La Rivista del Nuovo Cimento (2020) 43:187–227
https://doi.org/10.1007/s40766-020-00004-5

REVIEW PAPER

A mini-introduction to information theory

Edward Witten[1]

Received: 1 February 2020 / Accepted: 5 February 2020 / Published online: 23 March 2020

Abstract
This article consists of a very short introduction to classical and quantum information theory. Basic properties of the classical Shannon entropy and the quantum von Neumann entropy are described, along with related concepts such as classical and quantum relative entropy, conditional entropy, and mutual information. A few more detailed topics are considered in the quantum case.

Keywords Information theory · Quantum information theory · Entropy · von Neumann entropy

Contents

1 Introduction . . . 188
2 Classical information theory . . . 188
2.1 Shannon entropy . . . 188
2.2 Conditional entropy . . . 190
2.3 Relative entropy . . . 192
2.4 Monotonicity of relative entropy . . . 194
3 Quantum information theory: basic ingredients . . . 196
3.1 Density matrices . . . 196
3.2 Quantum entropy . . . 201
3.3 Concavity . . . 202
3.4 Conditional and relative quantum entropy . . . 204
3.5 Monotonicity of relative entropy . . . 207
3.6 Generalized measurements . . . 209
3.7 Quantum channels . . . 211
3.8 Thermodynamics and quantum channels . . . 213
4 More on quantum information theory . . . 214
4.1 Quantum teleportation and conditional entropy . . . 215
4.2 Quantum relative entropy and hypothesis testing . . . 219
4.3 Encoding classical information in a quantum state . . . 224
References . . . 226

✉ Edward Witten
witten@ias.edu

[1] School of Natural Sciences, Institute for Advanced Study Einstein Drive, Princeton, NJ 08540, USA

1 Introduction

This article is intended as a very short introduction to some basic aspects of classical and quantum information theory.[1]

Section 2 contains a very short introduction to classical information theory, focusing on the definition of Shannon entropy and related concepts such as conditional entropy, relative entropy, and mutual information. Section 3 describes the corresponding quantum concepts—the von Neumann entropy and the quantum conditional entropy, relative entropy, and mutual information. Section 4 is devoted to some more detailed topics in the quantum case, chosen to explore the extent to which the quantum concepts match the intuition that their names suggest.

In this article, we only consider topics that are fairly closely related to entropy. For other matters such as Bell's inequality, the reader will have to look elsewhere. There are several excellent introductory books on classical and quantum information theory, for instance [1–3]. Another excellent place to start is the lecture notes [4], especially chapter 10.

2 Classical information theory

2.1 Shannon entropy

We begin with a basic introduction to classical information theory. Suppose that one receives a message that consists of a string of symbols a or b, say

$$aababbaaaab\cdots \tag{2.1}$$

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

$$\begin{aligned}\frac{N!}{(pN)!((1-p)N)!} &\sim \frac{N^N}{(pN)^{pN}((1-p)N)^{(1-p)N}}\\ &= \frac{1}{p^{pN}(1-p)^{(1-p)N}} = 2^{NS}\end{aligned} \tag{2.2}$$

where S is the **Shannon entropy** per letter [5]

$$S = -p\log p - (1-p)\log(1-p). \tag{2.3}$$

(In information theory, one usually measures entropy in bits and uses logarithms in base 2.)

[1] The article is based on a lecture at the 2018 summer program Prospects in Theoretical Physics at the Institute for Advanced Study.

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

$$2^{NS} \tag{2.4}$$

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

$$NS. \tag{2.5}$$

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, $a_2, \ldots, a_k$, where the probability to observe a_i is p_i, for $i = 1, \ldots, k$. We write A for this probability distribution. In a long message with $N \gg 1$ letters, the symbol a_i will occur approximately Np_i times, and the number of such messages is asymptotically

$$\frac{N!}{(p_1 N)!(p_2 N)! \ldots (p_k N)!} \sim \frac{N^N}{\prod_{i=1}^k (p_i N)^{p_i N}} = 2^{NS_A} \tag{2.6}$$

where now the entropy per letter is

$$S_A = -\sum_{i=1}^{k} p_i \log p_i. \tag{2.7}$$

This is the general definition of the Shannon entropy of a probability distribution for a random variable 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

$$NS_A. \tag{2.8}$$

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

$$S_A \geq 0 \tag{2.9}$$

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

$$S_A = -\sum_{i=1}^{k} (1/k) \log(1/k) = \log k. \tag{2.10}$$

The reader can prove this by using the method of Lagrange multipliers to maximize $S_A = -\sum_i p_i \log p_i$ 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; then English grammar and syntax would dictate short range correlations. This situation was actually considered by Shannon in his original paper on this subject.) 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 letter.

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 ith symbol a_i is $-\log p_i$, so that the probability of the ith 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.

2.2 Conditional entropy

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[2] X whose possible values are $x_1, \ldots, x_k$. She sends the message over a noisy telephone connection, and what Bob receives is many copies of a random variable Y, drawn from an alphabet with letters $y_1, \ldots, y_r$. (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 $P_{X,Y}(x_i, y_j)$ is the probability that, in a given occurrence, Alice sends $X = x_i$ and Bob hears $Y = y_j$. The probability that Bob hears $Y = y_j$, summing over all choices of what Alice intended, is

$$P_Y(y_j) = \sum_i P_{X,Y}(x_i, y_j). \tag{2.11}$$

[2] Generically, a random variable will be denoted X, Y, Z, etc. The probability to observe $X = x$ is denoted $P_X(x)$, so if x_i, $i = 1, \ldots, n$ are the possible values of X, then $\sum_i P_X(x_i) = 1$. Similarly, if X, Y are two random variables, the probability to observe $X = x$, $Y = y$ will be denoted $P_{X,Y}(x, y)$.

If Bob does hear $Y = y_j$, his estimate of the probability that Alice sent x_i is the **conditional probability**

$$P_{X|Y}(x_i|y_j) = \frac{P_{X,Y}(x_i, y_j)}{P_Y(y_j)}. \tag{2.12}$$

From Bob's point of view, once he has heard $Y = y_j$, his estimate of the remaining entropy in Alice's signal is the Shannon entropy of the conditional probability distribution. This is

$$S_{X|Y=y_j} = -\sum_i P_{X|Y}(x_i|y_j) \log(P_{X|Y}(x_i|y_j)). \tag{2.13}$$

Averaging over all possible values of Y, the average remaining entropy, once Bob has heard Y, is

$$\begin{aligned}\sum_j P_Y(y_j) S_{X|Y=y_j} &= -\sum_j P_Y(y_j) \sum_i \frac{P_{X,Y}(x_i, y_j)}{P_Y(y_j)} \log\left(\frac{P_{X,Y}(x_i, y_j)}{P_Y(y_j)}\right) \\ &= -\sum_{i,j} P_{X,Y}(x_i, y_j) \log P_{X,Y}(x_i, y_j) + \sum_{i,j} P_{X,Y}(x_i, y_j) \log P_Y(y_j) \\ &= S_{XY} - S_Y. \end{aligned} \tag{2.14}$$

Here S_{XY} is the entropy of the joint distribution $P_{X,Y}(x_i, y_j)$ for the pair X, Y and S_Y is the entropy of the probability distribution $P_Y(y_j) = \sum_i P_{X,Y}(x_i, y_j)$ for Y only.

The left hand side of Eq. (2.14), which as we see equals $S_{XY} - S_Y$, is called the **conditional entropy** $S_{X|Y}$ or $S(X|Y)$; 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 $S_{X|Y=y_j}$ with positive coefficients, it is clearly positive:

$$S_{XY} - S_Y \geq 0. \tag{2.15}$$

(The analogous statement 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

$$I(X; Y) = S_X - S_{XY} + S_Y. \tag{2.16}$$

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, the relative entropy. This will complete our cast of characters.

2.3 Relative entropy

One can motivate the definition of relative entropy as follows. Suppose that we are observing a random variable X, for example the final state in the decays of a radioactive nucleus. We have a theory that predicts a probability distribution Q_X for the final state, say the prediction is that the probability to observe final state $X = x_i$, where i runs over a set of possible outcomes $\{1, 2, \ldots s\}$, is $q_i = Q_X(x_i)$. But maybe our theory is wrong and the decay is actually described by some different probability distribution P_X, such that the probability of $X = x_i$ is $p_i = P_X(x_i)$. 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_X, then after observing N decays, we will see outcome x_i approximately $p_i N$ times. Believing Q_X to be the correct distribution, we will judge the probability of what we have seen to be[3]

$$\mathcal{P} = \prod_{i=1}^{s} q_i^{p_i N} \frac{N!}{\prod_{j=1}^{s} (p_j N)!}. \tag{2.17}$$

We already calculated that for large N

$$\frac{N!}{\prod_{j=1}^{s} (p_j N)!} \sim 2^{-N \sum_i p_i \log p_i} \tag{2.18}$$

so

$$\mathcal{P} \sim 2^{-N \sum_i p_i (\log p_i - \log q_i)}. \tag{2.19}$$

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

$$S(P_X||Q_X) = \sum_i p_i (\log p_i - \log q_i). \tag{2.20}$$

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

$$NS(P_X||Q_X) \gg 1. \tag{2.21}$$

Suppose that one does an experiment and the data obtained agrees with the hypothesis Q less than would be expected 95% of the time, assuming that hypothesis Q is correct. Then one customarily says that hypothesis Q is excluded at 95% confidence. This way of saying things is a convenient shorthand for saying that 95% of the time, the data would have agreed with hypothesis Q better than it did, if the hypothesis

[3] Here $\frac{N!}{\prod_{j=1}^{s} (p_j N)!}$ is the number of sequences in which outcome x_i occurs $p_i N$ times, and $\prod_{i=1}^{s} q_i^{p_i N}$ is the probability of any specific such sequence, assuming that the initial hypothesis Q_X is correct.

is correct. With the same manner of speaking, we may say in the above-described situation that if hypothesis Q is incorrect and hypothesis P is correct, then a typical experimental outcome after N trials would exclude hypothesis Q with confidence level $1-\epsilon$, where ϵ would decay for large N as $2^{-NS(P_X||Q_X)}$. (Later we will more loosely say that the confidence in excluding the wrong hypothesis is controlled by $2^{-NS(P_X||Q_X)}$.) In this analysis, we have ignored noise in the observations. What we learned earlier about conditional entropy would give us a start in including the effects of noise.

