Information and Coding Theory **Autumn 2017**

Lecture 4: October 9, 2017

Lecturer: Madhur Tulsiani

1 More on mutual information

We consider a set of random variables in a particular relationship and its consequences for mutual information. An ordered tuple of random variables (X, Y, Z) is said to form a Markov chain, written as $X \to Y \to Z$, if X and Z are independent conditioned on Y. Here, we can think of Y as being sampled given the knowledge of X, and Z being sampled given the knowledge of Y (but not using the "history" about X).

Note that although the notation $X \to Y \to Z$ (and also the above description) makes it seem like this is only a Markov chain the forward order, the conditional independence definition implies that if $X \to Y \to Z$ is Markov chain, then so is $Z \to Y \to X$. This is sometimes to written as $X \leftrightarrow Y \leftrightarrow Z$ to clarify that the variables form a Markov chain in both forward and backward orders. The following inequality shows that information about the starting point cannot increase as we go further in a Markov chain.

Lemma 1.1 (Data Processing Inequality) *Let $X \to Y \to Z$ be a Markov chain. Then*

$$I(X;Y) \;\geq\; I(X;Z)\,.$$

Proof: It is perhaps useful to consider a useful special case first: let $Z = g(Y)$ be a function of Y. Then it is easy to see that $X \to Y \to g(Y)$ form a Markov chain. We can prove the inequality in this case by observing that conditioning on Y is the same as conditioning on $Y, g(Y)$.

$$\begin{aligned} I(X;Y) \;&=\; H(X) - H(X|Y) \\ &=\; H(X) - H(X|Y, g(Y)) \\ &\geq\; H(X) - H(X|g(Y)) \;=\; I(X;g(Y))\,. \end{aligned}$$

The first two lines of the above proof amounted to the fact that

$$I(X;Y) \;=\; I(X;(Y,g(Y)) \;=\; I(X;(Y,Z))\,.$$

However, this continues to be true in the general case, since

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

where the second term is zero due to the conditional independence. Hence, the proof for the general case is the same and we have

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

■

The special case $Z = g(Y)$ is also useful to define the concept of a "sufficient statistic", which is a function of Y that makes the data processing inequality tight.

Definition 1.2 *For random variables X and Y, a function $g(Y)$ is called a* sufficient statistic *(of Y) for X if $I(X;Y) = I(X;g(Y))$ i.e., $g(Y)$ contains all the relevant information about X.*

Exercise 1.3

$$X = \begin{cases} 1/4 & w.p\ 1/3 \\ 1/2 & w.p\ 1/3 \\ 3/4 & w.p\ 1/3 \end{cases}$$

Let Y be a sequence of n tosses of a coin with probability of heads given by X. Let $g(Y)$ be the number of heads in Y. Prove $I(X;Y) = I(X;g(Y))$.

2 Kullback Leibler divergence

The Kullback-Leibler divergence (KL-divergence), also known as relative entropy, is a measure of how different two distributions are. Note that here we will talk in terms of distributions instead of random variables, since this is how KL-divergence is most commonly expressed. It is of course easy to think of a random variable corresponding to a given distribution and vice-versa. We will use capital letters like $P(X)$ to denote a distribution for the random variable X and lowercase letters like $p(x)$ to denote the probability for a specific element x.

Let P and Q be two distributions on a universe U, then the KL-divergence between P and Q is defined as:

$$D(P||Q) := \sum_{x \in U} p(x) \log \left(\frac{p(x)}{q(x)} \right)$$

Let us consider a simple example.

Example 2.1 *Suppose $U = \{a,b,c\}$, and $p(a) = \frac{1}{3}$, $p(b) = \frac{1}{3}$, $p(c) = \frac{1}{3}$ and $q(a) = \frac{1}{2}$, $q(b) = \frac{1}{2}$, $q(c) = 0$. Then*

$$D(P||Q) = \frac{2}{3} \log \frac{2}{3} + \infty = \infty\,.$$

$$D(Q||P) = \log \frac{3}{2} + 0 = \log \frac{3}{2}.$$

The above example illustrates two important facts: $D(P||Q)$ and $D(Q||P)$ are not necessarily equal, and $D(P||Q)$ may be infinite. Even though the KL-divergence is not symmetric, it is often used as a measure of "dissimilarity" between two distribution. Towards this, we first prove that it is non-negative and is 0 if and only if $P = Q$.

