# Lecture 3

*Lecturer: Anup Rao* *Scribe: Prasang Upadhyaya*

## 1 Introduction

In the previous lecture we looked at the application of entropy to derive inequalities that involved counting. In this lecture we step back and introduce the concepts of *relative entropy* and *mutual information* that measure two kinds of relationship between two distributions over random variables.

## 2 Relative Entropy

The *relative entropy*, also known as the *Kullback-Leibler divergence*, between two probability distributions on a random variable is a measure of the distance between them. Formally, given two probability distributions $p(x)$ and $q(x)$ over a discrete random variable $X$, the relative entropy given by $D(p||q)$ is defined as follows:

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

In the definition above $0 \log \frac{0}{0} = 0 \log \frac{0}{q} = 0$ and $p \log \frac{1}{0} = \infty$.

As an example, consider a random variable $X$ with the law $q(x)$. We assume nothing about $q(x)$. Now consider a set $E \subseteq \mathcal{X}$ and define $p(x)$ to be the law of $X|_{X \in E}$. The divergence between $p$ and $q$:

$$\begin{aligned}
D(p||q) &= \sum_{x \in \mathcal{X}} Pr[X = x|_{X \in E}] \log \frac{Pr[X = x|_{X \in E}]}{Pr[X = x]} \\
&= \sum_{x \in E} Pr[X = x|_{X \in E}] \log \frac{Pr[X = x|_{X \in E}]}{Pr[X = x]} \text{ (Using } 0 \log 0 = 0\text{)} \\
&= \sum_{x \in E} Pr[X = x|_{X \in E}] \log \frac{Pr[X = x|_{X \in E}]}{Pr[X = x|_{X \in E}] Pr[X \in E]} \text{ (Using the chain rule)} \\
&= \sum_{x \in E} Pr[X = x|_{X \in E}] \log \frac{1}{Pr[X \in E]} \\
&= \log \frac{1}{Pr[X \in E]}
\end{aligned}$$

In the extreme case with $E = \mathcal{X}$, the two laws $p$ and $q$ are identical with a divergence of 0.

We will henceforth refer to relative entropy or Kullback-Leibler divergence as divergence

### 2.1 Properties of Divergence

1. Divergence is not symmetric. That is, $D(p||q) = D(q||p)$ is not necessarily true. For example, unlike $D(p||q)$, $D(q||p) = \infty$ in the example mentioned in the previous section, if $\exists x \in \mathcal{X} \setminus E : q(x) > 0$.

2. Divergence is always non-negative. This is because of the following:

$$\begin{aligned}
D(p||q) &= \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} \\
&= -\sum_{x\in\mathcal{X}} p(x)\log\frac{q(x)}{p(x)} \\
&= -\mathbb{E}\left[\log\frac{q}{p}\right] \\
&\geq -\log\left(\mathbb{E}\left[\frac{q}{p}\right]\right) \\
&= -\log\left(\sum_{x\in\mathcal{X}} p(x)\frac{q(x)}{p(x)}\right) \\
&= 0
\end{aligned}$$

The inequality is introduced due to the application of Jensen's inequality and the concavity of log.

3. Divergence is a convex function on the domain of probability distributions. Formally,

**Lemma 1** (Convexity of divergence). *Let $p_1, q_1$ and $p_2, q_2$ be probability distributions over a random variable $X$ and $\forall\lambda \in (0,1)$ define*

$$\begin{aligned}
p &= \lambda p_1 + (1-\lambda)p_2 \\
q &= \lambda q_1 + (1-\lambda)q_2
\end{aligned}$$

*Then, $D(p||q) \leq \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2)$.*

To prove the lemma, we shall use the log-sum inequality [1], which can be proved by reducing to Jensen's inequality:

**Proposition 2** (Log-sum Inequality). *If $a_1, \ldots, a_n, b_1, \ldots, b_n$ are non-negative numbers, then*

$$\sum_{i=1}^{n} a_i \log(1/b_i) \leq \left(\sum_{i=1}^{n} a_i\right)\log\left(\frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}\right)$$