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

Now we will use positivity of the relative entropy 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 $P_{X,Y}$, is defined by a possibly correlated joint probability distribution

$$P_{X,Y}(x_i, y_j). \tag{2.22}$$

Given such a joint probability distribution, the separate probability distributions for X and for Y are obtained by "integrating out" or summing over the other variable:

$$P_X(x_i) = \sum_j P_{X,Y}(x_i, y_j), \qquad P_Y(y_j) = \sum_i P_{X,Y}(x_i, y_j). \tag{2.23}$$

This is an important operation which will frequently recur. We define a second probability distribution for X, Y by ignoring the correlations between them:

$$Q_{X,Y}(x_i, y_j) = P_X(x_i) P_Y(y_j). \tag{2.24}$$

Now we calculate the relative entropy between these two distributions:

$$\begin{aligned} S(P_{X,Y}||Q_{X,Y}) &= \sum_{i,j} P_{X,Y}(x_i, y_j)(\log P_{X,Y}(x_i, y_j) - \log(P_X(x_i) P_Y(y_j))) \\ &= \sum_{i,j} P_{X,Y}(x_i, y_j)(\log P_{X,Y}(x_i, y_j) - \log P_X(x_i) - \log P_Y(y_j)) \\ &= S_X + S_Y - S_{XY} = I(X;Y). \end{aligned} \tag{2.25}$$

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

The property

$$S_X + S_Y - S_{XY} \geq 0 \tag{2.26}$$

is called **subadditivity** of entropy.

2.4 Monotonicity of relative entropy

Now there is one more very important property of relative entropy 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 $P_{X,Y}$ and $Q_{X,Y}$ be two probability distributions, described by functions $P_{X,Y}(x_i, y_j)$ and $Q_{X,Y}(x_i, y_j)$. If we start with a hypothesis $Q_{X,Y}$ for the joint probability, then after many trials in which we observe X and Y, our confidence that we are wrong (assuming that $P_{X,Y}$ is the correct answer) is determined by $S(P_{X,Y}||Q_{X,Y})$. 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

$$P_X(x_i) = \sum_j P_{X,Y}(x_i, y_j), \qquad Q_X(x_i) = \sum_j Q_{X,Y}(x_i, y_j). \tag{2.27}$$

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

$$S(P_{X,Y}||Q_{X,Y}) \geq S(P_X||Q_X). \tag{2.28}$$

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

$$(x_{i_1}, y_{i_1}), (x_{i_2}, y_{i_2}), \ldots, (x_{i_N}, y_{i_N}). \tag{2.29}$$

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.

Hopefully, the reader has found this explanation compelling, but it is also not difficult to give a proof in formulas. The inequality $S(P_{X,Y}||Q_{X,Y}) - S(P_X||Q_X) \geq 0$ can be written

$$\sum_{i,j} P_{X,Y}(x_i, y_j) \left(\log \left(\frac{P_{X,Y}(x_i, y_j)}{Q_{X,Y}(x_i, y_j)} \right) - \log \left(\frac{P_X(x_i)}{Q_X(x_i)} \right) \right) \geq 0. \tag{2.30}$$

Equivalently

$$\sum_i P_X(x_i) \sum_j \frac{P_{X,Y}(x_i, y_j)}{P_X(x_i)} \log \left(\frac{P_{X,Y}(x_i, y_j)/P_X(x_i)}{Q_{X,Y}(x_i, y_j)/Q_X(x_i)} \right) \geq 0. \tag{2.31}$$

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

$$\sum_i P_X(x_i) S(P_{Y|X=x_i} || Q_{Y|X=x_i}), \tag{2.32}$$

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

$$P_{Y|X=x_i}(y_j) = P_{X,Y}(x_i, y_j)/P_X(x_i), \quad Q_{Y|X=x_i}(y_j) = Q_{X,Y}(x_i, y_j)/Q_X(x_i). \tag{2.33}$$

So this establishes monotonicity of relative entropy.[4] An important special case is **strong subadditivity** of entropy. For this, we consider three random variables X, Y, Z. The combined system has a joint probability distribution $P_{X,Y,Z}(x_i, y_j, z_k)$. Alternatively, we could forget the correlations between X and YZ, defining a probability distribution $Q_{X,Y,Z}$ for the system XYZ by

$$Q_{X,Y,Z}(x_i, y_j, z_k) = P_X(x_i) P_{Y,Z}(y_j, z_k) \tag{2.34}$$

where as usual

$$P_X(x_i) = \sum_{j,k} P_{X,Y,Z}(x_i, y_j, z_k), \quad P_{Y,Z}(y_j, z_k) = \sum_i P_{X,Y,Z}(x_i, y_j, z_k). \tag{2.35}$$

The relative entropy is $S(P_{X,Y,Z} || Q_{X,Y,Z})$. But what if we only observe the subsystem XY? Then we replace $P_{X,Y,Z}$ and $Q_{X,Y,Z}$ by probability distributions $P_{X,Y}$, $Q_{X,Y}$ with

$$\begin{aligned} P_{X,Y}(x_i, y_j) &= \sum_k P_{X,Y,Z}(x_i, y_j, z_k), \\ Q_{X,Y}(x_i, y_j) &= \sum_k Q_{X,Y,Z}(x_i, y_j, z_k) = P_X(x_i) P_Y(y_j) \end{aligned} \tag{2.36}$$

and we can define the relative entropy $S(P_{X,Y} || Q_{X,Y})$. Monotonicity of relative entropy tells us that

$$S(P_{X,Y,Z} || Q_{X,Y,Z}) \geq S(P_{X,Y} || Q_{X,Y}). \tag{2.37}$$

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

$$S(P_{X,Y,Z} || Q_{X,Y,Z}) = I(X; YZ) = S_X - S_{XYZ} + S_{YZ} \tag{2.38}$$

[4] What we have described is not the most general statement of monotonicity of relative entropy in classical information theory. More generally, relative entropy is monotonic under an arbitrary stochastic map. We will not explain this here, though later we will explain the quantum analog (quantum relative entropy is monotonic in any quantum channel).

and

$$S(P_{X,Y}||Q_{X,Y}) = I(X;Y) = S_X - S_{XY} + S_Y. \tag{2.39}$$

So

$$S_X - S_{XYZ} + S_{YZ} \geq S_X - S_{XY} + S_Y \tag{2.40}$$

or

$$S_{XY} + S_{YZ} \geq S_Y + S_{XYZ}, \tag{2.41}$$

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

Equivalently, the comparison of Eqs. (2.38) and (2.39) gives

$$I(X;YZ) \geq I(X;Y), \tag{2.42}$$

which is called **monotonicity of mutual information**. The intuition is that what one learns about a random variable X by observing both Y and Z is at least as much as one could learn by observing Y only.

We conclude this mini-introduction to classical information theory with one last remark. We repeatedly made use of the ability to define a conditional probability distribution, conditional on some observation. This has no really close analog in the quantum mechanical case[5] and it is something of a miracle that many of the conclusions nonetheless have quantum mechanical analogs. The greatest miracle is strong subadditivity of quantum entropy.

3 Quantum information theory: basic ingredients

3.1 Density matrices

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 $\psi_B \in \mathcal{H}_B$:

[5] See, however, [6] for a partial substitute.

$$\psi_{AB} = \psi_A \otimes \psi_B. \tag{3.1}$$

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, predictions about the A system can be made by forgetting about the B system and using the state vector ψ_A. Indeed, 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

$$\langle \psi_{AB} | \mathcal{O}_A \otimes 1_B | \psi_{AB} \rangle = \langle \psi_A | \mathcal{O}_A | \psi_A \rangle \langle \psi_B | 1_B | \psi_B \rangle = \langle \psi_A | \mathcal{O}_A | \psi_A \rangle. \tag{3.2}$$

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 $\mathcal{H}_{AB}$ can be presented as an $N \times M$ matrix, for example in the 2×3 case

$$\psi_{AB} = \begin{pmatrix} * & * & * \\ * & * & * \end{pmatrix}. \tag{3.3}$$

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

$$\psi_{AB} \to U \psi_{AB} V \tag{3.4}$$

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

$$\psi_{AB} = \sum_i \sqrt{p_i} \psi_A^i \otimes \psi_B^i, \tag{3.5}$$

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

$$\langle \psi_A^i, \psi_A^j \rangle = \langle \psi_B^i, \psi_B^j \rangle = \delta^{ij} \tag{3.6}$$

and that $p_i > 0$. (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

$$\sum_i p_i = 1, \tag{3.7}$$

so we can think of the p_i as probabilities. Equation (3.5) is called the **Schmidt decomposition**.

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

$$\langle\psi_{AB}|\mathcal{O}_A\otimes 1_B|\psi_{AB}\rangle=\sum_{i,j}\sqrt{p_ip_j}\left\langle\psi_A^i|\mathcal{O}_A|\psi_A^j\right\rangle\left\langle\psi_B^i|1_B|\psi_B^j\right\rangle$$
$$=\sum_i p_i\left\langle\psi_A^i|\mathcal{O}_A|\psi_A^i\right\rangle. \tag{3.8}$$

This is the same as

$$\mathrm{Tr}_{\mathcal{H}_A}\,\rho_A\mathcal{O}_A, \tag{3.9}$$

where ρ_A is the **density matrix**

$$\rho_A=\sum_i p_i\left|\psi_A^i\right\rangle\left\langle\psi_A^i\right|. \tag{3.10}$$

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

$$\rho_A=\sum_i p_i\left|\psi_A^i\right\rangle\left\langle\psi_A^i\right| \tag{3.11}$$

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

$$\mathrm{Tr}_{\mathcal{H}_A}\,\rho_A=1. \tag{3.12}$$

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 of Eq. (3.11); 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

$$\psi_{AB}=\sum_i\sqrt{p_i}\psi_A^i\otimes\psi_B^i\in\mathcal{H}_A\otimes\mathcal{H}_B. \tag{3.13}$$

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.

The purification ψ_{AB} of a density matrix ρ_A is far from unique (even if the auxiliary system B is specified), because there is freedom in choosing the orthonormal states ψ_B^i in Eq. (3.13). However, any other set of orthonormal vectors in $\mathcal{H}_B$ can be obtained from a given choice ψ_B^i by a unitary transformation of $\mathcal{H}_B$, so we learn the following

important fact: any two purifications of the same density matrix ρ_A on system A by pure states of a bipartite system AB are equivalent under a unitary transformation of system B.

In general, a density matrix is just a nonnegative self-adjoint matrix ρ whose trace is 1. The above derivation shows that every such matrix is the density matrix of some bipartite pure state. The conditions satisfied by a density matrix are preserved under "mixing," that is under taking a linear combination with positive coefficients. So for example if ρ_1 and ρ_2 are density matrices, then so is $\rho = t\rho_1 + (1-t)\rho_2$, for $0 \leq t \leq 1$.

