Introduction To Information Theory

Edward Witten

PiTP 2018

We will start with a very short introduction to classical information theory (Shannon theory).

Suppose that you receive a message that consists of a string of symbols a or b, say

$$aababbaaaab \cdots$$

And let us suppose that a occurs with probability p, and b with probability $1 - p$.

How many bits of information can one extract from a long message of this kind, say with N letters?

For large N, the message will consist very nearly of pN occurrences of a and $(1-p)N$ occurrences of b. The number of such messages is

$$\begin{aligned} \frac{N!}{(pN)!((1-p)N)!} &\sim \frac{N^N}{(pN)^{pN}((1-p)N)^{(1-p)N}} \\ &= \frac{1}{p^{pN}(1-p)^{(1-p)N}} = 2^{NS} \end{aligned} \tag{1}$$

where S is the *Shannon entropy* per letter

$$S = -p \log p - (1-p)\log(1-p).$$

(The exponent is 2 because in information theory, one usually uses logarithms in base 2.)

The total number of messages of length N, given our knowledge of the relative probability of letters a and b, is roughly

$$2^{NS}$$

and so the number of bits of information one gains in actually observing such a message is

$$NS.$$

This is an asymptotic formula for large S, since we used only the leading term in Stirling's formula to estimate the number of possible messages, and we ignored fluctuations in the frequencies of the letters.

Suppose more generally that the message is taken from an alphabet with k letters a_1, a_2, $\cdots$, a_k, where the probability to observe a_i is p_i, for $i = 1, \cdots, k$. We write A for this probability distribution. In a long message with $N >> 1$ letters, the symbol a_i will occur approximately Np_i times, and the number of such messages is asymptotically

$$\frac{N!}{(p_1 N)^{p_1 N}(p_2 N)^{p_2 N} \cdots (p_k N)^{p_k N}} \sim \frac{N^N}{\prod_{i=1}^k (p_i N)^{p_i N}} = 2^{NS}$$

where now the entropy per letter is

$$S_A = -\sum_{i=1}^{k} p_i \log p_i .$$

This is the general definition of the Shannon entropy of a probability distribution for a random variable $\mathbf{a}$ that takes values $a_1, \ldots, a_k$ with probabilities $p_1, \ldots, p_k$. The number of bits of information that one can extract from a message with N symbols is again

$$NS_A$$

From the derivation, since the number 2^{NS_A} of possible messages is certainly at least 1, we have

$$S_A \geq 0$$

for any probability distribution. To get $S_A = 0$, there has to be only 1 possible message, meaning that one of the letters has probability 1 and the others have probability 0. The maximum possible entropy, for an alphabet with k letters, occurs if the p_i are all $1/k$ and is

$$S_A = -\sum_{i=1}^{k} (1/k) \log(1/k) = \log k.$$

(Exercise: prove this by using the method of Lagrange multipliers to maximize $-\sum_i p_i \log p_i$ with the constraint $\sum_i p_i = 1$.)

In engineering applications, NS_A is the number of bits to which a message with N letters can be compressed. In such applications, the message is typically not really random but contains information that one wishes to convey. However, in “lossless encoding,” the encoding program does not understand the message and treats it as random.

It is easy to imagine a situation in which one can make a better model by incorporating short range correlations between the letters. (For instance, the “letters” might be words in a message in the English language, and English grammar and syntax would dictate short range correlations.) A model incorporating such correlations would be a 1-dimensional classical spin chain of some kind with short range interactions. Estimating the entropy of a long message of N letters would be a problem in classical statistical mechanics. But in the ideal gas limit, in which we ignore correlations, the entropy of a long message is just NS where S is the entropy of a message consisting of only one word.

Even in the ideal gas model, we are making statements that are only natural in the limit of large N. To formalize the analogy with statistical mechanics, one could introduce a classical Hamiltonian H whose value for the i^{th} symbol a_i is $-\log p_i$, so that the probability of the i^{th} symbol in the thermodynamic ensemble is $2^{-H(a_i)} = p_i$. Notice then that in estimating the number of possible messages for large N, we ignored the difference between the canonical ensemble (defined by probabilities 2^{-H}) and the microcanonical ensemble (in which one specifies the precise numbers of occurrences of different letters). As is usual in statistical mechanics, the different ensembles are equivalent for large N. The equivalence between the different ensembles is important in classical and quantum information theory.

Now let us consider the following situation. Alice is trying to communicate with Bob, and she sends a message that consists of many letters, each being an instance of a random variable $\mathbf{x}$ whose possible values are $x_1, \cdots, x_k$. She sends the message over a noisy telephone connection, and what Bob receives is many copies of a random variable $\mathbf{y}$, drawn from an alphabet with letters $y_1, \cdots, y_r$. (Bob might confuse some of Alice's letters and misunderstand others.) How many bits of information does Bob gain after Alice has transmitted a message with N letters? To analyze this, let us suppose that $p(x_i, y_j)$ is the probability that, in a given occurrence, Alice sends $\mathbf{x} = x_i$ and Bob hears $\mathbf{y} = y_j$. The probability that Bob hears $\mathbf{y} = y_j$, summing over all choices of what Alice intended, is

$$p(y_j) = \sum_i p(x_i, y_j).$$

If Bob does hear $\mathbf{y} = y_j$, his estimate of the probability that Alice sent x_i is the *conditional probability*

$$p(x_i|y_j) = \frac{p(x_i, y_j)}{p(y_j)}.$$

From Bob's point of view, once he has heard $\mathbf{y} = y_j$, his estimate of the remaining entropy in Alice's signal is the Shannon entropy of the conditional probability distribution:

$$S_{X|\mathbf{y}=y_j} = -\sum_i p(x_i|y_j)\log(p(x_i|y_j)).$$

Averaging over all possible values of $\mathbf{y}$, the average remaining entropy, once Bob has heard $\mathbf{y}$, is

$$\begin{aligned}\sum_j p(y_j) S_{X|\mathbf{y}=y_j} &= -\sum_j p(y_j) \sum_i \frac{p(x_i, y_j)}{p(y_j)} \log\left(\frac{p(x_i, y_j)}{p(y_j)}\right) \\ &= -\sum_{i,j} p(x_i, y_j) \log p(x_i, y_j) + \sum_{i,j} p(x_i, y_j) \log p(y_j) \\ &= S_{XY} - S_Y.\end{aligned}$$

Here S_{XY} is the entropy of the joint distribution $p(x_i, y_j)$ for the pair $\mathbf{x}, \mathbf{y}$ and S_Y is the entropy of the probability distribution $p(y_j) = \sum_i p(x_i, y_j)$ for $\mathbf{y}$ only.

The difference $S_{XY} - S_Y$ is called the *conditional entropy* $S(X|Y)$; it is the entropy that remains in the probability distribution X once Y is known. Since it was obtained as a sum of ordinary entropies $S_{X|\mathbf{y}=y_j}$ with positive coefficients, it is clearly positive:

$$S_{XY} - S_Y \geq 0.$$

(This is *NOT* true quantum mechanically!) Since S_X is the total information content in Alice's message, and $S_{XY} - S_Y$ is the information content that Bob still does not have after observing Y, it follows that the information about X that Bob *does* gain when he receives Y is the difference or

$$I(X, Y) = S_X - S_{XY} + S_Y.$$

Here $I(X, Y)$ is called the *mutual information* between X and Y. It measures how much we learn about X by observing Y.

This interpretation convinces us that $I(X, Y)$ must be nonnegative. One can prove this directly but instead I want to deduce it from the properties of one more quantity – and this will essentially complete our cast of characters. I will motivate it as follows. Suppose we are observing a random variable $\mathbf{x}$, for example the final state in the decays of a radioactive nucleus. We have a theory that predicts a probability distribution Q for the final state, say the prediction is that the probability to observe final state i is q_i. But maybe our theory is wrong and the decay is actually described by some different probability distribution P, such that the probability of the i^{th} final state is p_i. After observing the decays of N atoms, how sure could we be that the initial hypothesis is wrong?

