Prepared for submission to JHEP

A brief overview of quantum information theory and related topics

Xiao-Liang Qi

Abstract: The note is updated Thursday 24th February, 2022.

Contents

1 Classical information theory

1.1 Shannon entropy

The amount of information in a message depends on its prior probability. If I expect "the sun rises from the east" with probability 1, then hearing this statement gives me no additional information. On the contrary, if I heard some rare event (say an earthquake is happening in 10 seconds) it contains a huge amount of information.

How to quantify this intuition? If a message has prior probability p, we can assume that the amount of information is a function of p, denoted as $I(p)$. If we consider two independent events (say the weather tomorrow and the stock market tomorrow), which has probability distribution p_i and q_a for possible outputs i, a, then the joint probability is $P_{ia} = p_i q_a$. We expect the amount of information in the message ia is just the sum of the two, so that the function $f(p)$ is required to satisfy

$$I(p_i q_a) = I(p_i) + I(q_a) \tag{1.1}$$

This equation suggests that I is a linear function of $\log p$. Without losing generality, we can define

$$I(p) = -\log p \tag{1.2}$$

The minus sign is chosen to be consistent with the intuition that rarer events contain higher information. Probability 1 events contains zero information. Since $p \leq 1$, $I(p)$ is non-negative.

For a probability distribution p_i, the average information amount is the Shannon entropy:

$$S(\{p_i\}) = \sum_i p_i I(p_i) = -\sum_i p_i \log p_i \tag{1.3}$$

Since the amount of information is related to the uncertainty (before we heard the message), the entropy is also a measure of uncertainty (how much we don't already know before getting the message). Historically, entropy as a measure of uncertainty was proposed by Boltzman in statistical physics.

1.2 Typical states

Consider a coin that has half probability up and half probability down when thrown. If we throw it 100 times we expect that roughly half of the time it is up. As we throw more and more, the percentage of times it is up should approach half more and more accurately. This is how we measure the probability by doing the experiment many times. To make this more precise, consider a coin with two states with probability p for up and $1-p$ for down. If we throw the coin N times, the probability that it has k up is

$$P_k = \binom{N}{k} p^k (1-p)^{N-k} \tag{1.4}$$

This probability distribution peaks sharply at $\frac{k}{N} = p$. To see this we assume $k \gg 1, N \gg 1$ and use the Stirling formula

$$-\log P_k = -k \log p - (N-k)\log(1-p) - N \log N + k \log k + (N-k)\log(N-k) \tag{1.5}$$

$$= N\left(q \log \frac{q}{p} + (1-q)\log\frac{1-q}{1-p}\right) \tag{1.6}$$

with $q = \frac{k}{N}$. This quantity is actually N times the relative entropy between two probability distributions $\{q, 1-q\}$ and $\{p, 1-p\}$, but we will postpone that discussion to later. The key thing to know is that P_k strongly peaks around $q = p$. If we expand $q = p + x$, for small x we get

$$-\log P_k \simeq \frac{N}{2p(1-p)} x^2 \tag{1.7}$$

Thus the width of the peak is $\propto N^{-1/2}$. If we take a range $|q-p| \leq \frac{c}{\sqrt{N}}$, or $|k - pN| \leq c\sqrt{N}$, the probability of k falling in this range can be arbitrarily close to 1 by choosing a bigger c. In other words, with a very high probability, if we throw a coin N times, we don't see all possible bit strings in the results, but see one of the typical strings satisfying $|k - pN| \leq c\sqrt{N}$. The number of typical string is roughly speaking

$$\binom{N}{k} \simeq e^{NS(\{p,1-p\})} \tag{1.8}$$

The discussion above was made for a coin with two states, but it can be generalized straightforwardly to multistate case. In general, entropy of a probability distribution tells us that the number of typical states in N copies is $e^{NS(\{p_i\})}$.

Since not all bit strings appear, we could store this data more efficiently if we know this probability distribution in advance. For example, if a book contains words w_i with probability p_i, $i = 1, 2, ..., M$, and the book has N words, then a typical state contains $p_i N$ word w_i, and we can store the book with $NS(\{p_i\}) / \log M$ code words (if the number of code word is also M).

1.3 Conditional entropy and mutual information

We will introduce some concepts that are related to entropy. For a joint probability distribution $P(x_i, y_j)$, we can define the conditional probability

$$P(y_j|x_i) = \frac{p(x_i, y_j)}{\sum_{y_j} p(x_i, y_j)} \equiv \frac{p(x_i, y_j)}{p(x_i)} \tag{1.9}$$

which is the probability distribution of y conditioned on a given value of x. The entropy of this probability distribution is

$$\begin{aligned} S(Y|x_i) = S\left[P(y_j|x_i)\right] &= -\sum_j P(y_j|x_i) \log P(y_j|x_i) \\ &= -\frac{1}{p(x_i)} \sum_j p(x_i, y_j) \log p(x_i, y_j) + \log p(x_i) \end{aligned} \tag{1.10}$$

If we average this quantity over x_i, we obtain

$$\begin{aligned} S(Y|X) \equiv \sum_i p(x_i) S(Y|x_i) &= -\sum_j p(x_i, y_j) \log p(x_i, y_j) + \sum_i p(x_i) \log p(x_i) \\ &= S(XY) - S(X) \end{aligned} \tag{1.11}$$

e can interpret this as information that are unknown in XY subtracting unknown information in X, which is equal to the remaining uncertainty if we already know X but does not know Y.

If X and Y are independent from each other, $S(XY) = S(X) + S(Y)$, and $S(Y|X) = S(Y)$, which means the uncertainty in Y is the same whether or not we know X. If we take the difference

$$I(X : Y) = S(Y) - S(Y|X) = S(Y) + S(X) - S(XY) \tag{1.12}$$

this measures how much knowing X can reduce the uncertainty in Y. Alternatively we can also say, it measures how much the amount of information in Y is reduced if we already know X. In other words, it measures how much we learn about Y by knowing X. Interestingly, $I(X : Y) = I(Y : X)$ is symmetric. This quantity is called the mutual information.

Mutual information is related to the classical channel capacity. A classical channel takes an input x_i and generates an output y_j with the probability distribution $p(y_j|x_i)$. If

Alice encode a message with probability distribution $p(x_i)$, the joint probability distribution is $p(x_i, y_j) = p(x_i)p(y_j|x_i)$. The channel capacity C is defined by the amount of information that can be sent through the channel, per each use of the channel. More precisely, if we take an input message m_I and encode it into a sequence $x_{i_1}x_{i_2}...x_{i_n}$ and fed into n copies of the channel. The output is a sequence $y_{i_1}y_{i_2}...y_{i_n}$. Then a decoding algorithm is needed to map it back to message m_I. If $I = 1, 2, ..., K$ and m_I can be transmitted faithfully, we say the channel has a capacity $C \geq \frac{\log K}{n}$. Take $n \to \infty$ and take an upperbound over encoding and decoding, we define $C = \lim_{n\to\infty} \sup \frac{\log K}{n}$.

To see how to compute C, we again take a two-state example, when $x_i = 0, 1$, $y_i = 0, 1$. If $x_{i_1}x_{i_2}...x_{i_n}$ contains k 0's, then y_i corresponding to these x are generated by the probability distribution $p(y_j|0)$. The 1's correspond to distribution $p(y_j|1)$. Consequently, the number of typical states for the bitstring $y_{i_1}y_{i_2}...y_{i_n}$ is

$$e^{kS(p(y_j|0))} \times e^{(N-k)S(p(y_j|1))} = e^{NH(Y|X)} \tag{1.13}$$

A message about $x_{i_1}x_{i_2}...x_{i_n}$ can be transmitted if typical states in the y bitstring are not totally random. Each given x string with k 0's corresponds to the ensemble of $e^{NS(Y|X)}$ states. The total number of y states with the same y counting is $e^{NS(Y)}$. Thus the number of different x string that can be possibly recovered from y is

$$e^{N(S(Y)-S(Y|X))} = e^{NI(X:Y)} \tag{1.14}$$

Note that $I(X : Y) \leq S(X)$, so that if we have an initial distribution $p(x_i)$, it corresponds to $e^{NS(X)}$ typical states and in general not all of them can be recovered from Y.

This discussion then suggest that the channel capacity is the supreme of mutual information over $p(x_i)$, which measures the maximal number of bits that can be transferred using this noisy channel.

$$C = \sup_{p(x_i)} I(X : Y) = \sup_{p(x_i)} (S(Y) - S(Y|X)) \tag{1.15}$$

C measures the maximal amount of information that can be transmitted *per use of the channel* with arbitrarily low error. One interesting feature of the classical channel capacity is that it is subadditive. If we take two identical channels and input a joint probability distribution $p(x_1, x_2)$, we can define the output distribution is

$$p(x_1, x_2, y_1, y_2) = \sum_{x_1,x_2} M(y_1|x_1)M(y_2|x_2)p(x_1, x_2) \tag{1.16}$$

The conditional entropy

$$S(Y_1, Y_2|X_1, X_2) = \sum_{x_1,x_2} p(x_1, x_2)S\left(M(y_1|x_1)M(y_2|x_2)\right) = S(Y_1|X_1) + S(Y_2|X_2) \tag{1.17}$$

Here $S(Y_1|X_1) = \sum_{x_1,x_2} p(x_1, x_2)S\left(M(y_1|x_1)\right)$ is the conditional entropy for the reduced input distribution $p(x_1) = \sum_{x_2} p(x_1, x_2)$. On the other hand, the output distribution satisfies

$$S(Y_1, Y_2) \leq S(Y_1) + S(Y_2) \tag{1.18}$$

Thus

$$I(X_1, X_2 : Y_1, Y_2) \leq I(X_1 : Y_1) + I(X_2 : Y_2) \tag{1.19}$$

This discussion can be generalized to more copies. Therefore if we take N copies of the channel and try to send message through a correlated code word choice, it won't be better than using an uncorrelated choice.

$$C = \frac{1}{N} \sup_{p(x_1,x_2,...,x_N)} I(X_1, X_2, ..., X_N|Y_1, Y_2, ..., Y_N) = \sup_{p(x_1)} I(X : Y) \tag{1.20}$$

As we will see later, the quantum channel case is very different.

