Information and Coding Theory **Autumn 2017**

Lecture 2: September 27, 2017

Lecturer: Madhur Tulsiani

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

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

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

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

This is called the *Chain Rule* for Entropy. Note that in the calculations above, we treat $Y \mid X = x$ as a random variable, with distribution given by $\mathbb{P}\left[Y = y\right] = p(y|x)$. Also note that $H(Y|X)$ is a simply a shorthand for the *expected* entropy of $(Y|X = x)$, with the expectation taken over the values for X.

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.

Proposition 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 $H(Y)$. Also, check that above inequality is tight only when X and Y are independent.

Exercise 1.3 *Show that $H(Y) = H(Y|X)$ if and only if X and Y are independent.*

Using induction, we can use the chain rule to show that the following also holds for a tuple of random variables $(X_1, \ldots, X_m)$.

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

Combining this with the fact that conditioning (on average) reduces the entropy, we get the following inequality which is referred to the sub-additivity property of entropy.

$$H(X_1, X_2, \ldots, X_m) \;\leq\; H(X_1) + H(X_2) + H(X_3) + \cdots + H(X_m)\,.$$

Sub-additivity of entropy is very useful in many applications to combinatorics and counting. However, we first use the chain rule to show that the upper bound on the expected code length of $H(X) + 1$ can be improved if we are communicating many symbols and encode a large block of them at once, rather than sending the code for one symbol at a time.

1.1 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 $l_i = \lceil \log(1/p_i) \rceil$).

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:

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

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 and independence:

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

Therefore, we get

$$\mathbb{E}\left[|C(X_1, \ldots, X_m)|\right] \leq m \cdot H(X) + 1 .$$

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

Theorem 1.4 (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.*

1.2 Bounding binomial sums

We use the subadditivity property to obtain an upper bound on the number of subsets of $[m] = \{1, \ldots, m\}$ of size at most k i.e., we need to bound size of the following set

$$S = \{(x_1, \ldots, x_m) \in \{0,1\}^m \mid x_1 + \cdots + x_m \leq k\} .$$

Of course we can write the following expression for the size of S

$$|S| = \binom{m}{0} + \binom{m}{1} + \cdots + \binom{m}{k},$$

but how much *is* the value of the above sum? We will estimate it in terms of the binary entropy function defined as

$$h(p) := p \cdot \log\left(\frac{1}{p}\right) + (1-p) \cdot \log\left(\frac{1}{1-p}\right)$$

Note that $h(p)$ is the entropy of a random variable X, which takes value 1 with probability p and 0 with value $1 - p$ (or vice-versa). This immediately tells us that the maximum possible value of $h(p)$ is 1, which is achieved at $p = 1/2$. The function $h(p)$ can also easily be shown to be concave: we don't need this property now but it will be useful for applications later.

Exercise 1.5 *Prove that the function $h(p)$ is concave, using $H(Y|X) \leq H(Y)$.*

We now return to the estimation problem. Let $(X_1, \ldots, X_m)$ be a uniformly distributed over S. Thus, we have that $H(X_1, \ldots, X_m) = \log |S|$. We can also use sub-additivity to say that

$$\log |S| = H(X_1, \ldots, X_m) \leq H(X_1) + \cdots + H(X_m) = m \cdot H(X_1),$$

where the last equality used the symmetry of the variables $X_1, \ldots, X_m$. 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 $H(X_1) = h(p)$. Also, we have that $X_1 + \cdots + X_m \leq k$, which gives by symmetry that $p = \mathbb{E}[x_1] \leq k/m$. Finally, we note that the function $h(p)$ is increasing for $p \leq 1/2$ (prove it!) and hence $H(X_1) = h(p) \leq h(k/n)$. This gives the bound

$$\log |S| \leq m \cdot h(k/m) \quad \Rightarrow \quad |S| \leq 2^{m \cdot h(k/m)}.$$

You can check that the upper bound obtained here is not too bad since the sum is approximately equal to $\frac{2^{m \cdot h(k/m)}}{\sqrt{2\pi \cdot k \cdot (1-k/n)}}$.

2 Shearer's Lemma and Combinatorial Applications

The sub-additivity property of entropy lets us bound the entropy of the tuple $(X_1, \ldots, X_m)$ in terms of the individual entropies $H(X_1), \ldots, H(X_m)$. Shearer's lemma can be viewed as a generalization of this statement which lets us obtain better bounds in case we can estimate the entropy of subsets of random variables containing more than one random variable.

