

Lecture 11

Entropy Quantities - Classical

We have already seen the Shannon entropy

$$H(\underline{X}) = - \sum_x p_{\underline{X}}(x) \log p_{\underline{X}}(x)$$

which can be thought of as the expectation of the information content $i(x) = -\log p_{\underline{X}}(x)$
 $\Rightarrow H(\underline{X}) = \mathbb{E}[i(\underline{X})]$. This is a quantified measure of the average "surprise" at learning the value of \underline{X} .

Properties:

- i) Positivity $H(\underline{X}) \geq 0$
- ii) Concavity $H(\lambda \underline{X}_1 + (1-\lambda) \underline{X}_2) \geq \lambda H(\underline{X}_1) + (1-\lambda) H(\underline{X}_2)$
 $0 \leq \lambda \leq 1$
- iii) Invariance under permutations: if we permute the values of \underline{X} , $H(\underline{X})$ is unchanged (it depends only on their probs, not their order).
- iv) Minimum value: $H(\underline{X}) = 0$ when \underline{X} is deterministic.
- v) Maximum value: $H(\underline{X}) = \log d$ when \underline{X} is uniform: $p_{\underline{X}}(x) = 1/d$ for all x , $|X| = d$.

From this basic function we can derive a number of important information quantities.

Joint entropy

(2)

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

When X and Y are independent, $p_{X,Y}(x,y) = p_X(x)p_Y(y)$, then $H(X, Y) = H(X) + H(Y)$. Otherwise, $H(X, Y) < H(X) + H(Y)$. (Subadditivity)

Conditional entropy

$$H(X|Y) \equiv - \sum_{x,y} p_Y(y) p_{X|Y}(x|y) \log p_{X|Y}(x|y)$$

Properties:

i) $H(X) \geq H(X|Y)$ (conditioning can only lower the entropy)

$$\text{ii) } H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

In fact, we can extend this to a chain rule:

$$H(X_1, \dots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \dots + H(X_n|X_1, \dots, X_{n-1})$$

$$\text{iii) } H(X|Y) \geq 0.$$

Mutual Information

$$I(X;Y) \equiv H(X) - H(X|Y)$$

Properties of mutual information

(3)

i) $I(X;Y) = I(Y;X)$

ii) $I(X;Y) \geq 0$.

We have already seen this quantity show up in Shannon's channel coding theorem; it can be interpreted, more or less, as the average information one gains about X from learning Y .

Relative Entropy

This quantity is like a "distance" between two prob. densities $p_{\Sigma_1}(x)$ and $p_{\Sigma_2}(x)$ — it vanishes when they are the same. But it is not a true distance measure, because it is not symmetric in the two distributions:

$$\begin{aligned} D(p_{\Sigma_1} \| p_{\Sigma_2}) &= \sum_x p_{\Sigma_1}(x) \log \left(\frac{p_{\Sigma_1}(x)}{p_{\Sigma_2}(x)} \right) \\ &= \sum_x p_{\Sigma_1}(x) (\log p_{\Sigma_1}(x) - \log p_{\Sigma_2}(x)) \\ &= \sum_x p_{\Sigma_1}(x) \log p_{\Sigma_2}(x) - H(\Sigma_1). \end{aligned}$$

$$D(p_{\Sigma_1} \| p_{\Sigma_2}) \geq 0 \text{ (not obvious!)}$$

$D(p_{\Sigma_1} \| p_{\Sigma_2})$ can diverge if $p_{\Sigma_2}(x) = 0$ while $p_{\Sigma_1}(x) \neq 0$. This actually has an interpretation in source coding, where $D(p_{\Sigma_1} \| p_{\Sigma_2})$ is the coding "penalty" one pays for using an incorrect distribution $p_{\Sigma_2}(x)$ for the correct one $p_{\Sigma_1}(x)$.

$D(P_{\bar{X}} \parallel P_{\bar{X}Y})$ is also related to the mutual information:

$$\begin{aligned} I(X; Y) &= D(P_{\bar{X}Y}(x, y) \parallel P_{\bar{X}}(x) P_Y(y)) \\ &= \sum_{xy} P_{\bar{X}Y}(x, y) \log \left(\frac{P_{\bar{X}Y}(x, y)}{P_{\bar{X}}(x) P_Y(y)} \right) \end{aligned}$$

$I(X; Y)$ is the "distance" of the density $P_{\bar{X}Y}(x, y)$ from the independent density $P_{\bar{X}}(x) P_Y(y)$ with the same marginals $P_{\bar{X}}(x)$ & $P_Y(y)$.

Conditional Mutual Information

We can condition both X and Y on a third random variable Z :

$$\begin{aligned} I(\bar{X}; Y|Z) &\equiv H(\bar{X}|Z) - H(\bar{X}|Y, Z) \\ &= H(Y|Z) - H(Y|\bar{X}, Z) \\ &= H(\bar{X}|Z) + H(Y|Z) - H(\bar{X}, Y|Z). \end{aligned}$$

i) $\bullet I(\bar{X}; Y|Z) \geq 0$. This is a condition called strong subadditivity. It is easy to prove classically; there is a quantum version that is quite hard to prove.

ii) $I(\bar{X}; Y|Z) = 0$ if X & Y are conditionally independent: $P_{\bar{X}Y|Z}(x, y|z) = P_{\bar{X}|Z}(x|z) P_{Y|Z}(y|z)$

Information Inequalities

These are important bounds and restrictions on these various entropic quantities:

