How to exploit Bell non-locality to advance communication technologies?

Lecture 4: Entropic notions

Nicolas Sangouard
Institut de Physique Theorique, CEA Paris Saclay
October 24/25, 2022

1 Introduction

We introduce basic entropic notions which are required when one wants to quantify uncertainty in quantum scenarios. These notions are extension of entropic notions in classical information theory, which has been introduced by Claude Shannon around 1940. By the way, information theory is sometimes called Shannon theory.

2 Shannon entropy

Consider a source of information emitting a message (also called sequence) in the form of a string of n random bits. We consider the case where each bit is independent and identically distributed. In particular, the bit 0 appears with probability p while the bit 1 appears with probability $1-p$. We assume $p \geq \frac{1}{2}$ without loss of generality. What is the expected surprisal per bit? How long a compressed string that conveys essentially the same information needs to be? What is the probability to guess in advance the message produced by the source?

Let us first consider the first question. A reasonable function to quantify surprisal, i.e. how surprised we are when we see the bit value 1 is $S(1) = -\log_2(1-p)$. If the outcome 1 has a low probability then we are very surprised when we see it happens and indeed, $S(1) \to \infty$ when $1-p \to 0$. Note that the surprisal is also additive for independent variables, i.e. if we observe two bit values, the total surprisal is the sum of individual surprisal. The expected surprisal per bit is simply $-p\log_2 p - (1-p)\log_2(1-p)$ which defines the entropy function. This function equals 1 for $p = 1/2$ and 0 for $p = 1$. In other words, the entropy tells us on average how surprised we are

to see the bit value.

Now, let us focus on the second question : Is it possible to compress the message to a shorter string that conveys essentially the same information? We are here considering Bernoulli random variables X_i that can take two possible values 0 and 1 with probability p and $1-p$ respectively. We then consider a collection of n such a variable $(X_1, \ldots, X_n)$ that are identical and independently distributed. In total, these random variables can take 2^n different values – 2 for each variables, but we want to separate all the sequences into two sets: the typical set (sequences that typically occur) and the untypical set (sequences that almost never occur). To do so, note that the probability of a given sequence with k ones first and then $n-k$ zeroes, e.g. the probability to get 1 for the first k values and and 0 for the remaining $n-k$ values $(\underbrace{1, \ldots, 1}_{k}, \underbrace{0, \ldots, 0}_{n-k})$ equals

$$p[(\underbrace{1, \ldots, 1}_{k}, \underbrace{0, \ldots, 0}_{n-k})] = p^{n-k}(1-p)^k. \tag{1}$$

The total probability that some sequence with k ones occurs (the probability to get k ones and $N-k$ zeroes in any order) is given by the binomial distribution

$$P(\#\text{ ones} = k) = \binom{n}{k} p^{n-k}(1-p)^k, \tag{2}$$

where $\binom{n}{k}$ counts all the different permutations. We plot an example of the binomial distribution in Fig. 1. One can see from the plot that the highest probabilities are centered in a small region around a number of ones equals to $(1-p)n$ (for large enough n a binomial distribution is well approximated by a Gaussian with mean $\mu = (1-p)n$ and variance $\sigma^2 = np(1-p)$). From this observation, a notion of typical set arises naturally: The typical set is made with all the sequences $(x_1, \ldots, x_n)$ having a total number of ones $\sum_i x_i$ inside a window centered at $(1-p)n$ of width $2\delta n$, that is

$$A^{(n)} = \{(x_1, \ldots, x_n), \text{such that} \, | \sum_i x_i - n(1-p)| \leq \delta n\} \tag{3}$$

The Hoeffding's bound ensures that the probability that the observed sequence falls in the typical set

$$P\Big((x_1, \ldots, x_n) \in A^{(n)}\Big) \geq 1 - e^{-2\delta^2 n}, \tag{4}$$

that is, it tends to 1 for large n. But what is the probability that a particular sequence from the typical set is realized? Obviously the extremal cases lay on