1.4 Relative entropy

As we discussed in Eq. (1.6), if we have a probability distribution $\{p, 1-p\}$, the chance that k out of N measurements return 0 is given by $P_k = e^{-NS(\{q,1-q\}|\{p,1-p\})}$ with

$$S(\{q, 1-q\} \,|\, \{p, 1-p\}) = q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p} \tag{1.21}$$

More generally, for two probability distribution $Q = \{q_i\}, P = \{p_i\}$ we have

$$S(Q|P) = \sum_i q_i \log \frac{q_i}{p_i} \tag{1.22}$$

$S(Q|P)$ measures how different Q is from P. A typical bit string from N copies of Q has the probability $P_N = e^{-NS(Q|P)}$ if the probability distribution were P. In other words, if we made the hypothesis that the probability distribution is P, and actually the distribution is Q, $S(Q|P)$ measures how fast we can find out that we are wrong. The relative entropy is also called Kullback–Leibler (KL) divergence.

To see that the relative entropy is non-negative, we can check that

$$\partial_{q_i} S(Q|P) = 1 + \log \frac{q_i}{p_i} \tag{1.23}$$

Note that q_i satisfies the constraint $\sum_i q_i = 1$, so that the minimum of $S(Q|P)$ is given by the requirement

$$\begin{aligned} \partial_{q_i} \left(S(Q|P) - \mu \left(\sum_i q_i - 1 \right) \right) = 0 \\ 1 + \log \frac{q_i}{p_i} = \mu \\ \Rightarrow \frac{q_i}{p_i} = e^{\mu - 1} \end{aligned} \tag{1.24}$$

Since q_i and p_i are both normalized, this requires $p_i = q_i,\ \mu = 1$. In addition,

$$\partial_{q_i} \partial_{q_j} S(Q|P) = \delta_{ij} q_i^{-1} \tag{1.25}$$

Thus $S(Q|P)$ is convex and has $q_i = p_i$ as the only minimum (with value 0).

Mutual information is actually a relative entropy. If we consider a joint probability distribution $P(x_i, y_j)$, with the reduced distribution $P(x_i) = \sum_{y_j} P(x_i, y_j)$ and $P(y_j) = \sum_{x_i} P(x_i, y_j)$, the relative entropy

$$S(P(x_i, y_j)|P(x_i)P(y_j)) = I(X : Y) \tag{1.26}$$

Physically, mutual information measures how different a probability distribution is from product distribution.

The relative entropy has a lot of important properties. One of the most important one is its monotonicity under a classical channel. If we have Q and P, and we have a classical channel defined by the conditional entropy $M(y_j|x_i) = M_{ji}$. Then the output distribution

$$\tilde{q}_j = \sum_i M_{ji} q_i, \ \tilde{p}_j = \sum_i M_{ji} p_i \tag{1.27}$$

has a relative entropy that is smaller or the same.

$$S(Mq|Mp) \leq S(q|p) \tag{1.28}$$

Intuitively, this means that applying M can only make it more difficult to distinguish the two distributions. (For example if $M_{ji} = M(y_j|x_i)$ is independent from the input x_i, it will erase all differences between P and Q.) To prove this inequality, we first consider a special case of it. Consider p as a joint probability distribution $p_{i\alpha} = p(x_i, y_\alpha)$ and the same for $q_{i\alpha} = q(x_i, y_\alpha)$. We want to compare the relative entropy $S(q|p)$ with that of the reduced distribution $p^x(i) = \sum_\alpha p_{i\alpha}$ and similar for q^x.

$$\begin{aligned} S(q|p) - S(q_x|p_x) &= \sum_{i\alpha} q_{i\alpha} \log \frac{q_{i\alpha}}{p_{i\alpha}} - \sum_{i\alpha} q_{i\alpha} \log \frac{\sum_\beta q_{i\beta}}{\sum_\gamma p_{i\gamma}} \\ &= \sum_{i\alpha} q_{i\alpha} \log \frac{q_{i\alpha}/q_i^x}{p_{i\alpha}/p_i^x} = \sum_i q_i^x \sum_\alpha q(\alpha|i) \log \frac{q(\alpha|i)}{p(\alpha|i)} \\ &\equiv \sum_i q_i^x S\left(q(y|x_i)|p(y|x_i)\right) \geq 0 \end{aligned} \tag{1.29}$$

where $q(\alpha|i) = q_{i\alpha}/q_i^x$ is the conditional entropy of y_α condition on $x = x_i$. This is an average of the relative entropy between condition probability distribution $S(q(y|x = x_i)|p(y|x = x_i))$, which is thus non-negative. Eq. (1.29) tells us that neglecting part of the variable (y) will only make it more difficult to distinguish the two distributions, which is reasonable. To see how this imply the more general monotonicity under classical channel, we can introduce a uniform random variable $r \in [0, 1]$ with uniform probability. Defining the following function

$$f(i, r) = j, \text{ if } r \in \left[\sum_{k=1}^{j-1} M_{ji}, \sum_{k=1}^{j} M_{ji}\right) \tag{1.30}$$

This makes sure that the probability $P(y_j|x_i) = \int_0^1 dr \delta_{f(i,r),j} = M_{ji}$. Since the mapping from (i, r) to (j, i, r) is one-to-one, no information is lost. If we define $\tilde{q}_{i,j,r} = q_i P(r) = q_i$ (since $P(r) = 1$) for $j = f(i, r)$, and zero otherwise, then

$$S(\tilde{q}|\tilde{p}) = S(q|p) \tag{1.31}$$

since this is just a (redundant) relabeling of the same data. Now if we forget i and r, the reduced distribution of $\tilde{q}$ and $\tilde{p}$ are

$$\sum_i \int_0^1 dr\tilde{q}\,(j,i,r) = \sum_i \int_0^1 dr q_i \delta_{f(i,r),j} = \sum_i M_{ji} q_i \tag{1.32}$$

Thus according to the result we have proven, we have

$$S\left(\sum_i M_{ji} q_i \middle| \sum_i M_{ji} p_i\right) \leq S\left(\tilde{q}|\tilde{p}\right) = S(q|p) \tag{1.33}$$

The monotonicity of relative entropy plays an important role in information theory, which has many useful consequences.

Monotonicity of mutual information. Since mutual information is a relative entropy, it is also monotonous under classical channel. In particular, if we take $p(x_i, y_j)$ and apply separate channel M_{ji} and N_{kl} to it to obtain $\tilde{p}(x_j, y_l) = \sum_{i,k} M_{ji} N_{lk} p(x_i, y_k)$, the mutual information can only decrease. $I(X:Y)[MNp] \leq I(X:Y)[p]$. As a special case of this, for any three variables $I(X:YZ) \geq I(X:Y)$, since forgetting Z is a special case of a channel.

Strong subadditivity. If we write $I(X:YZ) \geq I(X:Y)$ in term of Shannon entropy, we obtain

$$\begin{aligned} S(X) + S(YZ) - S(XYZ) &\geq S(X) + S(Y) - S(XY) \\ \Rightarrow S(YZ) + S(XY) &\geq S(Y) + S(XYZ) \end{aligned} \tag{1.34}$$

This is called strong subadditivity, which is a key property of entropy.

Joint convexity of relative entropy. The relative entropy is jointly convex:

$$S\left(\lambda q_1 + (1-\lambda)q_2 | \lambda p_1 + (1-\lambda)p_2\right) \leq \lambda S\left(q_1|p_1\right) + (1-\lambda) S\left(q_2|p_2\right) \tag{1.35}$$

for $\lambda \in [0,1]$. To prove this we can define a joint probability of $x_i, s = 1, 2$:

$$Q(x_i, 1) = \lambda q_1(x_i),\ Q(x_i, 2) = (1-\lambda) q_2(x_i) \tag{1.36}$$

and the same for $P(x_i, s)$. The relative entropy is

$$S(Q|P) = \sum_i q_1(x_i)\lambda \log \frac{q_1(x_i)}{p_1(x_i)} + \sum_i q_2(x_i)(1-\lambda)\log\frac{q_2(x_i)}{p_2(x_i)} = \lambda S\left(q_1|p_1\right) + (1-\lambda) S\left(q_2|p_2\right) \tag{1.37}$$

If we forget s variable, the induced probability distribution is

$$q(x_i) = \sum_{s=1,2} Q(x_i, s) = \lambda q_1(x_i) + (1-\lambda) q_2(x_i) \tag{1.38}$$

and similar for $p(x_i)$. Thus according to the monotonicity we obtain Eq. (1.35).