If there is more than one term in the expansion

$$\psi_{AB} = \sum_i \sqrt{p_i}\psi_A^i \otimes \psi_B^i \in \mathcal{H}_A \otimes \mathcal{H}_B, \tag{3.14}$$

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

$$\psi_{AB} = \psi_A \otimes \psi_B, \tag{3.15}$$

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 a multiple of the identity, we say that A is maximally mixed.

In the general case

$$\rho_A = \sum_i p_i |\psi_A^i\rangle\langle\psi_A^i| \tag{3.16}$$

one will describe all measurements of system A correctly if one says that system A is in the state ψ_A^i with probability p_i. However, one has to be careful here because the decomposition of Eq. (3.16) 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 this expansion. This means that if Alice prepares a quantum system to be in the pure state ψ_A^i with probability p_i, then there is no way to determine the p_i or the ψ_A^i by measurements, even if one is provided with many identical copies to measure. Any measurement of the system will depend only on $\rho_A = \sum_i p_i |\psi_A^i\rangle\langle\psi_A^i|$. There is no way to get additional information about how the system was prepared.

So far, when we have discussed a bipartite system AB, we have assumed that the combined system is in a pure state ψ_{AB}, and we have discussed density matrices ρ_A and ρ_B for systems A and B. More generally, we should allow for the possibility that the combined system AB is described to begin with by a density matrix ρ_{AB}. Consideration of this situation leads to the following very fundamental definition.

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 $P_{X,Y}(x_i, y_j)$ for a bipartite system XY, we "integrated out" Y to get a probability distribution for X only:

$$P_X(x_i) = \sum_j P_{X,Y}(x_i, y_j). \tag{3.17}$$

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 $|i\rangle_A$, $i = 1, \dots, n$ are an orthonormal basis of $\mathcal{H}_A$ and $|\alpha\rangle_B$, $\alpha = 1, \dots, m$ are an orthonormal basis of $\mathcal{H}_B$, then a density matrix for AB takes the general form

$$\rho_{AB} = \sum_{i,i',\alpha,\alpha'} c_{ii'\alpha\alpha'} |i\rangle_A \otimes |\alpha\rangle_B \, {}_A\langle i'| \otimes {}_B\langle \alpha'|. \tag{3.18}$$

The reduced density matrix for measurements of system A only is obtained by setting $\alpha = \alpha'$, replacing $|\alpha\rangle_B \, {}_B\langle\alpha|$ by its trace, which is 1, and summing:

$$\rho_A = \sum_{i,i',\alpha} c_{i,i',\alpha,\alpha} |i\rangle_A \, {}_A\langle i'|. \tag{3.19}$$

In other words, if we are going to measure system A only, we sum over all of the unobserved states of system B. This is usually written as a partial trace:

$$\rho_A = \mathrm{Tr}_{\mathcal{H}_B} \rho_{AB}, \tag{3.20}$$

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 $\mathcal{H}_A$)

$$\rho_B = \mathrm{Tr}_{\mathcal{H}_A} \rho_{AB}. \tag{3.21}$$

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 isolated atom interacting with passing photons. A photon might be scattered, or absorbed and reemitted, or might pass by without interacting with the atom. Regardless, after a certain time, the atom is again alone. After n photons have had the chance to interact with the atom, to give a pure state description, we need a joint wavefunction for the atom and all the outgoing 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 for describing the state of the system 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.

3.2 Quantum entropy

The **von Neumann entropy**[6] of a density matrix ρ_A is defined by a formula analogous to the Shannon entropy of a probability distribution:

$$S(\rho_A) = -\mathrm{Tr}\,\rho_A \log \rho_A. \tag{3.22}$$

As an immediate comment, we note that $S(\rho_A)$ is manifestly invariant under a unitary transformation

$$\rho_A \to U \rho_A U^{-1}. \tag{3.23}$$

Quantum conditional and relative entropy, which will be introduced in Sect. 3.4, are similarly invariant under a suitable class of unitaries.

By a unitary transformation, we can diagonalize ρ_A, putting it in the form

$$\rho_A = \sum_i p_i |\psi_A^i\rangle\langle\psi_A^i|, \tag{3.24}$$

with ψ_A^i being orthonormal and $p_i > 0$. Then in an obvious basis

$$\rho_A \log \rho_A = \begin{pmatrix} p_1 \log p_1 & & & \\ & p_2 \log p_2 & & \\ & & p_3 \log p_3 & \\ & & & \ddots \end{pmatrix} \tag{3.25}$$

and so

$$S(\rho_A) = -\sum_i p_i \log p_i, \tag{3.26}$$

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

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

$$S(\rho_A) \geq 0, \tag{3.27}$$

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

$$S(\rho_A) \leq \log k, \tag{3.28}$$

[6] The von Neumann entropy is the most important quantum entropy, but generalizations such as the Rényi entropies $S_\alpha(\rho_A) = \frac{1}{1-\alpha} \log \mathrm{Tr}\,\rho_A^\alpha$ can also be useful.

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

$$\rho_A = \frac{1}{k}\begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & \ddots \end{pmatrix}. \tag{3.29}$$

In this case, we say that 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

$$\psi_{AB} = \sum_i \sqrt{p_i}\psi_A^i \otimes \psi_B^i \in \mathcal{H}_A \otimes \mathcal{H}_B, \tag{3.30}$$

then the density matrices of systems A and B are

$$\rho_A = \sum_i p_i |\psi_A^i\rangle\langle\psi_A^i|, \tag{3.31}$$

and likewise

$$\rho_B = \sum_i p_i |\psi_B^i\rangle\langle\psi_B^i|. \tag{3.32}$$

The same constants p_i appear in each, so clearly

$$S(\rho_A) = S(\rho_B). \tag{3.33}$$

Thus a system A and a purifying system B always have the same entropy. Note that in this situation, since the combined system AB is in a pure state, its entropy S_{AB} vanishes.

3.3 Concavity

The von Neumann entropy—like its antecedents in classical thermodynamics and statistical mechanics—has the important property of **concavity**. Suppose that ρ_1 and ρ_2 are two density matrices, and set $\rho(t) = t\rho_1 + (1-t)\rho_2$, for $0 \leq t \leq 1$. We will write $\dot\rho(t)$, $\ddot\rho(t)$ for $\mathrm{d}\rho(t)/\mathrm{d}t$, $\mathrm{d}^2\rho(t)/\mathrm{d}t^2$. Then

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}S(\rho(t)) \leq 0. \tag{3.34}$$

To prove this, we first compute that[7]

$$\frac{\mathrm{d}}{\mathrm{d}t} S(\rho(t)) = -\mathrm{Tr}\, \dot{\rho} \log \rho. \tag{3.35}$$

Then as

$$\log \rho = \int_0^\infty \mathrm{d}s \left(\frac{1}{s+1} - \frac{1}{s+\rho(t)} \right) \tag{3.36}$$

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

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} S(\rho(t)) = -\int_0^\infty \mathrm{d}s \mathrm{Tr}\, \dot{\rho} \frac{1}{s+\rho(t)} \dot{\rho} \frac{1}{s+\rho(t)}. \tag{3.37}$$

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. Like any concave function, $S(\rho(t))$ has the property that the straight line connecting two points on its graph lies below the graph. Explicitly, this gives

$$t S(\rho_1) + (1-t) S(\rho_2) \leq S(t\rho_1 + (1-t)\rho_2) = S(\rho(t)). \tag{3.38}$$

More generally, let ρ_i, $i = 1, \ldots, n$ be density matrices and p_i, $i = 1, \ldots, n$ non-negative numbers with $\sum_i p_i = 1$. Then by induction starting with (3.38), or because this is a general property of concave functions, we have

$$\sum_i p_i S(\rho_i) \leq S(\rho), \quad \rho = \sum_i p_i \rho_i. \tag{3.39}$$

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 [7]:

$$\chi = S(\rho) - \sum_i p_i S(\rho_i). \tag{3.40}$$

An interesting special case is the following. Let ρ be any density matrix on a Hilbert space $\mathcal{H}$. Pick a basis of $\mathcal{H}$, and let ρ_D be the diagonal density matrix obtained in that

[7] For this, consider an arbitrary density matrix ρ and a first order perturbation $\rho \to \rho + \delta\rho$. After diagonalizing ρ, one observes that to first order in $\delta\rho$, the off-diagonal part of $\delta\rho$ does not contribute to the trace in the definition of $S(\rho+\delta\rho)$. Therefore, $S(\rho(t))$ can be differentiated assuming that ρ and $\dot{\rho}$ commute. So it suffices to check (3.35) for a diagonal family of density matrices $\rho(t) = \mathrm{diag}(\lambda_1(t), \lambda_2(t), \ldots, \lambda_n(t))$, with $\sum_i \lambda_i(t) = 1$. Another approach is to use (3.36) to substitute for $\log \rho(t)$ in the definition $S(\rho(t)) = -\mathrm{Tr}\, \rho(t) \log \rho(t)$. Differentiating with respect to t, observing that $\rho(t)$ commutes with $1/(s + \rho(t))$, and then integrating over s, one arrives at (3.35). In either approach, one uses that $\mathrm{Tr}\, \dot{\rho} = 0$ since $\mathrm{Tr}\, \rho(t) = 1$.

basis by dropping the off-diagonal matrix elements from ρ and keeping the diagonal ones. Let $\rho(t) = (1-t)\rho_D + t\rho$. We see that

$$\frac{\mathrm{d}}{\mathrm{d}t} S(\rho(t))\Big|_{t=0} = 0, \tag{3.41}$$

by virtue of (3.35), because $\rho(0)$ and $\log \rho(0)$ are diagonal while the diagonal matrix elements of $\mathrm{d}\rho/\mathrm{d}t$ vanish at $t = 0$. When we combine this with $\mathrm{d}^2 S(\rho(t))/\mathrm{d}t^2 \leq 0$, we get $S(\rho(1)) \leq S(\rho(0))$ or

$$S(\rho_D) \geq S(\rho). \tag{3.42}$$

Thus, dropping the off-diagonal part of a density matrix (in any basis) can only increase the entropy. Equation (3.42) is a strict inequality unless $\rho = \rho_D$, because Eq. (3.37) shows that $\frac{\mathrm{d}^2}{\mathrm{d}t^2} S(\rho(t))\Big|_{t=0}$ is strictly negative unless $\rho = \rho_D$.

