Shanonn and Von Neumann Entropies

Emad R. F. Boosari

January 9, 2024

1 Shannon's Entropy[1, 2]

The key concept of classical information theory is the *Shannon entropy*. Suppose we learn the value of a random variable X. The Shannon entropy of X quantifies how much information we gain, on average, when we learn the value of X. An alternative view is that the entropy of X measures the amount of *uncertainty* about X before we learn its value. These two views are complementary; we can view the entropy either as a measure of our uncertainty *before* we learn the value of X, or as a measure of how much information we have gained *after* we learn the value of X.

Example:

Consider you have a fair six-sided die. Before you roll the die, you have uncertainty about the outcome, and this uncertainty can be measured by the Shannon entropy.

- ***Uncertainty View (Before Rolling):***

  Before rolling the die, there are six possible outcomes (1, 2, 3, 4, 5, 6), and each outcome is equally likely in the case of a fair die. The uncertainty or lack of information about which specific outcome will occur is at its maximum. The Shannon entropy is high.

  As you haven't rolled the die yet, the entropy represents the uncertainty about the outcome.

  The formula for information content $I(x)$ of an event with probability $P(x)$ is given by:

$$I(x) = -\log_2 P(x) \tag{1}$$

  Before rolling, the uncertainty was high because any of the six outcomes was possible, and each outcome had a probability of $1/6$ (for a fair die).

$$I(\text{before rolling}) = -\log_2\left(\frac{1}{6}\right) = \log_2 6 \approx 2.585 \tag{2}$$

- ***Information Gain View (After Rolling):***

  Now, you roll the die and, let's say, it lands on 3. Before rolling, the uncertainty (entropy) was high because any of the six outcomes was possible. After rolling and learning the outcome (3), your uncertainty is resolved. You have gained information about the specific result. The Shannon entropy is now low, as there is little uncertainty left about the outcome.

  After rolling and learning the outcome (3), the uncertainty is resolved. The probability of getting a 3 is now 1 (because you observed it).

$$I(\text{after rolling}) = -\log_2(1) = 0 \tag{3}$$

  So, the information gained after rolling the die is:

$$\begin{aligned} \text{Information Gained} &= I(\text{before rolling}) - I(\text{after rolling}) && (4)\\ &= -\log_2\left(\frac{1}{6}\right) - 0 && (5)\\ &\approx 2.585 && (6) \end{aligned}$$

In this example, the Shannon entropy is a measure of the uncertainty about the die's outcome before rolling, and it also reflects how much information is gained (and thus, uncertainty reduced) after learning the specific outcome. The views are complementary: entropy measures uncertainty before an event (rolling the die), and the reduction in entropy measures the information gained after the event.

The first basic task of classical information theory is to quantify the information contained in a message. A message is a string of letters chosen from an alphabet

$$\mathcal{A} = a_1, a_2, \ldots, a_k. \tag{7}$$

We assume that the letters in the message are statistically independent and that the letter a_i occurs with a priori probability p_i, where

$$\sum_{i=1} p_i = 1 \tag{8}$$

Remarks:

The assumption that the letters are statistically independent has been made to simplify the discussion. In practice, this is not the case in many important examples. For instance, there are strong correlations between consecutive letters in an English text. However, the ideas developed

in this section can be extended to include more complicated situations with correlations. Thus, in what follows statistically independence of the letters will always be assumed and it should not be forgotten that the case of a real language (such as English) is somewhat different.

The Shannon entropy associated with the probability distribution $\{p_1, p_2, \ldots, p_k\}$ is defined by

$$H(p_1, p_2, \ldots, p_k) \equiv -\sum_{i=1}^{k} p_i \log p_i \tag{9}$$

Note that, here as in the rest of note, all the logarithms are base 2 unless otherwise indicated. We shall show that the Shannon entropy quantifies how much information we gain, on average, when we learn the value of a letter of the message. You may wonder what happens when $p_x = 0$, since $\log 0$ is undefined. Intuitively, an event which can never occur should not contribute to the entropy, so by convention we agree that $0 \log 0 \equiv 0$. More formally, note that $\lim_{x \to 0} x \log x = 0$, which provides further support for our convention.

Let us consider the special case $k = 2$ and define $p_1 = p$ (where $0 \leq p \leq 1$). Since $p_2 = 1 - p$, the Shannon binary entropy is a function of p alone and we can write

$$H_{bin}(p) \equiv -p \log p - (1 - p) \log(1 - p) \tag{10}$$

In the following we shall simply write $H(p)$ instead of $H_{bin}(p)$. The Shannon binary entropy $H(p)$ is plotted in Fig. (1): it is equal to zero when $p = 0$ or $p = 1$ and attains its maximum value $H = 1$ when $p = 1/2$. This is consistent with our interpretation of $H(p)$ as the average information content of each letter in the message. Indeed, information is a measure of our a *priori ignorance*. If we already know that we shall receive the letter a_1 with certainty ($p = 1$), then no information is gained from the reception of this letter. The same conclusion holds when $p = 0$ and we always receive a_2. If, on the other hand, both letters are equiprobable, our a *priori* ignorance is maximum and therefore when we receive a letter, we gain the maximum possible information $H(\frac{1}{2}) = 1$. In this case, we say that we have received one unit of information, known as a *bit*. Typically, we write the letters as binary digits; that is, $a_1 = 0$ and $a_2 = 1$.

Question:

Show that the Shannon entropy $H(p_1, p_2, \ldots, p_k)$ is maximum when $p_1 = \ldots = p_k = 1/k$.

