Lecture 13 Quantum information theory

Xiao Yuan

November 22, 2021

In this lecture, we study quantum information theory and introduce systematic ways for quantifying information of classical and quantum states.

1 Classical information theory

1.1 Uncertainty and entropy

What is information? Consider the password as an example. Suppose Alice has a password consists of 6 digits, does this password contain information? The answer is yes and no. On the one hand, if Alice does not know the password, such as she forget it, then knowing the password indeed gains information. However, if Alice already knows the password, telling her the password again does not increase any information. We can consider other similar examples, such as reading a book, listen to a lecture, download a file, etc. Therefore, intuitively speaking, information is defined with respect to unknownness or uncertainty. Denote an event by a random variable X with probability $p(X = x)$ or simply denoted as p_x. For the password example, it corresponds to the case of X being a 6-digit number. Then p_x is uniformly distributed and has maximal uncertainty if we have no information of the password, whereas it becomes a δ-function like distribution and has minimal uncertainty if we know the password. Then to quantify information, it is equivalent to quantify the uncertainty.

Shannon builds a beautiful theory and introduce the concept of entropy to quantify uncertainty or information. We first consider an abstract way of defining entropy. Suppose we can always use a real number/function I to quantify the uncertainty of a random event $X = E$ happens with probability $p \in [0, 1]$, it should satisfying the following conditions.

- The function $I(X)$ should only depend on the probability of E (instead of the value), i.e., $I(E) = I(p)$.
- The function is additive for independent events, that is $I(E \,\&\, F) = I(E) + I(F)$ or $I(pq) = I(p) + I(q)$ for two independent events E and F with probabilities $p \in [0, 1]$ and $q \in [0, 1]$, respectively. (We require the condition holds for any $p, q > 0$.)
- The function I is smooth.

We can show that the only function satisfying the above condition is $I(p) = k \log p$ for any constant k.

Proof. Since I is smooth, it has derivatives for all $p > 0$. Then

$$I'(p) = \lim_{\varepsilon \to 0^+} \frac{I(p+\varepsilon) - I(p)}{\varepsilon} = \lim_{\varepsilon \to 0^+} \frac{I(p(1+\varepsilon/p)) - I(p)}{\varepsilon} = \lim_{\varepsilon \to 0^+} \frac{I(1+\varepsilon/p)}{\varepsilon} = \frac{1}{p} \lim_{\varepsilon \to 0^+} \frac{I(1+\varepsilon)}{\varepsilon} \tag{1}$$

We note that $I(1) = 0$ and

$$I'(1) = \lim_{\varepsilon \to 0^+} \frac{I(1+\varepsilon) - I(1)}{\varepsilon} = \lim_{\varepsilon \to 0^+} \frac{I(1+\varepsilon)}{\varepsilon}. \tag{2}$$

Therefore we have

$$I'(p) = \frac{I'(1)}{p} \tag{3}$$

and hence

$$I(p) = \int_{x=1}^{p} I'(p) + I(1) = k \log p. \tag{4}$$

Here we have changed to the $\log_2$ basis and omitted 2. □

Since the uncertainty for each event E with probability p is $I(p) = k \log p$, the average uncertainty or *entropy* of the random variable X is

$$S(X) = S(\mathbf{p}) = \mathbb{E}(I(p)) = k \sum_x p_x \log p_x. \tag{5}$$

Now consider a binary distribution with probability $\mathbf{p} = (p, 1-p)$, the entropy is

$$S(\mathbf{p}) = k(p \log(p) + (1-p) \log(1-p)). \tag{6}$$

When we have a uniform distribution with $p = 1/2$, we have $S(\mathbf{p}) = -k$. Since a uniform distributed random variable has maximal uncertainty/entropy and we can regard a maximally random bit as a random unit, we set $S(\mathbf{p}) = -k = 1$ which leads to $k = -1$. We thus arrive at the definition of *Shannon entropy*

$$S(\mathbf{p}) = -\sum_j p_x \log p_x. \tag{7}$$

Again, the entropy $S(\mathbf{p})$ measures how random or how uncertain the event is. Information is gained when we become to know the event so that the randomness or uncertainty is eliminated. Therefore, the entropy $S(\mathbf{p})$ also measures the information gain of the random variable.

