Information and Coding Theory **Autumn 2014**

Lecture 2: October 2, 2014

Lecturer: Madhur Tulsiani Scribe: Shubhendu Trivedi

In the last class we have defined entropy for random variables, talked about prefix free codes, and also saw that we can always design a prefix free code whose expected length is the entropy plus one. Although we defined entropy, we did not really use it except that trying to show that it captures the notion of information in terms of the number of bits needed to communicate a message. The notion starts to make more sense when we look at the joint entropies instead.

1 Joint Entropy

We have two random variables X and Y. The joint distribution of the two random variables (X, Y) takes values (x, y) with probability $p(x, y)$. Merely by using the definition, we can write down the entropy of $Z = (X, Y)$ trivially. However what we are more interested in is seeing how the entropy of (X, Y), the joint entropy, relates to the individual entropies, which we work out below:

$$\begin{aligned}
H(X,Y) &= \sum_{x,y} p(x,y) \log \frac{1}{p(x,y)} \\
&= \sum_{x,y} p(x)p(y|x) \log \frac{1}{p(x)} + \sum_{x,y} p(x)p(y|x) \log \frac{1}{p(y|x)} \\
&= \sum_{x} p(x) \log \frac{1}{p(x)} \sum_{y} p(y|x) + \sum_{x,y} p(x)p(y|x) \log \frac{1}{p(y|x)} \\
&= H(X) + \sum_{x} p(x) H(Y|X=x) \\
&= H(X) + \mathop{\mathbb{E}}_{x} \left[H(Y|X=x)\right]
\end{aligned}$$

Denoting $\mathbb{E}_x \left[H(Y|X = x)\right]$ as $H(Y|X)$, this can simply be written as

$$H(X,Y) = H(X) + H(Y|X) \tag{1}$$

If we were to redo the calculations, we could similarly obtain:

$$H(X,Y) = H(Y) + H(X|Y) \tag{2}$$

This is called the *Chain Rule* for Entropy.

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

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

Now, let us calculate the following:

1. $H(X) = H(Y) = \frac{1}{3}\log 3 + \frac{2}{3}\log\frac{3}{2}$
2. $H(Y|X = 0) = 0$
3. $H(Y|X = 1) = \frac{1}{2}\log\frac{1}{\frac{1}{2}} + \frac{1}{2}\log\frac{1}{\frac{1}{2}} = 1$
4. $H(Y|X) = \frac{1}{3}\cdot 0 + \frac{2}{3}\cdot 1 = \frac{2}{3}$
5. $H(X, Y) = \frac{1}{3}\log 3 + \frac{1}{3}\log 3 + \frac{1}{3}\log 3 = \log 3$

From the above we see that:

$$H(Y) \geq H(Y|X)$$

this is actually *always* true and we prove this fact below.

Fact 1.2 $H(Y) \geq H(Y|X)$

Proof: We want to show that $H(Y|X) - H(Y) \leq 0$. Consider the quantity on the left hand side.

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

Now consider a random variable Z that takes value $\frac{p(x)p(y)}{p(x,y)}$ with probability $p(x, y)$. Then we can use Jensen's inequality to get:

$$\sum_{x,y} p(x,y)\left(\log\frac{p(x)p(y)}{p(x,y)}\right) \leq \log\left(\sum_{x,y}\frac{p(x)p(y)}{p(x,y)}p(x,y)\right) = \log(1) = 0\,.$$

■

Note however the fact that conditioning on X reduces the entropy of Y is only true *on average over all fixings of* X. In particular, in the above example we have $H(Y|X = 1) = 1 > H(Y)$. But $H(Y|X)$, which is an average over all fixings of X, is indeed smaller than $H(Y)$.

Exercise 1.3 *Consider* $H(p) = p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}$ *with* $0 \leq p \leq 1$. *Prove that* $H(p)$ *is concave using* $H(X|Y) \leq H(Y)$.

Let us have a quick look at some of the things we proved and considered:

- **Chain Rule:** $H(X, Y) = H(X) + H(Y|X)$. Using induction it can easily be shown that the following also holds:

$$H(X_1, X_2, \ldots, X_m) = H(X_1) + H(X_2|X_1) + H(X_3|X_1X_2) \ldots H(X_m|X_1 \ldots X_{m-1})$$

- **Conditioning on average reduces Entropy:** $H(Y) \geq H(Y|X)$.
  These two points also imply that entropy is **subadditive** i.e.

$$H(X_1, \ldots X_m) \leq H(X_1) + \cdots + H(X_m)$$

Using the above three facts, we can now start demonstrating some more interesting properties, but before that we would return to source coding once again.

2 Source Coding Theorem

We begin by recalling the Shannon Code. We considered a random variable X that took on values $a_1, a_2, \ldots, a_m$ with probabilities $p_1, p_2, \ldots, p_m$. We wanted to encode the values of X such that the expected number of bits needed is small. If $l_1, l_2, \ldots, l_m$ are the number of bits needed to encode $a_1, a_2, \ldots, a_m$, then we saw that a prefix free code exists iff:

$$\sum_{i=1}^{n} 2^{l_i} \leq 1$$

Furthermore, we saw that the expected length of the encoding is lower bounded by $H(X)$ and upper bounded by $H(X) + 1$ (a code as specified as above, the Shannon code may be constructed by setting $l_i = \lceil \log \frac{1}{p_i} \rceil$).