To show that the Shannon entropy $H(p_1, p_2, \ldots, p_k)$ is maximum when $p_1 = p_2 = \ldots = p_k = \frac{1}{k}$, we can use the concavity property of the logarithm function. The Shannon entropy is defined as:

$$H(p_1, p_2, \ldots, p_k) = -\sum_{i=1}^{k} p_i \log(p_i)$$

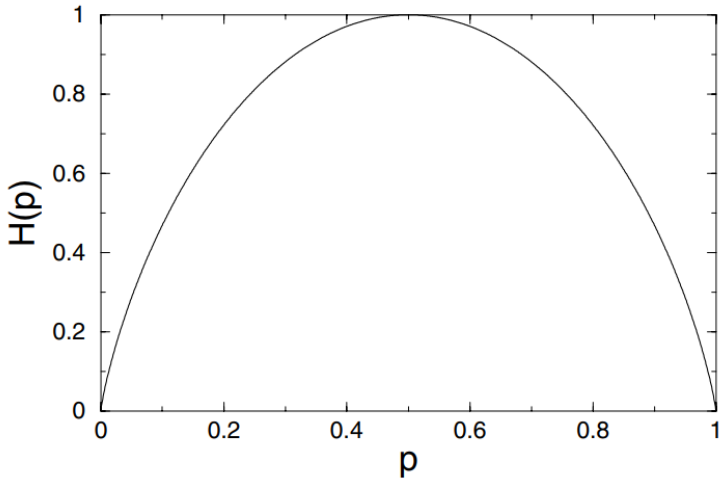

Figure 1: The Shannon binary entropy $H(p) = -p \log p - (1-p) \log(1-p)$.

The concavity of the logarithm implies that for any positive numbers $x_1, x_2, \ldots, x_k$ and non-negative numbers $p_1, p_2, \ldots, p_k$ such that $p_1 + p_2 + \ldots + p_k = 1$, the following inequality holds:

$$-\sum_{i=1}^{k} p_i \log(x_i) \geq -\log\left(\sum_{i=1}^{k} p_i x_i\right)$$

Now, let's consider the case where $x_1, x_2, \ldots, x_k$ are all equal to $\frac{1}{k}$ (i.e., $x_i = \frac{1}{k}$ for all i):

$$-\sum_{i=1}^{k} p_i \log\left(\frac{1}{k}\right) \geq -\log\left(\sum_{i=1}^{k} p_i \frac{1}{k}\right)$$

Simplifying, we get:

$$-\sum_{i=1}^{k} p_i \left(-\log(k)\right) \geq -\log\left(\frac{1}{k}\sum_{i=1}^{k} p_i\right)$$

$$\sum_{i=1}^{k} p_i \log(k) \geq \log(k)$$

Now, dividing both sides by $\log(k)$ (since $\log(k)$ is positive), we get:

$$\sum_{i=1}^{k} p_i \geq 1$$

The equality holds when $p_1 = p_2 = \ldots = p_k = \frac{1}{k}$. Therefore, we've shown that the Shannon entropy is maximized when $p_1 = p_2 = \ldots = p_k = \frac{1}{k}$.

1.1 Shannon's noiseless coding theorem

We now show that the Shannon entropy is a good measure of information. Let us consider the following fundamental problem: how much can a message be *compressed* while still obtaining essentially the same information? In other words, what are the minimal physical resources required in order to store a message without loosing its information content?

As an example, we consider a message written using an alphabet with four letters, $\mathcal{A} = a_1, a_2, a_3, a_4$. We assume that these letters occur with probabilities $p_1 = 1/2, p_2 = 1/4, p_3 = p_4 = 1/8$. To specify a letter out of four we need 2 bits of information. It is instead more convenient to encode the letters as follows:

$$a_1 \longrightarrow c_1 \equiv 0, \quad a_2 \longrightarrow c_2 \equiv 10, \quad a_3 \longrightarrow c_3 \equiv 110, \quad a_4 \longrightarrow c_4 \equiv 111 \tag{11}$$

To send one coded letter we need, on average,

$$\sum_{i=1}^{4} p_i l_i, \tag{12}$$

bits, where l_i is the length, in bits, of the coded letter c_i (we have $l_1 = 1$, $l_2 = 2$, $l_3 = l_4 = 3$). Since

$$\begin{aligned} \sum_{i=1}^{4} p_i l_i &= p_1 l_1 + p_2 l_2 + p_3 l_3 + p_4 l_4 \\ &= \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 = \frac{7}{4} < 2, \end{aligned} \tag{13}$$

we have compressed the information.

Attention:
Note that the good strategy, here as in any other useful compression code, is to encode the most probable strings in the shortest sequences and the less probable strings in the longest sequences.

Note that in the example considered above the optimal compression rate is $H = -\sum_{i=1} p_i log p_i = 7/4$. Since $\sum_i p_i l_i = 7/4 = H$, the optimal compression established by the Shannon's theorem has been attained.

Shannon proved that the optimal compression rate is given by the Shannon entropy. If Alice sends Bob a string of n letters taken from the alphabet $\mathcal{A} = a_1, \cdots, a_k$ and each letter a_i occurs with the a *priori* probability p_i, then, for large n, Alice can reliably communicate her message by sending only $nH(p_1, \ldots, p_k)$ bits of information. This is the content of the Shannon's noiseless coding theorem.

Shannon's noiseless coding theorem: *Given a message in which the letters have been chosen independently from the ensemble* $\mathcal{A} = \{a_1, \ldots, a_k\}$ *with a priori probabilities* $\{p_1, \ldots, p_k\}$*, there exists, asymptotically in the length of the message, an optimal and reliable code compressing the message to* $H(p_1, \ldots, p_k)$ *bits per letter.*

A proof of Shannon's theorem can be found in Cover and Thomas (1991).