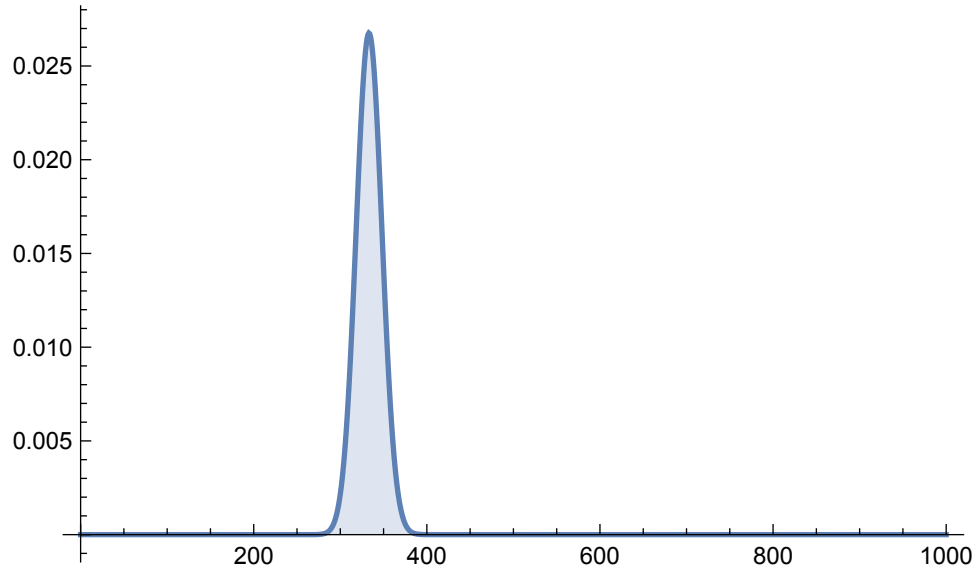

Figure 1: The binomial distribution $P(\#\text{ ones} = k)$ for $n = 1000$ and $p = 2/3$.

the boundary of the typical set, that is for $k = (1-p+\delta)n$ and $k = (1-p-\delta)n$ ($p^{n-k}(1-p)^k$ decreases as a function of k for $p \geq 1/2$). We have[1]

$$\begin{aligned}
(1-p)^{n(1-p+\delta)} p^{n(p-\delta)} \leq & p[(x_1, \ldots, x_N)_{\in A^{(n)}}] \leq (1-p)^{n(1-p-\delta)} p^{n(p+\delta)} \\
((1-p)^{(1-p)} p^p)^n \left(\frac{1-p}{p}\right)^{\delta n} \leq & p[(x_1, \ldots, x_N)_{\in A^{(n)}}] \\
& \leq ((1-p)^{(1-p)} p^p)^n \left(\frac{1-p}{p}\right)^{-\delta n} \\
2^{-(H(p)+\varepsilon)n} \leq & p[(x_1, \ldots, x_N)_{\in A^{(n)}}] \leq 2^{-(H(p)-\varepsilon)n}
\end{aligned} \tag{7}$$

with $\varepsilon = \delta \log_2 \left(\frac{p}{1-p}\right)$. The overall probability to end up in a typical sequence

[1]Here we used

$$\log_2 \left((1-p)^{(1-p)} p^p)^n\right) = n\left((1-p)\log_2(1-p) + p\log_2(p)\right) = -nH(p) \tag{5}$$

and since $\log_2 x = a \to x = 2^a$, we have

$$\left((1-p)^{(1-p)} p^p)^n\right) = 2^{-nH(p)}. \tag{6}$$

can be written as a function of ϵ as

$$P\Big((x_1,\ldots,x_n)\in A^{(n)}\Big)\geq 1-\exp\left(-2n\frac{\varepsilon^2}{\log_2^2\left(\frac{p}{1-p}\right)}\right)=1-e^{-\xi n\varepsilon^2}. \quad (8)$$

