\langle k|j\rangle \\ &= \sum_{j=1}^{K} |u_{mj}|^2\lambda_j \;=\; p_m, \end{aligned} \tag{2.37}$$

where the last line follows because the u_{mj} satisfy Eq.(2.36). That completes the proof. □

Theorem 5 tells us, for example, that for any density matrix we can always find an ensemble in which all the states have the same probability. This theorem also provides the justification for a claim we made in section 2.2.1: the entropy of the set of probabilities (the probability distribution) for an ensemble that generates a density matrix ρ is always greater than or equal to the von Neumann entropy of ρ.

2.3 Quantum Measurements and Information

The primary function of measurements is to extract information (reduce uncertainty). Now that we have quantitative measures of classical information, and also of our uncertainty about a quantum system, we can ask questions about the amount of information extracted by a quantum measurement. We will also see how the properties of quantum measurements differ in this regard from those of classical measurements.

2.3.1 Information theoretic properties

Information about the state of a quantum system

As we discussed in section 1.5, there are two kinds of information that a measurement can extract from a quantum system. The first concerns how much we know about the state of the system after we have made the measurement. We can now use the von Neumann entropy to quantify this information: we will define it as the average reduction in the von Neumann entropy of our state-of-knowledge, and denote it by $I_{\rm sys}$. Thus

$$I_{\rm sys} \equiv \langle \Delta S \rangle = S(\rho) - \sum_n p_n S(\tilde{\rho}_n), \tag{2.38}$$

where, as usual, ρ is the initial state of the system and $\tilde{\rho}_n$ are the possible final states. We will refer to $I_{\rm sys}$ as being the information that the measurement extracts about the *system* (as opposed to an ensemble in which the system may have been prepared. Note that if the measurement is *semi-classical*, meaning that the initial density matrix commutes with all the measurement operators, then $I_{\rm sys}$ is precisely the mutual information between the measurement result and the ensemble that contains the eigenstates of the density matrix.

It is worth to remembering in what follows, that if one discards the results of a measurement, then the state following the measurement is $\tilde{\rho} = \sum_n p_n \tilde{\rho}_n$. For semi-classical measurements one always has $\tilde{\rho} = \rho$, since the system itself is not affected by the measurement, and throwing away the result eliminates any information that the measurement obtained. However there are many quantum measurements for which

$$\sum_n p_n \tilde{\rho}_n \neq \rho. \tag{2.39}$$

This fact actually embodies a specific notion of the *disturbance* caused by a quantum measurement, and we will discuss this in section 2.3.2. We will also consider below two results regarding the relationship between $\sum_n p_n \tilde{\rho}_n$ and the initial state, ρ.

If a measurement is efficient (that is, we have full information about the result of the measurement) then intuition would suggest that we should never know less about the system after the measurement than we did beforehand. That is, I_{sys} should never be negative. This is in fact true:

Theorem 6. Efficient measurements, on average, always increase the observers information about the system. That is

$$\sum_n p_n S(\tilde{\rho}_n) \leq S(\rho), \tag{2.40}$$

where $\tilde{\rho}_n = A_n \rho A_n^\dagger / p_n$, $p_n = \text{Tr}[A_n^\dagger A_n \rho]$, and $\sum_n A_n^\dagger A_n = I$.

Proof. The following elegant proof is due to Fuchs [12]. First note that for any operator A, the operator $A^\dagger A$ has the same eigenvalues as $AA^\dagger$. This follows immediately from the polar decomposition theorem: Writing $A = PU$ we have $AA^\dagger = P^2$, as well as $A^\dagger A = UP^2U^\dagger = UAA^\dagger U^\dagger$, and a unitary transformation does not change eigenvalues. We now note that

$$\rho = \sqrt{\rho} I \sqrt{\rho} = \sum_n \sqrt{\rho} A_n^\dagger A_n \sqrt{\rho} = \sum_n p_n \sigma_n, \tag{2.41}$$

where σ_n are density matrices given by

$$\sigma_n = \frac{\sqrt{\rho} A_n^\dagger A_n \sqrt{\rho}}{\text{Tr}[A_n^\dagger A_n \rho]} = \frac{\sqrt{\rho} A_n^\dagger A_n \sqrt{\rho}}{p_n}. \tag{2.42}$$

From the concavity of the von Nuemann entropy we now have

$$S(\rho) \geq \sum_n p_n S(\sigma_n). \tag{2.43}$$

But σ_n has the same eigenvalues as $\tilde{\rho}_n$: defining $X_n = \sqrt{\rho} A_n$ we have $\sigma_n = X_n X_n^\dagger$ and $\tilde{\rho}_n = X^\dagger X$. Since the von Neumann entropy only depends on the eigenvalues, $S(\sigma_n) = S(\tilde{\rho}_n)$, and the result follows. □

The above proof also shows us that someone who measures a system can always hide this measurement from other observers by using feedback (that is, by applying a unitary operation to the system that depends on the measurement result):

Theorem 7. Conditional unitary operations can always be used, following a measurement, to return a system to its initial state, from the point of view of an observer who does not know the measurement outcome. That is, there always exist $\{U_n\}$ *such that*

$$\rho = \sum_n p_n U_n \tilde{\rho}_n U_n^\dagger, \tag{2.44}$$

where $\tilde{\rho}_n = A_n \rho A_n^\dagger / p_n$, $p_n = \text{Tr}[A_n^\dagger A_n \rho]$, and $\sum_n A_n^\dagger A_n = I$.

Proof. The proof in theorem 6 above shows us that $\rho = \sum_n p_n \sigma_n$. Now since $\sigma_n = X_n X_n^\dagger$ and $\tilde{\rho}_n = X_n^\dagger X_n$, the polar decomposition theorem applied to X (again as per the proof in theorem 6) shows us that $\sigma_n = U_n \tilde{\rho}_n U_n^\dagger$ for some U_n. □

The fact that the average entropy never increases for an efficient measurement is a result of a stronger fact regarding majorization. This stronger result implies that *any* concave function of the eigenvalues of the density matrix will never increase, on average, under an efficient measurement. The majorization relation in question is

$$\boldsymbol{\lambda}(\rho) \prec \sum_n p_n \boldsymbol{\lambda}(\tilde{\rho}_n). \tag{2.45}$$

This result can be obtained by modifying the proof of theorem 6 only a little, and we leave it as an exercise.