1.1.1 Examples of data compression

It is clear that an "asymptotic" data compression strategy; that is, a strategy based on the compression of long typical sequences is not practical: to compress a long n-letter message, we must accumulate all n letters before identifying the typical sequence and compressing it. Fortunately, there exist quite efficient methods to encode smaller strings of letters.

First Example:

Here we consider further examples. First of all, we apply the encoding (11) to a four-letter alphabet, with $p_1 = 0.9$, $p_2 = 0.05$, $p_3 = p_4 = 0.025$. The optimal compression is determined by $H(p_1, p_2, p_3, p_4) \approx 0.62$, while the code gives $\sum_i p_i l_i = 1.15$ and therefore data compression in this case, even though useful, is not optimal.

Let us apply the same code to the case in which the four letters are equiprobable, $p_i = \frac{1}{4}$ for $i = 1, \ldots, 4$. In this case, no compression is possible, because $H = 2$ and we send exactly two bits to specify a letter. Furthermore, if we try to apply the previous code, we obtain $\sum_i p_i l_i = 2.25 > 2$ and therefore the code is in this case detrimental to the efficiency of data transmission. of data transmission.

Second Example:

Let us consider the Huffman code, shown in Table (1.1.1). We consider a binary alphabet $\{0, 1\}$ and the encoding procedure is applied to strings four bits long. There are $2^4 = 16$ such strings $(0 \equiv 0000, 1 \equiv 0001, \cdots, 15 \equiv 1111)$. Let P_i denote the probability that the string i occurs, with $i = 0, \cdots, 15$. If we consider, for instance, the case with $p_0 = \frac{3}{4}$ and $p_1 = \frac{1}{4}$, we find that the best possible compression for a four-letter message is given by

$$4H(p_0, p_1) \approx 3.25, \tag{14}$$

while the Huffman code gives on average $\sum_{i=0}^{15} P_i l_i \approx 3.27$ bits, which is very close to the optimal

Message	Hoffman's Encoding	length of String	Probability	Probability
0000	10	$l_0 = 2$	P_0	p_0^4
0001	000	$l_1 = 3$	P_1	$p_0^3 p_1$
0010	001	$l_2 = 3$	P_2	$p_0^3 p_1$
0011	11000	$l_3 = 5$	P_3	$p_0^2 p_1^2$
0100	010	$l_4 = 3$	P_4	$p_0^3 p_1$
0101	11001	$l_5 = 5$	P_5	$p_0^2 p_1^2$
0110	11010	$l_6 = 5$	P_6	$p_0^2 p_1^2$
0111	1111000	$l_7 = 7$	P_7	$p_0 p_1^3$
1000	011	$l_8 = 3$	P_8	$p_0^3 p_1$
1001	11011	$l_9 = 5$	P_9	$p_0^2 p_1^2$
1010	11100	$l_{10} = 5$	P_{10}	$p_0^2 p_1^2$
1011	111111	$l_{11} = 6$	P_{11}	$p_0 p_1^3$
1100	11101	$l_{12} = 5$	P_{12}	$p_0^2 p_1^2$
1101	111110	$l_{13} = 6$	P_{13}	$p_0 p_1^3$
1110	111101	$l_{14} = 6$	P_{14}	$p_0 p_1^3$
1111	1111001	$l_{15} = 7$	P_{15}	p_1^4

Table 1: Data encodingby means of the Huffman code, with $p_0 = 3/4$, $p_1 = 1/4$.

value. This shows the power of data compression codes.

Remarks:

The enormous practical importance of data compression in fields such as telecommunication is self-evident. Data compression allows us to increase the transmission rate or the storage capacity of a computer. To achieve such results, we simply exploit the redundancies that any message contains: for instance, the letters of an (English) text are not equiprobable but appear with different frequencies. Shannon's theorem tells us that, as far as the letters of a message are not equiprobable, data compression is possible

2 Von Neumann's Entropy

The quantum analogue of the Shannon entropy is the von Neumann entropy. If a quantum system is described by the density matrix ρ, its von Neumann entropy $S(\rho)$ is defined as

$$S(\rho) \equiv -\mathrm{Tr}(\rho \log \rho) \tag{15}$$

To see the analogy with the Shannon entropy, let us consider the following situation: Alice has at her disposal an alphabet $\mathcal{A} = \{\rho_1, \rho_2, ..., \rho_k\}$, where the letters ρ_i are density matrices describing quantum states (pure or mixed). The letters are chosen at random with probabilities p_i, where $\sum_{i=1} p_i = 1$. Let us assume that Alice sends a letter (a quantum state) to Bob and that Bob only knows that the letter has been taken from the ensemble $\{\rho_i, p_i\}$. Thus, he describes this quantum system by means of the density matrix

$$\rho = \sum_{i=1}^{k} p_i \rho_i \tag{16}$$

Therefore

$$S(\rho) \equiv -\mathrm{Tr}(\rho \log \rho) = -\sum_{i=1}^{k} \lambda_i \log \lambda_i = H(\lambda_1, \cdots, \lambda_k) \tag{17}$$

where the λ_i are the eigenvalues of the density matrix ρ and $H(\lambda_1, \cdots, \lambda_k)$ is the Shannon entropy associated with the ensemble $\{\lambda_i\}$.