An alternative proof of Eq. (3.42), again using the inequality (3.39), is as follows. For an N state system, there are 2^N matrices that are diagonal matrices (in some chosen basis) with diagonal matrix elements that are all ± 1. Let U_i be any of these and set $\rho_i = U_i \rho U_i^{-1}$. Of course, ρ_i is also a density matrix, since U_i is unitary. The average of the ρ_i, over all 2^N choices of U_i, is the diagonal density matrix ρ_D. So Eq. (3.39) says that the average of $S(\rho_i)$ is less than or equal to $S(\rho_D)$. But $S(\rho_i)$ is independent of i and equal to $S(\rho)$, since the von Neumann entropy is invariant under conjugation by a unitary matrix such as U_i. So in fact the average of the $S(\rho_i)$ is just $S(\rho)$ and the inequality (3.39) becomes $S(\rho) \leq S(\rho_D)$.

Somewhat similarly to what we have explained here, concavity of the function $f(q) = -q \log q$ could have been used in the classical arguments in Sect. 2, though we circumvented this by using Stirling's formula instead.

3.4 Conditional and relative quantum entropy

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**

$$S(A|B) = S_{AB} - S_B. \tag{3.43}$$

This name is potentially misleading because there is not a good quantum notion of conditional probabilities. Unlike the classical case, quantum conditional entropy is not an entropy conditional on something. Nevertheless, in Sect. 4.1, we will discuss at least one sense in which quantum conditional entropy behaves in a way analogous to classical conditional entropy.

There is also a fundamental difference from the classical case: quantum mechanically, $S(A|B)$ 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 $S(A|B) < 0$.

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

$$I(A;B) = S_A - S_{AB} + S_B. \tag{3.44}$$

Here, however, we are more fortunate, and the quantum mutual information is non-negative:

$$I(A;B) \geq 0. \tag{3.45}$$

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

$$\rho_{AB} = \rho_A \otimes \rho_B. \tag{3.46}$$

Positivity of mutual information is also called subadditivity of entropy. To begin with, quantum mutual information is a formal definition and it is not obvious how it is related to information that one can gain about system A by observing system B. We will explore at least one aspect of this question in Sect. 4.3.

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

$$S_{AB} - S_B = S_C - S_{AC} \geq -S_A, \tag{3.47}$$

where the last step is positivity of mutual information. So

$$S(A|B) = S_{AB} - S_B \geq -S_A. \tag{3.48}$$

Reversing the roles of A and B in the derivation, we get the Araki-Lieb inequality [8]

$$S_{AB} \geq |S_A - S_B|. \tag{3.49}$$

It is saturated if $S_{AB} = 0$, which implies $S_B = S_A$. What has just been explained is 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 **quantum relative entropy** [9]. 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:

$$S(\rho||\sigma) = \mathrm{Tr}\rho(\log\rho - \log\sigma). \tag{3.50}$$

For now, this is just a formal definition, but we will learn in Sect. 4.2 that $S(\rho||\sigma)$ has the same interpretation quantum mechanically that it does classically: if one's

hypothesis is that a quantum system is described by a density matrix σ, and it is actually described by a different density matrix ρ, then to learn that one is wrong, one needs to observe N copies of the system where $NS(\rho||\sigma) >> 1$.

Just as classically, it turns out that $S(\rho||\sigma) \geq 0$ for all density matrices ρ, σ, with equality precisely if $\rho = \sigma$. To prove this, first diagonalize σ. In general ρ is not diagonal in the same basis. Let ρ_D be the diagonal density matrix obtained from ρ by dropping the off-diagonal matrix elements in the basis in which σ is diagonal, and keeping the diagonal ones. Since $\mathrm{Tr}\,\rho\log\sigma = \mathrm{Tr}\,\rho_D\log\sigma$, it follows directly from the definitions of von Neumann entropy and relative entropy that

$$S(\rho||\sigma) = S(\rho_D||\sigma) + S(\rho_D) - S(\rho). \tag{3.51}$$

This actually exhibits $S(\rho||\sigma)$ as the sum of two nonnegative terms. We showed in Eq. (3.42) that $S(\rho_D) - S(\rho) \geq 0$. As for $S(\rho_D||\sigma)$, it is nonnegative, because if $\sigma = \mathrm{diag}(q_1, \ldots, q_n)$, $\rho_D = \mathrm{diag}(p_1, \ldots, p_n)$, then

$$S(\rho_D||\sigma) = \sum_i p_i(\log p_i - \log q_i), \tag{3.52}$$

which can be interpreted as a classical relative entropy and so is nonnegative. To get equality in these statements, we need $\sigma = \rho_D$ and $\rho_D = \rho$, so $S(\rho||\sigma)$ vanishes only if $\rho = \sigma$.

Now we can use positivity of the relative entropy to prove that $I(A; B) \geq 0$ for any density matrix ρ_{AB}. Imitating the classical proof, we define

$$\sigma_{AB} = \rho_A \otimes \rho_B, \tag{3.53}$$

and we observe that

$$\log\sigma_{AB} = \log\rho_A \otimes 1_B + 1_A \otimes \log\rho_B, \tag{3.54}$$

so

$$\begin{aligned} S(\rho_{AB}||\sigma_{AB}) &= \mathrm{Tr}_{AB}\rho_{AB}(\log\rho_{AB} - \log\sigma_{AB}) \\ &= \mathrm{Tr}_{AB}\rho_{AB}(\log\rho_{AB} - \log\rho_A \otimes 1_B - 1_B \otimes \log\rho_B) \\ &= S_A + S_B - S_{AB} = I(A; B). \end{aligned} \tag{3.55}$$

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

The inequality (3.39) that expresses the concavity of the von Neumann entropy can be viewed as a special case of the positivity of mutual information. Let B be a quantum system with density matrices ρ_B^i and let C be an auxiliary system C with an orthonormal basis $|i\rangle_C$. Endow CB with the density matrix:

$$\rho_{CB} = \sum_i p_i |i\rangle_C\,{}_C\langle i| \otimes \rho_B^i. \tag{3.56}$$

The mutual information between C and B if the combined system is described by ρ_{CB} is readily computed to be

$$I(C;B) = S(\rho_B) - \sum_i p_i S\left(\rho_B^i\right), \tag{3.57}$$

so positivity of mutual information gives our inequality.

3.5 Monotonicity of relative entropy

So relative entropy is positive, just as it is classically. Do we dare to hope that relative entropy is also monotonic, as classically? Yes it is, as first proved by Lieb and Ruskai [10], using a lemma of Lieb [11]. How to prove monotonicity of quantum relative entropy will not be described here; this has been explored in a companion article [12], Sects. 3 and 4.

Monotonicity of quantum relative entropy is something of 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. We will explore some of this in Sect. 4.

There are different statements of monotonicity of relative entropy, but a very basic one (and actually the version proved in [10]) is monotonicity under partial trace. If AB is a bipartite system with two density matrices ρ_{AB} and σ_{AB}, then we can take a partial trace on B to get reduced density matrices on A:

$$\rho_A = \mathrm{Tr}_B \rho_{AB}, \quad \sigma_A = \mathrm{Tr}_B \sigma_{AB}. \tag{3.58}$$

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

$$S(\rho_{AB}||\sigma_{AB}) \geq S(\rho_A||\sigma_A). \tag{3.59}$$

(This is also called the Data Processing Inequality.)

By imitating what we said classically in Sect. 2, one can deduce strong subadditivity of quantum entropy from monotonicity of relative entropy. 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

$$\sigma_{ABC} = \rho_A \otimes \rho_{BC}. \tag{3.60}$$

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

$$\rho_{AB} = \mathrm{Tr}_C \rho_{ABC}, \quad \sigma_{AB} = \mathrm{Tr}_C \sigma_{ABC} = \rho_A \otimes \rho_B. \tag{3.61}$$

Monotonicity of relative entropy under partial trace says that

$$S(\rho_{ABC}||\sigma_{ABC}) \geq S(\rho_{AB}||\sigma_{AB}). \tag{3.62}$$

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

$$S(\rho_{ABC}||\sigma_{ABC}) = S(\rho_{ABC}||\rho_A \otimes \rho_{BC}) = I(A; BC) = S_A + S_{BC} - S_{ABC} \tag{3.63}$$

and similarly

$$S(\rho_{AB}||\sigma_{AB}) = S(\rho_{AB}||\rho_A \otimes \rho_B) = I(A; B) = S_A + S_B - S_{AB}. \tag{3.64}$$

So Eq. (3.62) becomes **monotonicity of mutual information**

$$I(A; BC) \geq I(A; B) \tag{3.65}$$

or equivalently **strong subadditivity** [10]

$$S_{AB} + S_{BC} \geq S_B + S_{ABC}. \tag{3.66}$$

All of 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

$$S_{CD} + S_{BC} \geq S_B + S_D. \tag{3.67}$$

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

$$S(C|D) + S(C|B) \geq 0. \tag{3.68}$$

(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 was explained in Sect. 2; one learns at least as much about system A by observing B and C as one could learn by observing B only. Quantum mechanically, it is just not clear *a priori* that the formal definition $I(A; B) = S_A - S_{AB} + S_B$ will lead to something consistent with that intuition. The rather subtle result of monotonicity of relative entropy [10] shows that it does.

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

3.6 Generalized measurements

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 orthogonal hermitian projection operators π_s, $s = 1, \ldots, k$ obeying

$$\sum_s \pi_s = 1, \quad \pi_s^2 = \pi_s, \quad \pi_s \pi_{s'} = 0, \quad s \neq s'. \tag{3.69}$$

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

$$p_s = \langle \psi | \pi_s | \psi \rangle. \tag{3.70}$$

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

$$p_s = \mathrm{Tr}_{\mathcal{H}} \, \pi_s \rho. \tag{3.71}$$

After the measurement is made, if outcome s has been found, the system can be described by a new density matrix

$$\rho_s = \frac{1}{p_s} \pi_s \rho \pi_s. \tag{3.72}$$

But Alice can make a more general type of measurement using an auxiliary system C (sometimes called an ancillary system) with Hilbert space $\mathcal{C}$. We suppose that $\mathcal{C}$ is k-dimensional with a basis of states $|s\rangle$, $s = 1, \ldots, k$. 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 $\psi \in \mathcal{H}$

$$U(|1\rangle \otimes \psi) = \sum_{s=1}^{k} |s\rangle \otimes E_s \psi \tag{3.73}$$

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

$$\sum_{s=1}^{k} E_s^\dagger E_s = 1, \tag{3.74}$$

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

$$\pi_s = |s\rangle\langle s| \otimes 1, \tag{3.75}$$

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

$$p_s = |E_s|\psi\rangle|^2 = \langle\psi|E_s^\dagger E_s|\psi\rangle. \tag{3.76}$$

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

$$p_s = \mathrm{Tr}\, E_s^\dagger E_s \rho. \tag{3.77}$$

The numbers 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 orthogonal projection operators; they are just nonnegative hermitian operators that add to 1. What we have described is a more general kind of quantum mechanical measurement of the original system. (In the jargon, the positive operators $E_s^\dagger E_s$ whose sum is 1 comprise a "positive operator-valued measure" or POVM.)

