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

Lecture 4: Data processing and Fano's inequalities; AEP

January 24, 2013

Lecturer: Venkatesan Guruswami *Scribe: Amit Datta*

1 Recap

- **KL Divergence** for two dist. p and q, the KL divergence is $D(p||q) = \mathbb{E}(\log \frac{p(x)}{p(y)})$
- **Gibbs' inequality** $D(p||q) \geq 0$, with equality holding if $p = q$
- If X, Y are correlated random variables, $I(X;Y) = D(p(x,y)||p(x)p(y))$

2 More viewpoints on KL Divergence

Three viewpoints were discussed in the previous lecture. "As if three weren't enough, here are two more"

2.4 A Lemma

Lemma 2.1. *If p is a distribution on the universe U, $H(p) = \log |U| - D(p||u)$, where u is the uniform distribution*

This lemma states partly what we already knew, that $H(X) \leq \log |support(X)|$ and the equality is achieved when X is distributed uniformly. When X is not so, the difference equals to the KL divergence between the distribution of X and a uniform distribution.

2.5 KL divergence and Chernoff Bound

Firstly, a brief introduction to the Chernoff Bound:

If a fair coin is tossed n times, on an average 'heads' will be observed $n/2$ times, and 'tails' $n/2$ times. However, $Pr[\text{seeing } (0.5+\epsilon) \text{ heads}] \leq 2^{-\frac{\epsilon^2 n}{4}}$. This bound can be rewritten using the KL divergence. In fact, this bound is tight:

$$\frac{2^{-nD(p||u)}}{n^2} \leq Pr[\text{seeing } pn \text{ heads}] \leq 2^{-nD(p||u)}$$

Given n i.i.d. random variables $X_1, X_2, ..., X_n$ drawn according to a distribution q over the universe $U = \{1, 2, ..., m\}$, the following holds:

$$\frac{2^{-D(p||q)}}{(n+1)^m} \leq Pr[\text{frequency of symbols we see are according to } p] \leq 2^{-D(p||q)}$$

where p is a probability distribution. The term measures the probability that there are exactly $p_i n$ i's for $i = 1, 2, \ldots, m$ among the n symbols.

3 Data Processing Inequality

Definition 3.1 (**Markov Chain**). *Three random variables X, Y, Z are said to form a Markov Chain, denoted by $X \to Y \to Z$ if the conditional distribution of Z depends only on Y and is independent of X.*

Example: $Z = g(Y)$, where $g()$ is some function.

Theorem 3.2. *If $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$*

Proof. The joint probability of x, y, z:

$$p(x, y, z) = p(x)p(y|x)p(z|x, y)$$

Since, Z is independent of X, we have $p(z|x, y) = p(z|y)$ and the joint probability becomes:

$$p(x, y, z) = p(x)p(y|x)p(z|y)$$

Now, we have the following observation:

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x)p(y|x)p(z|y)}{p(y)} = p(x|y)p(z|y)$$

i.e. X, Z are conditionally independent given Y.
Now, we expand $I(X;Y,Z)$ applying the Chain-Rule:

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

Again, expanding in a different order,

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

The second term on the R.H.S of the above equation is 0 since we concluded that X, Z are conditionally independent given Y.
So, we have:

$$I(X;Z) + I(X;Y|Z) = I(X;Y,Z) = I(X;Y) + 0 = I(X;Y)$$

Rearranging the above equation:

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

since $I(X;Y|Z) \geq 0$.

$\square$

Corollary 3.3. *If $X \to Y \to Z$, then $I(X;Y|Z) \leq I(X;Y)$*

Corollary 3.4. *If $X \to Y \to Z$, then $I(X;Y|g(Y)) \leq I(X;Y)$*

Recall that, in general, it is possible that $I(X;Y|Z) > I(X;Y)$

4 Fano's Inequality

Situation: We know a random variable Y and we want to guess the valus of a correlated r.v. X

Exercise 4.1. *If X is a function of Y, then the degree of surprise in X given Y is 0 and vice versa. Mathematically:*

$$X = g(Y) \Leftrightarrow H(X|Y) = 0$$

Fano's inequality is a quantitative version of the above.

Theorem 4.2. *Given Y and a function $g()$, which is used to estimate X, i.e. $\tilde{X} = g(Y)$, where the error of this estimation is given by $P_{err} = Pr[\tilde{X} \neq X]$, then*

$$h(P_{err}) + P_{err} \log(n-1) \geq H(X|Y)$$

where $n = |support(X)|$ and function $h()$ is defined as $h(x) = x \log \frac{1}{x} + (1-x) \log \frac{1}{1-x}$

Proof. Let $E = 1$ denote the event that there is an error in the estimation: $\tilde{X} \neq X$, so, $Pr[E = 1] = P_{err}$.