2.1 The von Neumann entropy satisfies the following properties:

1. For a **pure state**, $S(\rho) = 0$. Indeed, in this case only one eigenvalue of ρ is different from zero, say $\lambda_1 = 1$, so that $-\sum_i \lambda_i \log \lambda_i = -\lambda_1 \log \lambda_1 = 0$.

2. The entropy is not modified by a unitary change of basis; that is, $S(U\rho U^\dagger) = S(\rho)$. Actually $S(\rho)$ depends only on the eigenvalues of ρ, which are basis-independent. This property means that the von Neumann entropy is invariant under unitary temporal evolution.

3. If the density operator ρ acts on a N-dimensional Hilbert space, then $0 \geq S(\rho) \geq \log N$. It is easy to see that $S(\rho) \leq 0$ since $0 \geq \lambda_i \geq 1$ and therefore $-\lambda_i \log \lambda_i \leq 0$. To show that $S(\rho) \geq \log N$, we use $S(\rho) = H(\lambda_1, ..., \lambda_N)$ and remember that the Shannon entropy $H(\lambda_1, ..., \lambda_N)$ takes its maximum value $\log N when \lambda_1 = \cdots = \lambda_N = 1/N$. Hence, $S_{max} = -\frac{1}{N}\sum_{i=1} \log \frac{1}{N} = \log N$.

   The following examples give a flavour of the similarities and the differences between the von Neumann entropy and the Shannon entropy.

2.2 Example 1: source of orthogonal pure states

In the simplest case, Alice has at her disposal a source of two orthogonal pure states for a qubit. These states constitute a basis for the single qubit Hilbert space and we call them $|0\rangle$ and $|1\rangle$. The corresponding density matrices are $\rho_0 = |0\rangle\langle 0|$ and $\rho_1 = |1\rangle\langle 1|$. We assume that the source generates the states $|0\rangle$ or $|1\rangle$ with the a priori probabilities $p_0 = p$ and $p_1 = 1 - p$, respectively. Therefore, we can write

$$\rho = p_0 |0\rangle\langle 0| + p_1 |1\rangle\langle 1| = \begin{bmatrix} p_0 & 0 \\ 0 & p_1 \end{bmatrix} \tag{18}$$

and the von Neumann entropy is given by

$$\begin{aligned} S(\rho) &= -\mathrm{Tr}(\rho \log \rho) = -\mathrm{Tr}\Big(\begin{bmatrix} p_0 & 0 \\ 0 & p_1 \end{bmatrix} \begin{bmatrix} \log p_0 & 0 \\ 0 & \log p_1 \end{bmatrix}\Big) \\ &= -\mathrm{Tr}\Big(\begin{bmatrix} p_0 \log p_0 & 0 \\ 0 & p_1 \log p_1 \end{bmatrix}\Big) \\ &= -\big(p_0 \log p_0 + p_1 \log p_1\big) = H(p_0, p_1) \end{aligned} \tag{19}$$

Therefore, in this case, in which the letters of the alphabet correspond to orthogonal pure states, the von Neumann entropy coincides with the Shannon entropy. Thus, the situation is in practice classical, from the point of view of information theory. This is quite natural since orthogonal states are perfectly distinguishable.

2.3 Example 2: source of non-orthogonal pure states

Let us consider the case in which the pure states $|\tilde{0}\rangle$ and $|\tilde{1}\rangle$ generated by a source are not orthogonal. It is always possible to choose an appropriate basis set $\{|0\rangle, |1\rangle\}$ (see Fig. 2) so that

$$|\tilde{0}\rangle = \mathrm{Cos}\,|0\rangle + \mathrm{Sin}\,|1\rangle = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \qquad |\tilde{1}\rangle = \mathrm{Sin}\,|0\rangle + \mathrm{Cos}\,|1\rangle = \begin{bmatrix} \sin\theta \\ \cos\theta \end{bmatrix} \tag{20}$$

where we have defined $C \equiv \cos\theta$ and $S \equiv \sin\theta$. We consider, without any loss of generality, $0 \leq \theta \leq \frac{\pi}{4}$. Note that the inner product of these two states is in general non-zero and given by

$$\langle\tilde{0}|\tilde{1}\rangle = \sin(2\theta) \tag{21}$$

The density matrices corresponding to the states $|\tilde{0}\rangle$ and $|\tilde{1}\rangle$ read

$$\rho_0 = |\tilde{0}\rangle\langle\tilde{0}| = \begin{bmatrix} C^2 & CS \\ CS & S^2 \end{bmatrix}, \quad \rho_1 = |\tilde{1}\rangle\langle\tilde{1}| = \begin{bmatrix} S^2 & CS \\ CS & C^2 \end{bmatrix} \tag{22}$$

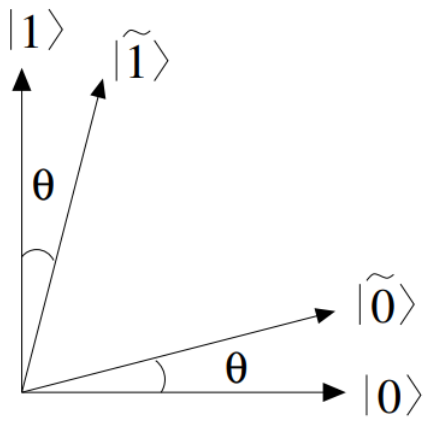

Figure 2: A representation of two non-orthogonal quantum states $|\tilde{0}\rangle$ and $|\tilde{1}\rangle$ in an appropriately chosen basis $\{|0\rangle, |1\rangle\}$ for a qubit.

