Unit 6: Properties of measures of information - 1

A Measures of information and their conditional variants

The cast of characters for our course is complete. This course will be about the following quantities:[1]

1. *Shannon Entropy.* For a discrete distribution P on $\mathcal{X}$, the entropy $H(P)$ is given by

$$H(P) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)}.$$

   We denote $H(P)$ by $H(X)$ when convenient.

2. *Joint and Conditional Entropy.* For (X, Y) with distribution P_{XY}, the joint entropy of (X, Y) is given by $H(X, Y) = H(\mathrm{P}_{XY})$. The conditional entropy of X given Y, denoted $H(X|Y)$ or $H(\mathrm{P}_{XY}|\mathrm{P}_Y)$, is given by $H(X|Y) = H(X, Y) - H(Y)$.

3. *Mutual Information.* Given a joint distribution P_{XY}, the mutual information between X and Y is given by $I(X \wedge Y) = H(X) - H(X|Y)$. We will use an alternative definition notation where we represent the mutual information as a function of the input distribution P_X and the channel $W = \mathrm{P}_{Y|X}$. Namely, we represent mutual information $I(X \wedge Y)$ as $I(P; W)$.

[1] We denote the pmf of distributions P and Q by p and q, respectively. All the logarithms in this course are to the base 2, and we shall follow the convention $0 \log 0 = 0$.

4. *Kullback-Leibler divergence.* Given discrete distributions P and Q on $\mathcal{X}$, the divergence $D(P\|Q)$ between P and Q is given by

$$D(P\|Q) = \begin{cases} \sum_{x\in\mathcal{X}} p(x)\log\frac{q(x)}{q(x)}, & \mathtt{supp}(P) \subset \mathtt{supp}(Q) \\ \infty, & \mathtt{supp}(P) \subsetneq \mathtt{supp}(Q), \end{cases}$$

where $\mathtt{supp}(P) = \{x : p(x) > 0\}$. Note that

$$H(P) = \log|\mathcal{X}| - D(P\|\mathrm{P}_{\mathtt{unif}}),$$

where $\mathrm{P}_{\mathtt{unif}}$ is the uniform distribution on $\mathcal{X}$. Also, $I(X \wedge Y) = D(\mathrm{P}_{XY}\|\mathrm{P}_X \times \mathrm{P}_Y)$ where $\mathrm{P}_X \times \mathrm{P}_Y$ denotes the independent distribution with marginals same as P_{XY}.

5. *Total variation distance.* Given discrete distributions P and Q on $\mathcal{X}$, the total variation distance $d(P,Q)$ between P and Q is given by

$$d(P,Q) = \frac{1}{2}\sum_{x\in\mathcal{X}} |p(x) - q(x)|.$$

We have only defined KL divergence and total variation distance for discrete distributions, but we can extend it to general distributions. For distributions with densities, we can simply replace p and q with densities f and g of P and Q, respectively, and summation by integration. In fact, these definitions can be made more general (if someone cares about it) as follows:

$$D(P\|Q) = \mathbb{E}_Q\left[\frac{p(X)}{q(X)}\log\frac{p(X)}{q(X)}\right], \quad d(P,Q) = \frac{1}{2}\mathbb{E}_Q\left[\left|\frac{p(X)}{q(X)} - 1\right|\right].$$

Thus, both definitions only depend on $p(x)/q(x)$. The more general definitions can be obtained by replacing the ratios by the "density of P with respect to Q" – this density is called the Radon-Nikodym derivative and is denoted by dP/dQ. You can look it up in a

book on probability theory.

In the notions above, we saw conditional entropy. It can alternatively be expressed as $H(X|Y) = \mathbb{E}\left[H(\mathrm{P}_{X|Y})\right]$, where the random variable inside expectation is the entropy of the conditional distribution $\mathrm{P}_{X|Y}$. In fact, we can similarly define conditional versions of all the quantities above. The *conditional mutual information* is given by

$$I(X \wedge Y|Z) = \mathbb{E}_{\mathrm{P}_Z}\left[I(\mathrm{P}_{X|Z}; \mathrm{P}_{X|Y,Z})\right],$$

where the expectation is taken over Z. Similarly, we can define *conditional divergence* between $\mathrm{P}_{Y|X}$ and $\mathrm{Q}_{Y|X}$ given P_X by

$$D(\mathrm{P}_{Y|X}\|\mathrm{Q}_{Y|X}|\mathrm{P}_X) = \mathbb{E}_{\mathrm{P}_X}\left[D(\mathrm{P}_{Y|X}\|\mathrm{Q}_{Y|X})\right].$$

Note that we could have used any distribution for taking expectation. We used P_X because it help us to show the so-called "chain rule for divergence."