For inefficient measurements the entropy reduction $\langle S(\rho) \rangle$ can be negative. This is because quantum measurements can disturb (that is, change the state) of a system. In general a measurement will disturb the system in a different way for each measurement result. If we lack full information about which result occurred, then we lack information about the induced disturbance, and this reduces our overall knowledge of the system following the measurement. This is best illustrated by a simple, and extreme, example. Consider a two-state system, with the basis states $|0\rangle$ and $|1\rangle$. We prepare the system in the state

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle). \tag{2.46}$$

Since the initial state is pure, we have complete knowledge prior to the measurement, and the initial von Neumann entropy is 0. We now make a measurement that projects the system into one of the states $|0\rangle$ or $|1\rangle$. The probabilities for the two outcomes are $p_0 = p_1 = 1/2$. But the catch is that we throw away all our information about the measurement result. After the measurement, because we do not know which result occurred, the system could be in state $|0\rangle$, or state $|1\rangle$, each with probability $1/2$. Our final density matrix is thus

$$\tilde{\rho} = \frac{1}{2}|0\rangle\langle 0| + |1\rangle\langle 1|. \tag{2.47}$$

This is the completely mixed state, so we have no information regarding the system after the measurement. The disturbance induced by the measurement, combined with our lack of knowledge regarding the result, was enough to completely mix the state.

Information about an encoded message

The second kind of information is relevant when a system has been prepared in a specific ensemble of states. In this case, the measurement provides information

about *which* of the states in the ensemble was actually prepared. As mentioned in section 1.5, this is useful in the context of a communication channel. To use a quantum system as a channel, an observer (the "sender") chooses a set of states $|\psi_j\rangle$ as an alphabet, and selects one of these states from a probability distribution $P(j)$. Another observer (the "recipient") then makes a measurement on the system to determine the message. From the point of view of the recipient the system is in the state

$$\rho = \sum_j P(j)|\psi_j\rangle\langle\psi_j|. \tag{2.48}$$

The result, m, of a quantum measurement on the system is connected to the value of j (the message) via a likelihood function $P(n|j)$. The likelihood function is determined by the quantum measurement and the set of states used in the encoding, as described in section 1.5. But once the likelihood function is determined, it describes a purely classical measurement on the classical variable j. Thus the information extracted about the message, which we will call I_{ens} (short for ensemble) is simply the mutual information:

$$I_{\text{ens}} = H[P(j)] - \sum_m P(m)H[P(j|m)] = H[J:M]. \tag{2.49}$$

On the right-hand side J denotes the random variable representing the message, and M the random variable representing the measurement result. When the encoding states are an orthogonal set, and the measurement operators commute with the density matrix, ρ, then the measurement reduces to a semi-classical measurement on the set of states labeled by j. It is simple to show that the result of this is

$$I_{\text{sys}} = \langle\Delta S\rangle = H[J:M] = I_{\text{ens}}, \qquad \text{(semi-classical measurement)}. \tag{2.50}$$

For a fixed density matrix, determining the measurement that maximizes I_{sys} is a nontrivial task (see, e.g. [31]), and the same is true for I_{ens}. For a given ensemble, the maximum possible value of I_{ens} has been given a special name — the *accessible information*. The maximum value of I_{sys} does not yet have a special name.

Bounds on information extraction

There are a number of useful bounds on I_{ens} and I_{sys}. The first is a bound on I_{sys} in terms of the entropy of the measurement results.

Theorem 8. [Bound on the entropy reduction] If a measurement is made on a system in state ρ, for which the final states are $\{\tilde{\rho}_n\}$, and the corresponding probabilities $\{p_n\}$, then

$$I_{\text{sys}} = S(\rho) - \sum_n p_n S(\tilde{\rho}_n) \leq H[\{p_n\}]. \tag{2.51}$$

Where the equality is achieved only for semi-classical measurements.

The proof of this result is exercises 3 and 4 at the end of this chapter. In Chapter 4 we will see that this bound is intimately connected with the second law of thermodynamics. This bound shows that quantum measurements are *less efficient*, from an information theoretic point of view, than classical measurements. The amount of information extracted about the quantum system is generally less than the randomness (the information) of the measurement results. That is, if we want to communicate the result of the measurement, we must use $H[\{p_n\}]$ nats in our message, but the amount this measurement tells us about the system is usually *less* than $H[\{p_n\}]$ nats. The above bound is also the origin of our claim in section 2.2.1, that the von Neumann entropy was the minimum possible entropy of the outcomes of a complete measurement on the system. To see this all we have to do is note that for a complete measurement all the final states are pure, so that $S(\rho_n) = 0$ for all n. In this case Eq.(2.51) becomes

$$H[\{p_n\}] \geq S(\rho). \tag{2.52}$$

The next set of bounds we consider are all special cases of the Schumacher-Westmoreland-Wootters (SWW) bound, a bound on the mutual information (I_{ens}) in terms of the initial ensemble and the final states that result from the measurement. This bound looks rather complex at first, and is most easily understood by introducing first one of the special cases: Holevo's bound.

Theorem 9. [Holevo's bound] If a system is prepared in the ensemble $\{\rho_j, p_j\}$, then

$$I_{\text{ens}} \leq S(\rho) - \sum_j p_j S(\rho_j), \tag{2.53}$$

where $\rho = \sum_j p_j \rho_j$ is the initial state. The equality is achieved if and only if $[\rho, \rho_j] = 0, \forall j$. The quantity χ that appears on the right-hand side is called the *Holevo χ-quantity* (pronounce "Kai quantity") for the ensemble.

Since the accessible information is defined as the maximum of I_{ens} over all possible measurements, Holevo's bound is a bound on the accessible information.

Recall that an incomplete (and possibly inefficient) measurement will in general leave a system in a mixed state. Further, as discussed in section 1.5, it will also leave the system in a new ensemble, and this ensemble is different for each measurement result. We are now ready to state the bound on the mutual information that we alluded to above:

Theorem 10. [The Schumacher-Westmoreland-Wootters Bound] A system is prepared in the ensemble $\{\rho_j, p_j\}$, whose Holevo χ-quantity we denote by χ. A measurement is then performed having N outcomes labeled by n, in which each outcome has probability p_n. If, on the n^{th} outcome, the measurement leaves the system in an ensemble whose Holevo χ-quantity is χ_n, then the mutual information is bounded by

$$I_{\text{ens}} \leq \chi - \sum_n p_n \chi_n. \tag{2.54}$$

The equality is achieved if and only if $[\rho, \rho_n] = 0, \forall n$, and the measurement is semi-classical. This result is also true for inefficient measurements, where in this case the measurement result is labeled by indices n, k, and the observer knows n but not k.