If the source generates the state $|\tilde{0}\rangle$ with probability p and the state $|\tilde{1}\rangle$ with probability $(1-p)$, the corresponding density matrix is

$$\rho = p\rho_0 + (1-p)\rho_1 = \begin{bmatrix} \sin^2\theta + p\cos 2\theta & \cos\theta\sin\theta \\ \cos\theta\sin\theta & \cos^2 - p\cos 2\theta \end{bmatrix}. \tag{23}$$

The eigenvalues of the density matrix are

$$\lambda_\pm = \frac{1}{2}\left(1 \pm \sqrt{1 + 4p(p-1)\cos^2 2\theta}\right) \tag{24}$$

They are represented in Fig. 3 as a function of the probability p and for different values of θ. We note that for $\theta = 0$ the states are orthogonal and the eigenvalues of the density matrix are p and $1-p$; namely, we recover the classical case. For the other values of θ the eigenvalues "repel" each other, as can be seen from Fig. 3. As we shall show in the next section, this has important consequences for quantum data compression.

Starting from the eigenvalues of the density matrix (**??**), it is easy to compute the von Neumann entropy

$$S(\rho) = -\lambda_+ \ln \lambda_+ - \lambda_- \ln \lambda_- \tag{25}$$

shown in Fig. 4. At $\theta = 0$, we recover the classical results since in this case $S(\rho) = H(p)$. If $\theta = \pi/4$, then $S(\rho) = 0$. Indeed, since in this case the states are identical, there is no transmission of information. As can be seen in Fig. 4, $S(\rho) \leq H(p)$ and it is possible to prove that this inequality has

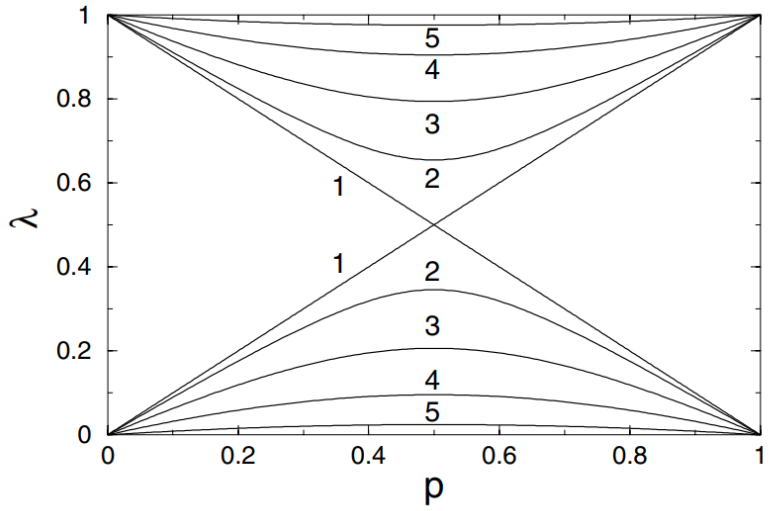

Figure 3: The eigenvalues of the density matrix (**??**) as a function of the probability p. The values of the angle θ are: $1 : \theta = 0$, 2: $\theta = 0.2\frac{\pi}{4}$, 3: $\theta = 0.4\frac{\pi}{4}$, 4: $\theta = 0.6\frac{\pi}{4}$ and $5 : \theta = 0.8\pi$ 4 . The value $\theta = 0$ corresponds to orthogonal states.

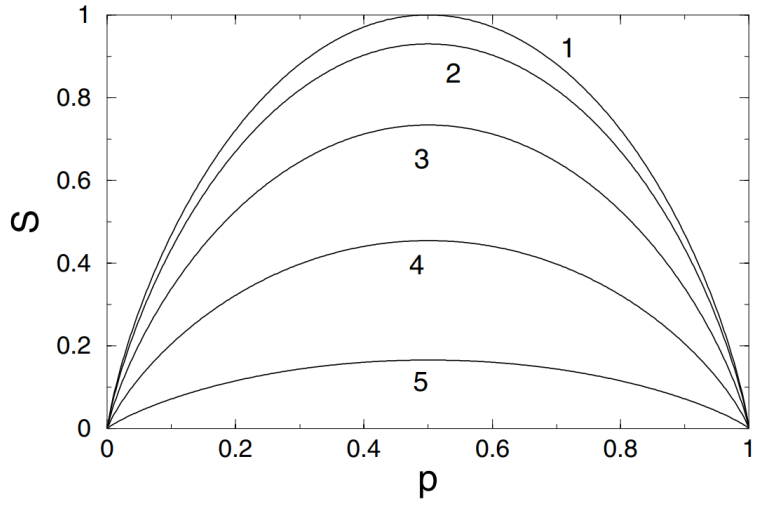

Figure 4: The Von Neumann entropy of the density matrix (**??**) as a function of the probability p. The numbers are associated with the same values of the angle θ as in the previous figure.

general validity. A qualitative interpretation follows from our understanding of entropy as a measure of our ignorance about the system. If the states are non-orthogonal, their similarity increases with their inner product $\langle\tilde{0}|\tilde{1}\rangle = sin2\theta$. Therefore, Bob obtains less information from the reception of a state taken from the ensemble $|\tilde{0}\rangle, |\tilde{1}\rangle$ since his a priori ignorance is smaller. In the limiting case $\theta = \pi/4$ the superposition of the states of the ensemble is unity; that is, the states are identical and there is no a priori ignorance about the system. Therefore, no information is transmitted in this case.