If the correct probability distribution is P, then after observing N decays, we will see outcome i approximately $p_i N$ times. We will judge the probability of what we have seen to be

$$\mathcal{P} = \prod_{i=1}^{N} q_i^{p_i N} \frac{N!}{\prod_{j=1}^{N}(p_j N)!}.$$

We already calculated that for large N

$$\frac{N!}{\prod_{j=1}^{N}(p_j N)!} \sim 2^{-N\sum_i p_i \log p_i}$$

so

$$\mathcal{P} \sim 2^{-N\sum_i p_i(\log p_i - \log q_i)}.$$

This is $2^{-NS(P||Q)}$ where the relative entropy (per observation) or Kullback-Liebler divergence is defined as

$$S(P||Q) = \sum_i p_i(\log p_i - \log q_i).$$

From the derivation, $S(P||Q)$ is clearly nonnegative, and zero only if $P = Q$, that is if the initial hypothesis is correct. If the initial hypothesis is wrong, we will be sure of this once

$$NS(P||Q) >> 1.$$

(Here we've ignored noise in the observations, which we could incorporate using what we learned in our discussion of conditional entropy.)

$S(P||Q)$ is an important measure of the difference between two probability distributions P and Q, but notice that it is asymmetric in P and Q. We broke the symmetry by assuming that Q was our initial hypothesis and P was the correct answer.

Now we will use this to prove positivity of the mutual information. We consider a pair of random variables **x**, **y** and we consider two different probability distributions. One, which we will call P, is defined by a possibly correlated joint probability distribution

$$p(x_i, y_j).$$

Given such a joint probability distribution, the separate probability distributions for **x** and for **y** are

$$p(x_i) = \sum_j p(x_i, y_j), \qquad p(y_j) = \sum_i p(x_i, y_j).$$

(I will always use this sort of notation for a reduced probability distribution in which some variable is "integrated out" or summed over.)

We define a second probability distribution Q for $\mathbf{x}$, $\mathbf{y}$ by ignoring the correlations between them:

$$q(x_i, y_j) = p(x_i)p(y_j).$$

Now we calculate the relative entropy between these two distributions:

$$\begin{aligned}
S(P||Q) =& \sum_{i,j} p(x_i, y_j)(\log p(x_i, y_j) - \log(p(x_i)p(y_j))) \\
=& \sum_{i,j} p(x_i, y_j)(\log p(x_i, y_j) - \log p(x_i) - \log p(y_j)) \\
=& S_X + S_Y - S_{XY} = I(X, Y).
\end{aligned}$$

Thus $I(X, Y) \geq 0$, with equality only if the two distributions are the same, meaning that $\mathbf{x}$ and $\mathbf{y}$ were uncorrelated to begin with.

The property

$$S_X + S_Y - S_{XY} \geq 0$$

is called *subadditivity* of entropy.

Now there is one more very important property of $S(P||Q)$ that I want to explain, and this will more or less conclude our introduction to classical information theory. Suppose that **x** and **y** are two random variables. Let P_{XY} and Q_{XY} be two probability distributions, described by functions $p(x_i, y_j)$ and $q(x_i, y_j)$. If we start with a hypothesis Q_{XY} for the joint probability, then after many trials in which we observe **x** and **y**, our confidence that we are wrong is determined by $S(P_{XY}||Q_{XY})$. But suppose that we only observe **x** and not **y**. The reduced distributions P_X and Q_X for X only are described by functions

$$p(x_i) = \sum_j p(x_i, y_j), \qquad q(x_i) = \sum_j q(x_i, y_j).$$

If we observe **x** only, then the confidence after many trials that the initial hypothesis is wrong is controlled by $S(P_X||Q_X)$.

It is harder to disprove the initial hypothesis if we observe only X, so

$$S(P_{XY}||Q_{XY}) \geq S(P_X||Q_X).$$

This is called *monotonicity of relative entropy*.

Concretely, if we observe a sequence $x_{i_1}, x_{i_2}, \ldots x_{i_N}$ in N trials, then to estimate how unlikely this is, we will imagine a sequence of y's that minimizes the unlikelihood of the joint sequence $(x_{i_1}, y_{i_1}), (x_{i_2}, y_{i_2}), \cdots (x_{i_N}, y_{i_N})$. An actual sequence of y's that we might observe can only be more unlikely than this. So observing y as well as x can only increase our estimate of how unlikely the outcome was, given the sequence of the x's. Thus, the relative entropy only goes down upon "integrating out" some variables and not observing them.

I think if you understand what I've said, you should regard it as a proof. However, I will also give a proof in formulas.

The inequality $S(P_{XY}||Q_{XY}) - S(P_X||Q_X) \geq 0$ can be written

$$\sum_{i,j} p(x_i, y_j) \left(\log\left(\frac{p(x_i, y_j)}{q(x_i, y_j)}\right) - \log\left(\frac{p(x_i)}{q(x_i)}\right) \right) \geq 0.$$

Equivalently

$$\sum_i p(x_i) \sum_j \frac{p(x_i, y_j)}{p(x_i)} \log\left(\frac{p(x_i, y_j)/p(x_i)}{q(x_i, y_j)/q(x_i)}\right) \geq 0.$$

The left hand side is a sum of positive terms, since it is

$$\sum_i p(x_i) S(P_{Y|\mathbf{x}=x_i}||Q_{Y|\mathbf{x}=x_i}),$$

where we define probability distributions $P_{Y|\mathbf{x}=x_i}$, $Q_{Y|\mathbf{x}=x_i}$ conditional on observing $\mathbf{x} = x_i$:

$$p(y_j)|_{\mathbf{x}=x_i} = p(x_i, y_j)/p(x_i), \quad q(y_j)|_{\mathbf{x}=x_i} = q(x_i, y_j)/q(x_i).$$

So this establishes monotonicity of relative entropy. An important special case is *strong subadditivity* of entropy. For this, we consider three random variables $\mathbf{x}, \mathbf{y}, \mathbf{z}$.The combined system has a joint probability distribution P_{XYZ} described by a function $p(x_i, y_j, z_k)$. Alternatively, we could forget the correlations between X and YZ, defining a probability distribution Q_{XYZ} for the system XYZ by

$$q(x_i, y_j, z_k) = p(x_i)p(y_j, z_k)$$

where as usual

$$p(x_i) = \sum_{j,k} p(x_i, y_j, z_k), \qquad p(y_j, z_k) = \sum_i p(x_i, y_j, z_k).$$

The relative entropy is $S(P_{XYZ}||Q_{XYZ})$.

But what if we only observe the subsystem XY? Then we replace P_{XYZ} and Q_{XYZ} by probability distributions P_{XY}, Q_{XY} with

$$p(x_i, y_j) = \sum_k p(x_i, y_j, z_k), \qquad q(x_i, y_j) = \sum_k q(x_i, y_j, z_k) = p(x_i)p(y_j)$$

and we can define the relative entropy $S(P_{XY}||Q_{XY})$.

Monotonicity of relative entropy tells us that

$$S(P_{XYZ}||Q_{XYZ}) \geq S(P_{XY}||Q_{XY}).$$

But the relation between relative entropy and mutual information that we discussed a moment ago gives

$$S(P_{XYZ}||Q_{XYZ}) = I(X, YZ) = S_X - S_{XYZ} + S_{YZ}$$

and

$$S(P_{XY}|Q_{XY}) = I(X, Y) = S_X - S_{XY} + S_Y.$$

So

$$S_X - S_{XYZ} + S_{YZ} \geq S_X - S_{XY} + S_Y$$

or

$$S_{XY} + S_{YZ} \geq S_X + S_{XYZ},$$

which is called *strong subadditivity*. Remarkably, the same statement turns out to be true in quantum mechanics, where it is both powerful and surprising.

One final comment before we get to the quantum mechanical case. We repeatedly made use of the ability to define a conditional probability distribution, conditional on some observation. There is not a good analog of this in the quantum mechanical case and it is a bit of a miracle that many of the conclusions nonetheless have quantum mechanical analogs. The most miraculous is strong subadditivity.

Now we turn to quantum information theory. Quantum mechanics always deals with probabilities, but the real quantum analog of a classical probability distribution is not a quantum state but a *density matrix*. Depending on one's view of quantum mechanics, one might believe that the whole universe is described by a quantum mechanical pure state that depends on all the available degrees of freedom. Even if this is true, one usually studies a subsystem that cannot be described by a pure state.