An important property of entropy is concavity,

$$S(a\mathbf{p} + (1-a)\mathbf{q}) \geq aS(\mathbf{p}) + (1-a)S(\mathbf{q}). \tag{8}$$

One way to understand it is as follows. Consider a random coin with head and tail probability a and $1-a$, respectively. When we get head, we generate a random variable with probability $\mathbf{p}$, otherwise we generate it with probability $\mathbf{q}$. Then, the left hand size of the above equation quantifies the total randomness/uncertainty of the coin and the random variable. While on the right hand size, it only contains the randomness of the random variable given known coin outcome, which is no larger than the total randomness.

1.2 Shannon's noiseless coding theorem

What is the operational meaning of the entropy? It actually exactly measures the asymptotic average number of bits we need to know the random variable. Consider two parties Alice and Bob, where Alice aims to tell Bob some information that he does not know. Suppose the information is encoded into independent and identically distributed (i.i.d.) random variables $X_1, X_2, \ldots, X_n$, where the probability of each X_i being 0 is p and being 1 is $1-p$. A naive strategy is of course to transfer all the n bits, which seems very inefficient if the $(p, 1-p)$ distribution is very biased. For example, if $p = 1/10$, we only have very rare random variables being 0s and we should be able to *compress* it quite well. There are various ways to construct different encoding/compression algorithm. Yet, Shannon's noiseless coding theorem tells us the limit of all encoding algorithms. Roughly speaking, Shannon's theorem says that *we need at least $nH(X)$ bits to transfer the n bits, and there is a protocol that achieves it for $n \to \infty$ with a failure probability $\delta \to 0^+$.*

The key idea is to use the concentration lemma or the typical space. Consider the probability distribution of $X_1, X_2, \ldots, X_n$, it is called a typical sequence if we have np number of 0s and $n(1-p)$ number of 1s with probability

$$p(x_1, x_2, \ldots, x_n) = p^{np}(1-p)^{n(1-p)} = 2^{-nH(p)}. \tag{9}$$

Here we have denoted $H(X) = H(p) = -p\log(p) - (1-p)\log(1-p)$. According to the concentration lemma or large number theorem, the number of 0s N_0 is not far from np. Specifically, for arbitrary error $\varepsilon > 0$ and failure probability $\delta > 0$, we can choose a sufficiently large n so that $P(|N_0 - np| \leq n\varepsilon) \leq \delta$. Therefore, with a small failure probability δ, we always have

$$2^{-nH(p+\varepsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-nH(p-\varepsilon)}. \tag{10}$$

Here we have assumed $p < 1/2$ (the analysis for $p > 1/2$ is similar), $p + \varepsilon \leq 1/2$ and $p - \varepsilon \geq 0$ by choosing a sufficiently small ε. Note that

$$H(p+\varepsilon) \leq H(p) + H'(p)\varepsilon \leq H(p) + H'(p-\varepsilon)\varepsilon,\ H(p-\varepsilon) \geq H(p) - H'(p-\varepsilon)\varepsilon \tag{11}$$

and denote $\tilde{\varepsilon} = H'(p-\varepsilon)\varepsilon$, we have

$$2^{-n(H(p)+\tilde{\varepsilon})} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(p)+\tilde{\varepsilon})}. \tag{12}$$

We thus call all sequences that satisfy the above equation ε-typical. We can see that when choosing sufficiently large n, all sequences are ε-typical except for an arbitrary small δ failure probability. Therefore, we only need to focus on these ε-typical sequences and encode them. Since there are at most $2^{n(H(p)+\tilde{\varepsilon})}$ different sequences, we only need $n(H(p) + \tilde{\varepsilon})$ bits or on average $H(p) + \tilde{\varepsilon}$ bit for each random variable. With the limit of $n \to \infty$, the optimal compression rate is $H(p)$.

