Quantum Information Theory

oliverobrien111

July 2021

1 Shannon entropy

Alternative definition is using the surprisal $\gamma(x) = -\log p(x)$ which measures how likely an event is. The Shannon entropy is the expected value of the surprisal. A good way of thinking about it is the information gained on average when you learn the value of X.

For a string chosen from the binary alphabet the number of distinct strings of length n is the binomial coefficient $\binom{n}{np}$ which is approximated by Strilings approximation $\log n! = n \log n - n$ to give

$$\log \binom{n}{np} \approx nH(p)$$

for the entropy function:

$$H(p) = -p \log p - (1-p) \log(1-p) \tag{1}$$

Therefore, the strings can be specified by a block code of $2^{nH(p)}$ letters. As $H(p) \leq 1$, the block code is shorter than the message. This reflects the fact that for large n the chance of an unlikely string like 1111111111... becomes very small, so can be left out of the block code. This generalises to n letters to give Shannon entropy:

$$H(X) = \sum_x -p(x) \log p(x) \tag{2}$$

1.1 Properties of H

$$H(X) = H(\{p(x)\}) = H(p_x)$$

Permutation invariant. If you have a π permutation of X then:

$$H(p_X \cdot \pi) = H(p_X) = -\sum_{x \in J} p(\pi(x)) \log p(\pi(x)) = -\sum_{x \in J} p(x) \log p(x)$$

$H(X) \geq 0$

1.2 Formal Proof

A typical sequence $\boldsymbol{u}$ is defined as one which for which the probability of occurrence satisfies:

$$2^{-n(H(X)+\epsilon} \leq p(\boldsymbol{u}) \leq 2^{-n(H(X)-\epsilon} \tag{3}$$

Typical sequence theorem states that the probability of getting any typical sequence can be made arbitrarily close to 1 for large enough n. Let $T_\epsilon^{(n)}$ be the set of typical sequences, then the probability of getting any typical sequence is bounded by:

$$2^{-n(H(X)+\epsilon}|T_\epsilon^{(n)}| \leq \sum_{\boldsymbol{u}\in T_\epsilon^{(n)}} p(\boldsymbol{u}) \leq 1$$

using the left hand inequality of 3. Therefore, $|T_\epsilon^{(n)}| \to 2^{nH(X)}$ as $\epsilon \to 0$.

Pick ϵ such that $R > H(X) + \epsilon$. Break set of possible sequences into typical set and its complement A_ϵ^n. Encode any value in A_ϵ^n to flag bit 0 and assign a codeword of length nR to elements in $T_\epsilon^{(n)}$ (which is possible as $|T_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon} < 2^{nR}$). Probability of failure is:

$$\sum_{\boldsymbol{u}\in A_\epsilon^n} p(\boldsymbol{u}) \tag{4}$$

which by typical sequence theorem can be made arbitrarily small (as the probability of being in $T_\epsilon^{(n)}$ tends to 1).

1.2.1 Converse

Pick a subset of the typical set S^n with $|S^n| = 2^{nR}$ with $R < H$. The probability of a sequence being in S^n is:

$$P(S^n) = \sum_{\boldsymbol{u}} p(\boldsymbol{u}) = 2^{nR}2^{-nH(X)}$$

($p(\boldsymbol{u}) = 2^{-nH(X)}$ in limit as $n \to \infty$). As $H(X) - R > 0$, this tends to 0 as n tends to infinity.

Intuitively it doesn't work as the number of sequences we missed grows exponentially as n heads to infinity.

Joint Entropy:

$$H(X,Y) = -\sum_{\boldsymbol{x},\boldsymbol{y}} p(x,y)\log p(x,y) \tag{5}$$

$$H(X,Y) = H(X) + H(Y)$$

Conditional Entropy (imagine X are the bits received over a noisy channel so this is the entropy of the source Y given the data received):

$$H(Y|X) = \sum_{\boldsymbol{x}} p(x)H(Y|X=x) = \sum_{\boldsymbol{xy}} p(x)p(y|x)\log p(y|x) \tag{6}$$

After every letter is received a new optimal code can be found that will specify the remaining string with $H(Y|X)$ bits per letter. Therefore, **chain rule**

$$H(X,Y) = H(Y|X) + H(X) \tag{7}$$

Relative Entropy (defines a sort of distance between two probability distributions but is not a metric as not symmetric and does not satisfy triangle inequality):

$$D(p||q) = \sum_{\boldsymbol{x}} p(x) \frac{p(x)}{q(x)} \tag{8}$$

only 0 for $p = q$. The above is only well-defined if $p << q$ (meaning that $q(x) = 0 \Rightarrow p(x) = 0$).