B Chain rules

Chain rule for entropy

We saw above that $H(X,Y) = H(X) + H(Y|X)$. For multiple rvs $X_1, ..., X_n$, we get

$$H(X_1, ..., X_n) = H(X_1) + H(X_2|X_1) + ... + H(X_n|X_1, ..., X_{n-1}) = \sum_{i=1} H(X_i|X_1, ..., X_{i-1}).$$

Note that the conditioning on multiple random variables can be defined by treating the tupple $(x_1, ...X_{i-1})$ as a single random variable.

Chain rule for mutual information

For random variables X_1, X_2, Y, we have

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

We can easily extend this proof to more than two random variables to get

$$I(X_1, ..., X_n \wedge Y) = \sum_{i=1}^{n} I(X_i \wedge Y|X_1, ..., X_{i-1}).$$

Chain rule for KL divergence

Consider two distributions P_{XY} and Q_{XY} on $\mathcal{X} \times \mathcal{Y}$. Then, assuming $\mathtt{supp}\mathrm{P}_{\mathtt{XY}} \subset \mathtt{supp}\mathrm{Q}_{\mathtt{XY}}$, we have

$$\begin{aligned} D(\mathrm{P}_{XY}\|\mathrm{Q}_{XY}) &= \mathbb{E}_{\mathrm{P}_{XY}}\left[\log \frac{p(X,Y)}{q(X,Y)}\right] \\ &= \mathbb{E}_{\mathrm{P}_{XY}}\left[\log \frac{p(X)}{q(X)} + \log \frac{p(Y|X)}{q(Y|X)}\right] \\ &= \mathbb{E}_{\mathrm{P}_{XY}}\left[\log \frac{p(X)}{q(X)}\right] + \mathbb{E}_{\mathrm{P}_{XY}}\left[\log \frac{p(Y|X)}{q(Y|X)}\right] \\ &= \mathbb{E}_{\mathrm{P}_{X}}\left[\log \frac{p(X)}{q(X)}\right] + \mathbb{E}_{\mathrm{P}_{XY}}\left[\log \frac{p(Y|X)}{q(Y|X)}\right] \\ &= D(\mathrm{P}_X\|\mathrm{Q}_X) + \mathbb{E}_{\mathrm{P}_X}\left[D(\mathrm{P}_{Y|X}\|\mathrm{Q}_{Y|X})\right] \\ &= D(\mathrm{P}_X\|\mathrm{Q}_X) + D(\mathrm{P}_{Y|X}\|\mathrm{Q}_{Y|X}|\mathrm{P}_X). \end{aligned}$$

Note that we have used a property of iterated expectation, which is a useful one to remember:

$$\mathbb{E}[Z] = \mathbb{E}[\mathbb{E}[Z|X]].$$

This chain rule can be extended to more than two random variables to get

$$D(\mathrm{P}_{X_1 \ldots X_n} \| \mathrm{Q}_{X_1 \ldots X_n}) = \sum_{i=1}^{n} D(\mathrm{P}_{X_i|X^{i-1}} \| \mathrm{Q}_{X_i|X^{i-1}} | \mathrm{P}_{X^{i-1}}),$$

where we have used the shorthand $X^{i-1} = (X_1, ..., X_{i-1})$

C Shape of information measure functions

C.1 Nonnegativity

We shall show that all the information measures considered above are nonnegative. Of course, all came out as an answer to a coding theorem, which can only be nonnegative. However, we now seek direct proofs of these facts.

The following simple inequality will be used.