Finally, we compute the size of the typical set

$$\begin{aligned}
|A^{(n)}| &= \sum_{(x_1,\ldots,x_n)\in A^{(n)}} 1 = \sum_{(x_1,\ldots,x_n)\in A^{(n)}} \frac{p[(x_1,\ldots,x_n)]}{p[(x_1,\ldots,x_n)]} \\
&\geq \sum_{(x_1,\ldots,x_n)\in A^{(n)}} \frac{p[(x_1,\ldots,x_n)]}{2^{-(H_{\rm bin}(p)-\varepsilon)n}} \\
&\geq \frac{1-e^{-\xi n\epsilon^2}}{2^{-(H_{\rm bin}(p)-\varepsilon)n}} = 2^{H_{\rm bin}(p)n}\,2^{-\epsilon n}(1-e^{-\xi n\epsilon^2}) \\
|A^{(n)}| &= \sum_{(x_1,\ldots,x_n)\in A^{(n)}} \frac{p[(x_1,\ldots,x_n)]}{p[(x_1,\ldots,x_n)]} \leq \frac{1}{2^{-(H_{\rm bin}(p)+\varepsilon)n}} = 2^{H_{\rm bin}(p)n}\,2^{\epsilon n}.
\end{aligned} \quad (9)$$

This means that to convey all the information carried by a long-n string, it is essentially sufficient to choose a block code that assigns a non-negative integer to each of the typical strings. These blocks need to distinguish about $2^{nH(p)}$ messages. We can specify any one of the message using a binary string with length $\approx nH(p) < n\ \forall p\neq 1/2$. If the source produces a sequence which is not in the typical set, the procedure fails, which happens with a vanishing probability. This result is called Shannon's source coding theorem.

Let us now focus on the third question. Consider the case with post-selection : If the source produces a sequence which is untypical (which happens with an exponentially small probability), the protocol aborts. If the sequence falls inside the typical set, the sequence is accepted. In this case, the highest guessing probability is simply obtained by betting on a sequence taking randomly from the typical set, as in the typical set, the sequences are essentially distributed uniformly. The probability to correctly guess the sequence is then given by $\approx 2^{-H(p)n}$. In case there is no post-selection, the same guess seems reasonable as the probability for the sequence to be untypical is vanishing ($\approx e^{-\xi n\varepsilon^2}$). However, the untypical sequence with only zeros provides the highest guessing probability p^n. This leads to the notion of min entropy $H_{\rm min}$ such that

$$\max_{(x_1,\ldots,x_n)} p[(x_1,\ldots,x_n)] = 2^{-H_{\rm min}(p)n}. \quad (10)$$

Obviously

$$H(p) \geq H_{\min}(p). \tag{11}$$

Note that we will mainly focus on Shannon entropy and forget about other entropic notions like min entropy in this course. This is not always perfectly correct, as we see for the guessing probability, but will be easier to convey the main ideas that are exploited in quantum information theory.

The notion of entropy applies much more generally. Consider for example an ensemble X^n of n-letter message $\vec{x} = x_1, ..., x_n$ in which each letter is generated independently by sampling from a probability distribution $X = \{x, p(x)\}$ where $x = 0, ..., k-1$ and $p(x)$ is the probability of the letter x. One can define a typical set of sequences $A_\varepsilon^{(n)}$ of size $|A_\varepsilon^{(n)}| \approx 2^{H(X)n}$ with a uniform distribution of sequences $p[(x_1, \ldots, x_n)_{\in A^{(n)}}] \approx 2^{-H(X)n}$ for which the probability to find the sequence $\vec{x}$ in the set is $P((x_1, \ldots, x_n) \in A_\varepsilon^{(n)}) \approx 1$.