3 Accessible information

We assume that Alice sends Bob a message whose letters are chosen independently from the alphabet $\mathcal{A} = \{a_1, \cdots, a_k\}$ with a priori probabilities $\{p_1, \cdots, p_k\}$. The letters of the alphabet are coded by quantum states that are not necessarily orthogonal. In this section we consider the following problem: how much information can Bob gain on the message by performing measurements on the quantum states received? This problem is non-trivial since non-orthogonal quantum states cannot be perfectly distinguished. It is important to emphasize that, this property lies at the heart of quantum cryptography. First of all, a few definitions are needed. If X is a random variable that takes the value x with probability $p(x)$ ($x \in a_1, \cdots, a_k$ and $p(x) \in p_1, \cdots, p_k$), then the Shannon entropy $H(p_1, ..., p_k)$ is also called $H(X)$ and we write

$$H(X) \equiv -\sum_x p(x) \log p(x) = -\sum_{i=1}^{k} p_i \log p_i \tag{26}$$

Note that $H(X)$ indicates a function not of X but of the information content of the random variable X.

3.1 Joint entropy:

the joint entropy of a pair of random variables X and Y having values x and y with probabilities $p(x)$ and $p(y)$, respectively, is defined by

$$H(X, Y) \equiv -\sum_{x,y} p(x, y) \log p(x, y) \tag{27}$$

where $p(x, y)$ is the probability that $X = x$ and $Y = y$.

3.2 Conditional entropy

: The conditional entropy $H(Y|X)$ is defi ned by

$$H(Y|X) \equiv H(X, Y) - H(X) \tag{28}$$

It is a measure of our residual ignorance about Y , provided we already know the value of X. Similarly, we can define $H(X|Y) \equiv H(X, Y) - H(Y)$. It is easy to show that

$$H(Y|X) = -\sum_{x,y} p(x, y) \log p(y|x) \tag{29}$$

where $p(y|x) = p(x,y)/p(x)$ is the probability that $Y = y$, provided $X = x$. Indeed,

$$\begin{aligned} H(X,Y) - H(X) &= -\sum_{x,y} p(x,y) \log p(x,y) + \sum_{x} p(x) \log p(x) && (30) \\ &= -\sum_{x,y} p(x,y) \log p(x) p(y|x) + \sum_{x,y} p(x,y) \log p(x) && (31) \\ &= -\sum_{x,y} p(x,y) log p(y|x), && (32) \end{aligned}$$

where we have used $\sum_y p(x,y) = p(x)$. Similarly, we obtain

$$H(X|Y) = -\sum_{x,y} p(x,y) \log p(y|x) \tag{33}$$

3.3 Mutual information:

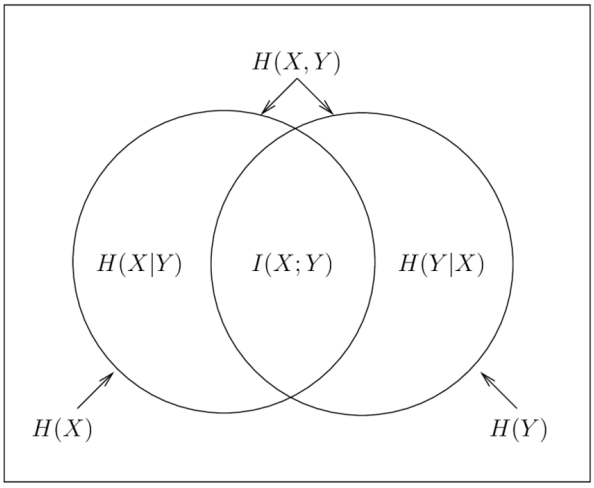

Figure 5: Mutual information $I(X : Y)$ measures the common information between two random variables, X and Y. The information of a random variable is measured with entropy H.

The mutual information $I(X : Y)$ is defined by

$$I(X : Y) \equiv H(X) + H(Y) - H(X,Y). \tag{34}$$

According to the Fig. 5, $I(X : Y)$ is a measure of how much information X and Y have in common. It can be easily shown that

$$I(X : Y) = -\sum x, y p(x,y) \frac{\log p(x)p(y)}{p(x,y)}. \tag{35}$$

From this expression it is clear that, if X and Y are independent, namely $p(x,y) = p(x)p(y)$, then $I(X:Y) = 0$. The mutual information is related to the conditional entropy as follows:

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

We note that, as is clear from its definition (34), the mutual information is symmetric:

$$I(X:Y) = I(Y:X) \tag{37}$$

Let us now return to the problem introduced at the beginning of this section. If X and Y denote the random variables associated with the letters generated by Alice and with Bob's measurement outcomes, respectively, then the **accessible information is defined as the maximum of $I(X:Y)$ over all possible measurement schemes**.

3.4 The Holevo bound:

The Holevo bound (proved by Holevo in 1973) establishes an upper bound on the accessible information.

Theorem: The Holevo Bound:

If Alice prepares a (mixed) state ρ_X chosen from the ensemble $\mathcal{A} = \{\rho_0, \cdots, \rho_k\}$ with a priori probabilities $\{p_1, \cdots, p_k\}$ and Bob performs a POVM measurement on that state, with POVM elements $\{F_1, \cdots, F_\ell\}$ and measurement outcome described by the random variable Y , then the mutual information $I(X:Y)$ is bounded as follows:

$$I(X:Y) \leq S(\rho) - \sum_{i=0}^{k} p_i S(\rho_i) \equiv \chi(\mathcal{E}) \tag{38}$$

where $\rho = \sum_{i=1}^{k} p_i \rho_i$ and $\chi(\mathcal{E})$ is known as the *Holevo information* of the ensemble $\mathcal{E} \equiv \rho_1, \cdots, \rho_k; p_1, \cdots, p_k$.