Lemma 1 (log-sum inequality)**.** *For nonnegative numbers* $\{a_i, b_i\}_{i=1^n}$*,*

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \sum_{i=1}^{n} a_i \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i},$$

with equality iff $a_i / \sum_j a_j = b_i / \sum_j b_j$ *for all* i*.*

Proof. The inequality is trivial if all a_i are 0, or if there exists an i such that $b_i = 0$ but $a_i \neq 0$. Otherwise, by rearranging the term, it suffices to show that

$$\sum_{i=1}^{n} a_i \log \frac{a_i / \sum_{j=1}^{n} a_j}{b_i / \sum_{j=1}^{n} b_j} \geq 0,$$

which holds if and only if

$$\sum_{i=1}^{n} a'_i \log \frac{a'_i}{b'_i} \geq 0,$$

where $a'_i = a_i / \sum_{j=1}^{n} a_i$ and $b'_i = b_i / \sum_{j=1}^{n} b_i$. Note that the previous inequality is simply our original inequality for the case when $\sum_{i=1}^{n} a_i = \sum_{i=1}^{n} b_i = 1$. Thus, without loss of generality we can assume that a_i and b_i constitute a pmf. Then, since $\ln x \leq x - 1$, $\log x \leq (x-1) \log e$, applying this inequality for $x = a_i/b_i$ we get

$$\sum_{i=1}^{n} a_i \log \frac{b_i}{a_i} \leq \log e \sum_{i=1}^{n} a_i \left(\frac{b_i}{a_i} - 1 \right) = 0,$$

which establishes the inequality.

Equality can hold only if equality holds for every instance of $\ln x \leq x - 1$ used in the proof, which happens only if $x = 1$. Thus, equality holds only if $a_i = b_i$ for every $i \in [n]$. □

In fact, this inequality is a speacial case of the so-called Jensen's inequality, which is a very powerful basic tool of applications beyond information theory. To state Jensen's inequality, we need to define a convex function. Roughly, speaking a convex function is a function which looks like a cup. Mathematically, a cup is an object where each chord lies above the curve. Formally,

Definition 2 (Convex and concave functions)**.** A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for every $x, y \in \mathbb{R}^n$ and $\theta \in [0, 1]$,

$$\theta f(x) + (1 - \theta) f(y) \geq f(\theta x + (1 - \theta) y).$$

A function $f : \mathbb{R}^n \to \mathbb{R}$ is concave (shaped like a cap) if $-f$ is convex, or equivalently,

$$\theta f(x) + (1 - \theta) f(y) \leq f(\theta x + (1 - \theta) y),$$

for all $x, y \in \mathbb{R}^n$ and $\theta \in [0,1]$.

Note the following interesting interpretation of the inequality characterizing convex functions: f is convex if and only if for every $\mathbb{R}^n$-valued binary rv X,

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

In fact, the same inequality must hold for every distribution (not necessarily only binary).

Lemma 3 (Jensen's inequality)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function and X be a rv taking values in $\mathbb{R}^n$. Then,*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

(The opposite inequality holds for concave functions.)

We omit the technical proof of this very believable fact.

Lemma 4. *The following hold for discrete distributions and random variables:*

1. *$D(P\|Q) \geq 0$ with equality iff $P = Q$.*
2. *$I(X \wedge Y) \geq 0$ with equality iff X and Y are independent.*
3. *$H(X) \geq 0$ with equality iff X is a constant.*

Proof. The first inequality is the same as the log-sum inequality; the second one follows from the first by noting that $I(X \wedge Y) = D(\mathrm{P}_{XY}\|\mathrm{P}_X\mathrm{P}_Y)$; and the third one follows from the second since $H(X) = I(X \wedge Y)$. Note that X is independent of X iff X is a constant random variable. □

C.2 Boundedness

Uniform distribution maximizes entropy

Lemma 5. *For a distribution P on $\mathcal{X}$, $H(P) \leq \log|\mathcal{X}|$ with equality if and only if $P = \mathtt{unif}(\mathcal{X})$.*

The proof follows from the nonnegativity of KL divergence since $D(P\|\texttt{unif}(\mathcal{X})) = \log|\mathcal{X}| - H(P)$.