According to Eq. (3.72), after Alice's measurement, if the outcome s has been found, then the combined system $\mathcal{C} \otimes \mathcal{H}$ can be described by the density matrix $\frac{1}{p_s}|s\rangle\langle s| \otimes E_s|\psi\rangle\langle\psi|E_s^\dagger$. Taking the trace over Alice's system, the original system, after the measurement, can then be described by the density matrix

$$\frac{1}{p_s} E_s|\psi\rangle\langle\psi|E_s^\dagger, \tag{3.78}$$

or more generally by $\frac{1}{p_s} E_s \rho E_s^\dagger$, if the original system was initially in a mixed state with density matrix ρ. If after acting with U, Alice simply discards the subsystem $\mathcal{C}$, or if this subsystem is inaccessible and we have no information about it, then at that point the original system can be described by the density matrix

$$\sum_s E_s|\psi\rangle\langle\psi|E_s^\dagger, \tag{3.79}$$

or more generally by $\sum_s E_s \rho E_s^\dagger$.

One can slightly generalize this construction as follows.[8] Suppose that the initial system actually had for its Hilbert space a direct sum $\mathcal{H} \oplus \mathcal{H}'$, but it is known that the initial state of the system is valued in $\mathcal{H}$, in other words the initial state ψ has the form $\chi \oplus 0$ with $\chi \in \mathcal{H}$, and 0 the zero vector in $\mathcal{H}'$. Then Alice couples $\mathcal{H} \oplus \mathcal{H}'$ to her auxiliary system C, so she describes the combined system by a Hilbert space $\mathcal{C} \otimes (\mathcal{H} \oplus \mathcal{H}')$. Now she picks U so that it maps a vector $|1\rangle \otimes (\chi \oplus 0)$ to $\sum_s |s\rangle \otimes$

[8] The following paragraph may be omitted on first reading. It is included to make possible a more general statement in Sect. 3.7.

$(0 \oplus E_s\chi)$, where E_s is a linear transformation $E_s : \mathcal{H} \to \mathcal{H}'$. (As before, Alice does not care what U does on other vectors.) After applying U, Alice makes a projective measurement using the same projection operators $\pi_s = |s\rangle\langle s| \otimes 1$ as before (of course, 1 is now the identity on $\mathcal{H} \oplus \mathcal{H}'$). The linear transformations E_s still obey Eq. (3.74), the probability of outcome s is still given by Eq. (3.77), and the density matrix after a measurement that gives outcome s is still given by Eq. (3.78).

3.7 Quantum channels

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

$$\rho \to U\rho U^{-1}. \tag{3.80}$$

As we remarked earlier (Eq. (3.23)), such unitary evolution preserves the von Neumann entropy of a density matrix, and similarly it preserves the relative entropy between two density matrices.

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

$$\widehat{\rho} = |1\rangle\langle 1| \otimes \rho \tag{3.81}$$

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

$$\widehat{\rho}' = U\widehat{\rho}U^{-1} = \sum_{s,s'=1}^{k} |s\rangle\langle s'| \otimes E_s\rho E_{s'}^\dagger. \tag{3.82}$$

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

$$\rho' = \mathrm{Tr}_{\mathcal{C}}\widehat{\rho}' = \sum_{s=1}^{k} E_s\rho E_s^\dagger. \tag{3.83}$$

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

$$\rho \to \sum_{s=1}^{k} E_s\rho E_s^\dagger, \quad \sum_s E_s^\dagger E_s = 1 \tag{3.84}$$

is called a "quantum channel," and the E_s are called Kraus operators. Unitary evolution is the special case in which there is only one Kraus operator.

The notion of a quantum channel is axiomatized in more complete treatments than we will give here.[9] The upshot of a general analysis is that the most general physically sensible evolution of a density matrix takes the form (3.84), provided one allows the generalization described at the end of Sect. 3.6 in which the E_s are linear transformations from one Hilbert space $\mathcal{H}$ to another Hilbert space $\mathcal{H}'$.

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

$$\rho' = \sum_s E_s \rho E_s^\dagger, \quad \sigma' = \sum_s E_s \sigma E_s^\dagger. \tag{3.85}$$

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

We conclude this section with some exercises to familiarize oneself with quantum channels.

1. Let ψ be any pure state of a given system. Find Kraus operators of a quantum channel that maps any density matrix ρ to $|\psi\rangle\langle\psi|$. (One way to implement this is to turn on a Hamiltonian for which ψ is the ground state, and wait until the system relaxes to its ground state by releasing energy to the environment.)

2. Find Kraus operators of a quantum channel that maps any density matrix for a given system (with finite-dimensional Hilbert space) to a maximally mixed one, a multiple of the identity. (This can arise as the outcome of sufficiently random interaction of the system with its environment.)

3. Do the same for a quantum channel that, in a given basis, maps any $k \times k$ density matrix $\rho = (\rho_{ij})$ to the corresponding diagonal density matrix $\rho_D = \mathrm{diag}(\rho_{11}, \rho_{22}, \ldots, \rho_{kk})$. (An idealized description of a physical realization is as follows. A cavity is probed by atoms. Denote as $|n\rangle$ the state of the cavity when it contains n photons. Suppose that n is unchanged when an atom passes through the cavity, but the final state of the atom depends on n. The probability to find the cavity in state $|n\rangle$ is unchanged by the interaction with a passing atom, so in the basis $\{|n\rangle\}$, the diagonal elements of the density matrix are unchanged. After many atoms have passed through the cavity, an observation of the atoms would reveal with high confidence the number of photons in the cavity. Therefore, tracing over the atomic states, the final density matrix of the cavity is diagonal in the basis $\{|n\rangle\}$. Regardless of what state the cavity begins in, it will end up with high probability in an eigenstate of the photon number operator, though one cannot say what the eigenvalue will be.)

[9] In the most general case, a quantum channel is a "completely positive trace-preserving" (CPTP) map from density matrices on one Hilbert space $\mathcal{H}$ to density matrices on another Hilbert space $\mathcal{H}'$.

4. Show that the composition of two quantum channels is a quantum channel. If the first channel has Kraus operators E_s, $s = 1, \ldots, p$, and the second has Kraus operators E'_t, $t = 1, \ldots, q$, what are the Kraus operators of the composite channel?

5. This and the next exercise involve quantum channels that map one Hilbert space to another. The goal is to show that natural operations that are well-motivated in other ways can also be viewed as special cases of the evolution described in Eq. (3.84). First, given a Hilbert space $\mathcal{H}$, construct a rather trivial quantum channel that maps density matrices on $\mathcal{H}$ to density matrices on a 1-dimensional Hilbert space $\mathcal{H}_0$. Note that, since a density matrix is hermitian, positive-definite, and of trace 1, there is a unique density matrix on $\mathcal{H}_0$, namely the unit density matrix 1. Thus, given a Hilbert space $\mathcal{H}$, find Kraus operators $E_s : \mathcal{H} \to \mathcal{H}_0$ for a quantum channel that maps any density matrix ρ on $\mathcal{H}$ to the density matrix 1 on $\mathcal{H}_0$. Once you have done this, show that a partial trace is a quantum channel in the following sense. If AB is a bipartite system with Hilbert space $\mathcal{H}_A \otimes \mathcal{H}_B$, find Kraus operators $E_s : \mathcal{H}_A \otimes \mathcal{H}_B \to \mathcal{H}_A$ that implement the partial trace $\rho_{AB} \to \rho_A = \mathrm{Tr}_B \rho_{AB}$. In other words, find operators $E_s : \mathcal{H}_A \otimes \mathcal{H}_B \to \mathcal{H}_A$, satisfying $\sum_s E_s^\dagger E_s = 1$ and $\sum_s E_s \rho_{AB} E_s^\dagger = \mathrm{Tr}_B\, \rho_{AB}$, for any ρ_{AB}.

6. Let A be a quantum system with Hilbert space $\mathcal{H}_A$, and let B be a second quantum system with Hilbert space $\mathcal{H}_B$ and some given density matrix ρ_B. Find Kraus operators $E_s : \mathcal{H}_A \to \mathcal{H}_A \otimes \mathcal{H}_B$ for a quantum channel that combines a quantum system A with some other system B by mapping any given density matrix ρ_A on A to the density matrix $\rho_A \otimes \rho_B$ on AB. (You might want to consider first the trivial case that $\mathcal{H}_A$ is 1-dimensional.) An example of this is what happens whenever a system A under study is combined with some experimental apparatus B, which has been initialized in the state ρ_B.

3.8 Thermodynamics and quantum channels

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

$$\sigma = \frac{1}{Z} \exp(-\beta H). \tag{3.86}$$

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

$$\begin{aligned} S(\rho||\sigma) &= \mathrm{Tr}\, \rho(\log \rho - \log \sigma) = -S(\rho) + \mathrm{Tr}\rho(\beta H + \log Z) \\ &= \beta(E(\rho) - TS(\rho)) + \log Z \end{aligned} \tag{3.87}$$

where the average energy computed in the density matrix ρ is

$$E(\rho) = \mathrm{Tr}\, \rho H. \tag{3.88}$$

We define the free energy

$$F(\rho) = E(\rho) - TS(\rho). \tag{3.89}$$

The log Z term in Eq. (3.87) is independent of ρ and gives a constant that ensures that $S(\sigma||\sigma) = 0$. So

$$S(\rho||\sigma) = \beta(F(\rho) - F(\sigma)). \tag{3.90}$$

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

$$S(\rho||\sigma) \geq S(\rho'||\sigma), \tag{3.91}$$

and therefore

$$F(\rho) \geq F(\rho'). \tag{3.92}$$

In other words, a quantum channel that preserves thermal equilibrium can only reduce the free energy. 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.

To specialize further, take the temperature $T = \infty$, $\beta = 0$. (This makes sense for a system with a finite-dimensional Hilbert space.) The thermal density matrix σ is then maximally mixed, a multiple of the identity. For $T \to \infty$, $F(\rho) \sim -TS(\rho)$. So in this case, reducing the free energy means increasing the entropy. Thus a quantum channel that maps a maximally mixed density matrix to itself can only increase the entropy. The condition that a channel maps a maximally mixed density matrix to itself is $\sum_s E_s E_s^\dagger = 1$. (A channel satisfying this condition is called unital. By contrast, the condition $\sum_s E_s^\dagger E_s = 1$ is satisfied by all quantum channels.)