Proof of Gibbs inequality

$$A = \{x \in J, p(x) > 0\}$$

$$D(p||q) = \sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} = -\sum_{x \in A} p(x) \log \frac{q(x)}{p(x)}$$

Define a random variable to take values $\log \frac{p(x)}{q(x)}$ with probablity $p(x)$ then.

$$-D(p||q) = \mathbb{E}_p(\log \frac{q(X)}{p(X)})$$

using jensens inequality

$$-D(p||q) \leq \log \mathbb{E}_p(\frac{q(X)}{p(X)}) = \log \sum_A p(x) \frac{q(X)}{p(X)} = \log \sum_A q(x) \leq \log \sum_J q(x) = 0$$

so

$$D(p||q) \geq 0$$

Mutual information:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{9}$$

If you were to fill out a venn diagram with overlapping circles $H(X)$ and $H(Y)$ the union would be $H(X,Y)$, the intersection would be $I(X,Y)$ and the remainder of the circle $H(X)$ would be $H(X|Y)$ as once you are given one of the values there is no longer any entropy coming from its circle.

Neat result. As $D(p||q) \geq 0$ and given $q(x) = \frac{1}{|J|}$ for alphabet J,

$$D(p||q) = \sum p(x) \log \frac{p(x)}{\frac{1}{|J|}} = -H(X) + \sum p(x) \log |J|$$

$$H(X) \leq \log |J|$$

Jensen's Inequality: (for concave functions)

$$\mathbb{E}(f(X)) \leq f(\mathbb{E}(X)) \tag{10}$$

Need to learn what concavity really means as used lots in quantum part of course. Important to know that $H(X)$ is concave.

Subadditivity Prove this with $D(p||q) \geq 0$ and jensens inequality. Do it on example sheet 1.

$$H(X,Y) \geq H(X) + H(Y) \tag{11}$$

5.28 of chapter 5 of Caltech I don't understand

Can use relative entropy as a parent quantity to get all the types of entropy from above. But only if we lift the restriction of the distributions having total probablity 1. For example if we take $q(x) = 1 \forall x$ then $D(P||Q) = -H(X)$. Also, $I(X:Y) = D(p(x,y)||p(x)p(y))$ and $H(Y|X) = D(p(x,y)||p(x))$. Prove and check these on example sheet.

1.3 Shannon's Noisy Channel Theorem

For certain codewords, their images after applying the channel map will represent disjoint subsets in the asymptotic limit. The typical number of sequences that will be received is $|T_n| \approx 2^{nH(Y|X)}$, whereas the size of the range is $2^{nH(Y)}$, so the maximum achievable rate (number of bits communicated per use of channel) is $\frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{nI(Y;X)}$.

If a message of length n is sent along a channel with an error rate of p. Then roughly np bits will flip, leading to $2^{nH(p)}$ typical output strings (it would be 2^{np} strings but we have to account for the encoding reducing the number of strings). In order for this input to be accurately distinguished from any other this "sphere" of possibilities must be distinct from the rest. Therefore, there must be at least $2^{nR}2^{nH(p)}$ possible output strings. Therefore, $R \leq 1 - H(p) = C$.

Can be shown that even picking random codewords gives the optimal rate in the asymptotic limit. If we adopt the decoding method of drawing a "Hamming sphere" of radius $2^{n(H(p)+\delta}$ around the received string and looking for a codeword within this radius. We would typically expect their to be at least one or our assumption about the error in the channel is wrong/we need a bigger delta. The chance of there being two can be calculated as the fraction of space occupied by the s"sphere" is:

$$\frac{2^{n(H(p)+\delta)}}{2^n} = 2^{-n(C-\delta)}$$

so the chance of one of the $2nR$ codewords lieing there is:

$$2^{-n(C-R-\delta)}$$

As δ can be as small as we like, we can pick R as close as we want to C and this will still vanish asymptotically. APPARENTLY THE AVERAGE IS TAKEN HERE BUT I DON'T SEE WHERE.

1.4 Shannon's Noisy Channel Coding Theorem - lectures

1.4.1 Discrete Memoryless Channel (DMC)

Action of each successive uses of $\mathfrak{N}$ is identical and independent to the previous use/the noise affecting each successive inputs in uncorrelated.

$$p(u^{(n)}|x^{(n)}) = \sum_{i=1}^{n} p(u_i|x_i)$$

Might as well restrict to only considering a single use of the channel. Can write channel matrix as $p_{ij} = p(y_i|x_i)$. The channel matrix is symmetric if the rows are permutations of each other.

1.4.2 Example - Memoryless Binary Symmetric Channel (m.b.s.c)

$$J_x = \{0, 1\} = J_y$$

Flips the bit with probability p. So we need an error-correcting code, e.g. the repetition code using three bits at once.

Rate: The encoding decoding pair is said to have a rate R if $|M|$ (number of possible messages) $= 2^{nR}$ for a given number of channel uses n.