Entropy growth. For a probability $q(x_i)$, we can consider it's relative entropy with the uniform distribution $p(x_i) = \frac{1}{M}$ when there are M states.

$$S(q|p) = -S(\{q(x_i)\}) - \sum_i q(x_i) \log \frac{1}{M} = \log M - S(\{q(x_i)\}) \tag{1.39}$$

Thus the entropy "deficit" comparing with the maximal entropy is a relative entropy. As a consequence, if we consider any classical channel N_{ji} which satisfies

$$\sum_i N_{ji} = 1 \tag{1.40}$$

(which is not always true), this channel preserves the maximal entropy state. Such channel applying to a generic q_i will only increase its entropy, since the relative entropy $S(Nq|Np) = S(Nq|p) = \log M - S(Nq) \leq \log M - S(q)$.

2 Quantum information theory

2.1 Relation between classical and quantum physics

There is an intrinsic relation between quantum mechanics and classical probability. The foundation of quantum mechanics is quantum state, unitary evolution and projective measurement. A quantum state $|\Psi\rangle$ is a vector in the Hilbert space, a complex linear space, modular a $U(1)$ phase. In other words, $|\Psi\rangle$ and $e^{i\theta}|\Psi\rangle$ is considered as the same state. Alternatively, we can also denote states by the projection operator $\rho = |\Psi\rangle\langle\Psi|$.

We assume the dynamics of the state is described by Schroedinger equation

$$i\frac{\partial}{\partial t}|\Psi(t)\rangle = H(t)|\Psi(t)\rangle \tag{2.1}$$

The key thing is that H is Hermitian, so that the time evolution is unitary. $|\Psi(t)\rangle = U(t)|\Psi(0)\rangle$ with $U^\dagger(t)U(t) = \mathbb{I}$ the identity matrix.

Projective measurements are defined by projectors P_n satisfying

$$\Pi_n\Pi_m = \delta_{nm}\Pi_n,\ \Pi_n = \Pi_n^\dagger,\ \sum_n \Pi_n = \mathbb{I} \tag{2.2}$$

In general each projector may have a rank higher than 1. If we have an orthonormal basis $|n\rangle$, then $P_n = |n\rangle\langle n|$ is a special set of projectors, which have rank 1.

A set of projectors define a quantum measurement. The wavefunction $|\Psi\rangle$ is related to measurement result by

$$P_n = \langle\Psi|\Pi_n|\Psi\rangle \tag{2.3}$$

which satisfies $P_n \geq 0,\ \sum_n P_n = 1$.

In probability theory, when we have multiple variables x_i, y_j we in general need to think about joint probability $P(x_i, y_j)$. In the same way, in quantum mechanics if we have multiple degrees of freedom such as different spins, or different (distinguishable) particles,

the Hilbert space has a direct product structure $\mathbb{H} = \mathbb{H}_1 \otimes \mathbb{H}_2$. If we are measuring the first degrees of freedom with projectors $\Pi_n : \mathbb{H}_1 \to \mathbb{H}_2$ in the first Hilbert space, in the bigger Hilbert space we can express the operator as $\Pi_n \otimes \mathbb{I}_2$. For such measurements, the probability

$$P_n = \langle\Psi| \Pi_n \otimes \mathbb{I}_2 |\Psi\rangle \tag{2.4}$$

Independent from Π_n, we see that P_n does not depend on the full $|\Psi\rangle$ but only depends on its reduction to the first Hilbert space $\langle\Psi| ... \otimes \mathbb{I}_2 |\Psi\rangle$. This defines the density operator

$$\rho_1 = \mathrm{tr}_2 |\Psi\rangle \langle\Psi| \tag{2.5}$$

which always satisfy $\mathrm{tr}\rho_1 = 1$ and $\langle\Phi| \rho_1 |\Phi\rangle \geq 0$ for all $|\Phi\rangle \in \mathbb{H}_1$. Thus ρ_1 is positive semi-definite. For the conceptual simplicity one could also forget about the pure state $|\Psi\rangle$ and start from general mixed state ρ. Unitary evolution $\rho \to U\rho U^\dagger$ preserves the normalization and positive semi-definiteness of ρ.

$$P_n = \mathrm{tr}\,(\Pi_n \rho) \tag{2.6}$$

is the general way how quantum mechanics is related to classical probability theory. It is helpful to introduce diagrammatic representation of quantum information concepts, as is shown in Fig. 1.

An alternative way of connecting quantum mechanics with classical probability is the simple statement that *a diagonal density operator ρ corresponds to a classical probabilistic mixture of pure states.* A diagonal density operator can be written as $\rho = \sum_n p_n |n\rangle \langle n|$ with $|n\rangle$ an orthogonal matrix. ρ corresponds to a classical probability p_n. However, it should be noted that this statement is meaningless if we only talk about a single state, since every state can be diagonalized. It becomes a meaningful statement if we consider a family of states which are all diagonal in the same basis. In that case if we are only interested in such states, we can say the physics reduces to classical physics.

It is interesting to note that these two point of views are related. For a state ρ with Hilbert space dimension d, and a set of projection operators $\Pi_n,\ n = 1, 2, .., M\ (M \leq d)$, we can define an ancilla with Hilbert space dimension M^k. Denote an orthonormal basis of the ancilla as $|a_1 a_2 ... a_k\rangle$ with $a_i = 1, 2, ..., M$. Denote a cyclic permutation operator R by

$$R |a_1 a_2 ... a_k\rangle = |a_1 + 1, a_2 + 1, ..., a_k + 1\rangle \tag{2.7}$$

where the addition is in the cyclic group, such that $M + 1 \sim 1$. R is a unitary operator. Now we can define an unitary operator acting on the system and ancilla

$$U = \sum_{n=1}^{M} \Pi_n \otimes R^n \tag{2.8}$$

This is unitary because

$$U^\dagger U = \sum_{n,m} \Pi_m \Pi_n \otimes R^{\dagger m} R^n = \sum_n \Pi_n \otimes \mathbb{I}_A = \mathbb{I}_S \otimes \mathbb{I}_A \tag{2.9}$$

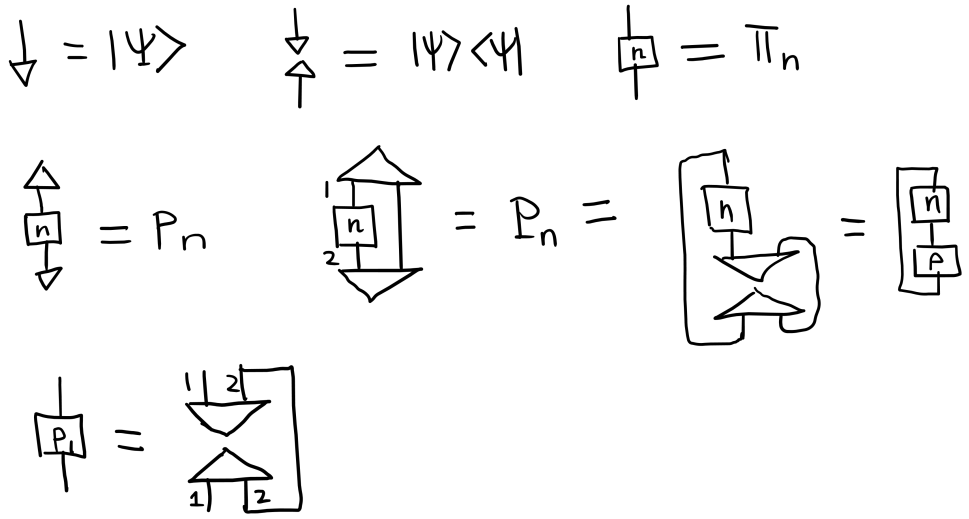

Figure 1. Diagrammatic representation of basic quantum information concepts. An open line pointing up represents a vector in $\mathbb{H}$. A line pointing down represents a vector index in $\mathbb{H}^*$. All internal lines are summed over, just like in Feynman diagrams. A triangle with a single line represents a ket or a bra. A box with two lines represents an operator. In the second and the third row we illustrated the definition of reduced density operator.

If we prepare the ancilla in the state $|00...0\rangle$ ($|0\rangle$ is the same as $|M\rangle$), then we obtain

$$\sigma_{SA} = U\rho \otimes |00...0\rangle \langle 00...0| U^\dagger = \sum_{n,m} \Pi_n \rho \Pi_m \otimes |nn...n\rangle \langle mm...m| \tag{2.10}$$

The reduced density operator of any one of the ancilla is

$$\rho_{A1} = \mathrm{tr}_{S,A_2A_3...A_k} (\sigma_{SA}) = \sum_n \mathrm{tr} (\Pi_n \rho) |n\rangle \langle n| \tag{2.11}$$

This unitary U followed by partial trace is the physical way to realize a quantum measurement (See. Fig. 2). As long as we "write down" the measurement result in more than one ancilla, if we only look at one of the ancilla system it will always have a diagonal density operator in this basis. (This is often referred to as "decoherence".) We can use this ancilla to measure different physical states ρ_1, ρ_2.... Since arbitrary states always lead to ρ_{A1} that are diagonal in the same basis, we can claim that the measurement maps quantum mechanics to classical probability theory.