**Proof** [of Lemma 1] Let $a_1(x) = \lambda p_1(x), a_2(x) = (1-\lambda)p_2(x)$ and $b_1(x) = \lambda q_1(x), b_2(x) = (1-\lambda)q_2(x)$. Then,

$$\begin{aligned}
D(p||q) &= \sum_x (\lambda p_1(x) + (1-\lambda)p_2(x))\log\frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} \\
&= \sum_x (a_1(x) + a_2(x))\log\frac{a_1(x) + a_2(x)}{b_1(x) + b_2(x)} \\
&\leq \sum_x \left(a_1(x)\log\frac{a_1(x)}{b_1(x)} + a_2(x)\log\frac{a_2(x)}{b_2(x)}\right) \quad \text{(Using the log-sum inequality)} \\
&= \sum_x \left(\lambda p_1(x)\log\frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda)p_2(x)\log\frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)}\right) \\
&= \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2)
\end{aligned}$$

■

## 2.2 Relationship of Divergence with Entropy

Intuitively, the entropy of a random variable $X$ with a probability distribution $p(x)$ is related to how much $p(x)$ diverges from the uniform distribution on the support of $X$. The more $p(x)$ diverges the lesser its entropy and vice versa. Formally,

$$\begin{aligned} H(X) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \\ &= \log |\mathcal{X}| - \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{\frac{1}{|\mathcal{X}|}} \\ &= \log |\mathcal{X}| - D(p||uniform) \end{aligned}$$

## 2.3 Conditional Divergence

Given the joint probability distributions $p(x, y)$ and $q(x, y)$of two discrete random variables $X$ and $Y$, the conditional divergence between two conditional probability distributions $p(y|x)$ and $q(y|x)$ is obtained by computing the divergence between $p$ and $q$ for all possible values of $x \in \mathcal{X}$ and then averaging over these values of $x$. Formally,

$$D(p(y|x)||q(y|x)) = \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log \frac{p(y|x)}{q(y|x)}$$

Given the above definition we can prove the following chain rule about divergence of joint probability distribution functions.

**Lemma 3** (Chain Rule)**.**

$$D\left(p(x, y)||q(x, y)\right) = D\left(p(x)||q(x)\right) + D\left(p(y|x)||q(y|x)\right)$$

**Proof**

$$\begin{aligned} D\left(p(x, y)||q(x, y)\right) &= \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{q(x, y)} \\ &= \sum_x \sum_y p(x)p(y|x) \log \frac{p(x)p(y|x)}{q(x)q(y|x)} \\ &= \sum_x \sum_y p(x)p(y|x) \log \frac{p(x)}{q(x)} + \sum_x \sum_y p(x)p(y|x) \log \frac{p(y|x)}{q(y|x)} \\ &= \sum_x p(x) \log \frac{p(x)}{q(x)} \sum_y p(y|x) + \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} \\ &= D\left(p(x)||q(x)\right) + D\left(p(y|x)||q(y|x)\right) \end{aligned}$$

■

# 3 Mutual Information

Mutual information is a measure of how correlated two random variables $X$ and $Y$ are such that the more independent the variables are the lesser is their mutual information. Formally,

$$\begin{aligned}
I(X \wedge Y) &= D(p(x,y)||p(x)p(y)) \\
&= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{-} \sum_{x,y} p(x,y) \log p(x) - \sum_{x,y} p(x,y) \log p(y) \\
&= -H(X,Y) + H(X) + H(Y) \\
&= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X)
\end{aligned}$$

Here $I(X \wedge Y)$ is the mutual information between $X$ and $Y$, $p(x,y)$ is the joint probability distribution, $p(x)$ and $p(y)$ are the marginal distributions of $X$ and $Y$.

As before we define the conditional mutual information when conditioned upon a third random variable $Z$ to be

$$\begin{aligned}
I(X \wedge Y|Z) &= \mathbb{E}_z[I(X \wedge Y|Z=z)] \\
&= H(X|Z) - H(Y|X,Z)
\end{aligned}$$

This leads us to the following chain rule.

