15-859: Information Theory and Applications in TCS **CMU: Spring 2013**

Lecture 3: KL-divergence and connections

January 22, 2013

Lecturer: Venkatesan Guruswami *Scribe: David Witmer*

1 Recap

Recall some important facts about entropy and mutual information from the previous lecture:

- $H(X,Y) = H(X) + H(Y|X)$
  $= H(Y) + H(X|Y)$
- $I(X;Y) = H(X) - H(X|Y)$
  $= H(Y) - H(Y|X)$
  $= H(X) + H(Y) - H(X,Y)$
- $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$
- $I(X;Y) = 0$ if X and Y are independent
- $I(X;Y) \geq 0$ or, equivalently, $H(X) \geq H(X|Y)$

Exercise 1.1 *Prove that $H(X|Y) = 0$ if and only if $X = g(Y)$ for some function g.*

2 More mutual information

2.1 Mutual information chain rule

We begin by proving the chain rule for mutual information.

Theorem 2.1 (Chain rule for mutual information)

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_1, X_2, \ldots, X_{i-1})$$

Proof.

$$\begin{aligned}
I(X_1, X_2, \ldots, X_n; Y) &= H(X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n|Y) \text{ by definition} \\
&= \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1}) - \sum_{i=1}^{n} H(X_i|Y, X_1, \ldots, X_{i-1}) \text{ by entropy chain rule} \\
&= \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1}) - H(X_i|Y, X_1, \ldots, X_{i-1}) \\
&= \sum_{i=1}^{n} I(X_i; Y|X_1, X_2, \ldots, X_{i-1}) \text{ by definition}
\end{aligned}$$

■

2.2 Comparing $I(X;Y)$ and $I(X;Y|Z)$

Next, we consider the question of whether $I(X;Y) \geq I(X;Y|Z)$ or $I(X;Y) \leq I(X;Y|Z)$ holds. It turns out that both cases are possible:

1. If $Z = Y$, $I(X;Y|Y) = H(X|Y) - H(X|Y) = 0$ so it is possible that $I(X;Y) > I(X;Y|Z)$.
2. Consider $Z = X \oplus Y$. Since X and Y are independent bits, $I(X;Y) = 0$. On the other hand, $H(X|Z) = H(X) = 1$ since X and Z are independent and $H(X|Y,Z) = 0$ since $X = Y \oplus Z$. This means that $I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = 1$, so it is possible that $I(X;Y) < I(X;Y|Z)$.

2.3 Mutual information and independence

If $X_1, X_2, \ldots, X_n$ are independent, it holds that

$$H(X_1, X_2, \ldots, X_n) = H(X_1) + H(X_2) + \cdots + H(X_n).$$

We will show a similar statement for mutual information:

Lemma 2.2 *If $X_1, X_2, \ldots, X_n$ are independent, then*

$$I(X_1, X_2, \ldots, X_n; Y) \geq \sum_{i=1}^{n} I(X_i; Y).$$

Proof.

$$\begin{aligned}
I(X_1, \ldots, X_n; Y) &= H(X_1, \ldots, X_n) + H(Y) - H(X_1, \ldots, X_n, Y) \\
&= \left(\sum_{i=1}^{n} H(X_i)\right) + H(Y) - H(X_1, \ldots, X_n, Y) \text{ by independence of } X_i\text{'s} \\
&= \left(\sum_{i=1}^{n} H(X_i)\right) + H(Y) \\
&\qquad - (H(Y) + H(X_1|Y) + H(X_2|X_1, Y) + \cdots + H(X_n|Y, X_1, \ldots, X_n - 1)) \\
&= \sum_{i=1}^{n} H(X_i) - H(X_i|Y, X_1, \ldots, X_{i-1}) \\
&\geq \sum_{i=1}^{n} H(X_i) - H(X_i|Y) \text{ by } H(X_i|Y, X_1, \ldots, X_{i-1}) \leq H(X_i|Y) \\
&= \sum_{i=1}^{n} I(X_i; Y)
\end{aligned}$$

■

Note that the inequality is necessary as equality does not hold in general. If X_1 and X_2 are independent bits and $Y = X_1 \oplus X_2$, we get that $I(X_1, X_2; Y) = 1 > 0 = I(X_1; Y) + I(X_2; Y)$.

3 Kullback-Leibler divergence

Kullback-Leibler divergence, also known as K-L divergence, relative entropy, or information divergence, acts like a distance measure between probability distributions on the same universe.