In summary, given the random variable $X = \{x, p(x)\}$, the following claims hold

- H(X) can be seen as the expected surprisal to see the value taken by X by sampling from the probability distribution $p(x)$
- nH(X) is the minimum number of bits needed to compress a n-long string obtained by sampling n times X which can take the value x with the probability $p(x)$
- $2^{-nH(X)}$ is the highest probability to correctly guess a typical n-long string obtained by sampling n times X from $p(x)$

3 Conditional entropy and mutual information

We consider the information source described before which samples from X. We have seen that the typical sequence $\vec{x}$ appears with probability $p(\vec{x})$ which is given by

$$p(\vec{x}) \approx 2^{-nH(X)} \tag{12}$$

where $H(X)$ is the expected surprisal about each letter. We can now consider two similar information sources sampling from

$$XY = \{(x, y), p(x, y)\}, \tag{13}$$

i.e. the pair (x, y) appears with probability $p(x, y)$. We use X to denote the marginal distribution defined as

$$X = \left\{ x, p(x) = \sum_y p(x, y) \right\} \tag{14}$$

and similarly for Y. If X and Y are correlated, then by reading a message generated from Y^n, I reduce my ignorance about a message generated from X^n. This should make it possible to compress the output of X further than if I did not have access to Y. We have $p(\vec{x}) \approx 2^{-nH(X)}$, $p(\vec{y}) \approx 2^{-nH(Y)}$ and $p(\vec{x}, \vec{y}) \approx 2^{-nH(XY)}$. Consequently, from the Bayes' law, we deduce

$$p(\vec{x}|\vec{y}) = \frac{p(\vec{x}, \vec{y})}{p(\vec{y})} \approx 2^{-nH(X|Y)} \tag{15}$$

where $H(X|Y) = H(XY) - H(Y) = -\sum_{x,y} p(x, y) \log_2 p(x|y)$ is the conditional entropy of X given Y. $H(X|Y)$ quantifies my remaining ignorance per letter about $\vec{x}$ once I know $\vec{y}$. This means that the number of possible values for $\vec{x}$ compatible with the known value of $\vec{y}$ is approximately $2^{nH(X|Y)}$. Hence, $H(X|Y)$ is the number of additional bits per letter needed to specify both $\vec{x}$ and $\vec{y}$ once $\vec{y}$ is known.

The information about $\vec{x}$ that I gain when I learn $\vec{y}$ is quantified by how much the number of bits per letter needed to specify $\vec{x}$ is reduced when $\vec{y}$ is known. Thus is

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) && (16) \\ &= H(X) + H(Y) - H(XY) && (17) \\ &= H(Y) - H(Y|X). && (18) \end{aligned}$$

$I(X;Y)$ is called the mutual information. It quantifies how X and Y are correlated. Let us specify some properties of the mutual information.

- If they are independent $p(x, y) = p(x)p(y)$ and

$$I(X;Y) = \left\langle \log_2 \left(\frac{p(x, y)}{p(x)p(y)} \right) \right\rangle = 0. \tag{19}$$

- It is symmetric under interchange of X and Y: I find out as much about X by learning Y as about Y by learning X.

- It is also non-negative

$$I(X;Y) \geq 0 \tag{20}$$

since learning Y never reduces my knowledge of X, that is

$$H(X) \geq H(X|Y) \geq 0. \tag{21}$$

- A further property is called strong subadditivity

$$I(X;YZ) \geq I(X;Y) \tag{22}$$

This is the eminently reasonable statement that the correlations of X with YZ are at least as strong as the correlations of X with Y alone.

4 Distributed source coding

Consider now the case where the joint distribution XY is sampled n times and where Alice receives the n-letter message $\vec{x}$ while Bob receives the n-letter message $\vec{y}$. Now Alice is to send a message to Bob which will enable Bob to determine $\vec{x}$ with a high success probability, and Alice wants to send as few bits to Bob as possible. This task is harder than in the previous scenario where we assumed that the encoder and the decoder share full knowledge of $\vec{y}$ and can choose their code for compressing $\vec{x}$ accordingly. It turns out, though, that even in this more challenging setting Alice can compress the message she sends to Bob down to $\approx nH(X|Y)$ bits, using a method called Slepian-Wolf coding.