So, we can say:

$$H(E) = h(P_{err})$$

Again, knowing X, Y completely determines the event E. Hence,

$$H(E|X,Y) = 0$$

Adding $H(X|Y)$ to both sides of the equation, we get:

$$H(E|X,Y) + H(X|Y) = H(X|Y)$$

Applying chain rule to compress the L.H.S

$$H(X,E|Y) = H(X|Y)$$

Applying chain rule again to the L.H.S., but in a different order:

$$H(E|Y) + H(X|E,Y) = H(X|Y)$$

Since conditioning can never increase entropy, $H(E|Y) \leq H(E)$. Applying this to the above equation:

$$H(X|Y) \leq H(E) + H(X|E,Y)$$

Since $H(E) = h(P_{err})$

$$H(X|Y) \leq h(P_{err}) + H(X|E,Y)$$

Now,

$$H(X|E,Y) = Pr[E=0]H(X|Y,E=0) + Pr[E=1]H(X|Y,E=1)$$

Given Y and $E = 0$, i.e. there is no error in estimating X from $g(Y)$, X is determined, implying $H(X|Y, E = 0) = 0$. $Pr[E = 1]$ is known to be P_{err}, and $H(X|Y, E = 1) \leq H(X)$ since conditioning can never increase entropy. Again $H(X) \leq \log n$ as $n = |support(X)|$. Additionally, knowing that $E = 1$, i.e., there *is* an error in the estimation, we can be certain that $X \neq g(Y)$. This reduces the maximum possible entropy of X conditioned on Y and $E = 1$, i.e., $H(X|Y, E = 1)$ to be at most $\log_2(n-1)$. So, we obtain:

$$H(X|Y) \leq h(P_{err}) + P_{err} \log_2(n-1)$$

as claimed. □

Exercise 4.3. *Analyze the optimal "maximum likelihood decoding" strategy.*

5 Asymptotic Equipartition Property (AEP)

First, we state the following law, since proof of AEP will require it:

Law 5.1 (Weak Law of Large Numbers). *Given n i.i.d. draws $\{Z_1, Z_2, ..., Z_n\}$ of a r.v. Z with $\mathbb{E}(Z) = \mu$,*

$$\forall\ \epsilon\ \exists\ n_0\ s.t.\ \forall n \geq n,\quad Pr\Big[\Big|\frac{Z_1 + Z_2 + ... + Z_n}{n} - \mu\Big| > \epsilon\Big] \leq \epsilon$$

Property 5.2. *If X is a random variable drawn from the distribution P and $X_1, X_2, ..., X_n$ are n i.i.d samples of X, then*

$$Pr[p(a_1, a_2, ..., a_n) \simeq 2^{-nH(X)}] \to 1$$

where $a_1, a_2, ..., a_n$ are values taken up by $X_1, X_2, ..., X_n$ respectively.

In other words, AEP states that "Almost all events are almost equally surprising".

Proof. AEP follows by applying the weak law of large numbers to the following variable:

$$Z = \log \frac{1}{p(a)} \quad \text{with probability } p(a)$$

Again:

$$\mathbb{E}(Z) = \sum_a p(a) \log \frac{1}{p(a)} = H(X)$$

Applying the weak law of large numbers to Z, we get:

$$Pr\Big[\Big|\frac{1}{n}\sum_{i=1}^{n} \log \frac{1}{p(a_i)} - H(X)\Big| > \epsilon\Big] \leq \epsilon$$

$$Pr\Big[\Big| - \frac{\log p(a_1, a_2, ..., a_n)}{n} - H(X)\Big| > \epsilon\Big] \leq \epsilon$$

$$Pr\Big[\Big|\frac{\log p(a_1, a_2, ..., a_n)}{n} + H(X)\Big| > \epsilon\Big] \leq \epsilon$$

$$Pr\Big[\Big|\frac{\log p(a_1, a_2, ..., a_n)}{n} + H(X)\Big| < \epsilon\Big] \geq 1 - \epsilon$$

$$Pr\Big[- \epsilon < \Big(\frac{\log p(a_1, a_2, ..., a_n)}{n} + H(X)\Big) < \epsilon\Big] \geq 1 - \epsilon$$

$$Pr\Big[-H(X)-\epsilon < \Big(\frac{\log p(a_1,a_2,...,a_n)}{n}\Big) < -H(X)+\epsilon\Big] \geq 1-\epsilon$$

$$Pr\Big[-n(H(X)+\epsilon) < \Big(\log p(a_1,a_2,...,a_n)\Big) < -n(H(X)-\epsilon)\Big] \geq 1-\epsilon$$

$$Pr\Big[2^{-n(H(X)+\epsilon)} < p(a_1,a_2,...,a_n) < 2^{-n(H(X)-\epsilon)}\Big] \geq 1-\epsilon \qquad \square$$

6 Postscript

The sequences whose probability are close to the $2^{-nH(X)}$ bound are the typical ones, and so we define the following set.

Definition 6.1 (Typical Set). *A typical set* $A_\epsilon^{(n)}$ *w.r.t.* $p(X)$ *is the set* $\{X_1, X_2, ..., X_n\} \in \Sigma^n$ *such that* $2^{-n(H(X)+\epsilon)} < p(a_1,a_2,...,a_n) < 2^{-n(H(X)-\epsilon)}$

The following is just a restatement of the AEP we proved above.

Lemma 6.2. *If* $a_1, a_2, ..., a_n$ *are drawn i.i.d. according to* X*, then* $Pr[(a_1,a_2,...,a_n) \in A_\epsilon^{(n)}] \geq 1-\epsilon$

A simple counting argument yields that the size of the typical set is $\approx 2^{H(X)n}$.

Lemma 6.3. $(1-\epsilon)2^{n(H(X)-\epsilon)} \leq |A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$