**Lemma 4** (Chain Rule)**.** $I(X,Z \wedge Y) = I(X \wedge Y) + I(Z \wedge Y|X)$

**Proof**

$$\begin{aligned}
I(X,Z \wedge Y) &= H(X,Z) - H(X,Z|Y) \\
&= H(X) + H(Z|X) - H(X|Y) - H(Z|X,Y) \\
&= (H(X) - H(X|Y)) + (H(Z|X) - H(Z|X,Y)) \\
&= I(X \wedge Y) + I(Z \wedge Y|X)
\end{aligned}$$

■

## 3.1 An Example

We now look at the effect of conditioning on Mutual information. We consider the following two examples.

**Example 1.** Let $X, Y, Z$ be uniform bits with zero parity. Now,

$$I(X \wedge Y|Z) = H(X|Z) - H(X|Y,Z) = 1 - 0 = 1$$

$H(X|Z) = 1$ since given $Z$, $X$ could be either of $\{0,1\}$ while given $Y, Z$, $X$ is already determined. Meanwhile,

$$I(X \wedge Y) = H(X) - H(X|Y) = 1 - 1 = 0$$

**Example 2.** Let $A, B, C$ be uniform random bits. Define $X = A, B$ and $Y = A, C$ and $Z = A$. Now,

$$I(X \wedge Y|Z) = H(X|Z) - H(X|Y,Z) = 1 - 1 = 0$$

while,

$$I(X \wedge Y) = H(X) - H(X|Y) = 2 - 1 = 1$$

Thus, unlike entropy, conditioning may decrease or increase the mutual information.

## 3.2 Properties of Mutual Information

**Lemma 5.** *If $X, Y$ are independent and $Z$ has an arbitrary probability distribution then,*

$$I(X, Y \wedge Z) \geq I(X \wedge Z) + I(Y \wedge Z)$$

**Proof**

$$\begin{aligned}
I(\{X, Y\} \wedge Z) &= I(X \wedge Z) + I(Y \wedge Z|X) \text{ (Using the chain rule)} \\
&= I(X \wedge Z) + H(Y|X) - H(Y|X, Z) \\
&= I(X \wedge Z) + H(Y) - H(Y|X, Z) \text{ ($X$ and $Y$ are independent)} \\
&\geq I(X \wedge Z) + H(Y) - H(Y|Z) \text{ (Conditioning can not increase entropy)} \\
&= I(X \wedge Z) + I(Y \wedge Z)
\end{aligned}$$

■

**Lemma 6.** *Let $(X, Y) \sim p(x, y)$ be the joint probability distribution of $X$ and $Y$. By the chain rule, $p(x, y) = p(x)p(y|x) = p(y)p(x|y)$. For clarity we represent $p(x)$ (resp. $p(y)$) by $\alpha$ and $p(y|x)$ (resp. $p(x|y)$) by $\pi$. The following holds:*

**Concavity in $p(x)$:** *For $i \in \{1, 2\}$, let $I_i(X, Y)$ be the mutual information for $(X, Y) \sim \alpha_i \pi$, respectively. For $\lambda_1, \lambda_2 \in [0, 1]$ such that $\lambda_1 + \lambda_2 = 1$, let $I(X \wedge Y)$ be the mutual information for $(X, Y) \sim \sum_i \lambda_i \alpha_i \pi$. Then,*

$$I(X \wedge Y) \geq \lambda_1 I_1(X \wedge Y) + \lambda_2 I_2(X \wedge Y)$$

**Convexity in $p(y|x)$:** *For $i \in \{1, 2\}$, let $I_i(X, Y)$ be the mutual information for $(X, Y) \sim \alpha \pi_i$, respectively. For $\lambda_1, \lambda_2 \in [0, 1]$ such that $\lambda_1 + \lambda_2 = 1$, let $I(X \wedge Y)$ be the mutual information for $(X, Y) \sim \sum_i \lambda_i \alpha \pi_i$. Then,*

$$I(X \wedge Y) \leq \lambda_1 I_1(X \wedge Y) + \lambda_2 I_2(X \wedge Y)$$

**Proof** We first prove the *convexity of $p(y|x|)$*: we will apply Lemma 1 and use the definition of mutual information in terms of divergence. Thus,