The proof of this theorem uses the strong subadditivity of the von Neumann entropy, and we give the details in Appendix D. The SWW bound can be understood easily, however, by realizing that the maximal amount of information that can be extracted by a subsequent measurement, after we have obtained the result n, is given by χ_n (Holevo's bound). So the average amount of information we can extract using a second measurement is $\sum_n p_n \chi_n$. Since the total amount of accessible information in the ensemble is χ, the first measurement must be able to extract no more than $\chi - \sum_n p_n \chi_n$.

In the measurement considered in theorem 10 above, the states in the initial ensemble are labeled by j, and the measurement results by n. By using the fact that the conditional probabilities for j and n satisfy $P(j|n)P(n) = P(n|j)P(j)$, it is possible to rewrite the SWW bound as

$$I_{\text{ens}} \leq \langle \Delta S(\rho) \rangle - \sum_j p_j \langle \Delta S(\rho_j) \rangle. \tag{2.55}$$

Here $\langle \Delta S(\rho) \rangle$ is, as usual, the average entropy reduction due to the measurement. The quantity $\langle \Delta S(\rho_j) \rangle$ is the entropy reduction that *would have* resulted from the measurement if the initial state had been the ensemble member ρ_i. For efficient measurements we know that $\langle \Delta S(\rho) \rangle$ is positive. In this case we are allowed to drop the last term, and the result is Hall's bound:

$$I_{\text{ens}} \leq \langle \Delta S(\rho) \rangle = I_{\text{sys}}, \qquad \text{(efficient measurements)}, \tag{2.56}$$

which says that for efficient measurements the mutual information is never greater than the average entropy reduction. For inefficient measurements, as we saw above, I_{sys} can be negative. Thus the last term in Eq.(2.55) is essential, because I_{ens} is always positive.

Information gain and initial uncertainty

In practice one tends to find that the more one knows about a system, the *less* information a measurement will extract. We will certainly see this behavior in many examples in later chapters. This is only a tendency, however, as counter examples can certainly be constructed. Because of this it is not immediately clear what fundamental property of measurements this tendency reflects. It is, in fact, the bound in Eq.(2.55) that provides the answer. Since I_{ens} is always nonnegative, Eq.(2.55) implies that

$$I_{\text{sys}}[\rho] = \langle \Delta S(\rho) \rangle \geq \sum_j p_j \langle \Delta S(\rho_j) \rangle = \sum_j p_j I_{\text{sys}}[\rho_j]. \tag{2.57}$$

But $\rho = \sum_j p_j \rho_j$, and thus Eq.(2.57) is nothing more than the statement that the average reduction in entropy is *concave* in the initial state. Choosing just two states in the ensemble, Eq.(2.57) becomes

$$I_{\rm sys}[p\rho_1 + (1-p)\rho_2] \geq pI_{\rm sys}[\rho_1] + (1-p)I_{\rm sys}[\rho_2]. \tag{2.58}$$

This tells us that if we take the average of two density matrices, a measurement made on the resulting state will reduce the entropy *more* than the reduction that it would give, on average, for the two individual density matrices. Of course, taking the average of two density matrices also results in a state that is more uncertain.

It is the concavity of $I_{\rm sys}$ that underlies the tendency of information gain to increase with initial uncertainty. A result of this concavity property is that there are special measurements for which $I_{\rm sys}$ *strictly* increases with the initial uncertainty. These are the measurements that are completely symmetric over the Hilbert space of the system. (That is, measurements whose set of measurement operators is invariant under all unitary transformations.) We wont give the details of the proof here, but it is simple to outline: the $I_{\rm sys}[\rho]$ for symmetric measurements is completely symmetric in the eigenvalues of ρ. Since $I_{\rm sys}[\rho]$ is also concave in ρ, this implies immediately that it is Schur-concave. This means that if $\rho \prec \sigma$, then $I_{\rm sys}[\rho] \geq I_{\rm sys}[\sigma]$.

Information, sequential measurements, and feedback

The information extracted by two classical measurements, made in sequence, is always less than or equal to the information that each would extract if it were the sole measurement. This is quite simple to show. Let us denote the unknown quantity by X, the result of the first measurement as the random variable Y, and the result of the second measurement as Z. The information extracted when both measurements are made in sequence is the mutual information between X and the joint set of outcomes labeled by the values of Y and Z. Another way to say this is that it is the mutual information between X and the random variable whose probability distribution is the joint probability distribution for Y and Z. We will write this as $I(X : Y, Z)$. Simply by rearranging the expression for $I(X : Y, Z)$, one can show that

$$I(X : Y, Z) = I(X : Y) + I(X : Z) - I(Y : Z). \tag{2.59}$$

The first two terms on the right-hand side are, respectively, the information that would be obtained about X if one made only the first measurement, and the corresponding information if one made only the second measurement. The final term is a measure of the correlations between Y and Z. Since the mutual information, and hence this final term, is always positive, we have

$$I(X : Y, Z) \leq I(X : Y) + I(X : Z). \tag{2.60}$$

The above inequality ceases to be true if we manipulate the state of the classical system between the measurements, in a manner that depends on the result of the

first measurement (that is, use feedback). Briefly, the reason for this is as follows. Performing a “feedback" process allows us to interchange the states of the classical system so that those states that we are most uncertain about following first measurement can be most easily distinguished by the second. Because we can do this for *every* outcome of the first measurement, it allows us to optimize the effect of the second measurement in a way that we could not do if we made the second measurement alone. Explicit examples of this will be described in Chapter 6, when we discuss feedback control and adaptive measurements.

Like a classical feedback process, the entropy reduction for general quantum measurements does not obey the inequality in Eq.(2.60). However, it is most likely that bare quantum measurements do. In fact, we can use Eq.(2.57) above to show that this is true when the initial state is proportional to the identity. We leave this as an exercise. The question of whether bare measurements satisfy this classical inequality for *all* initial states is an open problem. If the answer is affirmative, as seems very likely, then this would provide further motivation for regarding bare measurements as the quantum equivalent of classical measurements.

