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

Handout 4: Properties of the information measures - 1

Instructor: Shashank Vatedka

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. Please email the course instructor in case of any errors.*

4.1 Recap

- The minimum rate for fixed-length compression of a memoryless source with distribution p_X is equal to the entropy $H(X)$.
- The maximum rate of reliable communication over a DMC $p_{Y|X}$ is equal to the capacity $C = \max_{p_X} I(X;Y)$.
- The optimal error exponent for classifying between two sources p_s, p_g is equal to the KL divergence $D(p_s \| p_g)$.

The material covered in this handout is contained in Chapter 2 of the book by Cover and Thomas.

4.2 Relook of the information measures

Verify that we can write

$$\begin{aligned}
H(X) &= \mathbb{E}_X \log_2 \frac{1}{p_X(X)} \\
H(X,Y) &= \mathbb{E}_{XY} \log_2 \frac{1}{p_{XY}(X,Y)} \\
H(X|Y) &= \mathbb{E}_{XY} \log_2 \frac{1}{p_{X|Y}(X|Y)} \\
I(X;Y) &= \mathbb{E}_{XY} \log_2 \frac{p_{XY}(X,Y)}{p_X(X)p_Y(Y)}.
\end{aligned}$$

We have already seen in previous classes that

$$I(X;Y) = H(X) - H(X|Y).$$

4.3 Properties

4.3.1 Properties

Lemma 4.1. *Entropy is nonnegative.*

This follows from the fact that $-x \log x > 0$ for all $x \in (0,1)$.

4.3.2 Chain rules

In the homework, you will show that

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y). \tag{4.1}$$

and

$$H(X,Y|Z) = H(X|Z) + H(Y|X,Z). \tag{4.2}$$

This will be useful in proving a number of chain rules.

4.3.2.1 More variables

$$H(X_1,\dots,X_n) = \sum_{i=1}^{n} H(X_i|X_1,\dots,X_{i-1}). \tag{4.3}$$

This can be proved by repeatedly using (4.1) and (4.2).

$$\begin{aligned}
H(X_1,\dots,X_n) &= H(X_1) + H(X_2,\dots,X_n|X_1) && (4.4)\\
&= H(X_1) + H(X_2|X_1) + H(X_3,\dots,X_n|X_1,X_2) && (4.5)\\
&\vdots && (4.6)\\
&= H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_1,\dots X_{n-1}). && (4.7)
\end{aligned}$$

4.3.2.2 Chain rule of mutual information

The conditional mutual information is defined as

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z).$$

We can prove the following chain rule

$$I(X_1,\dots,X_n;Y) = \sum_{i=1}^{n} I(X_i;Y|X_1,\dots,X_{i-1}) \tag{4.8}$$

This follows by repeatedly using the chain rule of entropy for each of the conditional entropy terms.

$$\begin{aligned}
I(X_1,X_2;Y) &= H(X_1,X_2) - H(X_1,X_2|Y)\\
&= H(X_1) + H(X_2|X_1) - H(X_1|Y) - H(X_2|X_1,Y)\\
&= I(X_1;Y) + I(X_2;Y|X_1).
\end{aligned}$$

4.3.2.3 Chain rule for KL divergence

We can prove something similar here as well: For any pair of joint distributions p_{XY} and q_{XY} such that $p_{XY}(x,y)=0$ whenever $q_{XY}(x,y)=0$, we have

$$D(p_{XY}\|q_{XY}) = D(p_X\|q_X) + D(p_{Y|X}\|q_{Y|X}),$$

where we define

$$D(p_{Y|X}\|q_{Y|X}) \overset{\text{def}}{=} \sum_{x,y} p_{XY}(x,y)\log_2\frac{p_{XY}(x,y)}{q_{XY}(x,y)}.$$

This follows from definition.

$$\begin{aligned}
D(p_{XY}\|q_{XY}) &= \sum_{x,y} p_{XY}(x,y)\log_2 \frac{p_{XY}(x,y)}{q_{XY}(x,y)} \\
&= \sum_{x,y} p_{XY}(x,y)\log_2 \frac{p_X(x)p_{Y|X}(y|x)}{q_X(x)q_{Y|X}(y|x)} \\
&= \sum_{x,y} p_{XY}(x,y)\log_2 \frac{p_X(x)}{q_X(x)} + \sum_{x,y} p_{XY}(x,y)\log_2 \frac{p_{Y|X}(y|x)}{q_{Y|X}(y|x)} \\
&= \sum_{x} p_X(x)\log_2 \frac{p_X(x)}{q_X(x)} + \sum_{x,y} p_{XY}(x,y)\log_2 \frac{p_{Y|X}(y|x)}{q_{Y|X}(y|x)} \\
&= D(p_X\|q_X) + D(p_{Y|X}\|q_{Y|X}).
\end{aligned}$$

4.4 Convex sets and functions

A set $\mathcal{S} \subset \mathbb{R}^m$ is said to be convex if the line segment joining any two points within $\mathcal{S}$ also lies in $\mathcal{S}$. Formally, $\mathcal{S}$ is convex if

$$\alpha x_1 + (1-\alpha)x_2 \in \mathcal{S} \qquad \text{for all } x_1, x_2 \in \mathcal{S} \text{ and } \alpha \in [0,1].$$

Note that every closed interval $[a,b] \subset \mathbb{R}$ is convex.