We will now try to improve this upper bound and we will do so by considering multiple copies of X. The idea is that by amortizing the loss over many symbols, we can start to approach an expected length equal to the lower bound i.e. the entropy.

The design may be done as follows: Consider m copies of the random variable X, $\{X_1, \ldots, X_m \in U\}$ and a code $C : U^m \to \{0, 1\}^*$. Let $|U|^m = N$. Now, we know that:

$$H(X_1, \ldots, X_m) \leq \sum_{i=1}^{N} p_i \lceil \log \frac{1}{p_i} \rceil \leq H(X_1, \ldots, X_m) + 1 \tag{3}$$

Let us also assume that the m copies of X are drawn i.i.d. Using this assumption we try to work out the quantity $H(X_1, \ldots, X_m)$. Which may be expanded using the chain rule as:

$$\begin{aligned} H(X_1, \ldots, X_m) &= H(X_1) + H(X_2|X_1) + \cdots + H(X_m|X_1, \ldots, X_{m-1}) \\ &= H(X_1) + H(X_2) + \cdots + H(X_m) \\ &= mH(X) \end{aligned}$$

Therefore, equation 3 becomes:

$$mH(X) \leq \sum_{i=1}^{N} p_i \lceil \log \frac{1}{p_i} \rceil \leq mH(X) + 1 \tag{4}$$

Thus, we used $H(X) + \frac{1}{m}$ bits on average per copy of X. This leads us to the source coding theorem.

Theorem 2.1 (Fundamental Source Coding Theorem (Shannon)) *For all $\varepsilon > 0$ there exists a n_0 such that for all $n \geq n_0$ and given n copies of X, $X_1, \ldots, X_n$ sampled i.i.d., it is possible to communicate $(X_1, \ldots, x_n)$ using at most $H(X) + \varepsilon$ bits per copy on average.*

3 Some Appplications

3.1 An Application in Counting

We want to prove that for $k \leq n/2$:

$$\sum_{i=1}^{k} \binom{n}{i} \leq 2^{nH(\frac{k}{n})}$$

Proof: Let $\mathcal{F}$ be the set of all subsets of $[n]$ of size less than or equal to k. It is to be noted that $|\mathcal{F}| = \sum_{i \leq k} \binom{n}{i}$. Let X be a member of $\mathcal{F}$ picked at random. We can think of X as the random vector $X_1, \ldots, X_n \in \{0,1\}^n$ such that $\sum X_i \leq k$. Also we assume that $X_1, \ldots, X_n$ are uniformly distributed on $\mathcal{F}$. Now, let us try to compute:

$$\begin{aligned} H(X_1, \ldots, X_n) &\leq H(X_1) + \cdots + H(X_n) \\ &= nH(X_1) \\ \log |\mathcal{F}| &\leq nH(X_1) \\ |\mathcal{F}| &\leq 2^{nH(X_1)} \end{aligned}$$

Now since X_1 was an indicator variable, let us say that it takes value 1 with probability p and value 0 with probability $1 - p$. Then $H(X_1) = H(p)$. We note that the funcion $H(p)$ is increasing for $p \leq 1/2$ (prove it!) and hence $H(X_1) = H(p) \leq H(k/n)$. Thus we will have:

$$|\mathcal{F}| \leq 2^{nH(\frac{k}{n})}$$

■

3.2 A more interesting application

Consider a tripartite graph $G = (V, E)$ with vertex partitions A, B and C. Also, let n_1 be the number of edges between vertices in A and vertices in B, let n_2 be the number of edges between the partitions B and C and n_3 is the number of edges between A and C.

The question that we are interested in: What is the maximum number of triangles such a graph can have?

Clearly, if n is the number of triangles, then a trivial bound is $n \leq n_1 n_2 n_3$. We will now see that this bound can be significantly improved by using entropy.

Let (X, Y, Z) be the vertices of a randomly chosen triangle. With $X \in A$, $Y \in B$, $Z \in C$. Note that we are not choosing the vertices at random, but rather the triangles at random. Thus to be be able to bound n, we would have to look at the joint entropy $H(X, Y, Z)$.

$$\begin{aligned} \log n = H(X,Y,Z) &= H(X,Y) + H(Z|X,Y) \qquad (5) \\ &= \log n_1 + H(Z|X,Y) \qquad (6) \end{aligned}$$

Similarly we also know that:

$$\begin{aligned} \log n = H(X,Y,Z) &= H(Y,Z) + H(X|Y,Z) \qquad (7) \\ &= \log n_2 + H(X|Y,Z) \qquad (8) \end{aligned}$$

Adding the two equations from above, we have:

$$\begin{aligned} 2\log n = H(X,Y,Z) &= \log n_1 + \log n_2 + H(X|Y,Z) + H(Z|X,Y) \\ 2\log n &\leq \log n_1 + \log n_2 + H(Z) + H(X|Z) \\ 2\log n &\leq \log n_1 + \log n_2 + H(X,Z) \\ 2\log n &\leq \log n_1 + \log n_2 + \log n_3 \\ \log n^2 &\leq \log n_1 + \log n_2 + \log n_3 \\ n &\leq \sqrt{n_1 n_2 n_3} \end{aligned}$$

Which is a much sharper bound as compared to the trivial bound discussed a bit earlier. In fact, it is easy to check that this bound is attained by a complete tripartite graph.