A proof of this theorem can be found in Nielsen and Chuang [2]. Here, we shall limit ourselves to discuss the Holevo bound in a few concrete examples.

3.4.1 Example: two non-orthogonal pure states

If Alice sends Bob pure orthogonal quantum states drawn from the ensemble $\{|\psi_1\rangle, \cdots, |\psi_k\rangle\}$, then Bob can unambiguously distinguish these states by means of projective measurements described by the POVM elements (in this case, simple projectors) $\{F_1 = |\psi_1\rangle\langle\psi_1|, \cdots, F_k = |\psi_k\rangle\langle\psi_k|\}$. It is easy to check that $I(X:Y) = H(X)$ (we have $H(X|Y) = 0$) and therefore this case is no different from the transmission of classical information over a noiseless channel: if we send the letter a_x, we recover the

same letter; that is, $a_y = a_x$. The simplest example that cannot be reduced to classical information theory is that in which Alice sends Bob states generated by a source of non-orthogonal pure quantum states. We assume that the states $|\tilde{0}\rangle$ and $|\tilde{1}\rangle$, defined by Eq. (20), are generated with probabilities $p_0 = p$ and $p_1 = 1 - p$, respectively. Since the single letters are represented in this case by pure states, their von Neumann entropy is equal to zero:

$$S(\rho_0) = S(|\tilde{0}\rangle \langle\tilde{0}|) = 0, \tag{39}$$

$$S(\rho_1) = S(|\tilde{1}\rangle \langle\tilde{1}|)) = 0. \tag{40}$$

Therefore, the Holevo information $\chi(\mathcal{E})$ reduces to

$$\chi(\mathcal{E}) = S(\rho) \tag{41}$$

where $\rho = p\rho_0 + (1-p)\rho_1$. Hence, the Holevo bound gives

$$I(X:Y) \leq S(\rho) \tag{42}$$

A plot of $S(\rho)$ was already shown in Fig. 4. It reveals that, for nonorthogonal states ($\theta = 0$), $S(\rho) < H(X)$ and therefore $I(X:Y) < H(X)$. It is possible to show that this strict inequality also has general validity for mixed states $\{\rho_i\}$, provided they do not have orthogonal support (for orthogonal support, $I(X:Y) = H(X)$). It is instructive to consider the following special case: we assume that Bob performs a projective measurement on the received qubits along the direction $\hat{n}$ (that is, he measures $\hat{n}\boldsymbol{\sigma}$) and we show that in this case the Holevo bound is satisfied. For this purpose, we compute the mutual information. Bob's measurement along the direction $\hat{n}$ is described by the POVM elements (projectors)

$$F_0 = \frac{1}{2}(I + \hat{n}.\boldsymbol{\sigma}), \qquad F_1 = \frac{1}{2}(I - \hat{n}.\boldsymbol{\sigma}) \tag{43}$$

For instance, if $\hat{n} = (0,0,1)$, then $F_0 = |0\rangle\langle 0|$ and $F_1 = |1\rangle\langle 1|$. We compute the conditional probability

$$p(y|x) = Tr(\rho_x F_y), \qquad (x, y = 0, 1), \tag{44}$$

which is the probability that Bob's measurement gives outcome y, provided the state ρ_x was sent by Alice. For this purpose, we write down the Bloch sphere representation of the density matrices associated with the states $|\tilde{0}\rangle$ and $|\tilde{1}\rangle$

$$\rho_0 = |\tilde{0}\rangle\langle\tilde{0}| = \frac{1}{2}(I + \boldsymbol{r}_0.\boldsymbol{\sigma}), \qquad \rho_1 = |\tilde{1}\rangle\langle\tilde{1}| = \frac{1}{2}(I + \boldsymbol{r}_1.\boldsymbol{\sigma}), \tag{45}$$

where the Cartesian components of the Bloch vectorsr_0 and r_1 are given by

$$r_0 = (\sin 2\theta, 0, \cos 2\theta), \qquad r_1 = (\sin 2\theta, 0, -\cos 2\theta) \tag{46}$$

Taking into account that $Tr(\sigma_i) = 0$ and $Tr(\sigma_i\sigma_j) = 2\delta_{ij}$ for $i, j = x, y, z$, it is now straightforward to compute the conditional probabilities:

$$p(0|0) = Tr(\rho_0 F_0) = \frac{1}{2}(1 + \boldsymbol{r}_0.\hat{n}), \tag{47}$$

$$p(1|0) = Tr(\rho_0 F_1) = \frac{1}{2}(1 + \boldsymbol{r}_0.\hat{n}), \tag{48}$$

$$p(0|1) = Tr(\rho_1 F_0) = \frac{1}{2}(1 + \boldsymbol{r}_1.\hat{n}), \tag{49}$$

$$p(1|1) = Tr(\rho_1 F_1) = \frac{1}{2}(1 + \boldsymbol{r}_1.\hat{n}), \tag{50}$$

$$\tag{51}$$

If, for the sake of simplicity, we assume that the measurement direction lies in the (x, z) plane of the Bloch sphere; that is, $\hat{n} = (\sin\bar{\theta}, 0, \cos\bar{\theta})$ (see Fig. 6), we have

$$p(0|0) = \frac{1}{2}(1 + \cos(\bar{\theta} - 2\theta), \tag{52}$$

$$p(1|0) = \frac{1}{2}(1 - \cos(\bar{\theta} - 2\theta), \tag{53}$$

$$p(0|1) = \frac{1}{2}(1 - \cos(\bar{\theta} + 2\theta), \tag{54}$$

$$p(1|1) = \frac{1}{2}(1 + \cos(\bar{\theta} + 2\theta), \tag{55}$$

$$\tag{56}$$

We now compute $p(x, y) = p(x)p(y|x)$, where, as stated at the beginning of this subsection, we assume that the states $|\tilde{0}\rangle$ and $|\tilde{1}\rangle$ are generated with probabilities $p(X = 0) = p$ and $p(X = 1) = 1-p$, respectively. We thus have