Before receiving $\{\vec{x}, \vec{y}\}$ Alice and Bob agree to sort all the possible n-letter messages that Alice might receive into 2^{nR} possible bins of equal size, where the choice of bins is known to both Alice and Bob. When Alice receives $\vec{x}$, she sends nR bits to Bob, identifying the bin that contains $\vec{x}$. After Bob receives this message, he knows both $\vec{y}$ and the bin containing $\vec{x}$. If there is a unique message in that bin which can be correlated to $\vec{y}$, Bob is sure to decode correctly. The question is thus : how many bits nR does Alice needs to send to Bob so that there is a message in each bin uniquely correlated with $\vec{y}$? We know that $H(X|Y)$ bits are needed per letter to specify $\vec{x}$ given $\vec{y}$. We conclude that if Alice sends R bits to Bob per letter of the message $\vec{x}$, where

$$R \approx H(X|Y) \tag{23}$$

then the probability of a decoding error vanishes in the limit of large n. This is the basic idea of the Slepian-Wolf coding.

5 Von Neumman entropy

We have considered before a source that prepares messages of n letters, where each letter is taken independently from an ensemble $X = \{x, p(x)\}$. We have seen that the Shannon entropy $H(X)$ is the number of incompressible bits of information carried per letter. The correlations between two ensembles of letters X and Y are characterized by conditional probabilities $p(y|x)$. We have seen that the mutual information $I(X;Y)$ is the number of bits of information per letter about X that we can acquire by reading Y (or vice versa). We would like to generalize these considerations to quantum information. We now consider a source that prepares messages of n letters, but where each letter is chosen from an ensemble of quantum states. The signal alphabet consists of a set of quantum states $\{\rho(x)\}$, each occurring with a specified probability $p(x)$. The relevant density operator is given by

$$\rho = \sum_x p(x)\rho(x) \tag{24}$$

For any density operator like ρ, we define the Von Neumann entropy

$$H(\rho) = -\mathrm{tr}(\rho \log \rho). \tag{25}$$

Let $\{|a\rangle\}$ be the basis that diagonalizes ρ, that is $\rho = \sum_a \lambda_a |a\rangle\langle a|$ [2]. The set $\{\lambda_a\}$ is a probability distribution and the Von Neumann entropy of ρ is simply the Shannon entropy of this distribution

$$H(\rho) = H(\lambda_a). \tag{28}$$

We will argue that the Von Neumann entropy quantifies the incompressible information content of a quantum source (in the case where the signal states are pure) much as the Shannon entropy quantifies the information content of a classical source. It also quantifies the classical information content per letter of a pure-state ensemble (the maximum amount of information per

[2] We can show using the Taylor series of e^ρ and the property of the projector $|a\rangle\langle a|^2 = |a\rangle\langle a|$ that $\log \rho = \sum_a \log(\lambda_a)|a\rangle\langle a|$. We can then easily check that

$$-\mathrm{tr}(\rho \log \rho) = -\sum_c \langle c| \,(\sum_b \lambda_b |b\rangle\langle b|)(\sum_a \log(\lambda_a)|a\rangle\langle a|)|c\rangle \tag{26}$$

$$= -\sum_a \lambda_a \log(\lambda_a) = H(\lambda_a) \tag{27}$$

letter – in bits, not qubits – that we can gain about the preparation by making the best possible measurement). Before we give the details, we proceed with the notation and properties of the Von Neumman entropy.

6 Von Neumman entropy : properties

Let ρ_A the density operator of system A, we will sometimes use the notation $H(A) := H(\rho)$. Our convention is to denote quantum systems with A, B, C ... and classical probability distributions with X, Y, Z ...