Lemma 2.1 (Shearer's Lemma) *Let $\{X_1, \ldots, X_m\}$ be a set of random variables. For any $S \subset [m]$, let us denote $X_S = \{X_i : i \in S\}$. Let $\mathcal{F} \subseteq 2^{[m]}$ be a collection of subsets of $[m]$ with the property that for all $i \in [m]$, we have that $|\{S \in \mathcal{F} \mid S \ni i\}| \geq t$. Then*

$$t \cdot H(X_1, \ldots, X_m) \leq \sum_{S \in \mathcal{F}} H(X_S).$$

We will actually prove a more general version of the lemma which can be stated in terms of a distribution over subsets of $[m]$ such that for each $i \in [m]$, we have a lower bound on the probability that a random subset from the distribution includes i. The lemma below can easily be seen to imply the version above, by using the uniform distribution on the collection $\mathcal{F}$.

Lemma 2.2 (Shearer's Lemma: distribution version) *Let $\{X_1, \ldots, X_m\}$ be a set of random variables. For any $S \subset [m]$, let us denote $X_S = \{X_i : i \in S\}$. Let D be an arbitrary distribution on $2^{[m]}$ (set of all subsets of $[m]$) and let μ be such that $\forall i \in [n]$ $\mathbb{P}_{S \sim D}[i \in S] \geq \mu$. Then*

$$\mu \cdot H(X_1, \ldots, X_m) \leq \mathop{\mathbb{E}}_{S \sim D}[H(X_S)].$$

Before proving the lemma, let us consider a few applications of this to counting problems.

2.1 Bounding volumes using projections

Consider a set of points P in (say) three dimensions, such that the projections in the xy, yz and zx plain contain n_1, n_2 and n_3 points respectively. How many points can there be in the set P? Note that since many points in P can have the same projection on a plane, the numbers n_1, n_2 and n_3 can each be much smaller than $|P|$. However, since two different points cannot have the same projection in *all three* planes, we know that each triple of projections must determine a unique point. This gives

$$|P| \;\leq\; n_1 \cdot n_2 \cdot n_3 .$$

It turns out that we can significantly improve this bound using Shearer's lemma. Let (X, Y, Z) be a triple of random variables denoting the coordinates of a uniformly sampled point from P. Thus, we have that $H(X, Y, Z) = \log |P|$. Moreover, using Shearer's lemma, we also get that

$$2 \cdot H(X, Y, Z) \;\leq\; H(X, Y) + H(Y, Z) + H(Z, X) ,$$

since the family of pairs on the right includes each random variable twice. Also, since (X, Y) denotes the projection of a random point from P in the xy plane, and total number of projections is n_1, we get that $H(X, Y) \leq \log n_1$. Similarly, $H(Y, Z) \leq \log n_2$ and $H(Z, X) \leq \log n_3$. Combining these estimates gives

$$2 \cdot \log |P| \;\leq\; \log n_1 + \log n_2 + \log n_3 \quad \Rightarrow \quad |P| \;\leq\; \sqrt{n_1 \cdot n_2 \cdot n_3} .$$

Note that there is nothing special about three dimensions. One can also prove the following d-dimensional analogue using the same argument.

Proposition 2.3 *Let $P \subseteq \mathbb{R}^d$ be a finite set of points in d dimensions, and let $P_1, \ldots, P_d$ denote the set of projections orthogonal to each of the d coordinate axes. Then we have*

$$|P| \;\leq\; \left(\prod_{i=1}^{d} |P_i| \right)^{1/(d-1)} .$$

This can also be used to bound the volume of a body B in d dimensions in terms of the $(d-1)$-dimensional volumes of its projections. One can consider the body to be a union of axis parallel cubes, with a point at the center of each cube. Then, a limiting argument combined with the above estimate gives the following result known as the Loomis-Whitney inequality.

Proposition 2.4 (Loomis-Whitney inequality) *Let $B \subseteq \mathbb{R}^d$ be a measurable body and let $B_1, \ldots, B_d$ denote its projections orthogonal to each of the coordinate axes. Then, we have*

$$\mathsf{Vol}_d(B) \;\leq\; \left(\prod_{i=1}^{d} \mathsf{Vol}_{d-1}(B_i) \right)^{1/(d-1)} .$$

We will cover some more applications of entropy in the next lectures. Many more combinatorial applications can be found in the excellent notes by Radhakrishnan [Rad03] and Galvin [Gal14]. A generalization of Shearer's lemma was also used by Friedgut [Fri04] to derive many inequalities in geometry and analysis.

References

[Fri04] Ehud Friedgut, *Hypergraphs, entropy, and inequalities*, The American Mathematical Monthly **111** (2004), no. 9, 749–760. 7

[Gal14] David Galvin, *Three tutorial lectures on entropy and counting*, arXiv preprint arXiv:1406.7872 (2014). 7

[Rad03] Jaikumar Radhakrishnan, *Entropy and counting*, Computational mathematics, modelling and algorithms **146** (2003). 7