2.3.2 Quantifying Disturbance

Quantum measurements often cause a disturbance to the measured system. By this we do not mean that they merely change the observers state-of-knowlegde. All measurements, classical and quantum, necessarily change the observers state-of-knowledge as a result of the information they provide. What we mean by disturbance is that they cause a random change to some property of the system that increases the observers uncertainty in some way. Below we will discuss three distinct ways in which quantum measurements can be said to disturb a system. The second two are related, via context, to the two kinds of information, $I_{\rm sys}$ and $I_{\rm ens}$, discussed in the previous section. Further, each disturbance is in some way complementary to its related notion of information, in that one can identify situations in which there is a trade-off between the amount of information extracted, and the disturbance caused.

Disturbance to conjugate observables

The notion of disturbance that is taught in undergraduate physics classes is that induced in the momentum by a measurement of position, and vice versa. This is not a disturbance that increase the von Neumann entropy of our state-of-knowlegde, since in this case both the initial and final states can be pure. It is instead a disturbance that increases the Shannon entropy of the probability density for a *physical observable*.

One can show that measurements that extract information about an observable X will tend to reduce our information about any observable that does not commute with X. This is a result of an uncertainty relation that connects the variances of any two observables that do not commute. The derivation of this relation is quite short, and uses the Cauchy-Schwartz inequality. Let us assume that the commutator of two

observables, X and Y, is $[X, Y] = Z$, for some operator Z. For convenience we first define the operators $A = X - \langle X \rangle$ and $B = Y - \langle Y \rangle$. We then have

$$\begin{aligned}(\Delta X)^2(\Delta Y)^2 &= \langle A^2\rangle\langle B^2\rangle = (\langle\psi|A)(A|\psi\rangle)(\langle\psi|B)(B|\psi\rangle) \\ &\geq |\langle\psi|AB|\psi\rangle|^2 \\ &= \frac{1}{4i}(\langle\psi|[A,B]|\psi\rangle)^2 + \frac{1}{4}(\langle\psi|AB+BA|\psi\rangle)^2 \\ &\geq \frac{1}{4i}(\langle\psi|[X,Y]|\psi\rangle)^2 = \frac{1}{4i}\langle Z\rangle^2. \end{aligned} \tag{2.61}$$

Here the second line is given by the Cauchy-Schwartz inequality, that states that for two vectors $|a\rangle$ and $|b\rangle$, $\langle a|a\rangle\langle b|b\rangle \geq |\langle a|b\rangle|^2$. The third line is simply the fact that for any complex number z, $|z|^2 = (\mathrm{Re}[z])^2 + (\mathrm{Im}[z])^2$. The last line follows because the second term on the forth line is non-negative, and $[A, B] = [X, Y]$. For position and momentum this uncertainty relation is the famous Heisenberg uncertainty relation $\Delta x \Delta p \geq \hbar/2$.

The above uncertainty relation tells us that measurements that increase our knowledge of X enough that the final state has $\Delta X \leq \sqrt{\langle Z\rangle^2/(4i)}/\zeta$ for some constant ζ, must also leave us with an uncertainty for Y which is greater than ζ. This constitutes, loosely, a trade-off between information extracted about X and disturbance caused to Y, since the smaller ΔX, the larger the lower bound on ΔY. This notion of disturbance is relevant in a situation in which we are trying to control the values of one or more observables of a system, and this will be discussed further in Chapter 6.

Classically, an increase in an observers uncertainty is caused by a random change in the state that the observer does not have access to (usually called *noise*), and must therefore average over all the possible changes. Note that while the disturbance caused to Y is a result of the change induced in the state of the system by the quantum measurement, it is not induced because this change is random: if the initial state is pure, then on obtaining the measurement result we know exactly what change it has made to the system state. As we will see in Chapter 6, we can often interpret the increase of our uncertainty of Y as (classical) noise that has affected Y, but this is not always the case. A detailed discussion of this point is given in [32].

One can also obtain uncertainty relations that involve the Shannon entropy of two non-commuting observables, rather than merely the standard deviations. Consider two non-commuting observables, X and Y for a system of dimension N, and denote their respective eigenstates by $\{|x\rangle\}$ and $\{|y\rangle\}$, with $x, y = 1, \ldots, N$. The probability distributions for these two observables are given by

$$P_x = \langle x|\rho|x\rangle \quad \text{and} \quad P_y = \langle y|\rho|y\rangle, \tag{2.62}$$

where ρ is the state of the system. The sum of the entropies of the probability distributions for X and Y obey the Maassen-Uffink uncertainty relation

$$H[\{P_x\}] + H[\{P_y\}] \geq \max_{x,y} |\langle x|y\rangle|. \tag{2.63}$$

The proof of this is much more involved that the elementary uncertainty relation, Eq.(2.61), and is given in [33]. The Maassen-Uffink relation is different in kind from that in Eq.(2.61) because the lower bound does not depend on the state of the system. The lower bound in the elementary uncertainty relation depends on the state whenever $[X, Y]$ is not merely a number.

It is also possible to obtain entropic uncertainty relations for observables for infinite dimensional systems, and observables with a continuous spectrum. The Shannon entropy of a continuous random variable x, whose domain is $(-\infty, \infty)$, is defined as

$$H[P(x)] \equiv \int_{-\infty}^{\infty} P(x) \ln P(x) dx. \tag{2.64}$$

However, in this case the entropy can be negative. As a result, the entropy quantifies the relative difference in uncertainty between two distributions, rather than absolute uncertainty.

To state entropic uncertainty relations between observables of infinite dimensional systems, we must scale the observables so that their product is dimensionless. If we scale the position x and momentum p so that $[x, p] = i$, then the Hirschman-Beckner uncertainty relation is

$$H[P(x)] + H[P(p)] \geq \ln(2\pi). \tag{2.65}$$

The entropic uncertainty relations for angle and angular momentum are given in [34]. We note that entropic uncertainty relations have been extended to more general situations, that have applications to quantum communication, by Hall [35], Cerf *et al.* [36], and Renes and Boileau [37].

Noise added to a system: the entropy of disturbance

Semi-classical measurements, in providing information, reduce, on average, the entropy of the observers state-of-knowledge. They do not, however, affect the state of the system in the sense of applying forces to it. Because of this, if the observer is ignorant of the measurement outcome, then the measurement cannot change his or her state-of-knoweldge. This means that

$$\rho = \sum_n p_n \tilde{\rho}_n, \tag{2.66}$$

for all semi-classical measurements. Here, as usual, ρ is the initial state, and the $\tilde{\rho}_n$ are the final states that result from the measurement. This relation is simple to show using the fact that all the measurement operators commute with ρ for a semiclassical measurement (see section 1.3.1).