An example of a quantum channel that maps a maximally mixed density matrix to itself is the channel that maps any density matrix ρ to the corresponding diagonal density matrix ρ_D (in some chosen basis). The fact that the entropy can only increase under such a channel implies the inequality $S(\rho) \leq S(\rho_D)$ (Eq. (3.42)).

4 More on quantum information theory

From this point, one could pursue many different directions toward a deeper understanding of quantum information theory. This article will conclude with three topics that the author found helpful in gaining insight about the meaning of formal definitions such as quantum conditional entropy and quantum relative entropy. These concepts

were defined by formally imitating the corresponding classical definitions, and it is not really clear a priori what to expect of such formal definitions.

A secondary reason for the choice of topics is to help the reader appreciate the importance of monotonicity of quantum relative entropy—and its close cousin, strong subadditivity. At several points, we will have to invoke monotonicity of relative entropy to prove that quantities like quantum mutual information and quantum relative entropy that have been defined in a formal way do behave in a fashion suggested by their names.

The three topics that we will consider are quantum teleportation and conditional entropy, relative entropy and quantum hypothesis testing, and the use of a quantum state to encode classical information.

4.1 Quantum teleportation and conditional entropy

We start with **quantum teleportation** [13]. 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, it is too difficult 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

$$\Psi_{A_1B_1} = \frac{1}{\sqrt{2}}\left(|0\,0\rangle + |1\,1\rangle\right)_{A_1B_1}. \tag{4.1}$$

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_0A_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 the process, she also loses no information about A_0, since she had none before. But as we will see, after getting her measurement outcome, she can tell Bob what to do to recreate 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

$$\frac{1}{\sqrt{2}}\left(|0\,0\rangle \pm |1\,1\rangle\right)_{A_0A_1} \text{ and } \frac{1}{\sqrt{2}}\left(|0\,1\rangle \pm |1\,0\rangle\right)_{A_0A_1}. \tag{4.2}$$

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

$$\Psi_{A_0A_1B_1} = \frac{1}{\sqrt{2}}\left(\alpha|0\,0\,0\rangle + \alpha|0\,1\,1\rangle + \beta|1\,0\,0\rangle + \beta|1\,1\,1\rangle\right)_{A_0A_1B_1}. \tag{4.3}$$

Suppose that the outcome of Alice's measurement is to learn that A_0A_1 is in the state

$$\frac{1}{\sqrt{2}}(|0\,0\rangle - |1\,1\rangle)_{A_0A_1}. \tag{4.4}$$

After the measurement, B_1 will be in the state $(\alpha|0\rangle - \beta|1\rangle)_{B_1}$. Knowing this, Alice can tell Bob that he can recreate the initial state by acting on his qubit by

$$\Psi_{B_1} \to \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \Psi_{B_1} \tag{4.5}$$

in the basis $|0\rangle$, $|1\rangle$. The other cases are similar, as the reader can verify.

We will analyze 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 and Bob do not have access. We assume that R is maximally entangled with A_0 in a known state, say

$$\Psi_{RA_0} = \frac{1}{\sqrt{2}}\left(|0\,0\rangle + |1\,1\rangle\right)_{RA_0}. \tag{4.6}$$

In this version of the problem, Alice's goal is to manipulate her system A_0A_1 in some way, and then tell Bob what to do to his system $B = B_1$ so that in the end the system RB_1 will be in the same state

$$\Psi_{RB_1} = \frac{1}{\sqrt{2}}\left(|0\,0\rangle + |1\,1\rangle\right)_{RB_1} \tag{4.7}$$

that RA_0 was previously—with R never being touched. In this version of the problem, the combined system $RAB_1 = RA_0A_1B_1$ starts in a pure state $\Psi_{RAB_1} = \Psi_{RA_0} \otimes \Psi_{A_1B_1}$. 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}, so $S_{AB} = S_{A_0} = 1$ (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. Hence

$$S(A|B) = S_{AB} - S_B = 1 - 1 = 0. \tag{4.8}$$

It turns out that this is the key to quantum teleportation: teleportation, in a suitably generalized sense, is possible when and only when

$$S(A|B) \leq 0. \tag{4.9}$$

Let us explain first 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 neither one has access to R. For teleportation, Alice might measure her system A using some rank 1 orthogonal projection operators π_i. (If she makes a more general measurement, for example using projection operators of higher rank, the system RB does not end up in a known pure state and she will not be able to give appropriate instructions to Bob.) No matter what answer she gets, after the measurement, system A is in a pure state and therefore RB is also in a pure state χ_{RB}, generally entangled. For teleportation, Alice has to choose the π_i so that, no matter what outcome she gets, the density matrix ρ_R of R is the same as before. If this is so, then after her measurement, the state χ_{RB} of RB is a purification of the original ρ_R. Since she knows her measurement outcome, Alice knows which entangled state is χ_{RB} and can convey this information to Bob. Bob is then in possession of part B of a known purification χ_{RB} of system R. He makes in his lab a copy A' of Alice's original system A, initialized in a known pure state $\Omega_{A'}$, so now he has part $A'B$ of a known purification $\widetilde{\Psi}_{RA'B} = \Omega_{A'} \otimes \chi_{RB}$ of ρ_R. By a unitary transformation of system $A'B$, which Bob can implement in his lab, $\Psi_{RA'B}$ can be converted into any other pure state of $RA'B$ that purifies the same ρ_R. (This was explained following Eq. (3.13).) So Bob can convert $\widetilde{\Psi}_{RA'B}$ to a copy of the original Ψ_{RAB}.

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

$$S_{AB} = S_R. \tag{4.10}$$

Bob's density matrix at the beginning is

$$\rho_B = \mathrm{Tr}_{RA}\, \rho_{RAB} \tag{4.11}$$

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

$$S_B = S(\rho_B). \tag{4.12}$$

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

$$\rho_B^i = \frac{1}{p_i} \mathrm{Tr}_{RA}\, \pi_i \rho_{RAB}. \tag{4.13}$$

Note that

$$\rho_B = \sum_i p_i \rho_B^i, \tag{4.14}$$

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

$$S\left(\rho^i_B\right) = S_{AB}. \tag{4.15}$$

If all this is possible

$$S_{AB} = S\left(\rho^i_B\right) = \sum_i p_i S\left(\rho^i_B\right). \tag{4.16}$$

The concavity inequality (3.39) or equivalently positivity of the Holevo information (3.40) says that if $\rho_B = \sum_i p_i \rho^i_B$ then

$$S(\rho_B) \geq \sum_i p_i S\left(\rho^i_B\right). \tag{4.17}$$

So if teleportation can occur,

$$S_{AB} = \sum_i p_i S\left(\rho^i_B\right) \leq S(\rho_B) = S_B \tag{4.18}$$

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 [14]. (In this generality, what we are calling teleportation is known as **state merging**.) One has to consider the problem of teleporting not a single system but N copies of the system for large N. (This is a common device in quantum information theory. It is a rough analog of the fact that to get simple statements in the classical case in Sect. 2, we had to consider a long message, obtained by sampling N times from a probability distribution.) 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 condition $S(A|B) \leq 0$. 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 state merging is vanishing for $N \to \infty$). This statement actually has strong subadditivity as a corollary [14]. This approach to strong subadditivity has been described in sections 10.8-9 of [4].

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. We will find a quantum analog of this, but now involving qubits rather than classical bits. Suppose that $S(A|B) > 0$ and Alice nevertheless wants to share her state with Bob. Now we have to assume that Alice is capable of quantum communication, that is of sending a quantum system to Bob

while maintaining its quantum state, but that she wishes to minimize the amount of quantum communication she will need. 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, 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 $S(A|B)$ nonpositive and so make teleportation or state merging possible without any further quantum communication.

If $S(A|B)$ is negative, teleportation or state merging is possible to begin with and $-S(A|B)$ is the number of maximally entangled qubit pairs that Alice and Bob can be left with afterwards [14]. This may be seen as follows. Alice creates an auxiliary system $A'A''$, where A' consists of n qubits that are completely entangled with another set of n qubits that comprise system A''. Alice considers the problem of teleporting to Bob the combined system $\overline{A} = A''A$, while leaving A' untouched. Since $S(\overline{A}|B) = n + S(A|B)$, Alice observes that $S(\overline{A}|B) < 0$ provided $n < -S(A|B)$. Given this inequality, Alice can teleport $\overline{A} = A''A$ to Bob, keeping A' in reserve. At the end of this, Alice and Bob share n maximally entangled qubit pairs, namely Alice's system A' and Bob's copy of A''. This description is a shorthand; it is implicit that at each stage, we are free to replace the system under consideration by the tensor product of N copies of itself, for some large N. As a result, integrality of n is not an important constraint. A more precise statement of the conclusion is that for large N, after teleportation to Bob of part $A^{\otimes N}$ of a composite system $A^{\otimes N}B^{\otimes N}$, Alice and Bob can be left with up to $-NS(A|B)$ maximally entangled qubit pairs.

4.2 Quantum relative entropy and hypothesis testing

In a somewhat similar way, we can give a physical meaning to the relative entropy $S(\rho||\sigma)$ between two density matrices ρ, σ. Recall from Sect. 2.3 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

$$2^{-NS(P||Q)}. \tag{4.19}$$

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

$$2^{-NS(\rho||\sigma)}. \tag{4.20}$$

Let us first see that monotonicity of relative entropy implies that one cannot do better than that [15]. A measurement is a special case of a quantum channel, in the following sense. To measure a system X, one lets it interact quantum mechanically with some other system YC where Y is any quantum system and C is the measuring device. After they interact, one looks at the measuring device and forgets 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 with n distinguishable quantum states, this means that in a distinguished basis $|\alpha\rangle, \alpha = 1, \dots, n$, its density matrix β_C will have a diagonal form

$$\beta_C = \sum_\alpha b_\alpha |\alpha\rangle\langle\alpha|. \tag{4.21}$$

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 (or simply a quantum channel, since a partial trace can be viewed as a quantum channel) that maps ρ and σ into density matrices for C

$$\rho_C = \sum_\alpha r_\alpha |\alpha\rangle\langle\alpha| \quad \sigma_C = \sum_\alpha s_\alpha |\alpha\rangle\langle\alpha|, \tag{4.22}$$

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

$$2^{-NS_{\mathrm{cl}}(R||S)}, \tag{4.23}$$

where $S_{\mathrm{cl}}(R||S)$ is the classical relative entropy between R and S.