$$p(0, 0) = \frac{1}{2}p(1 + \cos(\bar{\theta} - 2\theta), \tag{57}$$

$$p(1, 0) = \frac{1}{2}p(1 - \cos(\bar{\theta} - 2\theta), \tag{58}$$

$$p(0, 1) = \frac{1}{2}(1 - p)(1 - \cos(\bar{\theta} + 2\theta), \tag{59}$$

$$p(1, 1) = \frac{1}{2}(1 - p)(1 + \cos(\bar{\theta} + 2\theta), \tag{60}$$

$$\tag{61}$$

Then we compute $p(y) = P_x p(x, y)$ and obtain

$$p(Y = 0) = \frac{1}{2}\Big[1 + p\cos(\bar{\theta} - 2\theta) - (1 - p)\cos(\bar{\theta} + 2\theta)\Big], \tag{62}$$

$$p(Y = 1) = \frac{1}{2}\Big[1 - p\cos(\bar{\theta} - 2\theta) - (1 - p)\cos(\bar{\theta} + 2\theta)\Big]. \tag{63}$$

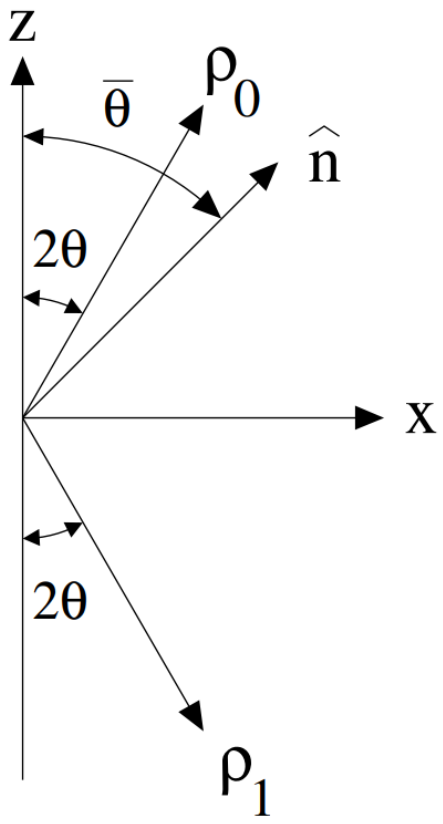

Figure 6: A geometric visualization of the Bloch sphere vectors ρ_0 and ρ_1 and of the measurement axis $\hat{\mathbf{n}}$.

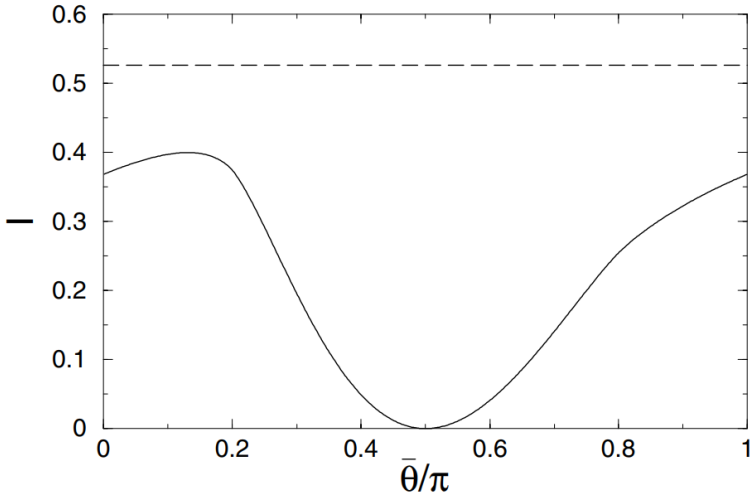

Figure 7: The mutual information $I(X:Y)$ for a message coded by means of the nonorthogonal states (5.163a–5.163b), with $\theta = \pi/10 and p = 0.8$. The angle $\bar{\theta}$ determines the measurement direction $\hat{n} = (sin\bar{\theta}, 0, \cos\bar{\theta})$. The dashed line shows the Holevo bound $\chi \approx 0.526$.

Finally, we insert the expressions derived for $p(x), p(y)$ and $p(x,y)$ into Eq. (5.201), obtaining the mutual information $I(X:Y)$. As an example, in Fig. 7 we show the mutual information $I(X:Y)$ for $\theta = \pi/10$ and $p = 0.8$. Within the chosen measurement scheme, the only free parameter that may be varied in order to maximize I is $\bar{\theta}$. The maximum value $I_{max} \equiv max_{\bar{\theta}} I(\bar{\theta}) \approx 0.40$ is attained for $\bar{\theta} \approx 0.14\pi$. We stress that this value is below the Holevo bound $\chi = S(\rho) \approx 0.526$. Of course, this value is also smaller than the classical bound $I(X:Y) \leq H(X) \approx 0.722$.

References

[1] Giuliano Benenti, Giulio Casati, Davide Rossini, and Giuliano Strini. *Principles of quantum computation and information: a comprehensive textbook*. World Scientific, 2019.

[2] Michael A Nielsen and Isaac L Chuang. Quantum computation and quantum information. *Phys. Today*, 54(2):60, 2001.