Information and Coding Theory **Winter 2025**

Lecture 3: January 14, 2025

Lecturer: Madhur Tulsiani

1 Mutual Information

The mutual information is a quantity which measures the amount of dependence between two random variables. Unlike correlation, which defines the random variables to take values in the same space, the mutual information can be defined for any two random variables. The mutual information between two random variables X and Y is defined by the formula

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

Using the chain rule for entropy, we can see that

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

We can use the first two expressions to observe that $I(X;Y) \geq 0$ and the last one to observe that $I(X;Y) = I(Y;X)$.

Example 1.1. *Consider the random variable (X,Y) with $X \vee Y = 1$, $X \in \{0,1\}$ and $Y \in \{0,1\}$ such that:*

$$(X,Y) = \begin{cases} 10 & \textit{w.p 1/3} \\ 01 & \textit{w.p 1/3} \\ 11 & \textit{w.p 1/3} \end{cases}$$

Then, we can calculate the entropy and mutual information as follows:

$$\begin{aligned} H(X) \;=\; H(Y) \;&=\; \frac{1}{3}\log 3 + \frac{2}{3}\log\frac{3}{2} \;=\; \log 3 - \frac{2}{3} \\ H(X,Y) \;&=\; \log 3 \\ I(X;Y) \;=\; H(X) + H(Y) - H(X,Y) \;&=\; \log 3 - \frac{4}{3} \end{aligned}$$

Conditioning on a third random variable Z, we can also define the conditional mutual information $I(X;Y|Z)$ as

$$\begin{aligned} I(X;Y|Z) \;&:=\; \mathop{\mathbb{E}}_{z}\left[I(X|Z=z;Y|Z=z)\right] \\ &=\; \mathop{\mathbb{E}}_{z}\left[H(X|Z=z) - H(X|Y,Z=z)\right] \\ &=\; H(X|Z) - H(X|Y,Z)\,. \end{aligned}$$

Consider the following example of three random variables.

Example 1.2. *Consider the random variable (X, Y, Z), $X \in \{0, 1\}$, $Y \in \{0, 1\}$ and $Z = X \oplus Y$ such that:*

$$(X, Y, Z) = \begin{cases} 000 & \textit{w.p 1/4} \\ 011 & \textit{w.p 1/4} \\ 101 & \textit{w.p 1/4} \\ 110 & \textit{w.p 1/4} \end{cases}$$

We can check that in this case, X, Y are independent and thus $I(X; Y) = 0$. However,

$$\begin{aligned} I(X : Y|Z) &= \mathop{\mathbb{E}}_{z}\left[I(X|Z = z; Y|Z = z)\right] \\ &= \frac{1}{2} I(X|Z = 0; Y|Z = 0) + \frac{1}{2} I(X|Z = 1; Y|Z = 1) \\ &= \frac{1}{2}\log 2 + \frac{1}{2}\log 2 \;=\; 1 \end{aligned}$$

The above example illustrates that unlike entropy, it is not true that conditioning (on average) decreases the mutual information. In the above example, while $I(X; Y) = 0$, we have $I(X; Y|Z) = 1$ which is in fact the maximum possible.

Recall that entropy provides theoretical limits on source coding, where the goal is to compress information when transmitting in a way such that whatever we send is received without any error. The concept of mutual information provides limits on transmission, when the transmission "channel" is noisy. We will discuss this in detail when we consider error-correcting codes, but it is instructive to consider the following example known as the "Binary Symmetric Chhannel".

Exercise 1.3. *Let X be a random variable supported on $\{0, 1\}$, and let Y be a "noisy" copy of X, which is equal to X with probability $1 - p$, and has the opposite value (0 is X is 1, and 1 if X is 0) with probability p. Calculate the maximum possible value of $I(X; Y)$ over all possible distributions for X. This is known as the* capacity *of the binary symmetric channel.*

As in the case of entropy, mutual information also obeys a chain rule.

Lemma 1.4. $I((X_1, \ldots, X_m); Y) = \sum_{i=1}^{m} I(X_i; Y|X_1, \ldots, X_{i-1})$

Proof: The chain rule for mutual information is a simple consequence of the chain rule for entropy. We have

$$\begin{aligned}
I((X_1,\ldots,X_m);Y) &= H(X_1,\ldots,X_m) - H(X_1,\ldots,X_m|Y) \\
&= \sum_{i=1}^{m} H(X_i|X_1,\ldots,X_{i-1}) - \sum_{i=1}^{m} H(X_i|Y,X_1,\ldots,X_{i-1}) \\
&= \sum_{i=1}^{m} \left[H(X_i|X_1,\ldots,X_{i-1}) - H(X_i|Y,X_1,\ldots,X_{i-1})\right] \\
&= \sum_{i=1}^{m} I(X_i;Y|X_1,\ldots,X_{i-1})
\end{aligned}$$

■

2 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 $X \to Y \to Z$, 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.

2.1 Data Processing Inequality

The following inequality shows that information about the starting point cannot increase as we go further in a Markov chain.

Lemma 2.1 (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 2.2. *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 2.3.

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

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

2.2 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 2.4 (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.