2.2 Quantum channel

If we consider two systems A, B and a product initial state $\rho_A \otimes \rho_B$, a unitary evolution generically entangle these two systems and lead to

$$\sigma_{AB} = U\rho_A \otimes \rho_B U^\dagger \tag{2.12}$$

If we only focus on the evolution of A subsystem, we can carry a partial trace over B and obtain

$$\sigma_A = \mathrm{tr}_B \left(U\rho_A \otimes \rho_B U^\dagger \right) \tag{2.13}$$

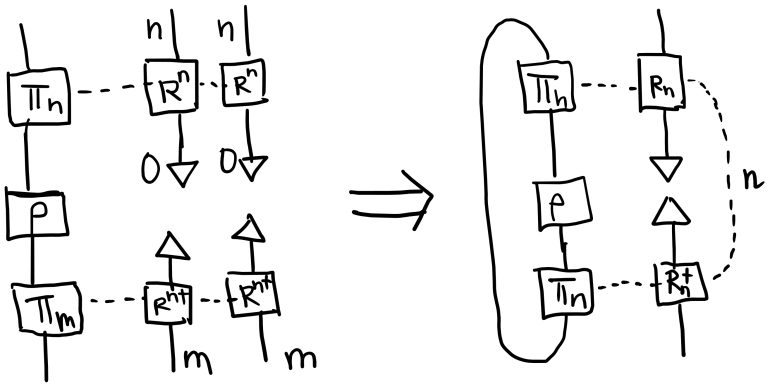

Figure 2. Diagrammatic representation of how a unitary followed by a partial trace can lead to a diagonal density operator of the ancilla.

This is a linear map from ρ_A to σ_A, which maps a density operator to another density operator. Such a map is called completely positive trace-preserving (CPTP) map. Most general CPTP map can always be viewed as a unitary in a bigger system followed by a partial trace. We could also always set ρ_B to be a pure state. If it's not pure, we can introduce a purification of ρ_B by introducing a system $\bar{B}$ with the same Hilbert space dimension as B. Denote $\rho_B = \sum_n p_n |n_B\rangle \langle n_B|$, we define a pure state

$$\left|\psi_{B\overline{B}}\right\rangle = \sum_n \sqrt{p_n} |n_B\rangle \left|n_{\overline{B}}\right\rangle \tag{2.14}$$

The reduced density operator of this state on B is ρ_B. We can take $\left|\psi_{B\overline{B}}\right\rangle \left\langle\psi_{B\overline{B}}\right|$ and apply $U \otimes \mathbb{I}_{\overline{B}}$. Tracing over $B\overline{B}$ gives the same σ_A. The action of U

$$U \otimes \mathbb{I}_{\overline{B}} |\psi_A\rangle \otimes \left|\psi_{B\overline{B}}\right\rangle = \left|\Phi_{AB\overline{B}}\right\rangle \tag{2.15}$$

can be viewed as a linear map from $\mathbb{H}_A$ to $\mathbb{H}_{AB\overline{B}}$, which preserves norm of each state. Such a linear map is called an isometry (Fig. 3). In general, each CPTP map is equivalent to an isometry followed by a partial trace. Mathematically, this is known as Stinespring's dilation theorem.

$$\mathcal{C}(\rho_A) = \mathrm{tr}_E \left(V \rho_A V^\dagger\right) \tag{2.16}$$

$$\text{with } V : \mathbb{H}_A \to \mathbb{H}_A \otimes \mathbb{H}_E \tag{2.17}$$

A CPTP map is also called a quantum channel, which is the quantum analog of a classical channel.

A CPTP map can be defined between two different Hilbert spaces. For example, we can view Eq. (2.13) as a linear map from ρ_B to σ_A. The measurement procedure we described earlier can be viewed as a quantum channel from $\mathbb{H}_S$ to $\mathbb{H}_{A_1}$. This is an example of a quantum-to-classical channel:

$$\mathcal{C}(\rho) = \sum_n |n\rangle \langle n| \,\mathrm{tr}_n (\Pi_n \rho) \tag{2.18}$$

A quantum state can be measured, and if the measurement result is n (with probability $p_n \mathrm{tr}_n (\Pi_n \rho)$), a state $\sigma_n = |n\rangle \langle n|$ is prepared.

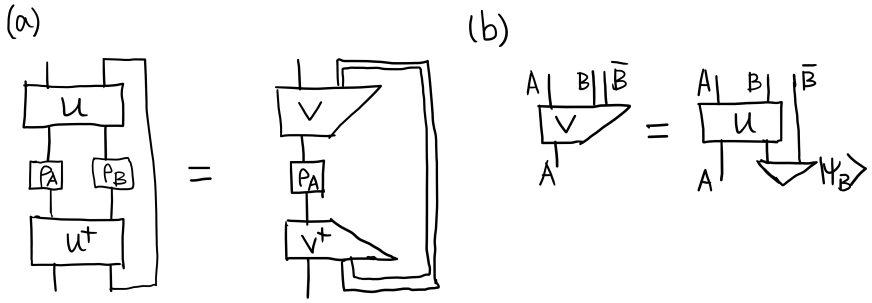

Figure 3. (a) Illustration of a general quantum channel induced from a unitary operator. A unitary acting on $\rho_A \otimes \rho_B$ followed by tracing over B is equivalent to an isometry V from A to $AB\overline{B}$. (b) Explicit definition of V, where $|\Psi_B\rangle$ is a purification of ρ_B.

2.3 Kraus operator and Lindbladian evolution

The relation of quantum channel with isometry in Eq. (2.16) allows us to take a complete basis $|\alpha\rangle$ in E, and define

$$K_\alpha = \langle\alpha| V \tag{2.19}$$

V maps A to AE and α projects the E part into a fixed state, so that K_α is a linear operator in $\mathbb{H}_A$. Eq. (2.16) can be rewritten as

$$\mathcal{C}(\rho_A) = \sum_\alpha K_\alpha \rho_A K_\alpha^\dagger \tag{2.20}$$

This decomposition can be taken for any quantum channel. The operators K_α are called Kraus operators. They satisfy

$$\sum_\alpha K_\alpha^\dagger K_\alpha = V^\dagger V = \mathbb{I}_A \tag{2.21}$$

which is required for preserving the norm of arbitrary input state ρ_A.

For a unitary evolution U, we can consider an infinitesimal evolution $U = 1 - iH\delta t$ with $\delta t \to 0$ and H a Hermitian matrix. Similarly, we can ask what is the infinitesimal form of quantum channel. Since an identity channel corresponds to $K_0 = \mathbb{I}$ as the only nonzero Kraus operator, it is natural to consider a nearly identity channel as corresponding to a K_0 close to identity and other $K_\alpha,\ \alpha \neq 0$ that are close to zero. If we want $\mathcal{C}(\rho_A) = \rho_A + O(\delta t)$, then $K_\alpha,\ \alpha \neq 0$ must be proportional to $\sqrt{\delta t}$. In contrast, $K_0 - \mathbb{I}$ is of order δt since it can have a nontrivial commutator with ρ_A:

$$K_0 = \mathbb{I} + (-iH + R)\delta t \tag{2.22}$$

$$K_\alpha = \sqrt{\delta t} L_\alpha,\ \alpha \neq 0 \tag{2.23}$$

$$\mathcal{C}(\rho_A) = \rho_A + \delta t \left(-i\left[H, \rho_A\right] + \{R, \rho_A\} + \sum_\alpha L_\alpha \rho_A L_\alpha^\dagger \right) + O(\delta t^2) \tag{2.24}$$

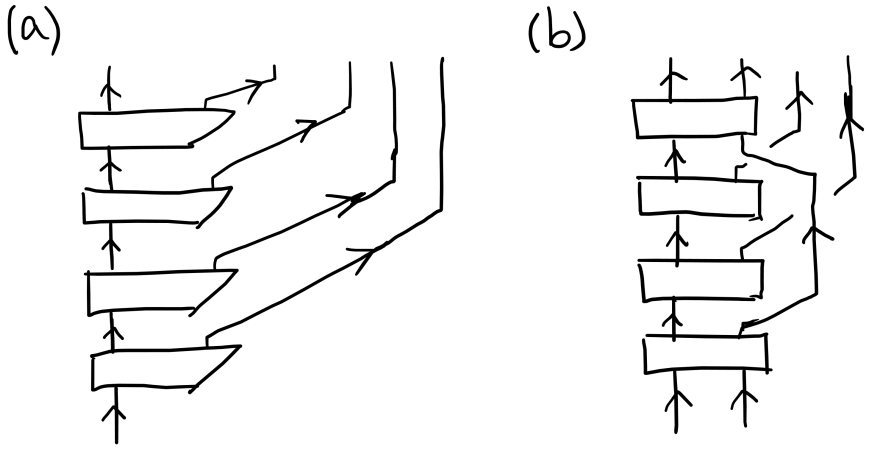

Figure 4. Illustration of a Lindbladian evolution (a) and a more general evolution that couples the system with bath (b).

Requiring the channel to preserve trace leads to the condition

$$\begin{aligned} \mathrm{tr}\left(\rho_A R\right) &= \frac{1}{2}\sum_{\alpha\neq 0}\mathrm{tr}\left(L_\alpha \rho_A L_\alpha^\dagger\right) \\ \Rightarrow R &= -\frac{1}{2}\sum_{\alpha\neq 0} L_\alpha^\dagger L_\alpha \end{aligned} \tag{2.25}$$

Thus we have

$$\mathcal{C}(\rho_A) - \rho_A = \delta t\left(-i\left[H,\rho_A\right] + \sum_{\alpha\neq 0}\left[\left[L_\alpha,\rho_A\right],L_\alpha^\dagger\right]\right) \tag{2.26}$$

or

$$\frac{d\rho_A}{dt} = -i\left[H,\rho_A\right] + \sum_{\alpha\neq 0}\left[\left[L_\alpha,\rho_A\right],L_\alpha^\dagger\right] \tag{2.27}$$

This is the Lindbladian equation, which can be used to describe the evolution of certain open systems. As is illustrated in Fig. 4, Lindbladian equation physically assumes that the system couples with the environment, and after every step, the environment qubit leaves the system and never returns, so that we can trace over it. This is called a Markovian approximation. For example, if the system is an atom and the environment is a photon system, and photons never return the system after being emitted, the evolution of the system can be described well by a Lindbladian evolution. In more general cases, a qubit can enter the environment and later return to the system, leading to a time evolution that is beyond Lindbladian.