If one applies forces to a system, dependent on the measurement result, then the final state, averaged over all measurement results, can have less entropy than the initial state. This is the heart of the process of feedback control — one learns the state of a system by making a measurement, and then applies forces that take the

system from each possible state to the desired state. The result is a deterministic mapping from a probability distribution over a range of states, to a single state. This takes the system from a state with high entropy to one with zero entropy (in fact, this entropy does not vanish from the universe, but is transferred to the control system, something that will be discussed in Chapter 4.)

We recall from section 1.3.3 that the definition of a quantum measurement includes both information extraction, and the application of forces. If we consider only measurements whose operators contain no unitary part (bare measurements), then the measurement operators are all positive. A fundamental fact about bare quantum measurements is revealed by Ando's theorem:

Theorem 11. [Ando [27], Bare measurements and entropy] Consider a bare measurement described by the set of positive measurement operators $\{\mathcal{P}_n\}$. If the initial state of the system is ρ, then it is always true that

$$S\left(\sum_n p_n \tilde{\rho}_n\right) \geq S(\rho), \tag{2.67}$$

where $\tilde{\rho}_n = \mathcal{P}_n \rho \mathcal{P}_n / \mathrm{Tr}[\mathcal{P}_n^2 \rho]$ is the final state for the n^{th} measurement result, and p_n is its probability.

Proof. This proof is remarkably simple, but does require that we set up some notation. We note first that the map from ρ to $\tilde{\rho} = \sum_n p_n \tilde{\rho}_n$ is linear, and it takes the identity to the identity: if $\rho = I$ then $\sum_n p_n \tilde{\rho}_n = \sum_n \mathcal{P}_n I \mathcal{P}_n = \sum_n \mathcal{P}_n^2 = I$. Let us denote this map by Θ, so that $\tilde{\rho} = \Theta(\rho)$. Since the entropy only depends on the eigenvalues of the density matrix, what we are really interested in is the map that takes $\boldsymbol{\lambda}(\rho)$ to $\boldsymbol{\mu}(\tilde{\rho})$. Let us denote by $\mathrm{Diag}[\mathbf{v}]$ the diagonal matrix that has the vector $\mathbf{v}$ as its diagonal. Conversely, let us write the vector that is the diagonal of a matrix M by $\mathbf{diag}(M)$. If U is the unitary that diagonalizes ρ, and V is the unitary that diagonalizes $\tilde{\rho}$, then $\mathrm{Diag}[\boldsymbol{\mu}] = U \rho U^\dagger$ and $\mathrm{Diag}[\boldsymbol{\lambda}] = V \rho V^\dagger$. With this notation, we can now write the map, Φ, that takes $\boldsymbol{\lambda}$ to $\boldsymbol{\mu}$ as

$$\boldsymbol{\mu} = \Phi(\boldsymbol{\lambda}) \equiv \mathbf{diag}[V \Theta(U^\dagger \mathrm{Diag}[\boldsymbol{\lambda}] U) V^\dagger]. \tag{2.68}$$

It is simple to verify from the above expression that Φ is linear, and thus $\boldsymbol{\mu} = M\boldsymbol{\lambda}$ for some matrix D. Each column of D is given by $\boldsymbol{\mu}$ when $\boldsymbol{\lambda}$ is chosen to be one of the elementary basis vectors. From the properties of the map Θ, it is not difficult to show that all the elements of D are positive, and that D is probability preserving and unital (see section 2.2.2 for the definitions of these terms). Thus D is doubly stochastic, $\boldsymbol{\lambda} \succ \boldsymbol{\mu}$, and the result follows. □

Theorem 11 tells us that if we discard the measurement results, a bare quantum measurement cannot decrease the entropy of a system, just like a classical measurement. This provides further motivation for regarding bare measurements as the quantum equivalent of classical measurements. It is also simple to verify that, quite

unlike classical measurements, bear quantum measurements can *increase* the uncertainty of a system. This reflects the fact that quantum measurements can cause disturbance; if they did not disturb the system, then averaging over the measurement results must leave the initial state unchanged. The disturbance caused by a quantum measurement can therefore be quantified by

$$S_{\mathrm{D}} \equiv S\left(\sum_n p_n \tilde{\rho}_n\right) - S(\rho), \tag{2.69}$$

which we will refer to as the *entropy of disturbance*.

Once we allow our measurement operators to contain a unitary part, then, like classical measurements with feedback, quantum measurements can deterministically reduce the entropy of a system. The notion of disturbance captured by the "entropy of disturbance" is relevant for feedback control. One of the primary purposes of feedback is to reduce the entropy of a system. The entropy of disturbance measures the amount by which a quantum measurement acts like noise driving a system, and thus affects the process of feedback control. We will return to this in Chapter 6.

Disturbance to encoded information

The third notion of disturbance is that of a change caused to the states of an ensemble in which the system has been prepared. To illustrate this consider a von Neumann measurement that projects a three-dimensional quantum system onto the basis $\{|0\rangle, |1\rangle, |2\rangle\}$. If we prepare the system in the ensemble that consists of these same three states, then the measurement does not not disturb the states at all; if the system is initially in ensemble state $|n\rangle$, then it will still be in this state after the measurement. The only change is to our state-of-knoweldge — before the measurement we do know which of the ensemble states the system has been prepared in, and after the measurement we do.

What happens if we prepare the system instead in one of the three states

$$|a\rangle = \frac{|0\rangle + |2\rangle}{\sqrt{2}}, \tag{2.70}$$

$$|b\rangle = \frac{|1\rangle + |2\rangle}{\sqrt{2}}, \tag{2.71}$$

$$|c\rangle = |2\rangle. \tag{2.72}$$

This time, if the initial state is $|a\rangle$ or $|b\rangle$, the measurement is guaranteed to change it. In fact, because the three states are not all mutually orthogonal, there is *no* complete measurement that will not change at least one of the states in the ensemble, for at least one outcome. In this case we cannot extract the accessible information without causing a disturbance at least some of the time.

To quantify the notion of disturbance to an ensemble, we must have a measure of the *distance* between two quantum states. Otherwise we cannot say how *far* a

measurement has moved the initial state of a system. While we will describe such distance measures next, we will not need the present notion of disturbance here. The interested reader can find further details regarding this kind of disturbance in the articles by Barnum [17] and Fuchs [38].