Geometric distribution maximizes entropy

Lemma 6. *Consider a distribution P on $\mathbb{N}$ with $\mathbb{E}_P[X] = \alpha$. Then,*

$$H(P) \leq \log(\alpha - 1) + \alpha \log \frac{\alpha}{\alpha - 1}$$

with equality if and only if P is the geometric distribution.

Proof. Let Q be the geometric distribution with expected value α. That is, $Q(i) = (1-p)p^{i-1}$, $i \in [n]$, with $p = (\alpha - 1)/\alpha$. Thus,

$$\begin{aligned} D(P\|Q) &= \sum_{i=1}^{\infty} P(i) \log \frac{1}{Q(i)} - H(P) \\ &= \sum_{i=1}^{\infty} P(i) \log \frac{1}{(1-p)p^{i-1}} - H(P) \\ &= \log \frac{p}{1-p} + \log \frac{1}{p} \sum_{i=1}^{\infty} iP(i) - H(P) \\ &= \log(\alpha - 1) + \alpha \log \frac{\alpha}{\alpha - 1} - H(P). \end{aligned}$$

The claimed bound follows from nonnegativity of KL divergence. □

C.3 Convexity/concavity

KL Divergence is Convex and Entropy is Concave

Lemma 7. *$D(P\|Q)$ is convex in the pair (P, Q).*

Proof. Consider pairs (P_1, Q_1) and (P_2, Q_2). Consider $\theta \in [0, 1]$ and the pair (P_θ, Q_θ) with

$P_\theta = \theta P_1 + (1-\theta)P_2$ and $Q_\theta = \theta Q_1 + (1-\theta)Q_2$. Then,

$$\begin{aligned} D(P_\theta \| Q_\theta) &= \sum_x P_\theta(x) \log \frac{P_\theta(x)}{Q_\theta(x)} \\ &= \sum_x \theta P_1(x) + (1-\theta)P_2(x) \log \frac{\theta P_1(x) + (1-\theta)P_2(x)}{\theta Q_1(x) + (1-\theta)Q_2(x)} \\ &\leq \sum_x \theta P_1(x) \log \frac{\theta P_1(x)}{\theta Q_1(x)} + (1-\theta)P_2(x) \log \frac{(1-\theta)P_2(x)}{(1-\theta)Q_2(x)} \\ &= \theta D(P_1 \| Q_1) + (1-\theta) D(P_2 \| Q_2), \end{aligned}$$

where the inequality is by the log-sum inequality. □

As a corollary, we get that for any fixed Q the function $D(P\|Q)$ is convex in P. Further, upon noting that $D(P\|\mathtt{unif}(\mathcal{X})) = \log|\mathcal{X}| - H(P)$, we get that $H(P)$ is concave in P.

Corollary 8. *$H(P)$ is a concave function of P.*

But we have seen this property already. We saw that $I(X \wedge Y) \geq 0$, which is the same as

$$H(Y) \geq H(Y|X), \quad \text{(Conditioning reduces entropy)}$$

namely, $H(P_Y) \geq \mathbb{E}\left[H(P_{Y|X})\right]$. Note that $P_Y = \mathbb{E}\left[P_{Y|X}\right]$. So, the fact that "conditioning reduces entropy" is the same as

$$H(\mathbb{E}\left[P_{Y|X}\right]) \geq \mathbb{E}\left[H(P_{Y|X})\right]. \quad \text{(concavity of entropy)}$$

Mutual Information is Concave and Convex!

Lemma 9. *The mutual information function $I(P;W)$ is concave in P for every fixed W.*

Proof. We will use the following simple properties of concave and linear functions (showing them will be assigned as a HW problem):

- Linear functions are convex and concave.

- Sum of concave functions is concave.
- Concave function of linear function is concave, i.e., the composition of concave function and linear function is concave.