Let us now specify some properties of the Von Neumann entropy.

- A pure state $\rho = |\phi\rangle\langle\phi|$ has $H(\rho) = 0$.
- It is unchanged by a change of unitary $H(U\rho U^\dagger) = H(\rho)$ simply because it only depends on the eigenvalues of ρ.
- If ρ has d non-zero eigenvalues, then $H(\rho) \leq \log d$ with equality when all the nonzero eigenvalues are equal. The entropy is maximized when the quantum state is maximally mixed.
- For $\lambda_1, \lambda_2, \ldots, \lambda_n \geq 0$ and $\lambda_1 + \lambda_2 + \ldots + \lambda_n = 1$,

$$H(\lambda_1\rho_1 + \ldots + \lambda_N\rho_n) \geq \lambda_1 H(\rho_1) + \ldots + \lambda_n H(\rho_n). \tag{29}$$

  The Von Neumann entropy is larger if we are more ignorant about how the state was prepared. This property is a consequence of the concavity of the log function.

- Consider a bipartite system AB in the state ρ_{AB}. Then

$$H(AB) \leq H(A) + H(B) \tag{30}$$

  (where $\rho_A = \mathrm{Tr}_B(\rho_{AB})$ and $\rho_B = \mathrm{Tr}_A(\rho_{AB})$), with equality only for $\rho_{AB} = \rho_A \otimes \rho_B$. Thus, entropy is additive for uncorrelated systems, but otherwise the entropy of the whole is less than the sum of the entropy of the parts. This property is the quantum generalization of subadditivity of Shannon entropy: $H(XY) \leq H(X) + H(Y)$.

- If the state ρ_{AB} of the bipartite system AB is pure, then

$$H(A) = H(B) \tag{31}$$

because ρ_A and ρ_B have the same nonzero eigenvalues[3].

- As in the classical case, we define the mutual information of two quantum systems as

$$I(A;B) = H(A) + H(B) - H(AB) \tag{32}$$

which is nonnegative because of the subadditivity of Von Neumann entropy, and zero only for a product state $\rho_{AB} = \rho_A \otimes \rho_B$.

7 Incompressible information content of a quantum source

We here define the quantum analog of Shannon's source coding theorem. Let's consider a long message consisting of n letters, where each letter is a pure quantum state chosen by sampling from the ensemble $\{|\phi(x)\rangle, p(x)\}$. If the states of this ensemble are mutually orthogonal, then the message might as well be classical; the interesting quantum case is when the states are not orthogonal and therefore not perfectly distinguishable. The density operator realized by this ensemble is

$$\rho = \sum_x p(x)|\phi(x)\rangle\langle\phi(x)| \tag{33}$$

and the entire n-letter message has the density operator

$$\rho^{\otimes n} = \rho \otimes \ldots \otimes \rho. \tag{34}$$

How redundant is the quantum information in this message? We would like to devise a quantum code allowing us to compress the message to a smaller Hilbert space, but without much compromising the fidelity of the message. The optimal compression that can be achieved was found by Schumacher. As you might guess, the message can be compressed to a Hilbert space $\mathcal{H}$ with

$$\dim\mathcal{H} \approx 2^{nH(\rho)} \tag{35}$$

with negligible loss of fidelity as when $n \to \infty$. In this sense, the Von Neumann entropy is the number of qubits of quantum information carried per letter of the message. Compression is always possible unless ρ is maximally

[3]This can be seen from the Schmidt decomposition of the bipartite state $\rho_{AB} = |\psi\rangle\langle\psi|$, i.e. $|\psi\rangle = \sum_k \lambda_k |\phi_k, \psi_k\rangle$ with $\{|\phi_k\rangle\}$ and $\{|\psi_k\rangle\}$ two orthonormal sets.

mixed, just as we can always compress a classical message unless the information source is uniformly random. This result provides a precise operational interpretation for Von Neumann entropy.

