Information and Coding Theory **Autumn 2022**

Lecture 5: January 26, 2021

Lecturer:

1 Inequalities for Markov chains

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 $X \rightarrow Y \rightarrow Z$, 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 $X \rightarrow Y \rightarrow Z$ (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 $X \rightarrow Y \rightarrow Z$ is Markov chain, then so is $Z \rightarrow Y \rightarrow X$. 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.

1.1 Data Processing Inequality

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 $X \rightarrow Y \rightarrow Z$ 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 $X \rightarrow Y \rightarrow g(Y)$ 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.

$$X = \begin{cases} p_1 & w.p.\ 1/2 \\ p_2 & w.p.\ 1/2 \end{cases}$$

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 $I(X;Y) = I(X;g(Y))$.

1.2 Fano's inequality

We first prove an important inequality that lets us understand how well can some "ground truth" random variable X be predicted based on some observed data Y. We state the inequality in the language of Markov chains, which we saw before in the context of data processing inequality. We will denote the Markov chain as $X \to Y \to \widehat{X}$. We can think of X as the choice of an unknown parameter from some finite set $\mathcal{X}$. We think of Y as the "data" generated from this, say a sequence independent samples. Finally, we think of $\widehat{X}$ as a "guess" for X, which depends only on the data. Fano's inequality is concerned with the probability of error in the guess, defined as $p_e = \mathbb{P}\left[\widehat{X} \neq X\right]$. We have the following statement

Lemma 1.4 (Fano's inequaity)**.** *Let $X \to Y \to \widehat{X}$ be a Markov chain, and let $p_e = \mathbb{P}\left[\hat{X} \neq X\right]$. Let $H_2(p_e)$ denote the binary entropy function computed at p_e. Then,*

$$H_2(p_e) + p_e \cdot \log\left(|\mathcal{X}| - 1\right) \;\geq\; H(X|\widehat{X}) \;\geq\; H(X|Y)\,.$$

Proof: We define a binary random variable, which indicates an error i.e

$$E := \begin{cases} 1 & \text{if } \widehat{X} \neq X \\ 0 & \text{if } \widehat{X} = X \end{cases}$$

The bound in the ineuality then follows from considering the undertainty that still remains after our prediction, i.e., the entroy $H(X, E|\widehat{X})$.

$$H(X, E|\widehat{X}) \ = \ H(X|\widehat{X}) + H(E|X, \widehat{X}) \ = \ H(X|\widehat{X}),$$

since $H(E|X, \widehat{X}) = 0$ (why?) Another way of computing this entropy is

$$\begin{aligned} H(X, E|\widehat{X}) &= H(E|\widehat{X}) + H(X|E, \widehat{X}) \\ &= H(E|\widehat{X}) + p_e \cdot H(X|E = 1, \widehat{X}) + (1 - p_e) \cdot H(X|E = 0, \widehat{X}) \\ &\leq H(E) + p_e \cdot H(X|E = 1, \widehat{X}) \\ &\leq H_2(p_e) + p_e \cdot \log\left(|\mathcal{X}| - 1\right) . \end{aligned}$$

Comparing the two expressions then proves the claim. ∎

Fano's inequality provides a useful way of lower bounding the error of a predictor, particularly in the case when $|\mathcal{X}| > 2$. As we will see later, in the case when $|\mathcal{X}| = 2$, we will be able to obtain better bounds using the concept of KL-divergence considered later.

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 $\mathcal{X}$, 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 $\mathcal{X} = \{a, b, c\}$, 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 $\mathcal{X}$. Then $D(P||Q) \geq 0$ with equality if and only if $P = Q$.*

Proof: Let $\text{Supp}(P) = \{x \mid p(x) > 0\}$. 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 $\text{Supp}(Q) = \mathcal{X}$. Using the fact the log is a (strictly) concave function, with Jensen inequality, we have:

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

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 $p(x) = q(x)\ \forall x \in \mathcal{X}$. ■

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,*

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

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. The first term in the RHS, which corresponds to the average number of bits used by the "wrong" encoding, is also referred to as cross entropy.

2.1 Convexity of KL-divergence

Before we consider applications, let us prove an important property of KL-divergence. We prove below that $D(P \parallel Q)$, when viewed as a function of the inputs P and Q, is jointly convext in both it's inputs i.e., it is convex in the input (P, Q) when viewed as a tuple.

Proposition 2.4. *Let P_1, P_2, Q_1, Q_2 be distributions on a finite universe $\mathcal{X}$, and let $\alpha \in [0,1]$. Then,*

$$D(\alpha \cdot P_1 + (1-\alpha) \cdot P_2 \parallel \alpha \cdot Q_1 + (1-\alpha) \cdot Q_2) \;\leq\; \alpha \cdot D(P_1 \parallel Q_1) + (1-\alpha) \cdot D(P_2 \parallel Q_2) \,.$$

Proof: For this proof, we will use an inequality called the log-sum inequality, the proof of which is left is an exercise. The inequality states that for $a_1, a_2, b_1, b_2 \geq 0$

$$(a_1 + a_2) \cdot \log\left(\frac{a_1 + a_2}{b_1 + b_2}\right) \;\leq\; a_1 \cdot \log\left(\frac{a_1}{b_1}\right) + a_2 \cdot \log\left(\frac{a_2}{b_2}\right)$$

Using the above inequality, we can bound the LHS as

$$\begin{aligned}
& D\left(\alpha \cdot P_1 + (1-\alpha) \cdot P_2 \parallel \alpha \cdot Q_1 + (1-\alpha) \cdot Q_2\right) \\
=\; & \sum_{x \in \mathcal{X}} \left(\alpha \cdot p_1(x) + (1-\alpha) \cdot p_2(x)\right) \cdot \log\left(\frac{\alpha \cdot p_1(x) + (1-\alpha) \cdot p_2(x)}{\alpha \cdot q_1(x) + (1-\alpha) \cdot q_2(x)}\right) \\
\leq\; & \sum_{x \in \mathcal{X}} \alpha \cdot p_1(x) \cdot \log\left(\frac{\alpha \cdot p_1(x)}{\alpha \cdot q_1(x)}\right) + (1-\alpha) \cdot p_2(x) \cdot \log\left(\frac{(1-\alpha) \cdot p_2(x)}{(1-\alpha) \cdot q_2(x)}\right) \\
=\; & \alpha \cdot D\left(P_1 \parallel Q_1\right) + (1-\alpha) \cdot D\left(P_2 \parallel Q_2\right) \,.
\end{aligned}$$

■

Exercise 2.5 (Log-sum inequality)**.** *Prove that for $a_1, a_2, b_1, b_2 \geq 0$*

$$(a_1 + a_2) \cdot \log\left(\frac{a_1 + a_2}{b_1 + b_2}\right) \;\leq\; a_1 \cdot \log\left(\frac{a_1}{b_1}\right) + a_2 \cdot \log\left(\frac{a_2}{b_2}\right) \,.$$