For an idealized case, let A be a subsystem of interest, with Hilbert space $\mathcal{H}_A$ and let B be everything else of relevance, or possibly all of the rest of the universe, with Hilbert space $\mathcal{H}_B$. The combined Hilbert space is the tensor product $\mathcal{H}_{AB} = \mathcal{H}_A \otimes \mathcal{H}_B$. The simple case is that a state vector ψ_{AB} of the combined system is the tensor product of a state vector $\psi_A \in \mathcal{H}_A$ and another state vector $\psi_B \in \mathcal{H}_B$:

$$\psi_{AB} = \psi_A \otimes \psi_B.$$

If ψ_{AB} is a unit vector, we can choose ψ_A and ψ_B to also be unit vectors. In the case of such a product state, measurements of the A system can be carried out by forgetting about the B system and using the state vector ψ_A. If $\mathcal{O}_A$ is any operator on $\mathcal{H}_A$, then the corresponding operator on $\mathcal{H}_{AB}$ is $\mathcal{O}_A \otimes 1_B$, and its expectation value in a factorized state $\psi_{AB} = \psi_A \otimes \psi_B$ is

$$\langle \psi_{AB} | \mathcal{O}_A \otimes 1_B | \psi_{AB} \rangle = \langle \psi_A | \mathcal{O}_A | \psi_A \rangle \langle \psi_B | 1_B | \psi_B \rangle = \langle \psi_A | \mathcal{O}_A | \psi_A \rangle.$$

However, a generic pure state $\psi_{AB} \in \mathcal{H}_{AB}$ is not a product state; instead it is "entangled." If $\mathcal{H}_A$ and $\mathcal{H}_B$ have dimensions N and M, then a generic state in $\mathcal{H}_{AB}$ is an $N \times M$ matrix, for example in the 2×3 case

$$\psi_{AB} = \begin{pmatrix} * & * & * \\ * & * & * \end{pmatrix}.$$

By unitary transformations on $\mathcal{H}_A$ and on $\mathcal{H}_B$, we can transform ψ_{AB} to

$$\psi_{AB} \to U\psi_{AB}V$$

where U and V are $N \times N$ and $M \times M$ unitaries. The canonical form of a matrix under that operation is a diagonal matrix, with positive numbers on the diagonal, and extra rows or columns of zeroes, for example

$$\begin{pmatrix} \sqrt{p_1} & 0 & 0 \\ 0 & \sqrt{p_2} & 0 \end{pmatrix}.$$

A slightly more invariant way to say this is that any pure state can be written

$$\psi_{AB} = \sum_i \sqrt{p_i} \psi_A^i \otimes \psi_B^i,$$

where we can assume that ψ_A^i and ψ_B^i are orthonormal,

$$\langle \psi_A^i, \psi_A^j \rangle = \langle \psi_B^i, \psi_B^j \rangle = \delta^{ij}$$

and that $p_i > 0$. This is called the Schmidt decomposition. (The ψ_A^i and ψ_B^i may not be bases of $\mathcal{H}_A$ or $\mathcal{H}_B$, because there may not be enough of them.) The condition for ψ_{AB} to be a unit vector is that

$$\sum_i p_i = 1,$$

so we can think of the p_i as probabilities.

What is the expectation value in such a state of an operator $\mathcal{O}_A$ that only acts on A? It is

$$\begin{aligned}\langle\psi_{AB}|\mathcal{O}_A\otimes 1_B|\psi_{AB}\rangle &= \sum_i p_i\langle\psi_A^i|\mathcal{O}_A|\psi_A^j\rangle\langle\psi_B^i|1_B|\psi_B^j\rangle\\ &= \sum_i p_i\langle\psi_A^i|\mathcal{O}_A|\psi_A^i\rangle.\end{aligned}$$

This is the same as

$$\mathrm{Tr}_{\mathcal{H}_A}\,\rho_A\mathcal{O}_A,$$

where ρ_A is the *density matrix*

$$\rho_A = \sum_i p_i|\psi_A^i\rangle\langle\psi_A^i|.$$

Thus, if we are only going to make measurements on system A, we do not need a wavefunction of the universe: it is sufficient to have a density matrix for system A.

From the definition,

$$\rho_A = \sum_i p_i |\psi_A^i\rangle\langle\psi_A^i|, \qquad (*)$$

we see that ρ_A is hermitian and positive semi-definite. Because $\sum_i p_i = 1$, ρ_A has trace 1:

$$\mathrm{Tr}_{\mathcal{H}_A}\, \rho_A = 1.$$

Conversely, every matrix with those properties can be "purified," meaning that it is the density matrix of some pure state on some "bipartite" (or two-part) system AB. For this, we first observe that any hermitian matrix ρ_A can be diagonalized, meaning that in a suitable basis it takes the form $(*)$; moreover, if $\rho_A \geq 0$, then the p_i are likewise positive (if one of the p_i vanishes, we omit it from the sum).

Having gotten this far, to realize ρ_A as a density matrix we simply introduce another Hilbert space $\mathcal{H}_B$ with orthonormal states ψ_B^i and observe that ρ_A is the density matrix of the pure state

$$\psi_{AB} = \sum_i \sqrt{p_i} \psi_A^i \otimes \psi_B^i \in \mathcal{H}_A \otimes \mathcal{H}_B.$$

Here ψ_{AB} is not unique (even after we choose B) but it is unique up to a unitary transformation of $\mathcal{H}_B$.

In this situation, ψ_{AB} is called a “purification” of the density matrix ρ_A. The existence of purifications is a nice property of quantum mechanics that has no classical analog: the classical analog of a density matrix is a probability distribution, and there is no notion of purifying a probability distribution.

If there is more than one term in the expansion

$$\psi_{AB} = \sum_i \sqrt{p_i} \psi_A^i \otimes \psi_B^i \in \mathcal{H}_A \otimes \mathcal{H}_B,$$

we say that systems A and B are entangled in the state ψ_{AB}. If there is only one term, the expansion reduces to

$$\psi_{AB} = \psi_A \otimes \psi_B,$$

an "unentangled" tensor product state. Then system A can be described by the pure state ψ_A and the density matrix is of rank 1:

$$\rho_A = |\psi_A\rangle\langle\psi_A|.$$

If ρ_A has rank higher than 1, we say that system A is in a mixed state. If ρ_A is invertible, we say that A is fully mixed (this will be the situation in quantum field theory) and if ρ_A is a multiple of the identity, we say that A is maximally mixed.

In the general case

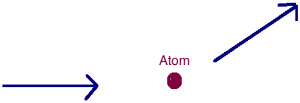
$$\rho_A = \sum_i p_i |\psi_A^i\rangle\langle\psi_A^i|, \qquad (*)$$

you will describe all measurements of system A correctly if you say that system A is in the state ψ_A^i with probability p_i. However, one has to be careful here because the decomposition $(*)$ is not unique. It is unique if the p_i are all distinct and one wants the number of terms in the expansion to be as small as possible, or equivalently if one wants the ψ_A^i to be orthonormal. But if one relaxes those conditions, then (except for a pure state) there are many ways to make the expansion $(*)$. This means that if Alice prepares a quantum system to be in the pure state ψ_i with probability p_i, then by measurements of that system, there is no way to determine the p_i or the ψ_i. Any measurement of the system will depend only on

$$\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i|.$$

Beyond that, there is no way to get information about how the system was prepared.

Before going on, perhaps I should give a simple example of a concrete situation in which it is impractical to not use density matrices. Consider an atom in a cavity illuminated by photons. A photon entering the cavity might be scattered, or it might be absorbed and reemitted:



After a certain time, the atom is again alone in the cavity. After the atom has interacted with n photons, to give a pure state description, we need a joint wavefunction for the atom and all n photons. The mathematical machinery gets bigger and bigger, even though (assuming we observe only the atom) the physical situation is not changing. By using a density matrix, we get a mathematical framework that does not change regardless of how many photons have interacted with the atom in the past (and what else those photons might have interacted with). All we need is a density matrix for the atom.

The von Neumann entropy of a density matrix ρ_A is defined by a formula analogous to the Shannon entropy of a probability distribution:

$$S(\rho_A) = -\mathrm{Tr}\,\rho_A \log \rho_A.$$

If

$$\rho_A = \sum_i p_i |\psi_A^i\rangle\langle\psi_A^i|,$$

with ψ_A^i being orthonormal, then

$$\rho_A \log \rho_A = \begin{pmatrix} p_1 \log p_1 & & & \\ & p_2 \log p_2 & & \\ & & p_3 \log p_3 & \\ & & & \ddots \end{pmatrix}$$

and so

$$S(\rho_A) = -\sum_i p_i \log p_i,$$

the same as the Shannon entropy of a probability distribution $\{p_i\}$.

