EE2340/EE5847: Information Sciences/Information Theory **2020**

Handout 5: Properties of Information Measures 2

Instructor: Shashank Vatedka

5.1 Convexity properties of information measures

The log-sum inequality that we studied in the last class will prove helpful in studying the convexity properties of information measures.

Let us first show that the KL divergence is convex.

Lemma 5.1. *$D(p\|q)$ is convex in (p, q).*

Proof. We need to show that for every $\alpha \in [0, 1]$ and pmfs (p_1, q_1) and (p_2, q_2),

$$D(\alpha p_1 + (1-\alpha)p_2 \| \alpha q_1 + (1-\alpha)q_2) \leqslant \alpha D(p_1\|q_1) + (1-\alpha)D(p_2\|q_2).$$

However,

$$\begin{aligned}
D(\alpha p_1 + (1-\alpha)p_2 \| \alpha q_1 + (1-\alpha)q_2) &= \sum_x \left((\alpha p_1(x) + (1-\alpha)p_2(x)) \log_2 \frac{(\alpha p_1(x) + (1-\alpha)p_2(x))}{(\alpha q_1(x) + (1-\alpha)q_2(x))} \right) \\
&\leqslant \sum_x \left(\alpha p_1(x) \log_2 \frac{\alpha p_1(x)}{\alpha q_1(x)} + (1-\alpha)p_2(x) \log_2 \frac{(1-\alpha)p_2(x)}{(1-\alpha)q_2(x)} \right) \\
&= \alpha \left(\sum_x p_1(x) \log_2 \frac{p_1(x)}{q_1(x)} \right) + (1-\alpha) \left(\sum_x p_2(x) \log_2 \frac{p_2(x)}{q_2(x)} \right) \\
&= \alpha D(p_1\|q_1) + (1-\alpha)D(p_2\|q_2)
\end{aligned}$$

where in the second step, we have used the log-sum inequality. ☐

Corollary 5.2. *The entropy $H(X)$ is a concave function of p_X.*

Proof. Let u denote the uniform distribution on $\mathcal{X}$ and p_1, p_2 be two distributions.

The trick is to write entropy as a KL divergence: $H(p_X) = \log|\mathcal{X}| - D(p_X\|u)$ (Verify!).

$$\begin{aligned}
H(\alpha p_1 + (1-\alpha)p_2) &= \log|\mathcal{X}| - D(\alpha p_1 + (1-\alpha)p_2 \| \alpha u + (1-\alpha)u) \\
&\geqslant \log|\mathcal{X}| - \alpha D(p_1\|u) - (1-\alpha)D(p_2\|u) \\
&= \alpha \left(\log|\mathcal{X}| - D(p_1\|u)\right) + (1-\alpha)\left(\log|\mathcal{X}| - D(p_2\|u)\right) \\
&= \alpha H(p_1) + (1-\alpha)H(p_2),
\end{aligned}$$

where in the second step we have used convexity of KL divergence. ☐

The above corollary means that maximizing the entropy is not a hopeless task. One can use any convex solver to do this.

Lemma 5.3. *$I(X;Y)$ is a concave function of p_X for fixed $p_{Y|X}$. It is a convex function of $p_{Y|X}$ for fixed p_X.*

Proof. To prove the first part, observe that

$$I(X;Y) = H(Y) - H(Y|X) = H(\sum_x p_{Y|X}(y|x)p_X(x)) + \sum_y \sum_x p_{Y|X}(y|x)p_X(x)\log_2 p_{Y|X}(y|x).$$

The first term is a concave function of p_X (from the previous corollary). The second term is linear in p_X for a fixed $p_{Y|X}$. Therefore, $I(X;Y)$ is concave in p_X for fixed $p_{Y|X}$.

To prove the second part, consider $p_{Y|X}$ and $q_{Y|X}$ and p_X. We have,

$$p_{XY}(x,y) = p_{Y|X}(y|x)p_X(x), \qquad q_{XY}(x,y) = q_{Y|X}(y|x)p_X(x)$$

and

$$p_Y(y) = \sum_x p_{Y|X}(y|x)p_X(x), \qquad q_Y(y) = \sum_x q_{Y|X}(y|x)p_X(x).$$

The mutual information can be written as

$$f(p_{Y|X}, p_X) = D(p_{XY} \| p_X p_Y).$$

Therefore, by linearity

$$f(\alpha p_{Y|X} + (1-\alpha)q_{Y|X}, p_X) = D(\alpha p_{XY} + (1-\alpha)q_{XY} \| \alpha p_X p_Y + (1-\alpha)p_X q_Y)$$

Using the property that KL divergence is convex,

$$f(\alpha p_{Y|X} + (1-\alpha)q_{Y|X}, p_X) \leqslant \alpha f(p_{Y|X}, p_X) + (1-\alpha)f(q_{Y|X}, p_X),$$

thus completing the proof. □

The above lemma implies that we can maximize $I(X;Y)$ with respect to p_X (as we usually do to find channel capacity) implying that there is a "best" distribution for a channel, and minimize $I(X;Y)$ with respect to $p_{Y|X}$ (implying that there is a "worst" channel for an input distribution).