8 Accessible information

How much can we learn from a measurement? Consider the following game: Alice prepares a quantum state drawn from the ensemble $\mathcal{E} = \{\rho(x), p(x)\}$ and sends the state to Bob. Bob knows this ensemble, but not the particular state that Alice chose to send. After receiving the state, Bob performs a POVM E with elements $\{E(y)\}$, hoping to find out as much as he can about what Alice sent. The improvement in Bob's knowledge achieved by the measurement is Bob's information gain, the mutual information

$$I(X;Y) = H(X) - H(X|Y). \tag{36}$$

Bob's best strategy (his optimal measurement) maximizes this information gain. The best information gain Bob can achieve is thus

$$\mathrm{Acc}(E) = \max_E I(X;Y) \tag{37}$$

which is a property of the ensemble $\mathcal{E}$ called the accessible information of $\mathcal{E}$. Though there is no simple general formula for the accessible information of an ensemble, we have a useful upper bound, called the Holevo bound. Let us derive it now.

Recall that quantum mutual information obeys monotonicity – if a quantum channel maps B to B', then $I(A;B) \geq I(A;B')$. We derive the Holevo bound by applying monotonicity of mutual information to the accessible information game. We will suppose that Alice records her chosen state in a classical register X and Bob likewise records his measurement outcome in another register Y, so that Bob's information gain is the mutual information $I(X;Y)$ of the two registers. After Alice's preparation of her system A, the joint state of XA

$$\rho_{XA} = \sum_x p(x)|x\rangle\langle x| \otimes \rho(x). \tag{38}$$

Bob's measurement is a quantum channel mapping A to AY according to

$$\rho(x) \to \sum_y M(y)\rho(x)M(y)^\dagger \otimes |y\rangle\langle y|, \tag{39}$$

where $M(y)^\dagger M(y) = E(y)$, yielding the state for XAY

$$\rho'_{XAY} = \sum_{x,y} p(x)|x\rangle\langle x| \otimes M(y)\rho(x)M(y)^\dagger \otimes |y\rangle\langle y| \tag{40}$$

Now we have

$$I(X;Y)_{\rho'} \leq I(X;AY)_{\rho'} \leq I(X;A)_{\rho}, \tag{41}$$

where the subscript indicates the state on which the mutual information is evaluated. The first inequality uses strong subadditivity in the state ρ' and the second uses monotonicity under the channel mapping ρ to ρ'. We have

$$\mathrm{Acc}(E) \leq \chi(E) := I(X;A)_{\rho}. \tag{42}$$

This is the Holevo bound. Note that [4]

$$\begin{aligned} H(XA) &= -\mathrm{tr}_{XA}\left(\sum_x p(x)|x\rangle\langle x| \otimes \rho(x) \log\left(\sum_x p(x)|x\rangle\langle x| \otimes \rho(x)\right)\right) \\ &= -\sum_x \mathrm{tr}_A p(x)\rho(x)\left(\log p(x) + \log \rho(x)\right) \\ &= H(X) + \sum_x p(x) H\left(\rho(x)\right) \end{aligned}$$

and therefore[5]

$$H(A|X) = H(XA) - H(X) = \sum_x p(x) H(\rho(x)). \tag{46}$$

[4]Another way to proof that $H(XA) = H(X) + \sum_x p(x) H\left(\rho(x)\right)$ is to consider the spectral decomposition of $\rho(x) = \sum_y \lambda_y^x |\phi_y^x\rangle\langle\phi_y^x|$. This provides an orthonormal basis $\{|x\rangle \otimes |\phi_y^x\rangle\langle\phi_y^x|\}$ for the joint Hilbert space XA in which ρ_{XA} is diagonal $\rho_{XA} = \sum_{xy} p(x)\lambda_y^x |x\rangle\langle x| \otimes |\phi_y^x\rangle\langle\phi_y^x|$, i.e. $p(x)\lambda_y^x$ are the eigenvalues of ρ_{XA}. This means that $H(XA) = -\sum_{xy} p(x)\lambda_y^x \log(p(x)\lambda_y^x) = -\sum_x p(x)\log(p(x)) - \sum_{xy} p(x)\lambda_y^x \log(\lambda_y^x) = H(X) - \sum_x p(x) H\left(\rho(x)\right)$.