This is the same as the relative entropy between ρ_C and σ_C:

$$S(\rho_C||\sigma_C) = S_{\mathrm{cl}}(R||S). \tag{4.24}$$

And monotonicity of relative entropy gives

$$S(\rho||\sigma) \geq S(\rho_C||\sigma_C). \tag{4.25}$$

So if we follow this procedure, then $S(\rho||\sigma)$ gives a bound on how well we can do:

$$2^{-NS_{\mathrm{cl}}(R||S)} \geq 2^{-NS(\rho||\sigma)}. \tag{4.26}$$

Actually, quantum mechanics allows us to do something more sophisticated than making N repeated measurements of the system of interest. We could more generally make a joint measurement on all N copies. Taking N copies replaces the Hilbert space $\mathcal{H}$ of the system under study by $\mathcal{H}^{\otimes N}$, and replaces the density matrices σ and ρ by $\sigma^{\otimes N}$ and $\rho^{\otimes N}$. All entropies and relative entropies are multiplied by N. A joint measurement on N copies would convert a density matrix $\sigma^{\otimes N}$ or $\rho^{\otimes N}$ to a probability distribution $S^{[N]}$ or $R^{[N]}$. We will not learn much from a single joint measurement on N copies, since it will just produce a random answer. But given NN' copies of the system, we could repeat N' times a joint measurement of N copies. The ability to distinguish $S^{[N]}$ from $R^{[N]}$ in N' tries is controlled for large N' by $2^{-N'S_{\mathrm{cl}}(R^{[N]}||S^{[N]})}$. The monotonicity of relative entropy gives $2^{-N'S_{\mathrm{cl}}(R^{[N]}||S^{[N]})} \geq 2^{-N'S(\rho^{\otimes N}||\sigma^{\otimes N})} = 2^{-\widehat{N}S(\rho||\sigma)}$, where

$\widehat{N} = NN'$. So also with such a more general procedure, the ability to disprove in $\widehat{N}$ trials an initial hypothesis σ for a system actually described by ρ is bounded by $2^{-\widehat{N}S(\rho||\sigma)}$.

In the limit of large $\widehat{N}$, it is actually possible to saturate this bound, as follows [16,17]. 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 many 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

$$S(\rho_D||\sigma) = S_{\rm cl}(R||S). \tag{4.27}$$

This is not very useful, because it is hard to compare $S(\rho_D||\sigma)$ to $S(\rho||\sigma)$. That is why it is necessary to consider a joint measurement on N copies, for large N, which makes possible an easier alternative to comparing $S(\rho_D||\sigma)$ to $S(\rho||\sigma)$, as we will see.

Let us recall the definition of relative entropy:

$$S\left(\rho^{\otimes N}||\sigma^{\otimes N}\right) = {\rm Tr}\,\rho^{\otimes N}\log\rho^{\otimes N} - {\rm Tr}\,\rho^{\otimes N}\log\sigma^{\otimes N}. \tag{4.28}$$

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

$$S\left(\rho^{\otimes N}||\sigma^{\otimes N}\right) - S\left(\left(\rho^{\otimes N}\right)_D||\sigma^{\otimes N}\right) = {\rm Tr}\rho^{\otimes N}\log\rho^{\otimes N} - {\rm Tr}\left(\rho^{\otimes N}\right)_D\log\left(\rho^{\otimes N}\right)_D. \tag{4.29}$$

Actually, there are many bases in which $\sigma^{\otimes N}$ is diagonal; it will be important to choose the right one in defining $(\rho^{\otimes N})_D$. For large N, and with the right choice of basis, we will be able to get a useful bound on the right hand side of Eq. (4.29).

Roughly speaking, there is simplification for large N because group theory can be used to simultaneously put $\rho^{\otimes N}$ and $\sigma^{\otimes N}$ in a block diagonal form with relatively small blocks. This will make possible the comparison we need. In more detail, the group S_N of permutations of N objects acts in an obvious way on $\mathcal{H}^{\otimes N}$. It commutes with the action on $\mathcal{H}^{\otimes N}$ of $U(k)$, the group of unitary transformations of the k-dimensional Hilbert space $\mathcal{H}$. Schur-Weyl duality gives the decomposition of $\mathcal{H}^{\otimes N}$ in irreducible representations of $S_N \times U(k)$. Every Young diagram Y with N boxes and at most k rows determines an irreducible representation λ_Y of S_N and an irreducible representation μ_Y of $U(k)$. The decomposition of $\mathcal{H}^{\otimes N}$ in irreducibles of $S_N \times U(k)$ is

$$\mathcal{H}^{\otimes N} = \oplus_Y \lambda_Y \otimes \mu_Y. \tag{4.30}$$

The λ_Y of distinct Y are non-isomorphic, and the same is true of the μ_Y. Let a_Y and b_Y be, respectively, the dimension of λ_Y and of μ_Y. The maximum value of b_Y is bounded[10] by a power of N:

$$b_{\max} \leq (N+1)^{k(k-1)/2}. \tag{4.31}$$

The important point will be that $b_{\max}$ grows only polynomially for $N \to \infty$, not exponentially. In contrast, the numbers a_Y can be exponentially large for large N.

Equation (4.30) gives a decomposition of $\mathcal{H}^{\otimes N}$ as the direct sum of subspaces of dimension $a_Y b_Y$. Since $\rho^{\otimes N}$ and $\sigma^{\otimes N}$ commute with S_N, they are block diagonal with respect to this decomposition. But more specifically, the fact that $\rho^{\otimes N}$ and $\sigma^{\otimes N}$ commute with S_N means that each $a_Y b_Y \times a_Y b_Y$ block is just the direct sum of a_Y identical blocks of size $b_Y \times b_Y$. So $\rho^{\otimes N}$ has a decomposition

$$\rho^{\otimes N} = \begin{pmatrix} p_1\rho_1 & & & \\ & p_2\rho_2 & & \\ & & p_3\rho_3 & \\ & & & \ddots \end{pmatrix} \tag{4.32}$$

in blocks of size $b_Y \otimes b_Y$, with each such block occurring a_Y times, for all possible Y. (The total number of blocks is $\sum_Y a_Y$.) The ρ_i are density matrices and the p_i are nonnegative numbers adding to 1. In the same basis, $\sigma^{\otimes N}$ has just the same sort of decomposition:

$$\sigma^{\otimes N} = \begin{pmatrix} q_1\sigma_1 & & & \\ & q_2\sigma_2 & & \\ & & q_3\sigma_3 & \\ & & & \ddots \end{pmatrix}. \tag{4.33}$$

We can furthermore make a unitary transformation in each block to diagonalize $\sigma^{\otimes N}$. This will generically not diagonalize $\rho^{\otimes N}$. But because $\rho^{\otimes N}$ is block diagonal with relatively small blocks, its entropy can be usefully compared with that of the diagonal density matrix $(\rho^{\otimes N})_D$ that is obtained by setting to 0 the off-diagonal matrix elements of $\rho^{\otimes N}$ in a basis in which $\sigma^{\otimes N}$ is diagonal within each block and keeping the diagonal ones:

[10] See Eq. (6.16) of [17]. One approach to this upper bound is as follows. In general, the highest weight of an irreducible representation of the group $SU(k)$ is a linear combination of certain fundamental weights with nonnegative integer coefficients a_i, $i = 1, \ldots, k-1$. In the case of a representation associated to a Young diagram with N boxes, the a_i are bounded by N. The dimension of an irreducible representation with highest weights $(a_1, a_2, \ldots, a_{k-1})$ is a polynomial in the a_i of total degree $k(k-1)/2$, so if all a_i are bounded by N, the dimension is bounded by a constant times $N^{k(k-1)/2}$. One way to prove that the dimension is a polynomial in the a_i of the stated degree is to use the Borel-Weil-Bott theorem. According to this theorem, a representation with highest weights $(a_1, a_2, \ldots, a_{k-1})$ can be realized as $H^0(F, \otimes_{i=1}^{k-1} \mathcal{L}_i^{a_i})$, where $F = SU(k)/U(1)^{k-1}$ is the flag manifold of the group $SU(k)$ and $\mathcal{L}_i \to F$ are certain holomorphic line bundles. Because F has complex dimension $k(k-1)/2$, the Riemann-Roch theorem says that the dimension of $H^0(F, \otimes_{i=1}^{k-1} \mathcal{L}_i^{a_i})$ is a polynomial in the a_i of that degree.

$$(\rho^{\otimes N})_D = \begin{pmatrix} p_1\rho_{1,D} & & & \\ & p_2\rho_{2,D} & & \\ & & p_3\rho_{3,D} & \\ & & & \ddots \end{pmatrix}. \tag{4.34}$$

One finds then

$$\mathrm{Tr}\rho^{\otimes N}\log\rho^{\otimes N} - \mathrm{Tr}(\rho^{\otimes N})_D \log\left(\left(\rho^{\otimes N}\right)_D\right) = \sum_i p_i(S(\rho_{iD}) - S(\rho_i)). \tag{4.35}$$

It is important that a potentially large term $\sum_i p_i \log p_i$ cancels out here. 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 above by $b_{\max} \sim N^{k(k-1)/2}$, and $\sum_i p_i = 1$, the right hand side[11] of Eq. (4.35) is bounded by $\log b_{\max} \sim \frac{1}{2}k(k-1)\log N$, which for large N is negligible compared to N.