An immediate consequence is that, just as for the Shannon entropy,

$$S(\rho_A) \geq 0,$$

with equality only for a pure state (one of the p's being 1 and the others 0). The formula $S(\rho_A) = -\sum_i p_i \log p_i$ also implies the same upper bound that we had classically for a system with k states

$$S(\rho_A) \leq \log k,$$

with equality only if ρ_A is a multiple of the identity:

$$\rho_A = \frac{1}{k} \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & \ddots \end{pmatrix}.$$

In this case, we say that the A is in a maximally mixed state. In fact, the von Neumann entropy has many properties analogous to the Shannon entropy, but the explanations required are usually more subtle and there are key differences.

Here is a nice property of the von Neumann entropy that does *not* have a classical analog. If a bipartite system AB is in a pure state

$$\psi_{AB} = \sum_i \sqrt{p_i} \psi_A^i \otimes \psi_B^i \in \mathcal{H}_A \otimes \mathcal{H}_B,$$

then the density matrices of systems A and B are

$$\rho_A = \sum_i p_i |\psi_A^i\rangle\langle\psi_A^i|,$$

and likewise

$$\rho_B = \sum_i p_i |\psi_B^i\rangle\langle\psi_B^i|.$$

The same constants p_i appear in each, so clearly

$$S(\rho_A) = S(\rho_B).$$

Thus a system A and a purifying system B always have the same entropy.

An important property of the von Neumann entropy is *concavity*. Suppose that ρ_1 and ρ_2 are two density matrices, and set $\rho(t) = t\rho_1 + (1-t)\rho_2$, for $0 \leq t \leq 1$. Then

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} S(\rho(t)) \leq 0.$$

To show this, we first compute that

$$\frac{\mathrm{d}}{\mathrm{d}t} S(\rho(t)) = -\mathrm{Tr}\,\dot{\rho}\log\rho.$$

(Exercise!) Then as

$$\log\rho = \int_0^\infty \mathrm{d}s \left(\frac{1}{s+1} - \frac{1}{s+\rho(t)}\right)$$

and $\ddot{\rho} = 0$, we have

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} S(\rho(t)) = -\int_0^\infty \mathrm{d}s \mathrm{Tr}\,\dot{\rho}\frac{1}{s+\rho(t)}\dot{\rho}\frac{1}{s+\rho(t)}.$$

The integrand is positive, as it is $\mathrm{Tr}\,B^2$, where B is the self-adjoint operator $(s+\rho(t))^{-1/2}\dot{\rho}(t)(s+\rho(t))^{-1/2}$. So $\frac{\mathrm{d}^2}{\mathrm{d}t^2} S(\rho(t)) \leq 0$.

In other words, the function $S(\rho(t))$ is concave. So the straight line connecting two points on its graph lies below the graph:

$$tS(\rho_1) + (1-t)S(\rho_2) \leq S(t\rho_1 + (1-t)\rho_2) = S(\rho(t)).$$

More generally, let ρ_i, $i = 1, \ldots, n$ be density matrices and p_i, $i = 1, \ldots, n$ nonnegative numbers with $\sum_i p_i = 1$. Then by induction from the above, or because this is a general property of concave functions, we have

$$\sum_i p_i S(\rho_i) \leq S(\rho), \quad \rho = \sum_i p_i \rho_i.$$

This may be described by saying that entropy can only increase under mixing. The nonnegative quantity that appears here is known as the Holevo information or Holevo χ:

$$\chi = S(\rho) - \sum_i p_i S(\rho_i).$$

Here is an example. Let ρ be any density matrix and (in some basis) let ρ_D be the corresponding diagonal density matrix. Let $\rho(t) = (1-t)\rho_D + t\rho$. Then

$$\frac{\mathrm{d}}{\mathrm{d}t} S(\rho(t))\bigg|_{t=0} = -\mathrm{Tr}\,\dot{\rho}(0)\log\rho_D = 0$$

since ρ_D is diagonal and $\dot{\rho}(0) = \mathrm{d}\rho/\mathrm{d}t|_{t=0}$ is purely off-diagonal. Also by concavity $\ddot{S}(\rho(t)) \le 0$. So $S(\rho(1)) \le S(\rho(0))$ or

$$S(\rho) \le S(\rho_D)$$

with equality only if $\rho = \rho_D$.

Just as for classical probability distributions, for density matrices we can always "integrate out" an unobserved system and get a reduced density matrix for a subsystem. Classically, given a joint probability distribution $p(x_i, y_j)$ for a bipartite system XY, we "integrated out" **y** to get a probability distribution for **x** only:

$$p(x_i) = \sum_j p(x_i, y_j).$$

The quantum analog of that is a partial trace. Suppose that AB is a bipartite system with Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B$ and a density matrix ρ_{AB}.

Concretely, if $|a_i\rangle$, $i = 1, \ldots, n$ are an orthonormal basis of $\mathcal{H}_A$ and $|b_\alpha\rangle$, $\alpha = 1, \ldots, m$ are an orthonormal basis of $\mathcal{H}_B$, then a density matrix for AB takes the general form

$$\rho_{AB} = \sum_{i,i',\alpha,\alpha'} c_{ii'\alpha\alpha'} |a_i\rangle \otimes |b_\alpha\rangle\langle a_{i'}| \otimes \langle b_{\alpha'}|.$$

The reduced density matrix for measurements of system A only is obtained by setting $\alpha = \alpha'$ and summing:

$$\rho_A = \sum_{i,i'\alpha} c_{i,i',\alpha,\alpha} |a_i\rangle\langle a_{i'}|.$$

This is usually written as a partial trace:

$$\rho_A = \mathrm{Tr}_{\mathcal{H}_B}\, \rho_{AB},$$

the idea being that one has "traced out" $\mathcal{H}_B$, leaving a density operator on $\mathcal{H}_A$. Likewise (summing over i to eliminate $\mathcal{H}_A$)

$$\rho_B = \mathrm{Tr}_{\mathcal{H}_A}\, \rho_{AB}.$$

It is now possible to formally imitate some of the other definitions that we made in the classical case. For example, if AB is a bipartite system, we define what is called quantum conditional entropy

$$S(A|B) = S_{AB} - S_B.$$

I need to warn you, though, that the name is a little deceptive because there is not a good quantum notion of conditional probabilities. A fundamental difference from the classical case is that $S(A|B)$ can be negative. In fact, suppose that system AB is in an entangled pure state. Then $S_{AB} = 0$ but as system B is in a mixed state, $S_B > 0$. So in this situation $S(A|B) < 0$. Nevertheless, one can give a reasonable physical interpretation to $S(A|B)$. I will come back to that when we discuss quantum teleportation.

Another classical definition that is worth imitating is the mutual information. Given a bipartite system AB with density matrix ρ_{AB}, the mutual information is defined just as it is classically:

$$I(A, B) = S_A - S_{AB} + S_B.$$

Here, however, we are more fortunate, and the quantum mutual information is nonnegative:

$$I(A, B) \geq 0.$$

Moreover, $I(A, B) = 0$ if and only if the density matrix factorizes, in the sense that

$$\rho_{AB} = \rho_A \otimes \rho_B.$$

Positivity of mutual information is also called subadditivity of entropy.

Before proving positivity of mutual information, I will explain an interesting corollary. Although conditional entropy $S(A|B)$ can be negative, the possibility of "purifying" a density matrix gives a lower bound on $S(A|B)$. Let C be such that ABC is in a pure state. Remember that in general if XY is in a pure state then $S_X = S_Y$. So if ABC is in a pure state then $S_{AB} = S_C$ and $S_B = S_{AC}$. Thus

$$S_{AB} - S_B = S_C - S_{AC} \geq -S_A,$$

where the last step is positivity of mutual information. So

$$S(A|B) = S_{AB} - S_B \geq -S_A.$$

This is the Araki-Lieb inequality; it is saturated if $S_{AB} = 0$ which implies $S_B = S_A$. This has been a typical argument exploiting the existence of purifications.