Maximum probability of error corresponding to C_n:

$$p_{err}^{(n)}(C_n) = P(\mathfrak{D}_n(Y^{(n)}) \neq m|X^{(n)} = \epsilon_n(m))$$

Achievable rate: A $R \in \mathbb{R}$ is said to be an achievable rate if there exists a sequence of codes $((C_n)_n)$ of rate R s.t. $p_{err}^{(n)} \to 0$ as $n \to \infty$.
Channel Capacity: Maximum rate of reliable transmission of information $C(\mathfrak{N}) = \sup\{R : \text{R is an achievable rate}\}$

Shannon's Theorem says that $C(\mathfrak{N}) = max_{\{p(x)\}} I(X : Y)$. Not going to do this proof as it is not very connected to the quantum proof.

For m.b.s.c

$I(X : Y) = H(Y) - H(Y|X) = H(Y) - (-(1-p)\log(1-p) - p\log p) = H(Y) - h(p) \leq \log|J_Y| - h(p) = 1 - h(p)$

Does there exist some distribution $\{p(x)\}$ for which $H(Y) = 1$ because if there is then this bound is saturated.

$$H(Y) = -\sum p(y)\log p(y)$$

$$p(y) = \sum_x p(x,y) = \sum_x p(x)p(y|x)$$

So look for $p(x)$ that makes $p(y)$ equiprobable. Try $p(x) = \frac{1}{2}$, and this indeed gives $H(Y) = 1$. Therefore $C(\mathfrak{N}) = 1 - h(p)$. (I am confused about this as the H(Y) only reaches its maximum for $p = 1/2$ when $h(p)$ is also 1 so surely this doesn't work? ask in office hours).

2 Quantum

Can associate many ensembles with the same state. You can use any $\{p_i, |\psi_i\rangle\}$ as long as $p_i \geq 0$, $\sum p_i = 1$ and $\langle\psi_i|\,|\psi_i\rangle = 1$. You can find an infinite number of these ensembles that give the same spectral decomposition (are identical) as the $|\psi_i\rangle$ need not even by orthogonal and can basically be anything as long as they have norm 1. e.g. $\rho = \frac{I}{d} = \sum_j \frac{1}{d}\,|e_j\rangle\langle e_j| = \sum_k \frac{1}{d}\,|\psi_k\rangle\langle\psi_k|$ so ensembles $\{\frac{1}{d}, |e_j\rangle\}$ and $\{\frac{1}{d}, |\psi_k\rangle\}$ are both equally valid.

Expectation of observable A in state ρ: $< A >=< A >_\rho = Tr(A\rho)$ which is a positive linear functional.

System of interest to us in the course is often a subsystem S of a composite system SE. The density matrix formalism provides a description of states of subsystems. Consider a comp. system AB then the underlying hilbert space is $\mathcal{H}_A \otimes \mathcal{H}_B$. If AB is in the state $\rho_{AB} \in \mathfrak{D}(\mathcal{H}_A \otimes \mathcal{H}_B) then state A is given by the reduce state : \rho_A = Tr_B \rho_{AB}. Consider orthonomal basis \{\ |i_A\rangle\}$ in $\mathcal{H}_A$ and $\{|\alpha_B\rangle\}$ in $\mathcal{H}_B$ then we have $\{|i_A\rangle \otimes |\alpha_B\rangle\}$ in $\mathcal{H}_A \otimes \mathcal{H}_B$. Can always write $A = \sum a_{ij}\,|i\rangle\langle j|$ with $a_{ij} = \langle i|\,A\,|j\rangle$.

$$\rho_{AB} = \sum_{i,j=1}^{d_A} \sum_{\alpha,\beta=1}^{d_B} r_{i\alpha,j\beta}\,|i_A\rangle\,|\alpha_B\rangle\,\langle j_A|\,\langle\beta_B|$$

$$\rho_A = Tr_B \rho_A B = Tr_B(\sum_{i,j=1}^{d_A} \sum_{\alpha,\beta=1}^{d_B} r_{i\alpha,j\beta}\,|i_A\rangle\,\langle j_A| \otimes |\alpha_B\rangle\,\langle\beta_B|) = \sum_{i,j=1}^{d_A} \sum_{\alpha=1}^{d_B} r_{i\alpha,j\alpha} |i_A\rangle\langle j_A|$$

2.1 Schmidt decomposition

Any state in space $H_A \otimes H_B$ can be described using coefficients of basis states of the form $|i_A\rangle \otimes |i_B\rangle$ where $|i_A\rangle$ for some set of basis eigenstates $|i_A\rangle$, $|i_B\rangle$. The schmidt rank is the number of positive Schmidt coefficents.

2.2 Purification

It is possible to convert a mixed state into a pure state by adding a purifying reference system R with Hilbert space H_R, and defining a pure state $|\psi_{AR}\rangle \in$

$H_A \otimes H_B$ such that:

$$\rho_A = Tr_R \left|\psi_{AR}\right\rangle \left\langle\psi_{AR}\right| = \sum_{i=1} \lambda_i^2 \left|i_A\right\rangle \left\langle i_A\right|$$