[5]Alternatively, we prove that for any state ρ of the form $\rho = \sum_i p_i \rho_i$ with $\{\rho_i\}$ having support on pairwise orthogonal subspaces $H(\rho) = H(p_i) + \sum_i p_i H(\rho_i)$ with $H(p_i)$ the Shannon entropy of the distribution $\{i, p_i\}$ and $H(\rho_i)$ the Von Neumann entropy of the distribution $\{i, \rho_i\}$. This can be shown by first considering the spectral decomposition of $\rho_i = \sum_j \lambda_j^{(i)} |j^{(i)}\rangle\langle j^{(i)}|$. We have $\rho = \sum_{i,j} p_i \lambda_j^{(i)} |j^{(i)}\rangle\langle j^{(i)}|$ and hence

$$\begin{aligned} H(\rho) &= -\sum_{ij} p_i \lambda_j^{(i)} \log(p_i \lambda_j^{(i)}) \\ &= -\sum_{ij} p_i \lambda_j^{(i)} (\log(p_i) + \log(\lambda_j^{(i)})) \\ &= -\sum_i p_i (\sum_j \lambda_j^{(i)}) \log(p_i) - \sum_i p_i \sum_j \lambda_j^{(i)} \log(\lambda_j^{(i)}) \\ &= -\sum_i p_i \log(p_i) - \sum_i p_i \sum_j \lambda_j^{(i)} \log(\lambda_j^{(i)}) \\ &= H(p_i) + \sum_i p_i H(\rho_i) \end{aligned} \tag{43}$$

Using $I(X;A) = H(A) - H(A|X)$, we find

$$\chi(E) = I(X;A) = H(\rho_A) - \sum_x p(x) H(\rho_A(x)). \tag{47}$$

Let us discuss a consequence of this bound. For the special case of an ensemble of pure states $\mathcal{E} = \{|\phi(x)\rangle, p(x)\}$, the Holevo bound becomes $\mathrm{Acc}(E) \leq H(\rho)$, where $\rho = \sum_x p(x)|\phi(x)\rangle\langle\phi(x)|$. Since the entropy for a quantum system with dimension d can be no larger than $\log d$, the Holevo bound asserts that Alice, by sending n qubits to Bob ($d = 2^n$) can convey no more than n bits of information. This is true even if Bob performs a sophisticated collective measurement on all the qubits at once, rather than measuring them one at a time.

Therefore, if Alice wants to convey classical information to Bob by sending qubits, she can do no better than treating the qubits as though they were classical, sending each qubit in one of the two orthogonal states $\{|0\rangle, |1\rangle\}$ to transmit one bit. This statement is not so obvious. Alice might try to stuff more classical information into a single qubit by sending a state chosen from a large alphabet of pure single-qubit signal states, distributed uniformly on the Bloch sphere. But the enlarged alphabet is of no use because as the number of possible signals increases the signals also become less distinguishable, and Bob is not able to extract the extra information Alice hoped to deposit in the qubit.

Applying this result to $\rho = \sum_x p(x)\,|x\rangle\langle x|\,\rho(x)$, we get

$$H(\rho) = H(p(x)) + \sum_x p_x H(|x\rangle\langle x| \otimes \rho(x)) \tag{44}$$

and since $H(|x\rangle\langle x| \otimes \rho(x)) = H(|x\rangle\langle x|) + H(\rho(x))$ and $H(|x\rangle\langle x|) = 0$, we get

$$H(\rho) = H(p(x)) + \sum_x p_x H(\rho(x)). \tag{45}$$