Just as in the classical case, to understand positivity of the mutual information, it helps to first define the relative entropy. Suppose that ρ and σ are two density matrices on the same Hilbert space $\mathcal{H}$.The relative entropy can be defined by imitating the classical formula:

$$S(\rho||\sigma) = \mathrm{Tr}\rho(\log\rho - \log\sigma).$$

$S(\rho||\sigma)$ turns out to have the same interpretation that it does classically: if your hypothesis is that a quantum system is described by σ, and it is actually described by ρ, then to learn that you are wrong, you need to observe N copies of the system where

$$NS(\rho||\sigma) >> 1.$$

Just as classically, it turns out that $S(\rho||\sigma) \geq 0$ with all ρ, σ, with equality precisely if $\rho = \sigma$. We will prove that in a moment, but first let us use it to prove that $I(A, B) \geq 0$ for any density matrix ρ_{AB}. Imitating the classical proof, we define

$$\sigma_{AB} = \rho_A \otimes \rho_B,$$

and we observe that

$$\log \sigma_{AB} = \log \rho_A \otimes 1_B + 1_A \otimes \log \rho_B,$$

so

$$\begin{aligned} S(\rho_{AB}||\sigma_{AB}) &= \mathrm{Tr}_{AB}\rho_{AB}(\log \rho_{AB} - \log \sigma_{AB}) \\ &= \mathrm{Tr}_{AB}\rho_{AB}(\log \rho_{AB} - \log \rho_A \otimes 1_B - 1_B \otimes \log \rho_B) \\ &= S_A + S_B - S_{AB} = I(A, B). \end{aligned}$$

So just as classically, positivity of the relative entropy implies positivity of the mutual information. (This is also called subadditivity of entropy.)

To prove positivity of the relative entropy, first observe that we can pick a basis in which σ is diagonal. If ρ is diagonal in the same basis, what we want reduces to the classical case. For if (say) $\sigma = \mathrm{diag}(q_1, \ldots, q_n)$, $\rho = \mathrm{diag}(p_1, \ldots, p_n)$, then

$$S(\rho||\sigma) = \sum_i p_i(\log p_i - \log q_i),$$

which can be interpreted as a classical relative entropy and so is nonnegative.

In the general case, let ρ_D be obtained from ρ by dropping off-diagonal matrix elements in the basis in which σ is diagonal. Then straight from the definitions we have

$$S(\rho||\sigma) = \mathrm{Tr}\rho(\log\rho - \log\sigma) = S(\rho_D||\sigma) + S(\rho_D) - S(\rho).$$

We just proved that $S(\rho_D||\sigma) \geq 0$, and earlier we used concavity to prove $S(\rho_D) - S(\rho) \geq 0$, so

$$S(\rho||\sigma) \geq 0,$$

with equality only if $\rho = \sigma$.

So relative entropy is positive, just as it is classically. Do we dare to hope that relative entropy is also monotonic, as classically? YES, as first proved by E. Lieb and M. B. Ruskai (1971). I consider this a miracle, because, as there is no such thing as a joint probability distribution for general quantum observables, the intuition behind the classical statement is not applicable in any obvious way. Rather, strong subadditivity is ultimately used to prove that quantities such as quantum conditional entropy and quantum relative entropy and quantum mutual information do have properties somewhat similar to the classical case.

There are different statements of monotonicity of relative entropy, but a basic one is monotonicity under partial trace. If AB is a bipartite system with two density matrices ρ_{AB} and σ_{AB}, then we can also take a partial trace on B to get reduced density matrices on A:

$$\rho_A = \mathrm{Tr}_B \rho_{AB}, \quad \sigma_A = \mathrm{Tr}_B \sigma_{AB}.$$

Monotonicity of relative entropy under partial trace is the statement that taking a partial trace can only reduce the relative entropy:

$$S(\rho_{AB}||\sigma_{AB}) \geq S(\rho_A||\sigma_A).$$

One of our main goals in lecture 2 will be to describe how this can be proved, so we will not discuss that now. I will just say that monotonicity of relative entropy has "strong subadditivity" as a corollary. We can prove it by imitating what we said classically. We consider a tripartite system ABC with density matrix ρ_{ABC}. There are reduced density matrices such as $\rho_A = \mathrm{Tr}_{BC}\rho_{ABC}$, $\rho_{BC} = \mathrm{Tr}_A\rho_{ABC}$, etc., and we define a second density matrix

$$\sigma_{ABC} = \rho_A \otimes \rho_{BC}.$$

The reduced density matrices of ρ_{ABC} and σ_{ABC}, obtained by tracing out C, are

$$\rho_{AB} = \mathrm{Tr}_C\rho_{ABC}, \qquad \sigma_{AB} = \mathrm{Tr}_C\sigma_{ABC} = \rho_A \otimes \rho_B.$$

Monotonicity of relative entropy under partial trace says that

$$S(\rho_{ABC}||\sigma_{ABC}) \geq S(\rho_{AB}||\sigma_{AB}). \qquad (*)$$

But (as in our discussion of positivity of mutual information)

$$S(\rho_{ABC}||\sigma_{ABC}) = S(\rho_{ABC}||\rho_A \otimes \rho_{BC}) = I(A, BC) = S_A + S_{BC} - S_{ABC}$$

and similarly

$$S(\rho_{AB}||\sigma_{AB}) = S(\rho_{AB}||\rho_A \otimes \rho_B) = I(A, B) = S_A + S_B - S_{AB}.$$

So $(*)$ becomes *monotonicity of mutual information*

$$I(A, BC) \geq I(A, B)$$

or equivalently *strong subadditivity*

$$S_{AB} + S_{BC} \geq S_B + S_{ABC}.$$

Note that these steps are the same as they were classically.

Using purifications, one can find various equivalent statements. If $ABCD$ is in a pure state then $S_{AB} = S_{CD}$, $S_{ABC} = S_D$ so the inequality becomes

$$S_{CD} + S_{BC} \geq S_B + S_D.$$

So for instance $S(C|D) = S_{CD} - S_D$ can be negative, or $S(C|B) = S_{BC} - S_B$ can be negative, but

$$S(C|D) + S(C|B) \geq 0.$$

(This is related to "monogamy of entanglement": a given qubit in C can be entangled with D, reducing S_{CD}, or with B, reducing S_{BC}, but not both.)

Classically, the intuition behind monotonicity of mutual information

$$I(A, BC) \geq I(A, B)$$

is that knowing both B and C will tell you at least as much about A as you would learn from knowing B only. Quantum mechanically, it is just not obvious that the formal definition $I(A, B) = S_A - S_{AB} + S_B$ fits that intuition. It takes the deep result of monotonicity of relative entropy – or strong subadditivity – to show that it does.

In general, strong subadditivity (or monotonicity of relative entropy) is the key to most of the interesting statements in quantum information theory. Most useful statements that are not more trivial are deduced from strong subadditivity.

Once we start using density matrices, there are a few more tools we should add to our toolkit. First let us discuss measurements. Textbooks begin with “projective measurements,” which involve projection onto orthogonal subspaces of a Hilbert space $\mathcal{H}$ of quantum states. We pick positive commuting hermitian projection operators π_s, $s = 1, \cdots, k$ obeying

$$\sum_s \pi_s = 1, \qquad \pi_s^2 = \pi_s, \qquad \pi_s \pi_{s'} = \pi_{s'} \pi_s.$$

A measurement of a state ψ involving these projection operators has outcome s with probability

$$p_s = \langle \psi | \pi_s | \psi \rangle.$$

These satisfy $\sum_s p_s = 1$ since $\sum_s \pi_s = 1$. If instead of a pure state ψ the system is described by a density matrix ρ, then the probability of outcome s is

$$p_s = \mathrm{Tr}_{\mathcal{H}}\, \pi_s \rho.$$

But Alice can make a more general type of measurement using an auxiliary system C with Hilbert space $\mathcal{C}$. We suppose that $\mathcal{C}$ is s-dimensional with a basis of states $|s\rangle$, $s = 1, \cdots, k$. Alice initializes $\mathcal{C}$ in the state $|1\rangle$. Then she acts on the combined system $\mathcal{C} \otimes \mathcal{H}$ with a unitary transformation U, which she achieves by suitably adjusting a time-dependent Hamiltonian. She chooses U so that for any $\psi \in \mathcal{H}$

$$U(|1\rangle \otimes \psi) = \sum_{s=1}^{k} |s\rangle \otimes E_s \psi$$

for some linear operators E_s. (She doesn't care about what U does on other states.) Unitarity of U implies that

$$\sum_{s=1}^{k} E_s^\dagger E_s = 1,$$

but otherwise the E_s are completely arbitrary.