I. The Fundamental Information Inequality (positivity of relative entropy).

Thm $D(P_{\Sigma} \| P_{\Sigma_2}) \geq 0$.

$$\begin{aligned}
 \text{Proof: } D(P_{\Sigma} \| P_{\Sigma_2}) &= \sum_x P_{\Sigma}(x) \log \left(\frac{P_{\Sigma}(x)}{P_{\Sigma_2}(x)} \right) \\
 &= -\frac{1}{\ln 2} \sum_x P_{\Sigma}(x) \ln \left(\frac{P_{\Sigma_2}(x)}{P_{\Sigma}(x)} \right) \\
 &\geq \frac{1}{\ln 2} \sum_x P_{\Sigma}(x) \left(1 - \frac{P_{\Sigma_2}(x)}{P_{\Sigma}(x)} \right) \quad \begin{matrix} \text{use} \\ \cancel{\text{if } x < 0} \end{matrix} \quad \begin{matrix} \ln x \leq x-1 \\ \forall x > 0. \end{matrix} \\
 &= \frac{1}{\ln 2} \left(\sum_x P_{\Sigma}(x) - \sum_x P_{\Sigma_2}(x) \right) \\
 &= \frac{1}{\ln 2} (1 - 1) = 0. \quad \square
 \end{aligned}$$

This can be used to simply prove a number of results, including $I(X; Y) \geq 0$ and $H(X) \geq H(X|Y)$.

II. The Data-Processing Inequality

Intuitively, this simply states that one cannot create new information (about a random variable) by processing existing information. Info can be lost by processing, but not amplified.

(6)

Thm Suppose we have 2 maps N_1 and N_2 given by $N_1 \equiv P_{Y|X}(y|x)$ and $N_2 \equiv P_{Z|Y}(z|y)$. (I.e., we pass X through N_1 to get Y , and pass Y through N_2 to get Z . This is a Markov chain.)

Then $\boxed{I(X; Y) \geq I(X; Z)}$

Proof: $P_{X, Z|Y}(x, z|y) = P_{Z|Y, X}(z|y, x) P_{X|Y}(x|y)$
 $= P_{Z|Y}(z|y) P_{X|Y}(x|y)$

So X & Z are conditionally independent through Y . This implies

$$I(X; Y, Z) = I(X; Y) + I(X; Z|Y) \stackrel{0}{=} I(X; Y).$$

But also

$$I(X; Y, Z) = I(X; Z) + I(X; Y|Z)$$

Since this \uparrow is $I(X; Y)$ and this \uparrow is ≥ 0 , then

$$I(X; Y) \geq I(X; Z). \quad \square$$

III. Fano's Inequality

This concerns the situation where X passes through a channel N to produce Y , and then Y is processed to produce \hat{X} an estimate of X . Then $p_e \equiv \text{prob}(X \neq \hat{X})$.

②

Thm $H(X|Y) \leq h(p_e) + p_e \log(|X|-1)$,

where $h(p)$ is the binary entropy

$$h(p) = -p \log p - (1-p) \log(1-p)$$

and X is the set of possible values of \hat{X} .

Proof: Let $E = \begin{cases} 0 & \hat{X} = \bar{X} \\ 1 & \hat{X} \neq \bar{X} \end{cases}$

$$\text{Then } H(E, \hat{X} | \bar{X}) = H(\hat{X} | \bar{X}) + H(E | \hat{X}, \bar{X}).$$

By the data processing inequality,

$$I(\hat{X}; Y) \geq I(\hat{X}; \bar{X}) \Rightarrow H(\hat{X} | \bar{X}) \geq H(\hat{X} | Y).$$

This implies

$$\begin{aligned} H(E, \hat{X} | \bar{X}) &= H(E | \bar{X}) + H(\hat{X} | E, \bar{X}) \\ &\leq H(E) + H(\hat{X} | E, \bar{X}) \\ &= h(p_e) + p_e H(\hat{X} | \bar{X}, E=1) \\ &\quad + (1-p_e) H(\hat{X} | \bar{X}, E=0) \\ &\leq h(p_e) + p_e \log(|X|-1). \end{aligned}$$

□

Quantum Entropy

③

We want to transfer as many of these concepts as we can from classical prob. densities and channels to quantum states and channels.

For a starting point: what is the Shannon entropy of a POVM?

$$p_{\bar{x}}(x) = \text{Tr}\{\Lambda_x \rho\} \quad 0 \leq \Lambda_x \leq I$$

$$\sum_x \Lambda_x = I.$$

$$H(\bar{x}) = - \sum_x \text{Tr}\{\Lambda_x \rho\} \log(\text{Tr}\{\Lambda_x \rho\}).$$

We can use this to define two classical/quantum info quantities:

A. Accessible information.

consider an ensemble $\{p_{\bar{x}}(x), p_x\} \in \mathcal{E}$

$$\Rightarrow \rho = \sum_x p_{\bar{x}}(x) \rho_x.$$

By doing a POVM $\{\Lambda_y\}$, how much can we learn about x ?

$$I_{\text{acc}}(\mathcal{E}) = \max_{\{\Lambda_y\}} I(x; Y)$$

where

$$P_{Y|\bar{x}}(y|x) = \text{Tr}\{\Lambda_y \rho_x\}.$$

B. Classical mutual information

$$I_c(\rho_{AB}) = \max_{\{\Lambda_x^A\}, \{\Lambda_y^B\}} I(x; Y)$$

$$\text{where } P_{\bar{x}, Y}(x, y) = \text{Tr}\{(\Lambda_x^A \otimes \Lambda_y^B) \rho^{AB}\}.$$