Recall that $I(P;W) = H(W \circ P) - H(W|P)$. The function $H(W|P)$ is a linear function of P, and therefore also a concave function. Further, since $W \circ P$ is a linear function of P and $H(\cdot)$ is a concave function, $H(W \circ P)$ is a concave function of P. Thus, by the properties seen earlier, $I(P;W)$ is a concave function of P. □

Lemma 10. *The mutual information function $I(P;W)$ is convex in W for every fixed P.*

Proof. Consider W_1 and W_2 and a $\theta \in [0,1]$. Then,

$$\begin{aligned}
&I(P; \theta W_1 + (1-\theta)W_2) \\
&= \sum_x P(x) \sum_y \theta W_1(y|x) + (1-\theta)W_2(y|x) \log \frac{\theta W_1(y|x) + (1-\theta)W_2(y|x)}{\theta(W_1 \circ P)(y) + (1-\theta)(W_2 \circ P)(y)} \\
&\leq \sum_x P(x) \sum_y \theta W_1(y|x) \log \frac{\theta W_1(y|x)}{\theta(W_1 \circ P)(y)} \\
&\qquad\qquad + \sum_x P(x) \sum_y (1-\theta) W_2(y|x) \log \frac{(1-\theta)W_2(y|x)}{(1-\theta)(W_2 \circ P)(y)} \\
&= \theta I(P; W_1) + (1-\theta) I(P; W_2),
\end{aligned}$$

where the inequality holds by the log-sum inequality. □

D Data Processing Inequality

Earlier we mentioned that it natural to expect any reasonable measure of distance to satisfy the data processing inequality. By triangular inequality, total variation distance satisfies the data processing inequality. Also, we mentioned that KL Divergence satisfies the data processing inequality. We now show this result.

Lemma 11. *For a fixed channel W and two distributions P and Q on the input of the channel, we have*

$$D(W \circ P \| W \circ Q) \leq D(P \| Q).$$

Proof. We have

$$\begin{aligned} D(W \circ P \| W \circ Q) &= \sum_y (W \circ P)(y) \log \frac{(W \circ P)(y)}{(W \circ Q)(y)} \\ &= \sum_y \sum_x P(x) W(y|x) \log \frac{\sum_x P(x) W(y|x)}{\sum_x Q(x) W(y|x)} \\ &\leq \sum_y \sum_x P(x) W(y|x) \log \frac{P(x) W(y|x)}{Q(x) W(y|x)} \\ &= \sum_x P(x) \sum_y W(y|x) \log \frac{P(x)}{Q(x)} \\ &= D(P \| Q), \end{aligned}$$

where the inequality above is by the log-sum inequality. □

A careful reader may have noticed that the only inequality we have used so far is the log-sum inequality, which itself is equivalent to the nonnegativity of the KL divergence. Thus, all the inequalities in this section are by the nonnegativity of KL divergence.

Next, we present a data processing inequality for mutual information which is obtained as a consequence of the more general form above. To present this inequality, we need to introduce the notion of a Markov chain.

Definition 12. Random variables X, Y, Z form a Markov chain if X and Z are independent when conditioned on Y, i.e., when $I(X \wedge Z|Y) = 0$ or, equivalently, $\mathrm{P}_{XYZ} = \mathrm{P}_{X|Y} \mathrm{P}_{Z|Y} \mathrm{P}_Y$. This definition extends naturally to multiple random variables: $X_1, ..., X_n$ form a Markov chain if $(X_1, ..., X_{i-1}), X_i, (X_{i+1}, ..., X_n)$ form a Markov chain for every $1 \leq i \leq n$. We use the notation $X_1 \multimap X_2 \multimap ... \multimap X_n$ to indicate that $X_1, ..., X_n$ form a Markov chain.

A specific example is when $Z = f(Y)$ for some function f. In this case, for every X

we have $X \multimap Y \multimap Z$. Heuristically, if $X \multimap Y \multimap Z$ holds, then X can contain no more information about Z than Y. The following result establishes this bound.

Lemma 13. *If $X \multimap Y \multimap Z$, then $I(X \wedge Z) \leq I(X \wedge Y)$. Equivalently, $H(X|Y) \leq H(X|Z)$.*

Proof. Instead of taking recourse to the data processing inequality for KL divergence, we present an alternative (perhaps more popular) proof. We have

$$I(X \wedge Z) \leq I(X \wedge Y, Z) = I(X \wedge Y) + I(X \wedge Z|Y) = I(X \wedge Y),$$

where the first inequality is easy to see and you will be asked to show it in a HW problem, and the final indentity holds since $X \multimap Y \multimap Z$. □