As a simple example of Lindbladian evolution, consider $L_\alpha = \sqrt{J}\left|\alpha\right\rangle\left\langle\alpha\right|$ with $\left|\alpha\right\rangle$ an orthonormal basis. $\sqrt{J}$ is an energy scale. For simplicity take the Hamiltonian $H = 0$. Then the Lindbladian equation is

$$\frac{d\rho}{dt} = J\sum_\alpha\left(\left|\alpha\right\rangle\left\langle\alpha\right|\left\langle\alpha\right|\rho\left|\alpha\right\rangle - \rho\right) \tag{2.28}$$

This equation describes an exponential decay of off-diagonal components of ρ:

$$\rho_{\alpha\beta}(t) = \rho_{\alpha\beta}(0)e^{-Jt} \text{ for } \alpha \neq \beta \tag{2.29}$$

$$\rho_{\alpha\alpha}(t) = \rho_{\alpha\alpha}(0) \tag{2.30}$$

This is a simplest model of decoherence, which describes how a quantum state (with all quantum phase information in the off-diagonals of density operator) evolves into a diagonal classical state by coupling to the bath.

2.4 More general measurements

In the same way how we formulate the projective measurement as a coupling with ancilla, we can consider more general unitary coupling with an ancilla. Without losing generality we can consider the ancilla with a pure state initial state (since otherwise we can always purify it), in which case we can define the coupling as an isometry V from the system $\mathbb{H}_S$ to the system and ancilla $\mathbb{H}_S \otimes \mathbb{H}_A$. Then we can carry a projective measurement on the ancilla (which physically requires to introduce another ancilla and do the special kinds of coupling shown in Fig. 2)). The outcome of this measurement is

$$p_n = \mathrm{tr}_{AS}\left[\Pi_{nA}\left(V\rho_S V^\dagger\right)\right] \tag{2.31}$$

In general, this probability is not the result of any direct projective measurement on ρ_S. We can express it in a form

$$p_n = \mathrm{tr}_S\left(\rho_S M_n\right) \tag{2.32}$$

$$M_n = V^\dagger \Pi_{nA} V \tag{2.33}$$

M_n are not projectors, but they are positive, and they satisfy

$$\sum_n M_n = \mathbb{I}_S \tag{2.34}$$

These are called positive operator valued measurement (POVM). Sometimes they are also called weak measurements. As a related concept, the mapping

$$\rho_S \to \Pi_{nA} V \rho_S V^\dagger \Pi_{nA} \equiv \mathcal{M}_n\left(\rho_A\right) \tag{2.35}$$

is a linear map that maps ρ_S to a positive operator but with a generically smaller trace. This is called a completely positive (CP) map. In this example, $\sum_n \mathcal{M}_n$ is a CPTP map.

2.5 Local operation and classical communication

A concept related to measurements is local operation and classical communication (LOCC). This is a set of operations that do not create quantum entanglement. In general, LOCC is defined between multiple parties, but here we will only discuss two-party case with two subsystems A and B. A local operation refers to quantum channels of the form $N_A \otimes \mathcal{N}_B$. A classical communication requires to first measure one subsystem, such as A, and send the classical result to B. After B receive this message, a local operation, i.e. quantum

channel can be applied to B. This operation is called a one-way LOCC $LOCC_{A\to B}$. We can decompose this process into two parts. Sending the information from A is a channel

$$\mathcal{N}_1 : \mathbb{H}_A \to \mathbb{H}_A \otimes \mathbb{H}_C \tag{2.36}$$

$$\mathcal{N}_1(\rho_A) = \sum_\alpha \mathcal{M}_\alpha(\rho_A) \otimes |\alpha\rangle\langle\alpha| \tag{2.37}$$

Here $\mathcal{M}_\alpha$ is a CP map, and $\sum_\alpha \mathcal{M}_\alpha$ is a CPTP map. This channel carries the POVM to A and record the result in C.[1]

The next step is to take the classical information in C and bring it to B. The influence of this message to B corresponds to a quantum channel

$$\mathcal{N}_2 : \mathbb{H}_B \otimes \mathbb{H}_C \to \mathbb{H}_B \tag{2.38}$$

$$\mathcal{N}_2(|\alpha\rangle\langle\alpha| \otimes \rho_B) \equiv \mathcal{N}_\alpha(\rho_B) \tag{2.39}$$

Here $\mathcal{N}_\alpha$ is a channel from B to B defined by restricting the initial state of C to the basis state. Combining the two steps we get the LOCC channel

$$\mathcal{N} = \mathcal{N}_2 \circ \mathcal{N}_1 = \sum_\alpha \mathcal{M}_\alpha \otimes \mathcal{N}_\alpha \tag{2.40}$$

with $\mathcal{M}_\alpha$ acting on A and $\mathcal{N}_\alpha$ acting on B. Any channel with this decomposition is a one-way LOCC.

If there can be two-way classical communication, we can define the class LOCC_r which denotes the channel that combines r rounds of back-and-forth classical communication. LOCC plays an important role in distinguishing classical correlation and quantum entanglement. LOCC cannot create quantum entanglement, since it can be realized by purely classical communication. For a review on LOCC, see [1].

2.6 von Neumann entropy and mutual information

Since each density operator ρ is diagonal in some basis, it is natural to define the Shannon entropy of this diagonal distribution as the entropy of ρ:

$$\rho = \sum_i p_i |i\rangle\langle i| \tag{2.41}$$

$$S(\rho) = -\sum_i p_i \log p_i = -\text{tr}(\rho \log \rho) \tag{2.42}$$

The second expression makes it explicit that the von Neumann entropy is invariant under unitary: $S(\rho) = S\left(U\rho U^\dagger\right)$. In analogy of the classical discussion of typical states, if we take N copies of the same state ρ,

$$\rho^{\otimes N} = \sum_{i_1 i_2 \ldots i_N} p_{i_1} p_{i_2} \ldots p_{i_N} |i_1 i_2 \ldots i_N\rangle\langle i_1 i_2 \ldots i_N| \tag{2.43}$$

[1] More precisely, the POVM operator M_α is $M_\alpha = \mathcal{M}_\alpha^\dagger(\mathbb{I}_A)$.

We can define the operator N_i as the number of $i_k = i$:

$$N_i = \sum_{i_1 i_2 ... i_N} \sum_{a=1}^{N} \delta_{i_a,i} \left| i_1 i_2 ... i_N \right\rangle \left\langle i_1 i_2 ... i_N \right| \tag{2.44}$$

Then we have

$$\mathrm{tr}\left(N_i \rho^{\otimes N}\right) = \sum_{n_i=0}^{N} \binom{N}{n_i} p_i^{n_i} \left(1 - p_i\right)^{N - n_i} n_i \tag{2.45}$$

In the same way as the classical case, one can show that $\rho^{\otimes N}$ is almost an eigenstate of $\frac{N_i}{N}$ with eigenvalue $\frac{N_i}{N} = p_i$. The number of typical states is the subspace defined by

$$\frac{N_i}{N} = p_i + O\left(N^{-1/2}\right), \ \forall i \tag{2.46}$$

which has the dimension $e^{NS(\rho) + O(\log N)}$.

The quantum analog of joint probability distribution is a state ρ living in the direct product Hilbert space $\mathbb{H}_A \otimes \mathbb{H}_B$. The quantum mutual information can be defined as

$$I(A:B) = S\left(\rho_A \otimes \rho_B\right) - S\left(\rho_{AB}\right) = S_A + S_B - S_{AB} \tag{2.47}$$

If there is an orthogonal basis $|n_A\rangle$ of A and $|n_B\rangle$ of B, such that

$$\rho_{AB} = \sum_{n_A, m_B} p_{nm} \left| n_A \right\rangle \left\langle n_A \right| \otimes \left| m_B \right\rangle \left\langle m_B \right| \tag{2.48}$$

then the quantum mutual information is the same as the classical one. However, in general this is not possible. This expression requires that the basis in which ρ_{AB} is diagonal is a product basis between A and B. In general, the eigenstates of ρ_{AB} are not product states.

One major difference with the classical case is that $I(A:B)$ can be bigger than S_A or S_B. For example if AB is in a pure state

$$\left| I_{AB} \right\rangle = \frac{1}{\sqrt{d}} \sum_{i=1}^{d} \left| i_A \right\rangle \left| i_B \right\rangle \tag{2.49}$$

for A, B having the same Hilbert space dimension d. For this state we have

$$S_{AB} = 0, \ S_A = S_B = \log d \Rightarrow I(A:B) = 2 \log d \tag{2.50}$$

If we write $I(A:B)$ in term of conditional entropy, we obtain

$$I(A:B) = S_A - S(A|B), \ S(A|B) \equiv S_{AB} - S_B \tag{2.51}$$

In quantum case, $S(A|B)$ can be negative, while classically the conditional entropy is always non-negative.