1.3 Conditional entropy and mutual information

Suppose we have two random variables X and Y, we can quantify their joint entropy as

$$S(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y). \tag{13}$$

We define the condition entropy as

$$S(X|Y) = S(X, Y) - S(Y), \tag{14}$$

which is the joint entropy of X and Y minus the entropy of Y. The mutual information is defined as

$$S(X : Y) = S(X) + S(Y) - S(X, Y) = S(X) - S(X|Y) = S(Y) - S(Y|X), \tag{15}$$

which measures information that is contained in both X and Y.

There are several important properties of the conditional entropy and mutual information, such as

- $S(X, Y) = S(Y, X)$, $S(X : Y) = S(Y : X)$.
- $S(X|Y)$, $S(Y|X)$, $S(X : Y) \geq 0$.
- Strong subadditivity: $S(X, Y, Z) + S(Y) \leq S(X, Y) + S(Y, Z)$ or equivalently $S(Z|X, Y) \leq S(Z|Y)$.

All these properties could be understood and remembered using the Venn diagram. Other important properties include the Chain rule and the data processing inequality (not required in this lecture). We refer to Nielsen & Chuang's book for more details and the proofs.

1.4 Relative entropy/(K-L divergence)

The relative entropy is somehow the *most* important entropy definition. It aims to measure the difference between two distributions $\mathbf{p}$ and $\mathbf{q}$ as

$$S(\mathbf{p}\|\mathbf{q}) = \sum_x p(x) \log(p(x)/q(x)) = \sum_x p(x)[I(q(x)) - I(p(x))]. \tag{16}$$

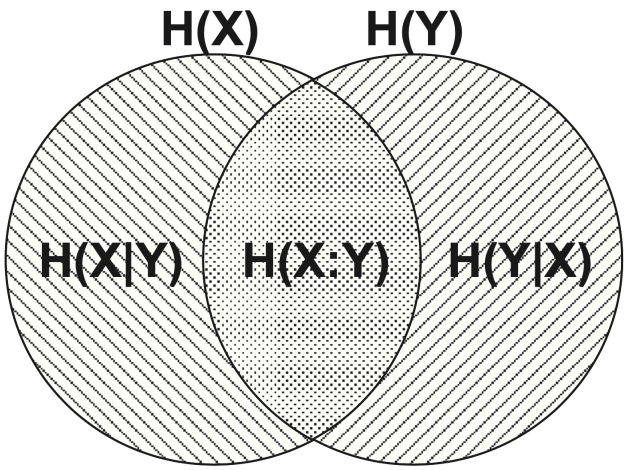

Figure 1: Venn diagram of the entropies. From Nielsen & Chuang's book.

(optional) How to understand this quantifier. Imagine that we have two random variables $X_{\mathbf{p}}$ and $\tilde{X}_{\mathbf{q}}$. Suppose we are given $X_{\mathbf{p}}$ which produces x with probability $p(x)$. When we do not have much samples of $X_{\mathbf{p}}$, we actually do not know the distribution, so as to the random variable. We may falsely thought that the random variable is $\tilde{X}_{\mathbf{q}}$ and, in this case, we falsely thought that we gained an amount of information $I(q(x))$. Therefore the relative entropy $S(\mathbf{p}\|\mathbf{q})$ roughly measures the average information discrepancy $I(q(x)) - I(p(x))$.

(optional) Another way to understand relative entropy is via hypothesis testing. Suppose we are given n samples of X, which are mostly ε-typical satisfying Eq. (12). Now we calculate the probability that those samples are produced by $\tilde{X}$, which is

$$p(\text{samples are } \tilde{X}|x_1, x_2, \ldots, x_n) = \binom{n}{N_0} q_0^{N_0}(1-q_0)^{N_1} \approx 2^{-nS(\mathbf{p}\|\mathbf{q})}. \tag{17}$$

Here we have used $N_0 \approx np_0$, $\binom{n}{N_0} \approx 2^{S(\mathbf{p})}$, and $q_0^{N_0}(1-q_0)^{N_1} \approx 2^{p_0 \log q_0 + p_1 \log q_1}$. Therefore, $S(\mathbf{p}\|\mathbf{q})$ is related to the probability that $\mathbf{q}$ fakes $\mathbf{p}$.