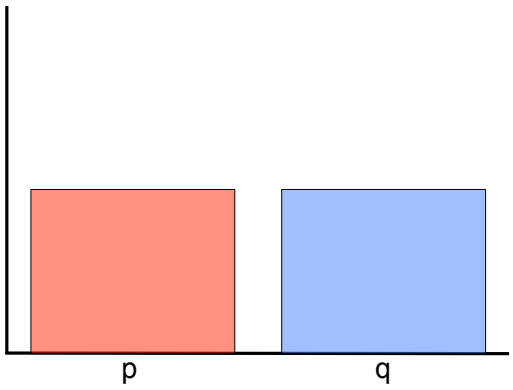
Definition 3.1 (K-L divergence) *Let p and q be distributions on the same set U. Then the K-L divergence of q from p, denoted $D(p||q)$, is defined to be*

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

Note that despite the intuition of divergence as a distance measure, $D(p||q) \neq D(q||p)$. It is the case, however, that if $p = q$, the $D(p||q) = 0$. At first glance, it is not even clear that divergence is nonnegative; we will prove this later.

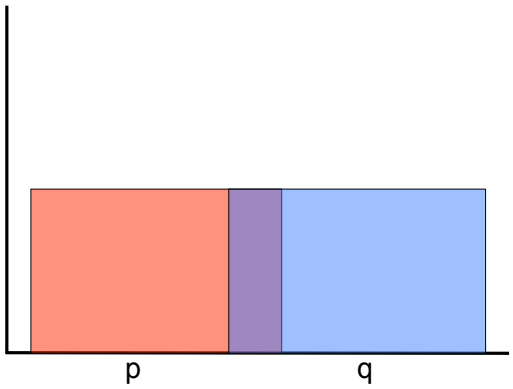
Note that if $D(p||q)$ is finite, then support$(p) \subseteq$ support(q). Consider the following examples:

1. support(p) and support(q) are disjoint



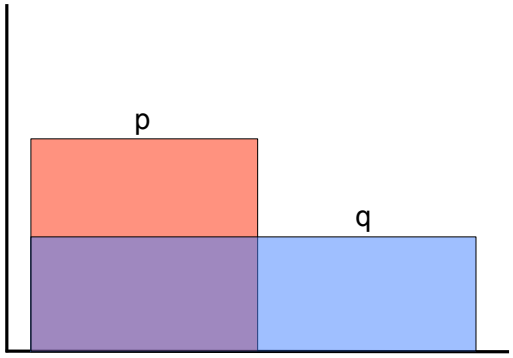
   In this case, both $D(p||q)$ and $D(q||p)$ are infinite.

2. support(p) and support(q) overlap, but neither is a subset of the other



   Again, both $D(p||q)$ and $D(q||p)$ are infinite.

3. support$(p) \subseteq$ support(q)



$D(p||q)$ is finite, but $D(q||p)$ is infinite.

3.1 Gibb's Inequality

Theorem 3.2 (Gibb's Inequality) *$D(p||q) \geq 0$ with equality if and only if $p = q$.*

Proof. The proof again uses concavity and Jensen's Inequality. We will show that $-D(p||q) \leq 0$.

$$\begin{aligned}
-D(p||q) &= \sum_x p(x) \log \frac{q(x)}{p(x)} \\
&= \mathbb{E}[\log Z] \text{ where } Z \text{ is an r.v. that takes value } \frac{q(x)}{p(x)} \text{ with probability } p(x) \\
&\leq \log \mathbb{E}[Z] \text{ by Jensen's Inequality} \\
&= \log 1 \text{ since } \mathbb{E}[Z] = \sum_x p(x) \frac{q(x)}{p(x)} = \sum_x q(x) = 1 \\
&= 0
\end{aligned}$$

Since $\log x$ is a strictly convex function, equality holds if and only if Z is constant, which occurs exactly when $p = q$. ■

3.2 Mutual information and divergence

Consider the joint distribution of two random variables X and Y. Let $p(x, y) = \Pr[X = x \wedge Y = y]$ and $p(x)$ and $p(y)$ be $\Pr[X = x]$ and $\Pr[Y = y]$, respectively. We can then relate divergence to mutual information.

Theorem 3.3 *Let X and Y be random variables. Then*

$$I(X;Y) = D(p(x,y)||p(x)p(y)).$$

Note that $p(x)p(y)$ is the probability that x and y are chosen from the product distribution of the marginal distributions for X and Y. If X and Y are independent, then $p(x, y) = p(x)p(y)$ and $D(p(x, y)||p(x)p(y)) = 0$.

Example 3.4 *Let $U = \{1, \ldots, 100\}$ and let $p(x, y)$ be the uniform distribution on $(1, 1), (2, 2), \ldots, (100, 100)$. Then $p(x) = \frac{1}{100}$ for all x, so $p(x)p(y) = \frac{1}{10000}$ for all x and y. We then have that*

$$D(p(x,y)||p(x)p(y)) = \sum_{x=y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = \sum_{i=1}^{100} \frac{1}{100} \log \frac{10000}{100} = \log 100.$$

Proof of Theorem 3.3:

$$\begin{aligned}
D(p(x,y)||p(x)p(y)) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(y)p(x|y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} \\
&= \sum_{x,y} p(x,y) \log \frac{1}{p(x)} - \sum_{x,y} p(x,y) \log \frac{1}{p(x|y)}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_x p(x) \log \frac{1}{p(x)} - \sum_y p(y) \log H(X|Y=y) \\
&= H(X) - H(X|Y) \\
&= I(X;Y)
\end{aligned}
$$

■

4 Some viewpoints on K-L divergence

In this section, we will describe three scenarios in which K-L divergence arises for the purpose of gaining intuition.