5.2 Data processing inequality

We say that X, Y, Z form a Markov chain, denoted $X - Y - Z$ if

$$p_{XYZ}(x,y,z) = p_X(x)p_{Y|X}(y|x)p_{Z|Y}(z|y).$$

- If Y denotes the encoding of X and Z is obtained by passing Y through a noisy channel, then $X - Y - Z$.
- Likewise, if Y is obtained by passing X through a channel and Z is obtained by decoding from Y, then $X - Y - Z$.
- $X - Y - Z$ also implies that

  $$p_{XYZ}(x,y,z) = p_Z(z)p_{Y|Z}(y|z)p_{X|Y}(x|y).$$

  To show the above, it suffices to prove that $p_{X|YZ}(x|yz) = p_{X|Y}(x|y)$. This can be shown using Bayes rule and the definition of $X - Y - Z$.

- If $X - Y - Z$, then X and Z are conditionally independent given Y, implying that

$$I(X; Z|Y) = 0.$$

Lemma 5.4 (Data processing inequality)**.** *If $X - Y - Z$, then $I(X;Y) \geqslant I(X;Z)$. In other words, further processing cannot increase mutual information.*

Proof. Using the chain rule of mutual information,

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

However,

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

since X, Z are conditionally independent given Y. Combining the above equations,

$$I(X; Y) \geqslant I(X; Z).$$

□

- The data processing inequality reveals that no amount of preprocessing or postprocessing may increase the capacity of a noisy channel.

5.3 Fano's inequality

Suppose that $X, \hat{X}, Y$ are jointly distributed random variables, where $\hat{X}$ is to be interpreted as an estimate of X from Y. More precisely, $X - Y - \hat{X}$ forms a Markov chain. For example, X could be a message, Y the received vector, and $\hat{X}$ the decoder's estimate of X. Define the probability of error

$$P_e \stackrel{\text{def}}{=} \Pr[X \neq \hat{X}].$$

Lemma 5.5 (Fano's inequality)**.**

$$H(X|Y) \leqslant H(X|\hat{X}) \leqslant H_2(P_e) + P_e \log_2 |\mathcal{X}|.$$

Proof. Let us define a new random variable

$$E = \begin{cases} 1, & \text{if } X \neq \hat{X} \\ 0, & \text{if } X = \hat{X}. \end{cases}$$

Using the chain rule of entropy,

$$H(X|\hat{X}) = H(X, E|\hat{X}) - H(E|X, \hat{X}) = H(X, E|\hat{X})$$

where the last step follows from the fact that E is a function of $(X, \hat{X})$ and therefore $H(E|X, \hat{X}) = 0$. Using the chain rule on $H(X, E|\hat{X})$, we have

$$H(X|\hat{X}) = H(E|\hat{X}) + H(X|E, \hat{X}) \leqslant H(E) + H(X|E, \hat{X}) = H_2(P_e) + H(X|E, \hat{X}).$$

The inequality above follows from the property that conditioning reduces entropy.

$$\begin{aligned} H(X|\hat{X}) &\leqslant H_2(P_e) + H(X|\hat{X}, E) \\ &= H_2(P_e) + H(X|\hat{X}, E = 0)\Pr[E = 0] + H(X|\hat{X}, E = 1)\Pr[E = 1] \\ &= H_2(P_e) + 0 \times \Pr[E = 0] + H(X|\hat{X}, E = 1)\Pr[E = 1] \end{aligned}$$

where the last step follows from the fact that if $E = 1$, then $\hat{X} = X$ and therefore $H(X|\hat{X}, E = 0) = 0$.

$$\begin{aligned} H(X|\hat{X}) &\leqslant H_2(P_e) + H(X|\hat{X}, E = 1)\Pr[E = 1] \\ &= H_2(P_e) + H(X|\hat{X}, E = 1)P_e \\ &\leqslant H_2(P_e) + H(X)P_e \\ &\leqslant H_2(P_e) + \log_2|\mathcal{X}| \times P_e \end{aligned} \tag{5.1}$$

Since $X - Y - \hat{X}$,

$$I(X; \hat{X}) \leqslant I(X; Y)$$

by the data processing inequality. By the definition of mutual information,

$$H(X) - H(X|\hat{X}) \leqslant H(X) - H(X|Y)$$

or

$$H(X|\hat{X}) \geqslant H(X|Y)$$

Using the above in (5.1),

$$H(X|Y) \leqslant H(X|\hat{X}) \leqslant H_2(P_e) + P_e \log_2|\mathcal{X}|,$$

completing the proof. □