2.7 Quantum relative entropy

The classical relative entropy also has a quantum counter part. If we consider two states ρ, σ that are commuting, we can diagonalize them as $\rho = \sum_i q_i |i\rangle \langle i|$, $\sigma = \sum_i p_i |i\rangle \langle i|$, and write the classical relative entropy

$$S(q|p) = \sum_i q_i \log \frac{q_i}{p_i} = \mathrm{tr}(\rho \log \rho - \rho \log \sigma) \tag{2.52}$$

We can generalize this definition to the quantum case when ρ, σ do not commute. In general

$$S(\rho|\sigma) = \mathrm{tr}(\rho \log \rho - \rho \log \sigma) \tag{2.53}$$

Just like the classical case, the relative entropy is non-negative, and monotonous under quantum channels. There are direct proofs of the monotonicity[2] which was based on the strong subadditivity of entropy[3]. Alternatively, we discuss the physical interpretation of relative entropy from the point of view of hypothesis testing. We will see that this interpretation implies monotonicity of the relative entropy. This discussion largely follows[4].

We first discuss the non-negativity of relative entropy. Denote $|n\rangle$ as a basis in which σ is diagonal, such that $\sigma = \sum_n p_n |n\rangle \langle n|$. Then

$$S(\rho|\sigma) = \mathrm{tr}(\rho \log \rho) - \sum_n \langle n| \rho |n\rangle \log p_n \tag{2.54}$$

In the second term, only diagonal terms of ρ appears. We can define a diagonal state

$$\rho_D \equiv \sum_n |n\rangle \langle n| \tilde{q}_n, \ \tilde{q}_n \equiv \langle n| \rho |n\rangle \tag{2.55}$$

then

$$S(\rho|\sigma) = S(\rho_D) - S(\rho) + S(\rho_D|\sigma) \tag{2.56}$$

Since ρ_D and σ commute with each other by definition, $S(\rho_D|\sigma)$ is a classical relative entropy which is non-negative. The remaining task is to show that $S(\rho_D) - S(\rho)$ is also non-negative. To see that, denote ρ in its diagonal basis as $\rho = \sum_m q_m |u_m\rangle \langle u_m|$. Then

$$\tilde{q}_n = \sum_m q_m |\langle n| u_m\rangle|^2 \tag{2.57}$$

$M_{nm} \equiv |\langle n| u_m\rangle|^2$ defines a classical channel, which satisfies $\sum_m M_{nm} = 1$. Thus as we discussed earlier, $S_{\max} - S(\rho)$ is equal to a classical relative entropy which decreases under this classical channel, which means $S(\rho_D) \geq S(\rho)$. Thus we have proven $S(\rho|\sigma) \geq 0$.

Now we discuss the relation of relative entropy and hypothesis testing. As we discussed in the classical case, if we conjecture that there is a classical state σ (which means the probability distribution given by the eigenvalues of σ), but actually the state is ρ_D (in the same sense), then after N measurements we conclude that the probability the state is actually σ is

$$P_N = e^{-NS(\rho_D|\sigma)} \tag{2.58}$$

Note that for ρ_D and σ, if we measure them in a different basis, in the same way as in Eq. (2.57) we see that the basis change corresponds to mapping the two ensembles corresponding to ρ_D and σ by the same classical channel, which can only decrease the relative entropy and make the two states more difficult to distinguish.

Compared with the classical case, in the quantum case when we have N copies of the system, we can carry a joint measurement by applying an operator in the N-copied Hilbert space. This is a new possibility that does not exist in the classical case. If the states commute with each other, so is $\rho_D^{\otimes N}$ and $\sigma^{\otimes N}$, such that the classical calculation above still apply. However, it should be noticed that $\sigma^{\otimes N}$ has a lot of degeneracies. For example if σ is 2×2, all eigenstates $|n_1 n_2 ... n_N\rangle$ with the same number of 0 and 1 corresponds to the same eigenvalue $p_0^k(1-p_0)^{N-k}$. Therefore we have a lot of freedom in choosing a diagonal basis for $\sigma^{\otimes N}$. Denote a basis choice as Π_I with I labeling the states in the N-copied Hilbert space $\mathbb{H}^{\otimes N}$, then for each choice of Π_I that makes σ diagonal, one can define

$$\rho_D^{(N)} = \sum_I \Pi_I \rho^{\otimes N} \Pi_I \tag{2.59}$$

Note that Π_I does not have to be a direct product of each copy, so that $\rho_D^{(N)}$ is not necessarily the product of N density operators. Naively, $\rho_D^{(N)}$ could have a very different entropy from $\rho^{\otimes N}$, but it turns out that we can reduce this difference by noticing the symmetry of $\rho^{\otimes N}$ and $\sigma^{\otimes N}$. In the N-copied Hilbert space we can consider the symmetry action of permutation group S^N and unitary group $U(d)$ (with d the Hilbert space dimension of each copy). The permutation acts naturally by permuting different replicas, and $U(d)$ refers to the same unitary operator acting on each replica. Thus in this action S^N commutes with $U(d)$ and the group is $S^N \times U(d)$. According to the Schur-Weyl duality, the Hilbert space $\mathbb{H}^{\otimes N}$ factorizes into a direct sum of different representations:

$$\mathbb{H}^{\otimes N} = \oplus_Y \mathbb{H}_Y^S \otimes \mathbb{H}_Y^U \tag{2.60}$$

Here the sum is over Young tableaux, which label the representation of S^N and $U(d)$. Since $\rho^{\otimes N}$ and $\sigma^{\otimes N}$ commute with all permutations, if we expand them in this decomposition of Hilbert space, they looks like

$$\sigma^{\otimes N} = \sum_Y \mathbb{I}_Y^S \otimes p_Y \sigma_Y \tag{2.61}$$

with $\sigma_Y \in \mathbb{H}_Y^U$. We have defined σ_Y to be normalized, so that there is a coefficient $p_Y \in [0,1]$ satisfying $\sum_Y p_Y = 1$. Similarly

$$\rho^{\otimes N} = \sum_Y \mathbb{I}_Y^S \otimes q_Y \rho_Y \tag{2.62}$$

This decomposition tells us a strong restriction on how non-commuting these two operators could be. Since both of them are block-diagonal already in this form, when we choose a diagonal basis of $\sigma^{\otimes N}$ we only need to apply projection to $\rho^{\otimes N}$ within each subspace. One can prove

$$S\left(\rho^{\otimes N}\middle|\sigma^{\otimes N}\right) = \sum_Y q_Y S\left(\rho_Y|\sigma_Y\right) + S\left(\{q_Y\}|\{p_Y\}\right) \tag{2.63}$$

The second term is the classical relative entropy between q_Y and p_Y probability distributions. Now we can choose a diagonal basis of σ_Y and take the diagonal elements of ρ_Y, denoted as ρ_{DY}. This will leave the second term invariant and decrease the first term.

$$S\left(\rho^{\otimes N}\middle|\sigma^{\otimes N}\right) = \sum_Y q_Y \left(S\left(\rho_{DY}\right) - S\left(\rho_Y\right)\right) + S\left(\rho_D^{(N)}\middle|\sigma^{\otimes N}\right) \tag{2.64}$$

$$S\left(\rho_D^{(N)}\middle|\sigma^{\otimes N}\right) = \sum_Y q_Y S\left(\rho_{DY}|\sigma_Y\right) + S\left(\{q_Y\}|\{p_Y\}\right) \tag{2.65}$$

Now the key point is that the entropy difference $S\left(\rho_{DY}\right) - S\left(\rho_Y\right)$ is bounded by the maximal entropy of this subspace:

$$\begin{aligned} & S\left(\rho_{DY}\right) - S\left(\rho_Y\right) \leq \log D_Y \\ & \Rightarrow S\left(\rho^{\otimes N}\middle|\sigma^{\otimes N}\right) \leq \max_Y \log D_Y + S\left(\rho_D^{(N)}\middle|\sigma^{\otimes N}\right) \end{aligned} \tag{2.66}$$

The maximal dimension D_Y scales polynomially with N:

$$D_Y \leq (N+1)^{d(d-1)/2} \tag{2.67}$$

For example for $d = 2$, the maximal dimensional representation has $SU(2)$ spin $N/2$, with dimension $N + 1$. Consequently we obtain

$$S\left(\rho_D^{(N)}\middle|\sigma^{\otimes N}\right) \leq S\left(\rho^{\otimes N}\middle|\sigma^{\otimes N}\right) \leq \frac{d(d-1)}{2}\log\left(N+1\right) + S\left(\rho_D^{(N)}\middle|\sigma^{\otimes N}\right) \tag{2.68}$$

$$\Rightarrow S\left(\rho|\sigma\right) = \lim_{N\to+\infty} \frac{1}{N} S\left(\rho_D^{(N)}\middle|\sigma^{\otimes N}\right) \tag{2.69}$$

The right-hand side controls the optimal distinguishability of the two states when we carry measurements on N copies. Thus we conclude that for large N, the probability of measured state ρ be mistaken as σ is at least

$$P_N = e^{-NS(\rho|\sigma)} \tag{2.70}$$

This result gives an operational definition of the relative entropy, and also implies its monotonicity under quantum channels. According to the dilation theorem, a quantum channel $\mathcal{N}$ can be viewed as an isometry followed by partial trace. Denote the isometry as $V : \mathbb{H}_A \to \mathbb{H}_A \otimes \mathbb{H}_E$, one can prove

$$S\left(V\rho V^\dagger\middle|V\sigma V^\dagger\right) = S\left(\rho|\sigma\right) \tag{2.71}$$

Now if we carry measurements on $\left(V\rho V^\dagger\right)^{\otimes N}$ and $\left(V\sigma V^\dagger\right)^{\otimes N}$, but restrict the measurements to A, then the probability P_N will only go up (since it is more difficult to distinguish the two states). Restricting the operators to (N-copies of) A is equivalent of computing the relative entropy between the reduced states $\mathrm{tr}_E\left(V\rho V^\dagger\right) = \mathcal{N}\left(\rho\right)$ and $\mathcal{N}\left(\sigma\right)$. Thus we obtain

$$S\left(\mathcal{N}\left(\rho\right)|\mathcal{N}\left(\sigma\right)\right) \leq S\left(\rho|\sigma\right) \tag{2.72}$$

for any quantum channel $\mathcal{N}$.