Then Alice makes a projective measurement of the system $\mathcal{C} \otimes \mathcal{H}$, using the commuting projection operators

$$\pi_s = |s\rangle\langle s| \otimes 1,$$

which have all the appropriate properties. The probability of outcome s is

$$p_s = |E_s|\psi\rangle|^2 = \langle\psi|E_s^\dagger E_s|\psi\rangle.$$

More generally, if the system $\mathcal{H}$ is described initially by a density matrix ρ, then the probability of outcome s is

$$p_s = \mathrm{Tr}\, E_s^\dagger E_s \rho.$$

The p_s are nonnegative because $E_s^\dagger E_s$ is nonnegative, and $\sum_s p_s = 1$ because $\sum_s E_s^\dagger E_s = 1$. But the $E_s^\dagger E_s$ are not commuting projection operators; they are just nonnegative hermitian operators that add to 1. What we've described is a more general kind of quantum mechanical measurement of the original system. (In the jargon, this is a "positive operator-valued measurement" or POVM.)

Now let us view this process from another point of view. How can a density matrix evolve? The usual Hamiltonian evolution of a state ψ is $\psi \to U\psi$ for a unitary operator U, and on the density matrix it corresponds to

$$\rho \to U\rho U^{-1}.$$

But let us consider Alice again with her extended system $\mathcal{C} \otimes \mathcal{H}$. She initializes the extended system with density matrix

$$\widehat{\rho} = |1\rangle\langle 1| \otimes \rho$$

where ρ is a density matrix on $\mathcal{H}$. Then she applies the same unitary U as before, mapping $\widehat{\rho}$ to

$$\widehat{\rho}' = U\widehat{\rho}U^{-1} = \sum_{s,s'=1}^{t} |s\rangle\langle s'| \otimes \; E_s\rho E_{s'}^{\dagger}.$$

The induced density matrix on the original system $\mathcal{H}$ is obtained by a partial trace and is:

$$\rho' = \mathrm{Tr}_{\mathcal{C}}\widehat{\rho}' = \sum_{s=1}^{k} E_s\rho E_s^{\dagger}.$$

We have found a more general way that density matrices can evolve. The operation

$$\rho \to \sum_{s=1}^{k} E_s \rho E_s^\dagger, \qquad \sum_s E_s^\dagger E_s = 1$$

is called a "quantum channel," and the E_s are called Kraus operators.

This is actually the most general physically sensible evolution of a density matrix (and a POVM is the most general possible measurement of a quantum system).

Now let ρ and σ be two different density matrices on $\mathcal{H}$. Let us ask what happens to the relative entropy $S(\rho||\sigma)$ when we apply a quantum channel, mapping ρ and σ to

$$\rho' = \sum_s E_s \rho E_s^\dagger, \qquad \sigma' = \sum_s E_s \sigma E_s^\dagger.$$

The first step of initialization, replacing ρ and σ by $|1\rangle\langle 1| \otimes \rho$ and $|1\rangle\langle 1| \otimes \sigma$, doesn't change anything. The second step, conjugating by a unitary matrix U, also doesn't change anything since the definition of relative entropy is invariant under conjugation. Finally the last step was a partial trace, which can only reduce the relative entropy. So relative entropy can only go down under a quantum channel:

$$S(\rho||\sigma) \geq S(\rho'||\sigma').$$

This is the most general statement of monotonicity of relative entropy.

As an example of this, let us suppose that σ is a thermal density matrix at some temperature $T = 1/\beta$

$$\sigma = \frac{1}{Z}\exp(-\beta H).$$

So $\log\sigma = -\beta H - \log Z$ and therefore the relative entropy between any density matrix ρ and σ is

$$\begin{aligned} S(\rho||\sigma) =& \mathrm{Tr}\,\rho(\log\rho - \log\sigma) = -S(\rho) + \mathrm{Tr}\rho(\beta H + \log Z) \\ =& \beta(E(\rho) - TS(\rho)) + \log Z \end{aligned} \tag{2}$$

where the average energy computed in the density matrix ρ is

$$E(\rho) = \mathrm{Tr}\,\rho H.$$

We define the free energy

$$F(\rho) = E(\rho) - TS(\rho).$$

Note that the $\log Z$ term is independent of ρ and gives a constant that ensures that $S(\sigma||\sigma) = 0$. So

$$S(\rho||\sigma) = \beta(F(\rho) - F(\sigma)).$$

Now consider any evolution of the system, that is any quantum channel, that preserves thermal equilibrium at temperature β. Thus, this channel maps σ to itself, but it maps ρ to a generally different density matrix ρ'. The relative entropy can only go down under a quantum channel, so

$$S(\rho||\sigma) \geq S(\rho'||\sigma),$$

and therefore

$$F(\rho) \geq F(\rho').$$

In other words, the free energy can only go down under a quantum channel that preserves thermal equilibrium. This is an aspect of the second law of thermodynamics. If you stir a system in a way that maps thermal equilibrium at temperature T to thermal equilibrium at the same temperature, then it moves any density matrix closer to thermal equilibrium at temperature T.

Exercises:

(1) Let ψ be an arbitrary pure state. Try to find Kraus operators of a quantum channel that maps any density matrix to $|\psi\rangle\langle\psi|$. Can you describe a physical realization?

(2) Same question for a quantum channel that maps any density matrix to a maximally mixed density matrix, a multiple of the identity.

(3) Same question for a quantum channel that maps any $\rho = (\rho_{ij})$ to the corresponding diagonal density matrix $\rho_D = \mathrm{diag}(\rho_{11}, \rho_{22}, \cdots, \rho_{nn})$.

Our next topic will be quantum teleportation. For a first example, imagine that Alice has in her possession a “qubit” A_0, a quantum system with a two-dimensional Hilbert space. Alice would like to help Bob create in his lab a qubit in a state identical to A_0. However, she doesn’t have the technology to actually send a qubit; she can only communicate by sending a classical message over the telephone. If Alice knows the state of her qubit, there is no problem: she tells Bob the state of her qubit and he creates one like it in his lab. If, however, Alice does not know the state of her qubit, she is out of luck. All she can do is make a measurement, which will give some information about the prior state of qubit A_0. She can tell Bob what she learns, but the measurement will destroy the remaining information about A_0 and it will never be possible for Bob to recreate A_0.

Suppose, however, that Alice and Bob have previously shared a qubit pair $A_1 B_1$ (Alice has A_1, Bob has B_1) in a known entangled state, for example

$$\Psi_{A_1 B_1} = \frac{1}{\sqrt{2}} \left(|0\,0\rangle + |1\,1\rangle \right).$$

Maybe Alice created this pair in her lab and then Bob took B_1 on the road with him, leaving A_1 in Alice's lab. In this case, Alice can solve the problem. To do so she makes a joint measurement of her system $A_0 A_1$ in a basis that is chosen so that no matter what the answer is, Alice learns nothing about the prior state of A_0. In that case, she also loses no information about A_0. But after getting her measurement outcome, she knows the full state of the system and she can tell Bob what to do to create A_0.

To see how this works, let us describe a specific measurement that Alice can make on A_0A_1 that will shed no light on the state of A_0. She can project A_0A_1 on the basis of four states

$$\frac{1}{\sqrt{2}}(|0\,0\rangle \pm |1\,1\rangle) \text{ and } \frac{1}{\sqrt{2}}(|0\,1\rangle \pm |1\,0\rangle).$$

To see the result of a measurement, suppose the unknown state of qubit A_0 is $\alpha|0\rangle + \beta|1\rangle$. So the initial state of $A_0A_1B_1$ is

$$\Psi_{A_0A_1B_1} = \frac{1}{\sqrt{2}}\left(\alpha|0\,0\,0\rangle + \alpha|0\,1\,1\rangle + \beta|1\,0\,0\rangle + \beta|1\,1\,1\rangle\right).$$

Suppose that the outcome of Alice's measurement is to learn that A_0A_1 is in the state $\frac{1}{\sqrt{2}}(|0\,0\rangle - |1\,1\rangle)$. After the measurement B_1 will be in the state $\alpha|0\rangle - \beta|1\rangle$. (Exercise!) Knowing this, Alice can tell Bob that he can recreate the initial state by acting on his qubit by

$$\Psi_{B_1} \to \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \Psi_{B_1}$$

in the basis $|0\rangle$, $|1\rangle$. The other cases are similar (Exercise!).