Lemma 2.2 *Let P and Q be distributions on a finite universe U. Then $D(P||Q) \geq 0$ with equality if and only if $P = Q$.*

Proof: Let $\text{Supp}(P) = \{x : p(x) > 0\}$. Then, we must have $\text{Supp}(P) \subseteq \text{Supp}(Q)$ if $D(P, Q) < \infty$. We can then assume without loss of generality that $\text{Supp}(Q) = U$. Using the fact the log is a concave function, with Jensen inequality, we have:

$$\begin{aligned}
D(P||Q) = \sum_{x \in U} p(x) \log \frac{p(x)}{q(x)} &= \sum_{x \in \text{Supp}(P)} p(x) \log \frac{p(x)}{q(x)} \\
&= - \sum_{x \in \text{Supp}(P)} p(x) \log \frac{q(x)}{p(x)} \\
&\geq - \log \left(\sum_{x \in \text{Supp}(P)} p(x) \cdot \frac{q(x)}{p(x)} \right) \\
&= - \log \left(\sum_{x \in \text{Supp}(P)} q(x) \right) \\
&\geq - \log 1 = 0 .
\end{aligned}$$

For the case when $D(P||Q) = 0$, we note that this implies $p(x) = p(x)\ \forall x \in \text{Supp}(P)$, which in turn gives that $p(x) = q(x)\ \forall x \in U$. ■

Like entropy and mutual information, we can also derive a chain rule for KL-divergence. Let $P(X, Y)$ and $Q(X, Y)$ be two distributions for a pair of variables X and Y. We then have the following expression for $D(P(X, Y)||Q(X, Y))$.

Proposition 2.3 (Chain rule for KL-divergence) *Let $P(X, Y)$ and $Q(X, Y)$ be two distributions for a pair of variables X and Y. Then,*

$$\begin{aligned}
D(P(X, Y) \parallel Q(X, Y)) &= D(P(X) \parallel Q(X)) + \mathop{\mathbb{E}}_{x \sim P} [D(P(Y|X = x) \parallel Q(Y|X = x))] \\
&= D(P(X) \parallel Q(X)) + D(P(Y|X) \parallel Q(Y|X))
\end{aligned}$$

Here $P(X)$ and $Q(X)$ denote the marginal distributions for the first variable, and $P(Y|X = x)$ denotes the conditional distribution of Y.

Proof: The proof follows from (by now) familiar manipulations of the terms inside the log function.

$$\begin{aligned} D(P(X,Y) \parallel Q(X,Y)) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)} \\ &= \sum_{x,y} p(x)p(y|x) \log \left(\frac{p(x)}{q(x)} \cdot \frac{p(y|x)}{q(y|x)} \right) \\ &= \sum_{x} p(x) \log \frac{p(x)}{q(x)} \sum_{y} p(y|x) + \sum_{x} p(x) \sum_{y} p(y|x) \log \frac{p(y|x)}{q(y|x)} \\ &= D(P(X) \parallel Q(X)) + \sum_{x} p(x) \cdot D(P(Y|X=x) \parallel Q(Y|X=x)) \\ &= D(P(X) \parallel Q(X)) + D(P(Y|X) \parallel Q(Y|X)) \end{aligned}$$

■

Note that if $P(X,Y) = P_1(X)P_2(Y)$ and $Q(X,Y) = Q_1(X)Q_2(Y)$, then $D(P||Q) = D(P_1||Q_1) + D(P_2||Q_2)$.

We note that KL-divergence also has an interesting interpretation in terms of source coding. Writing

$$D(P||Q) = \sum p(x) \log \frac{p(x)}{q(x)} = \sum p(x) \log \frac{1}{q(x)} - \sum p(x) \log \frac{1}{p(x)},$$

we can view this as the number of extra bits we use (on average) if we designed a code according to the distribution P, but used it to communicate outcomes of a random variable X distributed according to Q.

2.1 Total variation distance and Pinsker's inequality

We can now relate KL-divergence to some other notions of distance between two probability distributions.

Definition 2.4 *Let P and Q be two distributions on a finite universe U. Then the* total-variation distance *or* statistical distance *between P and Q is defined as*

$$\delta_{TV}(P,Q) = \frac{1}{2} \cdot \|P - Q\|_1 = \frac{1}{2} \cdot \sum_{x \in U} |p(x) - q(x)| .$$

*The quantity $\|P - Q\|_1$ is referred to as the ℓ_1-*distance *between P and Q.*

The total variation distance of P and Q represents the maximum probability with which any test can distinguish between the two distributions *given one random sample*. It may seem that the restriction to one sample severely limits the class of tests, but we can always think of an m-sample test for P and Q as getting one sample from one of the product distributions P^m or Q^m.