In the same way as the classical case, the monotonicity of quantum relative entropy implies the monotonicity of mutual information:

$$I(A:B) = S\left(\rho_{AB}|\rho_A \otimes \rho_B\right) \tag{2.73}$$

$$I(A:B)\left(\mathcal{N}_A \otimes \mathcal{N}_B\left(\rho_{AB}\right)\right) \leq I(A:B)\left(\rho_{AB}\right) \tag{2.74}$$

In particular, it implies the strong subadditivity of entropy:

$$\begin{aligned} &I(A:BC) \geq I(A:B) \\ \Rightarrow\ &S_{BC} + S_{AB} \geq S_{ABC} + S_B \end{aligned} \tag{2.75}$$

If we consider N copies of ρ_{AB} and carry joint measurements by applying POVM $M_{A\alpha} \in \mathbb{H}_A^{\otimes N}$ and $M_{B\beta} \in \mathbb{H}_B^{\otimes N}$[2], the measurement leads to the classical joint probability distribution $P_{\alpha\beta}$, which has a classical mutual information $I_c^{(N)}(A:B)$. The monotonicity implies $I_c^{(N)}(A:B) \leq NI(A:B)$. Note that even if we optimize over all POVM, $\frac{1}{N}I_c^{(N)}(A:B)$ may not reach $I(A:B)$. The classical mutual information always satisfy $\frac{1}{N}I_c^{(N)}(A:B) \leq S_A$ (and also S_B) but $I(A:B) > S_A$ is possible. For example for an EPR pair state $I(A:B) = 2\log 2$ while the classical MI is upper bounded by $\log 2$. On the other hand, the discussion above suggests that the relative entropy $I(A:B) = \frac{1}{2}S\left(\rho_{AB}^{\otimes N}\middle|\rho_A^{\otimes N} \otimes \rho_B^{\otimes N}\right)$ is close to classical relative entropy. This suggests that the optimal measurement (the one in the diagonal basis of $\rho_A^{\otimes N} \otimes \rho_B^{\otimes N}$ as we discussed above) cannot factorize into measurements in $\mathbb{H}_A^{\otimes N}$ and $\mathbb{H}_B^{\otimes B}$. In other words, the optimal measurements must be in entangled basis if $I(A:B) > S_A$ or $I(A:B) > S_B$. This is a consequence of the intrinsic nonlocality of quantum mechanics.

We can also obtain the joint convexity of relative entropy in the same way as the classical case:

$$S\left(p\rho_1 + (1-p)\rho_2|p\sigma_1 + (1-p)\sigma_2\right) \leq pS\left(\rho_1|\sigma_1\right) + (1-p)S\left(\rho_2|\sigma_2\right) \tag{2.76}$$

In the quantum language we can introduce an ancilla and construct the state $\pi = p\rho_1 \otimes |1\rangle\langle 1| + (1-p)\rho_2 \otimes |2\rangle\langle 2|$ and $\eta = p\sigma_1 \otimes |1\rangle\langle 1| + (1-p)\sigma_2 \otimes |2\rangle\langle 2|$. The convexity follows from the monotonicity upon tracing over the ancilla.

2.8 Holevo information

When we have a quantum channel, one natural question is how much classical information can be transmitted through this quantum channel. For classical information in the probability distribution $p(x_i) = p_i$, we can define states $|i\rangle\langle i|$ and encode this classical information in a diagonal density operator $\rho = \sum_i p_i |i\rangle\langle i|$. A quantum channel brings this state to

$$\sigma = \mathcal{C}\left(\rho\right) = \sum_i p_i \mathcal{C}\left(|i\rangle\langle i|\right) \equiv \sum_i p_i \sigma_i \tag{2.77}$$

[2]Here I mean the operator acts in this Hilbert space. More precisely we should say $O_A \in \mathbb{H}_A^{\otimes N} \otimes \mathbb{H}_A^{*\otimes N}$.

How much information is transferred depends on σ_i. For example, σ_i could be all the same, in which case there is no information transferred. To measure the amount of classical information transferred, one need to apply a POVM and obtain

$$P(\alpha|i) = \mathrm{tr}\left(M_\alpha \sigma_i\right) \tag{2.78}$$

Just like in the classical channel capacity discussion, the information transferred through the channel can be measured by the classical mutual information

$$I_{\mathcal{C}} = \sup_{M_\alpha} \sum_{p_i} I[P(\alpha|i)p_i] \tag{2.79}$$

We can define an auxiliary state

$$\pi = \sum_i p_i \left|i\right\rangle \left\langle i\right| \otimes \sigma_i \tag{2.80}$$

where the $\left|i\right\rangle \left\langle i\right|$ lives in an additional copy of the input state. This is obtained by copying the input classical information before sending it to $\mathcal{C}$. The probability distribution $P(\alpha|i)p_i$ corresponds to the diagonal state

$$\begin{aligned} \eta &= \sum_i p_i \left|i\right\rangle \left\langle i\right| \otimes \mathcal{M}\left(\sigma_i\right) \\ &= \sum_i p_i \mathrm{tr}\left(M_\alpha \sigma_i\right) \left|i\right\rangle \left\langle i\right| \otimes \left|\alpha\right\rangle \left\langle \alpha\right| \end{aligned} \tag{2.81}$$

Thus the mutual information in η is upper bounded by that in π:

$$I\left[P(\alpha|i)p_i\right] = I_\eta \leq I_\pi \tag{2.82}$$

One can show that

$$I_\pi = S\left(\sum_i p_i \sigma_i\right) - \sum_i p_i S\left(\sigma_i\right) \equiv \chi \tag{2.83}$$

This is known as the Holevo χ quantity, which provides an upper bound of the classical information that can be transmitted.

As a special case, if $\sigma_i = \left|\psi_i\right\rangle \left\langle \psi_i\right|$ is a pure state, we have $\chi = S\left(\sum_i p_i \left|\psi_i\right\rangle \left\langle \psi_i\right|\right)$. From a different point view, this means that for any mixed state that is an ensemble of pure states $\sigma = \sum_i p_i \left|\psi_i\right\rangle \left\langle \psi_i\right|$, there is a way to encode classical information of the amount $S(\sigma)$ if we can pick the pure state in the ensemble.

Similar to the discussion about classical channel capacity, it is natural to ask what happens when we use a quantum channel N times in parallel. Denote the input Hilbert space of each channel as $\mathbb{H}_{\text{in}}$, In general we can choose a set of basis states $\left|I\right\rangle$ in $\mathbb{H}_{\text{in}}^{\otimes N}$ and a probability distribugtion p_I. Then the Holevo information is

$$\chi^{(N)} = S\left(\sum_I p_I \mathcal{C}^{\otimes N}\left(\left|I\right\rangle \left\langle I\right|\right)\right) - \sum_I p_I S\left(\mathcal{C}^{\otimes N}\left(\left|I\right\rangle \left\langle I\right|\right)\right) \tag{2.84}$$

In general, $\frac{1}{N}\chi^{(N)}$ could increase as a function of N. However, if we assume that $|I\rangle$ are all product states of the N copies, *i.e.*

$$|I\rangle = \prod_{a=1}^{N} |i_a\rangle \tag{2.85}$$

Then we have

$$\chi^{(N)} = S\left(\sum_I p_I \otimes_{a=1}^N \sigma_{i_a}\right) - \sum_I p_I \left(\sum_{a=1}^N S\left(\sigma_{i_a}\right)\right) \tag{2.86}$$

Here $\sigma_{i_a} = \mathcal{C}\left(|i_a\rangle\langle i_a|\right)$. Denote the reduced probability distribution of $p_I = p_{i_1 i_2 \ldots i_N}$ as p_{i_a}, the reduced state of $\sum_I p_I \otimes_{a=1}^N \sigma_a$ on a-th Hilbert space is $\sum_{i_a} p_{i_a}\sigma_{i_a}$. Thus due to strong subadditivity we have

$$S\left(\sum_I p_I \otimes_{a=1}^N \sigma_{i_a}\right) \leq \sum_{a=1}^N S\left(\sum_{i_a} p_{i_a} S\left(\sigma_{i_a}\right)\right) \tag{2.87}$$

This implies that

$$\chi^{(N)} \leq \sum_{a=1}^N \chi_a \tag{2.88}$$

Thus for product state, the Holevo information is subadditive. In other words, for transmitting information, it is not useful to use classically correlated code words (but it is useful to use quantum entangled code word states $|I\rangle$).