2.4 Distinguishing Quantum States

Consider the following scenario: Alice prepares a quantum system in one of two predetermined states, and does not tell Bob which one. Bob then makes a measurement on the system to determine which of the two states the system is in. The question of how well Bob can achieve this goal is motivated by two quite different situations. The first is that in which Alice is trying to communicate a message to Bob, and in this case she would want to choose the states so that Bob could distinguish them as reliably as possible.

The second situation is that in which Alice is trying to control the state of a system. In this case there is a specific "target" state that Alice would ideally like the system to be in. In the presence of noise her control algorithm will not be able to achieve the target state perfectly, but will be able to get close to it. The questions are, how close, and what do we mean by close anyway? To answer the second, we mean that the effect of the system on the rest of the world is similar to the effect it has when it is in the target state. While the exact nature of this effect will depend on the situation, we can obtain a good general-purpose measure of the similarity of two states by considering the similarity of the results of measurements performed on them. This is directly related to the question of how well an observer, Bob, can distinguish those states, and this is the original question we posed above.

To distinguish the two states ρ and σ, an observer must make a measurement on the system, and decide upon one of the two states based on the result of the measurement. This is, of course, a special case of the more general situation we discussed in section 1.5, in which an observer extracts information about which of a set of N states have been prepared. It would be natural, therefore, to measure the distinguishability of two states as the mutual information between the measurement result and the variable denoting which of the states had been prepared, maximized over all measurements. Unfortunately this measure of distinguishability does not provide a closed-form expression in terms of ρ and σ.

The trace distance

Fortunately there is a measure of the "distance" between two quantum states that has a simple form, *and* which is directly related to a specific notion of how surely two states can be distinguished. Given two states described by the density matrices ρ and σ, this measure is given by

$$D(\rho, \sigma) \equiv \frac{1}{2}\mathrm{Tr}[|\rho - \sigma|], \tag{2.73}$$

and is called the *trace distance*. The reason that the factor of a half is included in the definition of the trace distance is so that its value lies between zero and unity.

There are two sensible notions of the "difference" between two quantum states to which the trace distance corresponds. The first, which is most relevant to problems in quantum control, is as follows. Let us say we make a measurement that has M outcomes, described by the measurement operators $\{A_m\}$. If the system is in state ρ, then each outcome has probability $p_m = \text{Tr}[A_m^\dagger A_m \rho]$, otherwise these probabilities are $q_m = \text{Tr}[A_m^\dagger A_m \sigma]$. It is the measurement result that constitutes the realized impact of the system on the rest of the world. Thus a reasonable measure of how similar the two states are is the absolute value of the difference between p_m and q_m, summed over all m. It turns out that the trace distance is half the maximum possible value of this sum, where the maximum is taken over all possible measurements. The proof of this is actually quite short:

Theorem 12. Given two density matrices, ρ and σ, and a measurement $\mathcal{M}$, described by the operators $\{A_m\}$, then

$$D(\rho,\sigma) \equiv \frac{1}{2} Tr[|\rho - \sigma|] = \frac{1}{2} \max_{\mathcal{M}} \sum_m |p_m - q_m| \tag{2.74}$$

where $p_m = \text{Tr}[A_m^\dagger A_m \rho]$ and $q_m = \text{Tr}[A_m^\dagger A_m \sigma]$. The maximum is taken over all possible measurements $\mathcal{M}$.

Proof. We need to note first that the operator $X = \rho - \sigma$ can always be written as the difference of two positive operators, each of which has support on orthogonal subspaces. (The support of an operator is the space spanned by those of its eigenvectors with non-zero eigenvalues.) To see this we note that since X is Hermitian all its eigenvalues are real. If the eigenvectors and eigenvalues of X are, respectively $\{|x_n\rangle\}$ and $\{x_n\}$, then $X = x_n |x_n\rangle\langle x_n|$. If we denote the positive eigenvalues of X as y_n, and the negative eigenvalues as z_n, we can define the positive operators $P \equiv y_n |y_n\rangle\langle y_n|$ and $Q \equiv -z_n |z_n\rangle\langle z_n|$. Then P and Q have support on orthogonal spaces ($PQ = 0$), and $X = P - Q$. This also means that $|\rho - \sigma| = P + Q$.

We now show that the trace distance is an upper bound on the sum in Eq.(2.74). We have

$$\begin{aligned}
\sum_m |p_m - q_m| &= \sum_m |\text{Tr}[A_m^\dagger A_m (\rho - \sigma)]| \\
&= \sum_m |\text{Tr}[A_m^\dagger A_m P] - \text{Tr}[A_m^\dagger A_m Q]| \\
&\leq \sum_m \text{Tr}[A_m^\dagger A_m P] + \text{Tr}[A_m^\dagger A_m Q] \\
&= \text{Tr}\left[\left(\sum_m A_m^\dagger A_m\right)|\rho - \sigma|\right] \\
&= \text{Tr}\left[|\rho - \sigma|\right] = 2D(\rho,\sigma).
\end{aligned} \tag{2.75}$$

All we have to do now is to find a measurement for which equality is reached in Eq.(2.74). We leave this as an exercise. □

The second sensible measure of distinguishability to which the trace distance corresponds is the minimum probability that Bob will make an error when trying to determine whether the state is ρ or σ. To decide which state Alice has prepared, Bob makes a measurement, and uses Bayesian inference, as described in section 1.5, to determine the probability that the system was initially prepared in each of the states. This depends upon the prior, which in this case is given by the probabilities with which Alice chose to prepare each of the states. For our purposes here, Alice choses each state with equal probability. Once Bob has determined the probabilities for each of the two states, his best guess is simply the one that is most likely. By averaging over all the measurement outcomes, we can calculate the total probability that Bob's best guess is wrong, and this is the "probability of error". Some measurements will give a smaller probability of error that others. Assuming that Alice prepares each of the two states with equal probability the expression for the error probability (which we leave as an exercise) is

$$P_{\text{err}} = \frac{1}{2}\sum_m \min(p_m, q_m), \tag{2.76}$$

where $\min(p_m, q_m)$ denotes the smaller of the two values p_m and q_m. It turns out that we can rewrite P_{err} as

$$P_{\text{err}} = \frac{1}{2} - \frac{1}{4}\sum_m |p_m - q_m|. \tag{2.77}$$

But we now know from theorem 12 that the minimum value of this, minimized over all measurements, is

$$\min_{\mathcal{M}} P_{\text{err}} = \frac{1}{2}[1 - D(\rho, \sigma)]. \tag{2.78}$$

Note that when the probability of error is $1/2$, Bob has no information about which of the two states the system is in.

