Theorist's Toolkit **Spring 2019**

Lecture 8: Applications of Information Theory

Scribe: Jon Hillery 5/20/2019

8.1 Basic Definitions

Definition 8.1. *For a random variable X with distribution $p : X \to [0,1]$, we define the* ***entropy*** *of the distribution as*

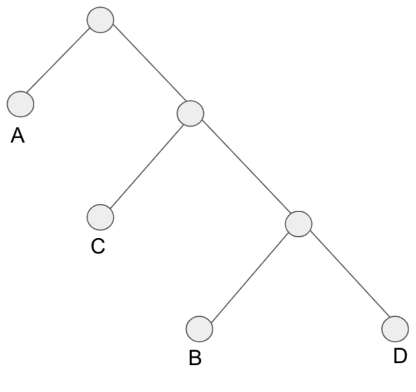
$$H(X) := -\sum_{x \in X} p(x) \log_2 p(x) = -\sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$$

The correct interpretation of this definition is as the "average number of bits" needed to store information about our distribution, which should become more clear through examples. In fact, we see that

$$H(X) = \mathbb{E}(\log_2 \frac{1}{p(x)})$$

and from this representation we can see that since $0 \leq p(x) \leq 1$ for all x, $H(X) \geq 0$ (which makes sense since we should always need some bits). Another important thing to note is that the entropy of X is independent of the values of the random variables; we only care about the distribution of probabilities.

Example 8.2. One familiar example (hopefully) is Huffman encoding. Suppose we have the distribution is $\{A, B, C, D\}$ with respective frequencies $10, 1, 2, 1$, i.e. the probability distribution is $p(A) = \frac{5}{8}, p(B) = \frac{1}{16}, p(C) = \frac{3}{16}, p(D) = \frac{1}{8}$. The Huffman encoding tree is



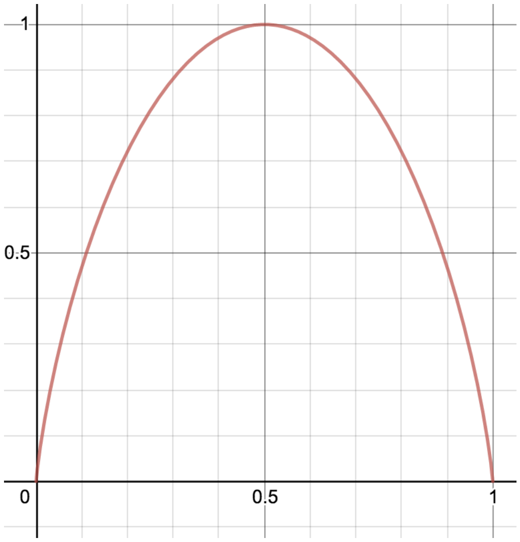
(diagram created with Google Slides) and so we can see the average number of bits needed is $1 \cdot \frac{5}{8} + 2 \cdot \frac{3}{16} + 3 \cdot \frac{1}{16} + 3 \cdot \frac{1}{8} \approx 1.5625$. The entropy of this distribution is $H(X) = -(\frac{5}{8} \log \frac{5}{8} + \frac{1}{16} \log \frac{1}{16} + \frac{3}{16} \log \frac{3}{16} + \frac{1}{8} \log \frac{1}{8}) \approx 1.502$.

Note that the entropy is less than the Huffman bits. This is because even though the Huffman encoding is optimal, it is by nature discrete whereas the entropy is simply an average. However, we do have that the entropy will lower bound the average number of bits we need in our discrete representations.

Example 8.3. Let $X = \{a, b\}$ with probabilities p and $(1 - p)$ respectively. Then

$$H(X) = p \log(\frac{1}{p}) + (1 - p) \log(\frac{1}{1 - p})$$

. As we vary p, this looks like:



(graph taken from Desmos). This fits our intuition, since we would expect that the more "unbalanced" our distribution is, the more certainty we have about the outcome, and thus the fewer number of bits we need to express it on average.

Example 8.4. If we have a uniform distribution on n events (i.e. the probability of each is $\frac{1}{n}$), then

$$H(X) = \sum_{x=1}^{n} \frac{1}{n} \log n = n \cdot \frac{1}{n} \log n = \log n$$

We will show that in fact the uniform distribution is the highest amount of entropy you can have from n events. This should make sense since we have to "allocate information equally", i.e. we start with no information about which event is more likely to happen.

Definition 8.5. *Given a joint distribution (X, Y), the **joint entropy** is*

$$H(X, Y) := \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{1}{p(x, y)} = \mathbb{E}(\log \frac{1}{p(x, y)})$$

Example 8.6. If X and Y are independent, i.e. $p(x, y) = p(x)p(y)$, then

$$H(X, Y) = -\mathbb{E}(\log p(x, y)) = -\mathbb{E}(\log(p(x)p(y))) = -\mathbb{E}(\log p(x) + \log p(y)) = -\mathbb{E}(\log p(x)) - \mathbb{E}(\log p(y)) = H(X) + H(Y)$$

This makes sense since we would expect that if the distributions are independent then the distribution of one shouldn't affect the "disorder" or "average information" of the other.