2.9 Quantum channel capacity

In analogy with the classical channel capacity, we are interested in the amount of quantum information that can be transmitted through a quantum channel. In the classical case, the channel capacity is C if we can (faithfully) transmit C bits of classical information per each use of the channel. In classical case, the quantum channel capacity is defined as the maximal number of qubits that can be transmitted errorlessly per each use of the channel. More precisely, consider an arbitrary M qubit state $|\psi\rangle \in \mathbb{H}_I$. For a channel $\mathcal{C} : \mathbb{H}_A \to \mathbb{H}_B$, the quantum information in $|\psi\rangle$ can be transmitted accurately with N copies of the channel, if there exists an encoding map $\mathcal{E} : \mathbb{H}_I \to \mathbb{H}_A^{\otimes N}$ and a decoding map $\mathcal{D} : \mathbb{H}_B^{\otimes N} \to \mathbb{H}_I$, such that

$$\mathcal{D} \circ \mathcal{C}^{\otimes N} \circ \mathcal{E}(|\psi\rangle) = |\psi\rangle \tag{2.89}$$

Since we require this to be true for arbitrary $|\psi\rangle$, this is equivalent to the requirement

$$\mathcal{D} \circ \mathcal{C}^{\otimes N} \circ \mathcal{E} = \mathbb{I} \tag{2.90}$$

The quantum channel capacity is the upper limit of $\frac{M}{N}$:

$$Q\left(\mathcal{C}\right) = \lim_{N\to\infty} \sup_{\mathcal{D},\mathcal{E}} \frac{M}{N} \log 2 \tag{2.91}$$

This definition is illustrated in Fig. 5 (a). (There shouldn't be $\log 2$ if we are measuring the channal capacity in term of qubits, but we add it because it is more convenient for relating this quantity to von Neumann entropy, since we have been using natural log in the definition of the latter.)

We would want to have a more explicit formula, ideally N independent, for quantum channel capacity. This is an open question. In the following we will discuss some more explicit formula, although it is still N-dependent. First, if we apply the channel $\mathcal{D}\circ\mathcal{C}^{\otimes N}\circ\mathcal{E}$ to a maximally entangled state between I and an auxiliary system R, the identity condition translates to the statement that after applying $\mathcal{D}\circ\mathcal{C}^{\otimes N}\circ\mathcal{E}$ the state of IR is still maximally entangled. (It's ok if it's a different maximally entangled state, since that just means the quantum channel is applying a unitary to I, which can be easily absorbed by redefinition of $\mathcal{D}$.) The encoding map is an isometry which preserves information. The key question is whether $\mathcal{C}^{\otimes N}$ also preserves the quantum information.

To obtain a more explicit formula for quantum channel capacity, we consider the setup in Fig. 5 (c). The dilation of channel $\mathcal{C}$, denoted as $V_{\mathcal{C}}$, is applied to a state $|\psi_{RA}\rangle$. Really we should be discussing N copies of A, B and $\mathcal{C}$ but we leave out the N for simplicity. We will discuss the N copy case later. The state $\mathcal{E}|IR\rangle$ in (b) is a special case of $|\psi_{RA}\rangle$. The requirement of information preservation is that one can find a decoding channel $\mathcal{D}$ applying to B which maps the state of RBE to $\rho_{R\tilde{A}E} = |\psi_{RA}\rangle\langle\psi_{RA}|\otimes\rho_E$. Now remember that the monotonicity of mutual information requires

$$I(R:A) \geq I(R:B) \geq I(R:\tilde{A}) \tag{2.92}$$

so if we want $I(R:A) = I(R:\tilde{A})$, the necessary condition is $I(R:A) = I(R:B)$. Conversely, if $I(R:A) = I(R:B)$, it implies that

$$S(A) = S(B) - S(RB) = S(B) - S(E) \tag{2.93}$$

where we have used the fact that RA together is a purestate. Thus

$$S(R) = S(A) = S(RE) - S(E) \Rightarrow I(R:E) = 0 \tag{2.94}$$

This equation says that if the mutual information did not decrease when A is mapped to B, then the environment E does not know anything about R. The quantity on the right-hand side can be written as

$$S(A) = S(B) - S(RB) = -S(R|B) \tag{2.95}$$

with $S(R|B)$ the conditional entropy. In classical case the conditional entropy is always positive, but in quantum case it can be negative. It actually must be negative in order for the equation above to hold. The quantity $-S(R|B)$ is important below, which is named as coherent information, denoted as

$$I_c(A:B)_{\rho_A} = S(\mathcal{C}(\rho_A)) - S\left(\mathcal{C}\left(|\psi_{RA}\rangle\langle\psi_{RA}|\right)\right) \tag{2.96}$$

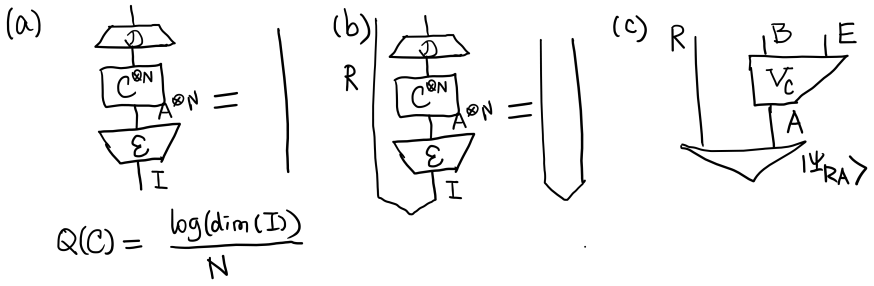

Figure 5. (a) Definition of the quantum channel capacity. (b) For an input state in A that is purified to $|\psi_{RA}\rangle$, we consider the dilation of the channel $\mathcal{C}$ to an isometry from A to BE.

It should be noted that the coherent information only depends on ρ_A, since different $|\psi_{RA}\rangle$ that corresponds to the same ρ_A can always be related by a unitary acting on R, which does not change $S(B)$ or $S(RB)$.

Because $I(R:E) = 0$ we have $\rho_{RE} = \rho_R \otimes \rho_E$. A state like this can be purified by separately purifying ρ_R and ρ_E. We denote the purified state as

$$|\psi_{REB_1B_2}\rangle = |\psi_{RB_1}\rangle \otimes |\psi_{B_2E}\rangle \tag{2.97}$$

where B_1 and B_2 are two independent systems. Since all different purifications of a state are related by unitaries, we could simply take B_1 to have the same size as A and take $|\psi_{RB_1}\rangle = |\psi_A\rangle$. For the same reason, there exists a unitary U_B acting on B such that

$$|\psi_{RB_1}\rangle \otimes |\psi_{EB_2}\rangle = U_B |\psi_{RB}\rangle \tag{2.98}$$

Thus we have proven that there exists a channel $\mathcal{D}$, obtained by applying U_B and tracing over B_2, which recovers the state $|\psi_{RA}\rangle$.

The discussion above can be easily generalized to the case of N copies. For a general state of RA^N, the condition of recovery is

$$\begin{aligned} &I\left(R:A^N\right) = I\left(R:B^N\right) \\ &\text{or } S(A^N) = I_c\left(A^N:B^N\right)_{\rho_{A^N}} \end{aligned} \tag{2.99}$$

Now if we consider the case in Fig. 5 (b), $S(A) = S(I) = M \log 2$ is the number of qubits. Thus

$$Q\left(\mathcal{C}\right) = \sup_{\mathcal{E}} \frac{S(I)}{N} = \sup_{|\psi_{\mathcal{E}}\rangle} \frac{1}{N} I_c\left(A^N:B^N\right)_{\rho_{A^N}} \tag{2.100}$$

The supreme is taken for all states of the form $\mathcal{E}|IR\rangle$. Thus we have an upper bound if we replace it by a supreme over all possible states:

$$Q\left(\mathcal{C}\right) \leq \lim_{N\to\infty} \sup_{\rho_{A^N}} \frac{1}{N} I_c\left(A^N:B^N\right)_{\rho_{A^N}} \tag{2.101}$$

This upper bound can actually be achieved. Let's assume that the state $|\psi_{RA^N}\rangle$ is one that maximizes I_c in the N copy case. Then we can take another large number M and look at

NM copies with the state $|\psi_{RA^N}\rangle^{\otimes M}$. In large M, such a state can be approximated by a different state $\left|\tilde{\psi}_{\tilde{R}A^{NM}}\right\rangle$, in which $\rho_{\tilde{R}}$ is maximally mixed, with a Hilbert space dimension $\dim_{\tilde{R}} \simeq e^{MS_R}$. Thus this state has the form of $|\psi_{\mathcal{E}}\rangle = \mathcal{E}\,|IR\rangle$. Of course, to be more precise we need to quantify how close are these two states, and prove that the small deviation corresponds to a small change in the coherent information. We won't go into these details here.

2.10 Quantum teleportation

2.11 Quantum error correction

References

[1] E. Chitambar, D. Leung, L. Mančinska, M. Ozols, and A. Winter, *Everything you always wanted to know about locc (but were afraid to ask)*, *Communications in Mathematical Physics* **328** (2014), no. 1 303–326.

[2] G. Lindblad, *Completely positive maps and entropy inequalities*, *Communications in Mathematical Physics* **40** (1975), no. 2 147–151.

[3] E. H. Lieb and M. B. Ruskai, *Proof of the strong subadditivity of quantum-mechanical entropy*, *Les rencontres physiciens-mathématiciens de Strasbourg-RCP25* **19** (1973) 36–55.

[4] E. Witten, *A mini-introduction to information theory*, *La Rivista del Nuovo Cimento* **43** (2020), no. 4 187–227.