To be a proper measure of distance (a metric) between two points x and y in some space, a measure $d(x, y)$ must satisfy three properties. It must be equal to zero if and only if $x = y$, it must be symmetric in its arguments (that is, $d(x, y) = d(y, x)$), and it must satisfy the triangle inequality, $d(x, y) \leq d(x, a) + d(a, y)$, where a is any third point in the space. It is clear that $D(\rho, \sigma)$ satisfies the first two. We give the proof of the third property in Appendix D.

The trace distance has a very nice relationship to the Bloch sphere. A summary of the Bloch sphere representation of a two-state system, also known as a qubit, is as follows. A density matrix for a two-state system can be represented by a three dimensional vector, $\mathbf{a}$, via the relation

$$\rho = \frac{1}{2}[I + \mathbf{a} \cdot \boldsymbol{\sigma}] = \frac{1}{2}[I + a_x\sigma_x + a_y\sigma_y + a_z\sigma_z]. \tag{2.79}$$

The vector $\mathbf{a}$ is called the *Bloch vector* for ρ. Here the σ_i are the Pauli spin matrices:

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_z = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{2.80}$$

The Pauli spin matrices all have zero trace. Their commutation relations are $[\sigma_i, \sigma_j] = 2i\sigma_k$ (where the indices i, j, k are in cyclic order) and $[\sigma_i, \sigma_j] = -2i\sigma_k$ (where they are in anti-cyclic order). For pure states the Bloch vector has unit length (and thus lies on the surface of a unit sphere), and for mixed states the Bloch vector has length $0 \leq \mathbf{a} < 1$. The elements of the Bloch vector are the expectation values of the three spin directions: $a_i = \text{Tr}[\sigma_i \rho]$.

The trace distance between two two-state density matrices, ρ and σ, whose Bloch vectors are, respectively, $\mathbf{a}$ and $\mathbf{b}$, is

$$D(\rho, \sigma) = \frac{|\mathbf{a} - \mathbf{b}|}{2}. \tag{2.81}$$

That is, half the geometrical distance between the two Bloch vectors.

The fidelity

There is another measure of the similarity between two quantum states, called the *Fidelity*. This is often used in quantum information theory in place of the trace distance because it has a simpler algebraic form. Further, a connection can be made between the trace distance and the fidelity. The fidelity between two density matrices is defined as

$$F(\rho, \sigma) \equiv \text{Tr}\left[\sqrt{\sigma^{1/2}\rho\sigma^{1/2}}\right]. \tag{2.82}$$

While not obvious, the fidelity is, in fact, symmetric in its two arguments. Its value lies between zero and unity, being zero if and only if ρ and σ have orthogonal supports, and unity if and only if $\rho = \sigma$. If σ is the pure state $|\psi\rangle$, then the fidelity reduces to the very simple form

$$F(\rho, |\psi\rangle) = \sqrt{\langle\psi|\rho|\psi\rangle}. \tag{2.83}$$

The connection between the fidelity and the trace distance is that, for every state ρ and σ,

$$1 - F(\rho, \sigma) \leq D(\rho, \sigma) \leq \sqrt{1 - [F(\rho, \sigma)]^2}. \tag{2.84}$$

When one of the states is pure, the lower inequality may also be replaced with

$$1 - [F(\rho, |\psi\rangle)]^2 \leq D(\rho, |\psi\rangle). \tag{2.85}$$

The proofs of the above properties of the fidelity may be found in [21], and the references at the end of this chapter.

2.5 Fidelity of Quantum Operations

For the purposes of controlling a quantum system, one may want to characterize how close the state of the system is to some "target" state, as discussed above. Further, one may wish to control not only the state of a system, but the entire mapping from a set of initial states to a set of final states. That is, one may wish to ensure that a system undergoes a specified unitary transformation. To determine how close the actual evolution is to the target unitary transformation, one needs a measure of the similarity of evolutions.

A transformation that specifies how all the initial states of a system will map to the final states is called a *quantum operation*. We can distinguish two kinds of quantum operations. Those that are probabilistic (those that depend upon the outcome of a measurement) and those that are deterministic (those for which a given input state always gives the same output state).

In the most general case, for a deterministic quantum operation, our system may interact with another "auxiliary" system, which is subsequently discarded. As we saw in Chapter 1, the effect of discarding the auxiliary system is the same as that of averaging over the results of a von Neumann measurement made on it. This, in turn, is the same as averaging over the results of some generalized measurement made on the primary system. Thus, if our system starts in the state ρ, the most general deterministic quantum operation may be written as

$$\varepsilon(\rho) = \sum_m A_m \rho A_m^\dagger, \tag{2.86}$$

where $\sum_m A_m^\dagger A_m = 1$. (For a probabilistic quantum operation, the most general case is that of an inefficient measurement, as described in section 1.4).

The standard measure of the similarity of two quantum operations is the fidelity between their output states, averaged over all pure input states. The average is, of course, taken with respect to the Harr measure. One is usually interested in engineering evolutions that are unitary operations. These are the most useful, because any other kind of operation involves discarding an auxiliary system, resulting in a loss of information. If we assume that our target evolution is given by the unitary operator U, and the actual evolution is given by the map ε, then the average fidelity is

$$\overline{F}(\varepsilon, U) = \int \langle\psi|U^\dagger \varepsilon(\rho) U|\psi\rangle \; d|\psi\rangle, \tag{2.87}$$

where $\rho = |\psi\rangle\langle\psi|$, and $\int d|\psi\rangle$ denotes integration with respect to the Haar measure (this measure is discussed in Appendix C).

Most often one needs to evaluate the fidelity of an operation that is simulated numerically, or implemented experimentally. In the above form the fidelity would be prohibitively difficult to calculate, since it involves obtaining a continuum of output states. Due to the fact that the space of operations is finite dimensional, it is possible

to rewrite the fidelity as a sum over a discrete set of initial states, and this makes numerical evaluation feasible.