(optional) Finally, the relative entropy measures the average extra information we need to transfer X if we used $\mathbf{q}$ for the distribution. For any event x happens with probability $p(x)$, we on average need $-\log p(x)$ bits to transfer it. Thus, if we used $\mathbf{q}$ for X, we need to transfer

$$-\sum_x p(x) \log q(x) = S(\mathbf{p}) + S(\mathbf{p}\|\mathbf{q}). \tag{18}$$

The power of the relative entropy is that it defines all other entropic measures. We will see this for the more general quantum case.

2 Quantum information theory

2.1 Quantum entropy measures

The Von Neumann entropy of ρ is defined as

$$S(\rho) = -\mathrm{Tr}[\rho \log \rho]. \tag{19}$$

Given a spectral decomposition of $\rho = \sum_j \lambda_j |\psi_j\rangle \langle\psi_j|$, the Von Neumann entropy is

$$S(\rho) = -\sum_j \lambda_j \log \lambda_j. \tag{20}$$

Therefore, the quantum Von Neumann entropy is simply the Shannon entropy of the diagonal elements.

The conditional entropy and mutual information is defined the same. For a bipartite quantum state ρ^{AB},

$$S(A:B) = S(A) + S(B) - S(A,B) = S(A) - S(A|B) = S(B) - S(B|A). \tag{21}$$

Important properties of the above entropies are

- $S(\rho)$, $S(A:B)$ are non-negative, whereas $S(A|B)$ or $S(B|A)$ could be negative.
- For the Von Neumann entropy, we have
  - $S(\rho) \in [0, \log d]$ for qudit states.
  - $S(\rho \otimes \sigma) = S(\rho) + S(\sigma)$.
  - $S(A) = S(B)$ for pure states ψ^{AB}.
  - $S(\sum_j p_j |j\rangle\langle j| \otimes \rho_j) = S(\mathbf{p}) + \sum_j p_j S(\rho_j)$.
  - Mixture of states
    $$\sum_j p_j S(\rho_j) \leq S(\sum_j p_j \rho_j) \leq S(\mathbf{p}) + \sum_j p_j S(\rho_j), \tag{22}$$
    The first equal sign achieves when all ρ_j are the same. The second equality satisfies iff ρ_j are in orthogonal subspaces. We will shortly show its proof.
- Subadditivity: $|S(A) - S(B)| \leq S(A,B) \leq S(A) + S(B)$.
- Strong Subadditivity: $S(A,B,C) + S(B) \leq S(A,B) + S(B,C)$ or equivalently $S(C|A,B) \leq S(C|B)$. The proof of strong subadditivity is not required in this lecture.

2.2 Quantum relative entropy

Suppose $\text{supp}(\sigma) \subseteq \text{supp}(\rho)$[1], we define the quantum relative entropy as

$$S(\rho\|\sigma) = \text{Tr}[\rho \log \rho] - \text{Tr}[\rho \log \sigma], \tag{23}$$

otherwise $S(\rho\|\sigma) = \infty$.

Three important theorem of the quantum relative entropy are

Theorem 1. *The quantum relative entropy is always non-negative*

$$S(\rho\|\sigma) \geq 0 \tag{24}$$

with equal sign iff $\rho = \sigma$.

Theorem 2. *The quantum relative entropy is monotonic under quantum operations*

$$S(\rho\|\sigma) \geq S(\mathcal{E}(\rho)\|\mathcal{E}(\sigma)) \tag{25}$$

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

Theorem 3. *The quantum relative entropy is jointly convex*

$$S(\sum_j p_j \rho \| \sum_j p_j \sigma) \leq \sum_j p_j S(\rho_j \| \sigma_j) \tag{26}$$

Interestingly, even though monotonicity or joint convexity are not easy to prove, we can prove one of them using the other theorem quite easily.