Combining this with Eqs. (4.27) and (4.29), we see that for large N, a measurement that projects onto 1-dimensional eigenspaces of σ_i within each block maps the density matrices $\rho^{\otimes N}$ and $\sigma^{\otimes N}$ to classical probability distributions $R^{[N]}$ and $S^{[N]}$ such that the quantum relative entropy $S(\rho^{\otimes N}||\sigma^{\otimes N})$ and the classical relative entropy $S(R^{[N]}||S^{[N]})$ are asymptotically equal. To be more precise, $S(\rho^{\otimes N}||\sigma^{\otimes N}) = NS(\rho||\sigma)$ is of order N for large N, and differs from $S(R^{[N]}||S^{[N]})$ by at most a constant times $\log N$. In other words

$$S(\rho||\sigma) = \frac{1}{N}S(\rho^{\otimes N}||\sigma^{\otimes N}) = \frac{1}{N}S\left(R^{[N]}||S^{[N]}\right) + \mathcal{O}\left(\frac{\log N}{N}\right). \tag{4.36}$$

Once we have identified a measurement that converts the quantum relative entropy (for N copies of the original system) to a classical relative entropy, we take many copies again and invoke the analysis of classical relative entropy in Sect. 2.3. In more detail, consider a composite system consisting of N copies of the original system. Suppose that we observe N' copies of this composite system (making NN' copies of the original system), for very large N'. On each copy of the composite system, we make the above-described measurement. This means that we sample N' times from the classical probability distribution $S^{[N]}$ (if the original hypothesis σ was correct) or $R^{[N]}$ (if the original system was actually described by ρ). According to the classical analysis in Sect. 2.3, the ability to distinguish between $R^{[N]}$ and $S^{[N]}$ in N' trials is controlled by $2^{-N'S(R^{[N]}||S^{[N]})}$. According to Eq. (4.36), this is asymptotically the same as $2^{-N'S(\rho^{\otimes N}||\sigma^{\otimes N})} = 2^{-NN'S(\rho||\sigma)}$. In short, we learn that after a suitable measurement on $\widehat{N} = NN'$ copies of the original system, we can distinguish between the hypotheses σ and ρ with a power

$$2^{-\widehat{N}S(\rho||\sigma)}, \tag{4.37}$$

[11] The right hand side is actually positive because of the inequality (3.42).

saturating the upper bound (4.26) (with the total number of trials now being $\widehat{N}$ rather than N). In the exponent, there are errors of order $N' \log N$ (from the logarithmic correction in (4.36)) and $N \log N'$ (coming from the fact that the classical analysis of Sect. 2.3, which for instance used only the leading term in Stirling's formula, has corrections of relative order $\frac{1}{N'} \log N'$).

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. A noteworthy fact [16] is that the measurement that must be made on the composite system to accomplish this depends only on σ (the initial hypothesis) and not on ρ (the unknown answer).

At the outset, we assumed monotonicity of relative entropy and deduced from it an upper bound (4.20) on how well one can distinguish two density matrices in N trials. Actually, now that we know that the upper bound is attainable, one can reverse the argument and show that this upper bound implies monotonicity of relative entropy. Suppose that AB is a bipartite system with density matrices ρ_{AB}, σ_{AB} that we want to distinguish by a measurement. One thing that we can do is to forget system B and just make measurements on A. The above argument shows that, after taking N copies, the reduced density matrices $\rho_A = \mathrm{Tr}_B\, \rho_{AB}, \sigma_A = \mathrm{Tr}_B\, \sigma_{AB}$ can be distinguished at the rate $2^{-NS(\rho_A||\sigma_A)}$. But since measurements of subsystem A are a special case of measurements of AB, this implies that ρ_{AB} and σ_{AB} can be distinguished at the rate $2^{-NS(\rho_A||\sigma_A)}$. If therefore we know the bound (4.20), which says that ρ_{AB} and σ_{AB} cannot be distinguished at a faster rather than $2^{-NS(\rho_{AB}||\sigma_{AB})}$, then the monotonicity inequality $S(\rho_{AB}||\sigma_{AB}) \geq S(\rho_A||\sigma_A)$ follows. In [18], monotonicity of relative entropy has been proved by giving an independent proof of the upper bound on how well two density matrices can be distinguished.

4.3 Encoding classical information in a quantum state

Finally, we will address the following question: how many bits of information can Alice send to Bob by sending him a quantum system X with a k-dimensional Hilbert space $\mathcal{H}$? (See [4], especially section 10.6, for more on this and related topics.)

One thing Alice can do is to send one of k orthogonal basis vectors in $\mathcal{H}$. Bob can find which one she sent by making a measurement. So in that way Alice can send $\log k$ 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 that takes the values x_i, $i = 1, \dots, n$ with probability p_i. When the value is x_i, she writes down this fact in her notebook C and creates a density matrix ρ_X^i on system X. If $|i\rangle$ is the state of the notebook when Alice has written the value x_i, then on the combined system CX, Alice has created the density matrix

$$\rho_{CX} = \sum_i p_i |i\rangle\langle i| \otimes \rho_X^i \tag{4.38}$$

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 ρ_{CX} of the system CX is the one (Eq. (3.56)) that was used earlier in discussing the entropy inequality for mixing. It is sometimes called a classical-quantum density matrix. The reduced density matrix of X is $\rho_X = \mathrm{Tr}_C\, \rho_{CX} = \sum_i p_i \rho_X^i$. As before, the mutual information between C and X is the Holevo information

$$I(C; X) = S(\rho_X) - \sum_i p_i S(\rho_X^i). \tag{4.39}$$

Since $S(\rho_X^i) \geq 0$ and $S(\rho_X) \leq \log k$, it follows that

$$I(C; X) \leq \log k. \tag{4.40}$$

If we knew that quantum mutual information has a similar interpretation to classical mutual information, we would stop here and say that since $I(C; X) \leq \log k$, at most $\log k$ bits of information about the contents of Alice's notebook have been encoded in X. However, we aim to demonstrate that quantum mutual information behaves like classical mutual information, at least in this respect, not to assume it. As we will see, what we want is precisely what monotonicity of mutual information says, in the present context.

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 unitary transformation or more general quantum channel and then reads C'. The combined operation is a quantum channel. As in our discussion of relative entropy, the outcome of the channel is a density matrix of the form

$$\rho_{C'} = \sum_{\alpha=1}^{r} q_\alpha |\alpha\rangle\langle\alpha|, \tag{4.41}$$

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 $I(C; C')$.

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. In more detail, the initial description of the combined system $CXYC'$ is by the tensor product of a density matrix ρ_{CX} for CX and a density matrix $\rho_{YC'}$ for YC'. As one can deduce immediately from the definitions, the mutual information between C and XYC' if the full system $CXYC'$ is described by $\rho_{CX} \otimes \rho_{YC'}$ is the same as the mutual information between C and X if the subsystem CX is described by ρ_{CX}. Bob then acts on XYC' with a unitary transformation, or

maybe a more general quantum channel, which can only reduce the mutual information. Then he takes a partial trace over XY, which also can only reduce the mutual information, since monotonicity of mutual information under partial trace tells us that

$$I(C; XYC') \geq I(C; C'). \tag{4.42}$$

So

$$\log k \geq I(C; X) = I(C; XYC')_{\text{before}} \geq I(C; XYC')_{\text{after}} \geq I(C; C')_{\text{after}}, \tag{4.43}$$

where "before" and "after" mean before and after Bob's manipulations. More briefly, any way that Bob processes the signal he receives can only reduce the mutual information. Thus Alice cannot encode more than $\log k$ bits of classical information in an k-dimensional quantum state, though it takes strong subadditivity (or its equivalents) to prove this.

The problem that we have discussed also has a more symmetrical variant. In this version, Alice and Bob share a bipartite state AB; Alice has access to A and Bob has access to B. The system is initially described by a density matrix ρ_{AB}. Alice makes a generalized measurement of A and Bob makes a generalized measurement of B. What is the maximum amount of information that Alice's results may give her about Bob's measurements, and vice-versa? An upper bound is given by the mutual information $I(A; B)$ in the initial density matrix ρ_{AB}. Alice's measurements amount to a quantum channel mapping her system A to her measurement apparatus C; Bob's measurements amount to a quantum channel mapping his system B to his measurement apparatus C'. The mutual information between their measurement outcomes is simply the mutual information $I(C; C')$ in the final state. Monotonicity of mutual information in any quantum channel says that this can only be less than the initial $I(A; B)$.

A more subtle issue is the extent to which these upper bounds can be saturated. For an introduction to such questions see [4], section 10.6.

Acknowledgements Research supported in part by NSF Grant PHY-1606531. I thank N. Arkani-Hamed, J. Cotler, B. Czech, M. Headrick, and R. Witten for discussions. I also thank M. Hayashi, as well as the referees, for some explanations and helpful criticisms and for a careful reading of the manuscript.

References

1. M.A. Nielsen, I.L. Chuang, *Quantum Computation And Quantum Information* (Cambridge University Press, Cambridge, 2000)
2. T.M. Cover, J.A. Thomas, *Elements of Information Theory*, 2nd edn. (Wiley, New York, 2006)
3. M.M. Wilde, *Quantum Information Theory*, 2nd edn. (Cambridge University Press, Cambridge, 2017)
4. J. Preskill, Lecture notes (2019). http://www.theory.caltech.edu/~preskill/ph219/index.html#lecture
5. C.E. Shannon, A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379-423–623-656 (1918)
6. M.F. Leifer, R.W. Spekkens, Towards a formulation of quantum theory as a causally neutral theory of Bayesian inference. Phys. Rev. A **88**, 052130 (2013). arXiv:1107.5849
7. A.S. Holevo, Bounds for the quantity of information transmitted by a quantum communication channel. Probl. Inf. Transm. **9**, 177–83 (1973)

8. H. Araki, E.H. Lieb, Entropy inequalities. Commun. Math. Phys. **18**, 160–70 (1970)
9. H. Umegaki, Conditional expectation in an operator algebra. Kodai Math. Sem. Rep. **14**, 59–85 (1962)
10. E.H. Lieb, M.B. Ruskai, Proof of the strong subadditivity of quantum mechanical entropy. J. Math. Phys. **14**, 1938 (1973)
11. E.H. Lieb, Convex trace functions and the Wigner–Yanase–Dyson conjecture. Adv. Math. **11**, 267–88 (1973)
12. E. Witten, Notes on some entanglement properties of quantum field theory. Rev. Mod. Phys. **90**, 045003 (2018). arXiv:1803.04993
13. C.H. Bennett, G. Brassard, C. Crépeau, R. Jozsa, A. Peres, W.K. Wootters, Teleporting an unknown quantum state via dual classical and Einstein–Podolsky–Rosen channels. Phys. Rev. Lett. **70**, 1895–9 (1993)
14. M. Horodecki, J. Oppenheim, A. Winter, Quantum state merging and negative information. Commun. Math. Phys. **269**, 107–36 (2007). arXiv:quant-ph/0512247
15. F. Hiai, D. Petz, The proper formula for relative entropy and its asymptotics in quantum probability. Commun. Math. Phys. **143**, 99–114 (1991)
16. M. Hayashi, Asymptotics of quantum relative entropy from representation theoretical viewpoint. J. Phys. A **34**, 3413–20 (2001)
17. M. Hayashi, *A Group Theoretic Approach to Quantum Information* (Springer, New York, 2017)
18. I. Bjelakovic, R. Siegmund-Schultze, Quantum Stein's Lemma Revisited, Inequalities For Quantum Entropies, and a Concavity Theorem of Lieb (2012). arXiv:quant-ph/0307170