Definition 8.7. *The **conditional entropy** of Y conditioned on X is defined as*

$$H(Y|X) := \sum_{x \in X} p(x) H(Y|X = x) = -\mathbb{E} \log p(Y|X)$$

Example 8.8. Suppose we have flip two weighted coins with the first having $p(H) = \frac{1}{3}$, $p(T) = \frac{2}{3}$ and the second coin is fair if the first result is H and $p(H) = \frac{1}{4}$, $p(T) = \frac{3}{4}$ if the result is T. The total entropy is then

$$H(Z) = -(\frac{1}{3} \cdot \frac{1}{2} \log(\frac{1}{3} \cdot \frac{1}{2}) + \frac{1}{3} \cdot \frac{1}{2} \log(\frac{1}{3} \cdot \frac{1}{2}) + \frac{2}{3} \cdot \frac{1}{4} \log(\frac{2}{3} \cdot \frac{1}{4}) + \frac{2}{3} \cdot \frac{3}{4} \log(\frac{2}{3} \cdot \frac{3}{4}))$$

Notice that this can be re-arranged as

$$-(\frac{1}{3} \log \frac{1}{3} + \frac{2}{3} \log \frac{2}{3}) + (\frac{1}{3}(-(\frac{1}{2} \log \frac{1}{2} + \frac{1}{2} \log \frac{1}{2})) + \frac{2}{3}(-\frac{1}{4} - \frac{3}{4} \log \frac{3}{4})) = H(X) + H(Y|X)$$

In fact, we can do this in general as we will see in the next section.

8.2 Important Properties

Theorem 8.9 (The Chain Rule)**.** $H(X, Y) = H(X) + H(Y|X)$

Proof.

$$H(X, Y) = -\mathbb{E}(\log p(X, Y)) = -\mathbb{E}(\log(p(Y|X)p(X))) = -\mathbb{E}(\log p(Y|X) + \log p(X))) =$$

$$= -\mathbb{E}(\log p(Y|X)) - \mathbb{E}(\log p(X)) = H(Y|X) + H(X)$$

□

Theorem 8.10 (Jensen's Inequality)**.** *If ϕ is a concave function, then for positive weights a_i,*

$$\phi(\frac{\sum a_i x_i}{\sum a_i}) \geq \frac{\sum a_i \phi(x_i)}{a_i}$$

with equality iff the x_i are equal.

We will need this later, but state it without proof, as it is a well-known tool in probability.

Theorem 8.11. *$H(Y) \geq H(Y|X)$.*

This makes sense since conditioning "reduces the amount of disorder", i.e. gives you more information.

Proof.

$$H(Y|X) - H(Y) = -\mathbb{E}(\log p(Y|X)) + \mathbb{E}(\log p(Y)) = \mathbb{E}(\log \frac{p(Y)}{p(Y|X)}) = \mathbb{E}(\log \frac{p(X)p(Y)}{p(X, Y)})$$

By Jensen's Inequality (since log is a concave function), we have that

$$\mathbb{E}(\log \frac{p(X)p(Y)}{p(X,Y)}) \leq \log \mathbb{E}(\frac{p(X)p(Y)}{p(X,Y)}) = \log(\sum_{x,y} p(x,y)\frac{p(x)p(y)}{p(x,y)}) = \log(\sum_{x,y} p(x)p(y)) = \log(1) = 0$$

and therefore

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

□

Corollary 8.12. *$H(X,Y) \leq H(X) + H(Y)$ (since $H(X,Y) = H(X) + H(Y|X) \leq H(X) + H(Y)$).*

Theorem 8.13. *$H(X) \leq \log(|\mathcal{X}|)$, where $\mathcal{X}$ is the set of events that occur with nonzero probability (the **support**), with equality iff X is a uniform distribution.*

Proof. $H(X) = \sum_{x \in X} p(x) \log(\frac{1}{p(x)}) \leq \log \sum_{x \in X} p(x)\frac{1}{p(x)}$ by Jensen's inequality. This quantity is just $\log(|\mathcal{X}|)$, completing the proof (equality follows from the equality condition of Jensen's inequality). □

8.3 Applications

8.3.1 Tripartite Triangles

Suppose we have a tripartite graph of partitions A, B, C and we want to have a bound on the number of triangles T (cycles with one vertex in each partition set) in the graph given the number of edges n_1 between A and B, n_2 between B and C, and n_3 between A and C. Clearly $T \leq n_1 n_2 n_3$, but can we do better?

Pick the triangles uniformly. Since the distribution is uniform we have

$$\log T = H(X, Y, Z)$$

and by the chain rule

$$\log T = H(X,Y) + H(Z|X,Y) \leq n_1 + H(Z|X,Y)$$

and similarly

$$\log T \leq n_2 + H(X|Y,Z)$$

Adding these, we get

$$2 \log T \leq n_1 + n_2 + H(Z|X,Y) + H(X|Y,Z) \leq n_1 + n_2 + H(Z) + H(Z|Z) = n_1 + n_2 + n_3$$

and thus

$$T \leq \sqrt{n_1 n_2 n_3}$$

8.3.2 Directed Vees and Triangles

Let G be a directed graph. Can we related the number of triangles to the number of vees (a vertex and two directed edges leaving it, but they are allowed to be the same edge)? Our intuition says that the number of vees should be at least the number of triangles, but this is hard to show.

Again, take a uniformly random triangle, giving

$$\log T = H(X, Y, Z) = H(X) + H(Y, Z|X) = H(X) + H(Y|X) + H(Z|X, Y) \leq$$

$$\leq H(X) + H(Y|X) + H(Z|Y) = H(X) + 2H(Y|X)$$

by symmetry.

Now, let's create a distribution on the vee's (a, b, c) based on the triangle distribution:

$$Pr(a, b, c) = Pr[X = a] \cdot Pr[Y = b|X = a] \cdot Pr[Y = c|X = a]$$

The entropy of this distribution is

$$H(A, B, C) = H(A) + H(B|A) + H(C|A) = H(X) + 2H(Y|X)$$

and since the uniform distribution has the highest entropy, we have that

$$\log V \geq H(X) + 2H(Y|X)$$

but this is the same bound we had for $\log T$, giving

$$\log V \geq \log T \implies V \geq T$$

8.4 References