A function defined on a convex set $\mathcal{S}$, $f : \mathcal{S} \to \mathbb{R}$ is convex if for all $x, y \in \mathcal{S}$ and $\alpha \in [0,1]$, we have

$$f(\alpha x + (1-\alpha)y) \leqslant \alpha f(x) + (1-\alpha)f(y).$$

The function is strictly convex if equality holds only for $\alpha = 0, 1$.

A function f is concave if $-f$ is convex.

Lemma 4.2. *A twice-differentiable function $f : \mathbb{R} \to \mathbb{R}$ has nonnegative second derivative on $[a,b]$ if and only if it is convex in $[a,b]$. If the second derivative is strictly positive in the interval, then it is also strictly convex and vice versa.*

Proof. To show that the second derivative being nonnegative implies convexity, see Theorem 2.6.1 in Cover and Thomas.

For the other way around, recall from your calculus course that

$$f''(x) = \lim_{t\to 0} \frac{f(x+t) + f(x-t) - 2f(x)}{t^2}.$$

It therefore suffices to show that $\frac{f(x+t)+f(x-t)-2f(x)}{t^2} \geqslant 0$ for all $t > 0$.

Now since f is convex,

$$f(x) = f\left(\frac{x+t}{2} + \frac{x-t}{2}\right) \leqslant \frac{1}{2}f(x+t) + \frac{1}{2}f(x-t).$$

Rearranging, we get that

$$f(x+t) + f(x-t) - 2f(x) \geqslant 0.$$

This completes the proof. If the function is strictly convex, then the inequality is strict, and therefore the second derivative is positive. □

A very useful inequality for convex functions is the following:

Lemma 4.3 (Jensen's inequality)**.** *For any convex function $f : \mathbb{R} \to \mathbb{R}$ and random variable X, we have*

$$\mathbb{E}f(x) \geqslant f(\mathbb{E}X).$$

If f is strictly convex, then equality implies that X is a constant.

Proof. The Cover and Thomas book gives a proof specific to discrete random variables using mathematical induction. Here is a shorter proof:

Let $c + \alpha x$ denote the tangent to $f(x)$ at the point $\mathbb{E}X$. I claim that the following is true (Why?):

$$c + \alpha x \leqslant f(x).$$

Using this,

$$\mathbb{E}f(X) \geqslant \mathbb{E}(a + bX) = a + b\mathbb{E}X \tag{4.9}$$

However, the line intersects $f(x)$ for $x = \mathbb{E}X$, and hence $a + \mathbb{E}X = f(\mathbb{E}X)$. This completes the proof.

Equality in (4.9) implies that $f(x) = a + bx$ for all x having nonzero probability (or nonzero density for continuous rvs). If the random variable is not a constant, then it means that the second derivative of f is zero and hence it is not strictly convex. □

Jensen's inequality is the basis for many other results in information theory and statistics.

4.4.0.1 Nonnegativity of KL divergence

Lemma 4.4. *$D(p\|q) \geqslant 0$ with equality iff $p(x) = q(x)$ for all x such that $p(x) > 0$.*

Proof. Let $\mathcal{S}$ denote the set of all x such that $p(x) > 0$. This is called the support of p. The main observation here is that $\log_2(x)$ is a concave function of x, and we can use Jensen's inequality. To start with, consider

$$-D(p\|q) = -\sum_{x\in\mathcal{S}} p(x)\log_2\frac{p(x)}{q(x)} = \sum_{x\in\mathcal{S}} p(x)\log_2\frac{q(x)}{p(x)}$$

Note that the last term can be written as $\mathbb{E}_p \log_2 \frac{q(X)}{p(X)}$, where $X \sim p$. Using Jensen's inequality,

$$\begin{aligned}
-D(p\|q) &\leqslant \log_2\left(\sum_{x\in\mathcal{S}} p(x)\frac{q(x)}{p(x)}\right) \\
&= \log_2\left(\sum_{x\in\mathcal{S}} q(x)\right) \\
&= \log_2(1) = 0.
\end{aligned}$$

Rearranging, we get $D(p\|q) \geqslant 0$. □

4.4.0.2 Implications of nonnegativity of KL divergence

- $$I(X;Y) \geqslant 0.$$
  And equality holds in the above iff X, Y are independent. This follows from the fact that $I(X;Y) = D(p_{XY} \| p_X p_Y)$.