I want to explain a generalization, but first it is useful to formalize in a different way the idea that Alice is trying to teleport an arbitrary unknown quantum state. For this, we add another system R, to which Alice does not have access. We assume that R is maximally entangled with A_0 in a known state, say

$$\Psi_{RA_0} = \frac{1}{\sqrt{2}} \left(|0\,0\rangle + |1\,1\rangle \right).$$

In this version of the problem, Alice's goal is to manipulate her system $A_0 A_1$ in some way, and then tell Bob what to do so that in the end the system RB_1 will be in the same state

$$\Psi_{RB_1} = \frac{1}{\sqrt{2}} \left(|0\,0\rangle + |1\,1\rangle \right)$$

that RA_0 was previously – with R never being touched. In this version of the problem, the combined system $RAB = RA_0A_1B_1$ starts in a pure state $\Psi_{RAB} = \Psi_{RA_0} \otimes \Psi_{A_1B_1}$. The solution of this version of the problem is the same as the other one: Alice makes the same measurements and sends the same instructions as before.

We can understand better what is happening if we take a look at the *conditional entropy* of the system $AB = A_0A_1B_1$. Since A_1B_1 is in a pure state, it does not contribute to S_{AB} and $S_{AB} = S_{A_0} = 1$ (A_0 is maximally mixed, since it is maximally entangled with R). Also $S_B = 1$ since $B = B_1$ is maximally entangled with A_1. So

$$S(A|B) = S_{AB} - S_B = 1 - 1 = 0.$$

It turns out that this is the key to quantum teleportation: teleportation is possible when and only when

$$S(A|B) \leq 0.$$

Let me explain why this is a necessary condition. We start with an arbitrary system RAB in a pure state Ψ_{RAB}; Alice has access to A, Bob has access to B, and no one has access to R. For teleportation, Alice will measure her system A using some rank 1 orthogonal projection operators π_i. No matter what answer she gets, after the measurement, system A is in a pure state and therefore RB is also in a pure, generally entangled state. For teleportation, Alice has to choose the π_i so that, no matter what the outcome she gets, the density matrix ρ_R of R is the same as before. If this is so, then after her measurement, system RB is in an entangled pure state with ρ_R unchanged. Any two such states can be converted into each other by a unitary transformation of system B (Exercise!) which Bob can implement. Since she knows her measurement outcome, Alice knows which entangled state RB is in and knows what instructions to give Bob.

But do projection operators of Alice's system with the necessary properties exist? The initial state Ψ_{ABR} is pure so it has

$$S_{AB} = S_R.$$

Bob's density matrix at the beginning is

$$\rho_B = \mathrm{Tr}_{RA}\,\rho_{RAB}$$

where ρ_{RAB} is the initial pure state density matrix. By definition

$$S_B = S(\rho_B).$$

If Alice gets measurement outcome i, then Bob's density matrix after the measurement is

$$\rho_B^i = \frac{1}{p_i}\mathrm{Tr}_{RA}\,\pi_i\rho_{RAB}.$$

Note that

$$\rho_B = \sum_i p_i\rho_B^i,$$

since $\sum_i \pi_i = 1$.

After the measurement, since A is in a pure state, RB is also in a pure state Ψ^{i}_{RB}, so $S(\rho^{i}_{B}) = S_R$. But by hypothesis, the measurement did not change ρ_R, so S_R is unchanged and so equals the original S_{AB}. Hence

$$S(\rho^{i}_{B}) = S_{AB}.$$

If all this is possible

$$S_{AB} = S(\rho^{i}_{B}) = \sum_i p_i S(\rho^{i}_{B}).$$

But earlier we found, using concavity, a general entropy inequality for mixing; if as here $\rho_B = \sum_i p_i \rho^{i}_{B}$ then

$$S(\rho_B) \geq \sum_i p_i S(\rho^{i}_{B}).$$

So if teleportation can occur,

$$S_{AB} = \sum_i p_i S(\rho^{i}_{B}) \leq S(\rho_B) = S_B$$

and hence $S(A|B) = S_{AB} - S_B \leq 0$.

Actually, $S(A|B) \leq 0$ is sufficient as well as necessary for teleportation, in the following sense. One has to consider the problem of teleporting not a single system but N copies of the system for large N. (This is a standard simplifying device in information theory, which unfortunately we haven't had time to explore. It generalizes the idea we started with of a long classical message.) So one takes N copies of system RAB for large N, thus replacing RAB by $R^{\otimes N}A^{\otimes N}B^{\otimes N}$. This multiplies all the entropies by N, so it preserves the conditions $S(A|B) \leq 0$. Now Alice tries to achieve teleportation by making a complete projective measurement on her system $A^{\otimes N}$. It is very hard to find an explicit set of projection operators π_i with the right properties, but it turns out, remarkably, that for large N, a random choice will work (in the sense that with a probability approaching 1, the error in teleportation is vanishing for $N \to \infty$). This statement actually has strong subadditivity as a corollary. (This is not the approach that I will take tomorrow.) What I've described is the "state merging" protocol of Horodecki, Oppenheim, and Winter; see also Preskill's notes.

We actually can now give a good explanation of the meaning of quantum conditional entropy $S(A|B)$. Remember that classically $S(A|B)$ measures how many additional bits of information Alice has to send to Bob after he has already received B, so that he will have full knowledge of A. Quantum mechanically, suppose that $S(A|B) > 0$ and Alice nevertheless wants to share her state with Bob, minimizing the quantum communication required. To do this, she first creates some maximally entangled qubit pairs and sends half of each pair to Bob. Each time she sends Bob half of a pair, S_{AB} is unchanged but S_B goes up by 1 (check!), so $S(A|B) = S_{AB} - S_B$ goes down by 1. So $S(A|B)$, if positive, is the number of such qubits that Alice must send to Bob to make $S(A|B)$ negative and so make teleportation possible.

If $S(A|B)$ is negative, teleportation is possible to begin with and $-S(A|B)$ is the number of entangled qubit pairs that Alice and Bob can be left with after teleportation.

Now we are going to address the following question: how many bits of information can Alice send to Bob by sending him a quantum system X with an N-dimensional Hilbert space $\mathcal{H}$? One thing Alice can do is to send one of N orthogonal basis vectors in $\mathcal{H}$. Bob can find which one she sent by making a measurement. So in that way Alice can send $\log N$ bits of information. We will see that in fact it is not possible to do better.

We suppose that Alice wants to encode a random variable $\mathbf{x}$ that takes the values x_i, $i = 1, \ldots, n$ with probability p_i. When $\mathbf{x} = x_i$, she writes down this fact in her notebook C and creates a density matrix ρ_i on system X. If $|i\rangle$ is the state of the notebook when Alice has written that $\mathbf{x} = x_i$, then on the combined system CX, Alice has created the density matrix

$$\widehat{\rho} = \sum_i p_i |i\rangle\langle i| \otimes \rho_i$$

Then Alice sends the system X to Bob. Bob's task is to somehow extract information by making a measurement.

Before worrying about what Bob can do, let us observe that the density matrix $\widehat{\rho}$ of the system CX is the one that I mentioned earlier in discussing the entropy inequality for mixing, so the mutual information between C and X is

$$I(C,X) = S(\rho) - \sum_i p_i S(\rho_i).$$

Since $S(\rho_i) \geq 0$ and $S(\rho) \leq \log N$, it follows that

$$I(C,X) \leq \log N.$$

If we knew that quantum mutual information has a similar interpretation to classical mutual information, we would stop here and say that since $I(C,X) \leq \log N$, at most $\log N$ bits of information about the contents of Alice's notebook have been encoded in X. The problem with this is that, *a priori*, quantum mutual information is a formal definition. Our present goal is to show that this formal definition does behave as we might hope.

What can Bob do on receiving system X? The best he can do is to combine it with some other system which may include a quantum system Y and a measuring apparatus C'. He acts on the combined system XYC' with some sort of quantum channel. Then he forgets X and Y and looks at C' – that is, he takes a partial trace over X and Y, another quantum channel. The output of the resulting quantum channel is a density matrix of the form

$$\rho_{C'} = \sum_{\alpha=1}^{r} q_\alpha |\alpha\rangle\langle\alpha|,$$

where $|\alpha\rangle$ are distinguished states of C' – the states that one reads in a classical sense. The outcome of Bob's measurement is a probability distribution $\{q_\alpha\}$ for a random variable whose values are labeled by α. What Bob learns about the contents of Alice's notebook is the classical mutual information between Alice's probability distribution $\{p_i\}$ and Bob's probability distribution $\{q_\alpha\}$. Differently put, what Bob learns is the mutual information $I_{CC'}$.