4.1 Source coding

Say we design a Shannon code for the distribution $q = (q_1, \ldots, q_n)$, but the actual distribution we end up encoding is $p = (p_1, \ldots, p_2)$. Then the increase in the expected length of the encoding is the K-L divergence of the two distributions.

Let L be the expected length of the encoding. We would like to have $L \approx H(p)$, but our source code is a Shannon code based on q, so we instead get

$$
\begin{aligned}
L &= \sum_{i=1}^n p_i \left\lceil \log \frac{1}{q_i} \right\rceil \\
&\geq \sum_{i=1}^n p_i \log \frac{1}{q_i} \\
&= \sum_{i=1}^n p_i \log \frac{1}{p_i} + p_i \log \frac{p_i}{q_i} \\
&= H(p) + D(p||q).
\end{aligned}
$$

Similarly, we can upper bound L:

$$
\begin{aligned}
L &= \sum_{i=1}^n p_i \left\lceil \log \frac{1}{q_i} \right\rceil \\
&\leq \sum_{i=1}^n p_i \left(\log \frac{1}{q_i} + 1 \right) \\
&= \sum_{i=1}^n p_i \log \frac{1}{q_i} + \sum_{i=1}^n p_i \\
&= H(p) + D(p||q) + 1.
\end{aligned}
$$

So the extra expected length required is between $D(p||q)$ and $D(p||q) + 1$.

4.2 Rejection sampling

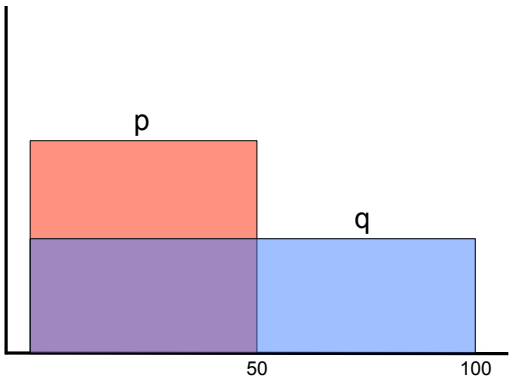
Let p and q be two distributions on U. We have an oracle that outputs random samples according to q, but we want random samples from p.

Setting We are given a sequence $x_1, x_2, \ldots$ of i.i.d. samples drawn according to q.

Goal Output i^* such that x_{i^*} is distributed according to p.

We first give an example, then state a more general result without proof.

Example 4.1 *Consider these distributions p and q:*



in which p is uniform over $\{1, \ldots, 50\}$ and q is uniform over $\{1, \ldots, 100\}$. We can output the first i such that $x_i \leq 50$ as i^* because conditioned on $x_i \leq 50$, x_i is uniform over $\{1, \ldots, 50\}$. Then we have that $\mathbb{E}[i^*] = 2$ and $\mathbb{E}[\log i^*] = 1$. Note that if instead p is uniform over $\{1, \ldots, 100\}$ and q is uniform over $\{1, \ldots, 50\}$, we cannot simulate p in this model.

More generally, the following theorem:

Theorem 4.2 *There is a strategy to output an i^* satisfying the above condition that achieves*

$$\mathbb{E}[\log i^*] \lesssim D(p||q).$$

4.3 Compressing communication protocols

Consider the following situation:

Setting We have two parties, A and B, that have shared randomness. A knows a distribution p and B knows a distribution q.

Goal Communicate x drawn from p to B in the minimum possible number of bits, taking advantage of the fact that B knows q.

We can solve this problem by sending $D(p||q)$ bits between A and B:

Theorem 4.3 *There is an interactive protocol between A and B with expected number of bits of communication approximately equal to $D(p||q)$ such that, in the end,*

1. *A outputs a distributed according to p.*
2. *B outputs b such that for all x, $\Pr[b = x | a = x] \geq 1 - \varepsilon$ for some small ε.*

Note that rejection sampling example from the previous section also deals with compressing communication protocols.

5 Data processing inequality

Consider random variables X and Y and a deterministic function g of Y. We can think of g as a function "processing" the data given by Y. The data processing inequality says that we can never process information to create more information:

Theorem 5.1 (Data processing inequality)

$$I(X;Y) \geq I(X;g(Y))$$

We conclude with a joke:

Joke 5.2 *Consider the information obtained by attending lectures for some class. A student could read scribe notes instead of attending lectures. However, setting g to be the function mapping the information from lectures to scribe notes, the data processing inequality says that the student will always get at least as much information from attending lectures as he or she will get from reading scribe notes.*