Let $f : U \to \{0,1\}$ be any classifier, which given one sample $x \in U$, outputs 1 is the guess is that the sample came from P, and 0 if the guess is that it came from Q. The difference in its behavior over the two distributions can be measured by the quantity (which can be thought of as the rate of true positive minus the rate of false positive) $|\mathbb{E}_P[f(x)] - \mathbb{E}_Q[f(x)]|$. The following lemma bounds this in terms of the total variation distance.

Lemma 2.5 *Let P, Q be any distributions on U. Let $f : U \to [0, B]$. Then*

$$\left|\mathop{\mathbb{E}}_{P}[f(x)] - \mathop{\mathbb{E}}_{Q}[f(x)]\right| \;\leq\; \frac{B}{2} \cdot \|P - Q\|_1 \;=\; B \cdot \delta_{TV}(P, Q)\,.$$

Proof:

$$\begin{aligned}
\left|\mathop{\mathbb{E}}_{P}[f(x)] - \mathop{\mathbb{E}}_{Q}[f(x)]\right| &= \left|\sum_{x\in U} p(x)\cdot f(x) - \sum_{x\in U} q(x)\cdot f(x)\right| \\
&= \left|\sum_{x\in U} (p(x) - q(x))\cdot f(x)\right| \\
&= \left|\sum_{x\in U} (p(x) - q(x))\cdot \left(f(x) - \frac{B}{2}\right) + \frac{B}{2}\cdot\left(\sum_{x\in U} p(x) - q(x)\right)\right| \\
&\leq \sum_{x\in U} |p(x) - q(x)| \cdot \left|f(x) - \frac{B}{2}\right| \\
&\leq \frac{B}{2}\cdot\|P - Q\|_1
\end{aligned}$$

■

Exercise 2.6 *Prove that the above inequality is tight. What is the optimal classifier f?*

In many applications, we want to actually bound the ℓ_1-distance between P and Q but it's easier to analyze the KL-divergence. The following inequality helps relate the two.

Lemma 2.7 (Pinsker's inequality) *Let P and Q be two distributions defined on a universe U. Then*

$$D(P \parallel Q) \;\geq\; \frac{1}{2\ln 2}\cdot\|P - Q\|_1^2\,.$$

We will first consider a special case when $U = \{0, 1\}$ and P, Q are distributions as below

$$P \;=\; \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases} \qquad \text{and} \qquad Q \;=\; \begin{cases} 1 & \text{w.p. } q \\ 0 & \text{w.p. } 1-q \end{cases}$$

In this case, we have

$$D(P\|Q) \;=\; p \cdot \log\left(\frac{p}{q}\right) + (1-p) \cdot \log\left(\frac{1-p}{1-q}\right) \qquad \text{and} \qquad \|P - Q\|_1 \;=\; 2 \cdot |p - q| \;.$$

We will first prove Pinsker's inequality for this special case.

Proposition 2.8 (Pinsker's inequality for $U = \{0, 1\}$) *Let P and Q be distributions as above. Then,*

$$p \cdot \log\left(\frac{p}{q}\right) + (1-p) \cdot \log\left(\frac{1-p}{1-q}\right) \;\geq\; \frac{2}{\ln 2} \cdot (p-q)^2 \,.$$

Proof: Let

$$f(p, q) \;:=\; p \cdot \log\left(\frac{p}{q}\right) + (1-p) \cdot \log\left(\frac{1-p}{1-q}\right) - \frac{2}{\ln 2} \cdot (p-q)^2 \,.$$

We have,

$$\frac{\partial f}{\partial q} = -\frac{(p-q)}{\ln 2}\left(\frac{1}{q(1-q)} - 4\right) \,.$$

Since $\frac{1}{q(1-q)} - 4 \geq 0$ for all q, we have that $\frac{\partial f}{\partial q} \leq 0$ when $q \leq p$ and $\frac{\partial f}{\partial q} \geq 0$ when $q \geq p$. Moreover, $f(p, q) = \infty$ when $q = 0$ and $f(p, q) = 0$ when $q = p$. Thus, the function achieves its minimum value at $q = p$ and is always non-negative, which proves the desired inequality. ■