To present the alternative expression for the fidelity, we need first to define a number of operators. Let us denote a basis for an N-dimensional system as $|n\rangle$, $n = 0, \ldots, N-1$. First we define an operators X that shifts each of the basis states "up by one":

$$X|n\rangle = |n+1\rangle, \qquad n = 0, \ldots, N-1, \tag{2.88}$$

where it is understood that $n = N$ is identified with $n = 0$. We also define the operator Z by

$$Z|n\rangle = e^{2\pi i n/N}|n\rangle, \qquad n = 0, \ldots, N-1. \tag{2.89}$$

Using these operators we define a set of N^2 unitary operators by

$$V_{jk} = X^j Z^k \,, \quad j, k = 0, 1, \ldots N-1. \tag{2.90}$$

Using this set of operators, the expression for the average fidelity is

$$\overline{F}(\varepsilon, U) = \frac{\sum_{j=1}^{N}\sum_{k=1}^{N} \mathrm{Tr}[U V_{jk}^{\dagger} U^{\dagger} \varepsilon(V_{jk})]}{N^2(N+1)}. \tag{2.91}$$

To evaluate this expression, one simulates the actual evolution of a system (the map ε) for the set of N^2 initial density matrices given by V_{jk}. Of course, since these matrices are unitary, they are not real initial states. Nevertheless one can still use them as initial states when performing the simulation, and this provides the "output states" $\varepsilon(V_{jk})$. One then substitutes these output states into Eq.(2.91).

History and Further Reading

Some of the results presented in this chapter have a long history, and some are very recent. The concept of entropy first arose in statistical mechanics, and the information theoretic quantity that Shannon discovered turned out to have the same form as thermodynamic entropy. The fact that Shannon's entropy is identical to the entropy of statistical mechanics is not coincidental, but fundamental. The essential connection between information theory and statistical mechanics will be discussed in Chapter 4. Further details regarding information theory can be found in, for example, "Elements of Information Theory" by Cover and Thomas [39].

The powerful and highly non-trivial subadditivity property of the von Neumann entropy has a fairly long history. Lanford and Robinson [40] conjectured this property in 1968, after Robinson and Ruelle [41] noted the importance of subadditivity of the classical entropy for statistical mechanics. It was finally proven in 1973 by Lieb and Ruskai [42, 43], using Lieb's theorem which was obtained in the same year [44]. Lieb's original proof of his eponymous theorem was rather involved, and it was some time before simple proofs of strong subadditivity were obtained. The first were given by Narnhofer and Thiring in 1985 [45], and Petz in 1986 [46] (a

straightforward presentation is given in Nielsen and Petz [47]). The simplest proof to date was obtained by Ruskai in 2006 [24].

The classification theorem for ensembles was obtained independently by Jaynes [48], and Hughston, Josza, and Wootters [49]. The majorazation result regarding the ensemble probabilities was obtained by Nielsen [29].

The intuitive (but nontrivial) result that efficient measurements never increase the average von Neumann entropy was first obtained by Ozawa in 1986 [50]. The more general result regarding majorization, and thus all concave functions, was obtained by Nielsen in 2000 [51] (published 2001). The very simple proof that we use here was obtained by Fuchs, and appears in Fuchs and Jacobs [12]). In his beautiful paper in 2001 Nielsen also re-derives in a simple and unified way a number of majorization inequalities, one of which gives the bound whose derivation we include as exercise 3, which was originally obtained by Lanford and Robinson [40].

The famous Holevo bound on the accessible information was obtained by Holevo in 1973 [52]. The stronger bound on the mutual information (theorem 10) was obtained for incomplete measurements by Schumacher, Westmoreland and Wootters (SWW) [13] in 1996. The extension to inefficient measurements is quite straightforward and was done in [53]. Barchielli and Lupieri also derived the general form of the SWW theorem, using more sophisticated methods [54].

The entropic uncertainty relation between position and momentum was proved by Beckner in 1975 [55]. The result that the disturbance from bare measurements is always nonnegative is due to Ando [27].

The minimum error probability was first obtained by Helstrom, and the connection with the trace distance was made by Fuchs. Proofs can be found in [56, 38]. The inequalities connecting the trace distance with the fidelity were obtained by Fuchs and van der Graph [57]. Properties of the trace distance and fidelity may be found in [58, 57, 59, 21].

Exercises

1. Show that for two independent random variables X and Y, the entropy of their joint probability distribution is the sum of their individual entropies.

2. Show that $\langle \Delta S \rangle$ reduces to the mutual information for a classical measurement.

3. We are going to show that if $\rho = \sum_i p_i \rho_i$, then

$$S(\rho) \leq H[\{p_i\}] + \sum_i p_i S(\rho_i). \tag{2.92}$$

   i) First consider the case when all the ρ_i are pure states. Show that for this case, theorem 4 implies that

$$S(\rho) \leq H[\{p_i\}].$$

(For pure states this result is the same as Eq.(2.92), because in this case $S(\rho_i) = 0, \forall i$.)

ii) Now consider the case in which the ρ_i are mixed. First we write all of the ρ_i in terms of their eigenbases, so that for each ρ_i we have $\rho_i = \sum_j \lambda_{ij} |v_{ij}\rangle\langle v_{ij}|$. This means that the state is given by

$$\rho = \sum_i \sum_j p_i \lambda_{ij} |v_{ij}\rangle\langle v_{ij}|.$$

Now use the result in i), and the fact that for each i, $\sum_j \lambda_{ij} = 1$, to show that Eq.(2.92) is true.

4. Use the inequality proved in exercise 2, above, and the trick used to prove theorem 6, to prove theorem 8.

5. Show that the classification theorem for ensembles (theorem 4) shows that any two pure-state ensembles that generate the same density matrix, are related by a unitary transformation.

6. Use the concavity relation for I_{sys} (Eq.(2.57)) to show that for bare measurements on an initially completely mixed state, I_{sys} satisfies the classical relation for sequential measurements (Eq.(2.60)).

7. Complete the proof of theorem 12.

8. Show that the probability of error, when trying to distinguish, from a single sample, two probability distributions given by $\{p_m\}$ and $\{q_m\}$, is given by Eq.(2.76). Show that this may also be written in the form given in Eq.(2.77).

9. Complete the proof of theorem 4.

10. Prove the inequality in Eq.(2.45) by using the method in the proof of theorem 6.

11. Rewrite Eq.(2.54) in the form given by Eq.(2.55).

12. *Further exercises to be added...*