Quantum relative entropy is important since it defines all other entropies.

[1]The support of a hermitian matrix A is defined to be the vectors $|\psi\rangle$ satisfying $\langle v|A|v\rangle \neq 0$

- Von Neumann entropy: consider $\sigma = \mathbb{I}_d/d$, we have $S(\rho\|\mathbb{I}_d/d) = \log d - S(\rho)$ or equivalently

$$S(\rho) = \log d - S(\rho\|\mathbb{I}_d/d).$$

- Conditional entropy: for ρ^{AB}, consider $\sigma^{AB} = \mathbb{I}_d^A/d \otimes \rho^B$, we have $S(\rho^{AB}\|\sigma^{AB}) = -\mathrm{Tr}[\rho^{AB}\log(\mathbb{I}_d^A/d \otimes \rho^B)] - S(\rho^{AB}) = \log d + S(\rho^B) - S(\rho^{AB}) = \log d - S(A|B)$ or equivalently

$$S(A|B) = \log d - S(\rho^{AB}\|\mathbb{I}_d^A/d \otimes \rho^B).$$

- Mutual information: for ρ^{AB}, consider $\sigma^{AB} = \rho^A \otimes \rho^B$, we have $S(\rho^{AB}\|\sigma^{AB}) = -\mathrm{Tr}[\rho^{AB}\log(\rho^A \otimes \rho^B)] - S(\rho^{AB}) = S(\rho^A) + S(\rho^B) - S(\rho^{AB}) = S(A:B)$ or equivalently

$$S(A:B) = S(\rho^{AB}\|\rho^A \otimes \rho^B).$$

Using the above definitions, we can now prove several interesting properties such as Eq. (22), subadditivity, and strong subadditivity.

First, the second inequality of subadditivity could be easily shown by noticing that $S(A:B) = S(A) + S(B) - S(A,B) = S(\rho^{AB}\|\rho^A \otimes \rho^B) \geq 0$. For the first one, we only need to consider a purification ψ^{ABC} of ρ^{AB}, apply $S(C) + S(A) \geq S(A,C)$, and notice $S(C) = S(A,B)$, $S(A,C) = S(B)$.

To prove the first inequality of Eq. (22), we first consider the state $\rho^{AB} = \sum_j p_j |j\rangle\langle j| \otimes \rho_j$, then $S(A) = S(\mathbf{p})$ and $S(B) = S(\sum_j p_j\rho_j)$, $S(A,B) = S(\mathbf{p}) + \sum_j p_j S(\rho_j)$. Applying subadditivity, we have $S(A) + S(B) - S(A,B) = S(\sum_j p_j\rho_j) - \sum_j p_j S(\rho_j) \geq 0$. We can similarly prove it using the joint convexity of quantum relative entropy. The proof of the second inequality and strong subadditivity could be found from Nielsen & Chuang's book.

2.3 Schumachers quantum noiseless channel coding theorem

The Von Neumann entropy plays a similar role for quantum state compression. Schumachers quantum noiseless channel coding theorem guarantees that for any $\varepsilon, \delta > 0$, we can choose a sufficiently large n, an encoding channel $\mathcal{E}$ such that the dimension of $\mathcal{E}(\rho^{\otimes n})$ is no larger than $2^{n(S(\rho)+\varepsilon)}$, and a decoding channel $\mathcal{D}$ such that

$$F(\mathcal{D}\circ\mathcal{E}(\rho^{\otimes n}), \rho^{\otimes n}) \geq 1 - \delta, \tag{27}$$

for any state ρ. We can thus compress n copies of ρ into $n(S(\rho)+\varepsilon)$ qubits and recover them with negligible error. The proof is very similar to the one for Shannon's noiseless coding theorem.