To analyze this, we note that before Bob does anything, $I(C, X)$ is the same as $I(C, XYC')$ because YC' (Bob's auxiliary quantum system Y and his measuring apparatus C') is not coupled to CX. Bob then acts on XYC' with a quantum channel, which can only reduce $I(C, XYC')$, and then he takes a partial trace over XY, which also can only reduce the mutual information since monotonicity of mutual information tells us that

$$I(C, XYC') \geq I(C, C').$$

So

$$\log N \geq I(C, X) = I(C, XYC')_{\text{before}} \geq I(C, XYC')_{\text{after}} \geq I(C, C')_{\text{after}},$$

where "before" and "after" mean before and after Bob's manipulations. Thus Alice cannot encode more than $\log N$ bits of classical information in an N-dimensional quantum state, though it takes strong subadditivity (or its equivalents) to prove this.

Finally, we can give a physical meaning to the relative entropy $S(\rho||\sigma)$ between two density matrices ρ, σ. Recall that classically, if we believe a random variable is governed by a probability distribution Q but it is actually governed by a probability distribution P, then after N trials the ability to disprove the wrong hypothesis is controlled by

$$2^{-NS(P||Q)}.$$

A similar statement holds quantum mechanically: if our initial hypothesis is that a quantum system X has density matrix σ, and the actual answer is ρ, then after N trials with an optimal measurement used to test the initial hypothesis, the confidence that the initial hypothesis was wrong is controlled in the same sense by

$$2^{-NS(\rho||\sigma)}.$$

Let us first see that monotonicity of relative entropy implies that one cannot do better than that. A measurement is a special case of a quantum channel, in the following sense. To measure a system X, you let it interact quantum mechanically with some other system YC where Y is any quantum system and C is the measuring device. After they interact, you look at the measuring device and forget the rest. Forgetting the rest is a partial trace that maps a density matrix β_{XYC} to $\beta_C = \mathrm{Tr}_{XY}\beta_{XYC}$. If C is a good measuring device, this means that in a distinguished basis $|\alpha\rangle$ its density matrix β_C will have a diagonal form

$$\beta_C = \sum_\alpha b_\alpha |\alpha\rangle\langle\alpha|.$$

The "measurement" converts the original density matrix into the probability distribution $\{b_\alpha\}$.

So when we try to distinguish ρ from σ, we use a quantum channel plus partial trace that maps ρ and σ into density matrices for C

$$\rho_C = \sum_\alpha r_\alpha |\alpha\rangle\langle\alpha| \qquad \sigma_C = \sum_\alpha s_\alpha |\alpha\rangle\langle\alpha|,$$

and thereby into classical probability distributions $R = \{r_\alpha\}$ and $S = \{s_\alpha\}$. The only way to learn that ρ and σ are different is by observing that R and S are different, a process controlled by

$$2^{-NS_{\rm cl}(R||S)},$$

where $S_{\rm cl}(R||S)$ is the classical relative entropy between R and S. This is the same as the relative entropy between ρ_C and σ_C:

$$S(\rho_C||\sigma_C) = S_{\rm cl}(R||S).$$

And monotonocity of relative entropy gives

$$S(\rho||\sigma) \geq S(\rho_C||\sigma_C).$$

So finally $S(\rho||\sigma)$ gives an upper bound on how well we can do:

$$2^{-NS_{\rm cl}(R||S)} \geq 2^{-NS(\rho||\sigma)}.$$

In the limit of large N, it is actually possible to saturate this bound, as follows.

If ρ is diagonal in the same basis in which σ is diagonal, then by making a measurement that involves projecting on 1-dimensional eigenspaces of σ, we could convert the density matrices ρ, σ into classical probability distributions R, S with $S(\rho||\sigma) = S_{\rm cl}(R||S)$. The quantum problem would be equivalent to a classical problem, even without taking N copies. As usual the subtlety comes because the matrices are not simultaneously diagonal.By dropping from ρ the off-diagonal matrix elements in some basis in which σ is diagonal, we can always construct a diagonal density matrix ρ_D. Then a measurement projecting on 1-dimensional eigenspaces of σ will give probability distributions R, S satisfying

$$S(\rho_D||\sigma) = S_{\rm cl}(R||S).$$

But it is hard to compare $S(\rho_D||\sigma)$ to $S(\rho||\sigma)$. That is why it is necessary to take N copies, which does make it possible to compare $S(\rho_D||\sigma)$ to $S(\rho||\sigma)$, as we will see.

Taking N copies replaces the Hilbert space $\mathcal{H}$ of system X by $\mathcal{H}^{\otimes N}$, and replaces the density matrices ρ, ρ_D, and σ by $\rho^{\otimes N}$, $\rho_D^{\otimes N}$, and $\sigma^{\otimes N}$. Let us recall the definition of relative entropy:

$$S(\rho^{\otimes N}||\sigma^{\otimes N}) = \mathrm{Tr}\,\rho^{\otimes N}\log\rho^{\otimes N} - \mathrm{Tr}\,\rho^{\otimes N}\log\sigma^{\otimes N}.$$

The second term $\mathrm{Tr}\,\rho^{\otimes N}\log\sigma^{\otimes N}$ is unchanged if we replace $\rho^{\otimes N}$ by its counterpart $\rho_D^{\otimes N}$ that is diagonal in the same basis as $\sigma^{\otimes N}$. So

$$S(\rho^{\otimes N}|\sigma^{\otimes N}) - S(\rho_D^{\otimes N}|\sigma^{\otimes N}) = \mathrm{Tr}\rho^{\otimes N}\log\rho^{\otimes N} - \mathrm{Tr}\rho_D^{\otimes N}\log\rho_D^{\otimes N}.$$

The reason that we will get simplification for large N is that $\rho^{\otimes N}$ commutes with the group of permutations of the N factors of $\mathcal{H}^N$. Therefore, $\rho^{\otimes N}$ is block diagonal in a basis of irreducible representations of the permutation group. Although $\mathcal{H}^{\otimes N}$ has an exponentially large dimension k^N for large N (where $k = \dim\mathcal{H}$), the irreducible representations have much smaller dimensions, with an upper bound of the form aN^b, where a and b depend on k but not on N.

In a basis of irreducible representations of the permutation group, $\rho^{\otimes N}$ is block diagonal

$$\rho^{\otimes N} = \begin{pmatrix} p_1\rho_1 & & & \\ & p_2\rho_2 & & \\ & & p_3\rho_3 & \\ & & & \ddots \end{pmatrix}$$

where in the i^{th} block ρ_i is a density matrix on the i^{th} irreducible representation, and $\sum_i p_i = 1$. In a basis in which $\sigma^{\otimes N}$ is diagonalized within each block, $\rho_D^{\otimes N}$ is obtained by replacing each of the ρ_i with is diagonal version $\rho_{i,D}$:

$$\rho^{\otimes N} = \begin{pmatrix} p_1\rho_{1,D} & & & \\ & p_2\rho_{2,D} & & \\ & & p_3\rho_{3,D} & \\ & & & \ddots \end{pmatrix}$$

One finds then

$$\mathrm{Tr}\rho^{\otimes N}\log\rho^{\otimes N}-\mathrm{Tr}\rho_D^{\otimes N}\log\rho_D^{\otimes N}=\sum_i p_i(S(\rho_{iD})-S(\rho_i)). \qquad (*)$$

Any density matrix on an n-dimensional space has an entropy S bounded by $0\leq S\leq\log n$. Because the sizes of the blocks are bounded by aN^b, and $\sum_i p_i=1$, the right hand side of $(*)$ is bounded in absolute value by $\log(aN^b)$, which for large N is negligible compared to N.

Putting these facts together, for large N, a measurement that projects onto 1-dimensional eigenspaces of σ within each block is good enough to saturate the bound

$$2^{-NS_{\mathrm{cl}}(R||S)} \geq 2^{-NS(\rho||\sigma)}$$

within an error of order $\log N$ in the exponent. This confirms that quantum relative entropy has the same interpretation as classical relative entropy: it controls the ability to show, by a measurement, that an initial hypothesis is incorrect.

(I followed a paper by M. Hayashi.)