- $$D(p_{Y|X} \| q_{Y|X}) \geqslant 0$$
  with equality iff $p_{Y|X}(y|x) = q_{Y|X}(y|x)$ for all (x, y) such that $p_{Y|X}(y|x) > 0$.

- $$I(X;Y|Z) \geqslant 0$$

- **Conditioning reduces entropy**
  $$H(Y|X) \leqslant H(Y)$$
  with equality iff X and Y are independent. This follows from $I(X;Y) = H(Y) - H(Y|X)$.

- $$H(X) \leqslant \log_2 |\mathcal{X}|.$$
  Consider $q(x) = \frac{1}{|\mathcal{X}|}$ for all x and $p(x) = p_X(x)$. Simplifying $D(p\|q)$, we get
  $$0 \leqslant D(p\|q) = \log_2 |\mathcal{X}| - H(X).$$

- $$H(X_1, \ldots, X_n) \leqslant \sum_{i=1}^{n} H(X_i).$$

4.4.0.3 Log-sum inequality

Lemma 4.5. *Suppose that $\alpha_1, \ldots, \alpha_k$ and $\beta_1, \ldots, \beta_k$ are nonnegative numbers such that $\alpha_i > 0$ whenever $\beta_i > 0$. Then,*
$$\sum_i \alpha_i \log_2 \frac{\alpha_i}{\beta_i} \geqslant \left(\sum_i \alpha_i \right) \log_2 \frac{(\sum_i \alpha_i)}{(\sum_i \beta_i)}$$
and equality holds if and only if $\alpha_i = \beta_i$ for all i such that $\alpha_i > 0$.

Proof. We will again use Jensen's inequality on the strictly convex function $x \log x$. Start with the LHS.
$$\begin{aligned}
\sum_{i=1}^{k} \alpha_{i=1}^{k} \log_2 \frac{\alpha_i}{\beta_i} &= \sum_{i=1}^{k} \beta_i \frac{\alpha_i}{\beta_i} \log_2 \frac{\alpha_i}{\beta_i} \\
&= (\sum_{j=1}^{k} \beta_j) \left(\sum_{i=1}^{k} \frac{\beta_i}{\sum_{j=1}^{k} \beta_j} \frac{\alpha_i}{\beta_i} \log_2 \frac{\alpha_i}{\beta_i} \right)
\end{aligned}$$

If we set $p(i) = \beta_i / \sum_{j=1}^{k} \beta_j$, and $t_i = \alpha_i / \beta_i$, then the above becomes
$$\sum_{i=1}^{k} \alpha_{i=1}^{k} \log_2 \frac{\alpha_i}{\beta_i} = (\sum_{j=1}^{k} \beta_j) \sum_{i=1}^{k} p(i)(t_i \log_2 t_i)$$

$$
\begin{aligned}
&\geqslant (\sum_{j=1}^{k} \beta_j)(\sum_{i=1}^{k} p(i)t_i) \log_2(\sum_{i=1}^{k} p(i)t_i) \\
&\geqslant (\sum_{j=1}^{k} \beta_j) \left(\sum_{i=1}^{k} \frac{\beta_i}{\sum_{j=1}^{k} \beta_j} \frac{\alpha_i}{\beta_i} \right) \log_2 \left(\sum_{i=1}^{k} \frac{\beta_i}{\sum_{j=1}^{k} \beta_j} \frac{\alpha_i}{\beta_i} \right) \\
&= \left(\sum_i \alpha_i \right) \log_2 \frac{(\sum_i \alpha_i)}{\left(\sum_j \beta_j\right)}
\end{aligned}
$$

□

This can be generalized easily to the continuous case by replacing summations with integrals in the proofs (assuming that they all exist).

Lemma 4.6. *Let* f, g *be nonnegative integrable functions on* $\mathbb{R}$ *with* $\int_{x=-\infty}^{\infty} f(x)dx > 0$ *and* $\int_{x=-\infty}^{\infty} g(x)dx > 0$*. Additionally,* $f(x) = 0$ *whenever* $g(x) = 0$*. Then,*

$$
\int_{x=-\infty}^{\infty} f(x) \log_2 \frac{f(x)}{g(x)} dx \geqslant \left(\int_{x=-\infty}^{\infty} f(x)dx \right) \log_2 \frac{\left(\int_{x=-\infty}^{\infty} f(x)dx \right)}{\left(\int_{x=-\infty}^{\infty} g(x)dx \right)}.
$$