$$\begin{aligned}
I(X \wedge Y) &= D\left(\lambda_1 \alpha \pi_1 + \lambda_2 \alpha \pi_2 \,||\, \left(\sum_y \lambda_1 \alpha \pi_1 + \lambda_2 \alpha \pi_2\right)\left(\sum_x \lambda_1 \alpha \pi_1 + \lambda_2 \alpha \pi_2\right)\right) \\
&= D\left(\lambda_1 \alpha \pi_1 + \lambda_2 \alpha \pi_2 \,||\, \left(\lambda_1 \alpha \sum_y \pi_1 + \lambda_2 \alpha \sum_y \pi_2\right)\left(\sum_x \lambda_1 \alpha \pi_1 + \lambda_2 \alpha \pi_2\right)\right) \\
&= D\left(\lambda_1 \alpha \pi_1 + \lambda_2 \alpha \pi_2 \,||\, \alpha \sum_x \lambda_1 \alpha \pi_1 + \alpha \lambda_2 \alpha \pi_2\right) \\
&= D\left(\lambda_1 \alpha \pi_1 + \lambda_2 \alpha \pi_2 \,||\, \lambda_1 \sum_y \alpha \pi_1 \sum_x \alpha \pi_1 + \lambda_2 \sum_y \alpha \pi_1 \alpha \pi_2\right) \\
&\leq \lambda_1 D\left(\alpha \pi_1 \,||\, \left(\sum_y \alpha \pi_1\right)\left(\sum_x \alpha \pi_1\right)\right) + \lambda_2 D\left(\alpha \pi_2 \,||\, \left(\sum_y \alpha \pi_2\right)\left(\sum_x \alpha \pi_2\right)\right) \\
&= \lambda_1 I_1(X \wedge Y) + \lambda_2 I_2(X \wedge Y)
\end{aligned}$$

Here we used the fact that $\sum_y \pi_i = 1$ and used Lemma 1 to introduce the inequality.

We now prove the *concavity of* $p(x)$. We first simplify the LHS and the RHS.

$$
\begin{aligned}
I(X \wedge Y) &= \sum_{x,y} (\lambda_1\alpha_1\pi + \lambda_2\alpha_2\pi) \log \frac{\lambda_1\alpha_1\pi + \lambda_2\alpha_2\pi}{\left(\sum_y \lambda_1\alpha_1\pi + \lambda_2\alpha_2\pi\right)\left(\sum_x \lambda_1\alpha_1\pi + \lambda_2\alpha_2\pi\right)} \\
&= \sum_{x,y} (\lambda_1\alpha_1\pi + \lambda_2\alpha_2\pi) \log \frac{\pi}{\left(\sum_x \lambda_1\alpha_1\pi + \lambda_2\alpha_2\pi\right)} \\
&= \sum_{x,y} (\lambda_1\alpha_1\pi + \lambda_2\alpha_2\pi) \log \pi - \sum_{x,y} \left( \sum_{i\in\{0,1\}} \lambda_i\alpha_i\pi \right) \log \left( \sum_x \sum_{i\in\{0,1\}} \lambda_i\alpha_i\pi \right) \\
\lambda_1 I_1(X \wedge Y) + \lambda_2 I_2(X \wedge Y) &= \sum_{x,y} \sum_{i\in\{0,1\}} \lambda_i\alpha_i\pi \log \frac{\alpha_i\pi}{\left(\sum_y \alpha_i\pi\right)\left(\sum_x \alpha_i\pi\right)} \\
&= \sum_{x,y} (\lambda_1\alpha_1\pi + \lambda_2\alpha_2\pi) \log \pi - \sum_{x,y} \sum_{i\in\{0,1\}} \lambda_i\alpha_i\pi \log \left( \sum_x \alpha_i\pi \right)
\end{aligned}
$$

Thus, to prove that $LHS \geq RHS$ we need to prove that,

$$
\left( \sum_{i\in\{0,1\}} \lambda_i\alpha_i\pi \right) \log \left( \sum_x \sum_{i\in\{0,1\}} \lambda_i\alpha_i\pi \right) \leq \sum_{i\in\{0,1\}} \lambda_i\alpha_i\pi \log \left( \sum_x \alpha_i\pi \right)
$$

that follows directly from the application of the log-sum inequality [1]
■

## References


[1] Thomas M. Cover and Joy A. Thomas. *Elements of information theory.* Wiley-Interscience, New York, NY, USA, 1991.