COS597D: Information Theory in Computer Science	September 19, 2011
Lecture 1	
Lecturer: Mark Braverman	*Scribe: Mark Braverman*[*]

1 Introduction

Information theory is the study of a broad variety of topics having to do with quantifying the amount of information carried by a random variable or collection of random variables, and reasoning about this information. It gives us tools to define and reason about fundamental quantities in a broad spectrum of disciplines. Information has been traditionally studied in the context of communication theory and in physics (statistical mechanics, quantum information). However, has many important applications in other fields, such as economics, mathematics, statistics, and, as we shall see in this course, in theoretical computer science.

In this course, we shall see a quick primer on basic concepts in information theory, before switching gears and getting a taste for a few domains in discrete mathematics and computer science where these concepts have been used to prove beautiful results.

2 How to measure information?

There are several ways in which one might try to measure the information of a variable X.

- We could measure how much space it takes to store X. Note that this definition only makes sense if X is a random variable. If X is a random variable, it has some distribution, and we can calculate the amount of memory it takes to store X *on average*.

  For example, if X_1 is a uniformly random n-bit string, it takes n-bits of storage to write down the value of X_1. If X_2 is such that there is a string a such that $\Pr[X_2 = a] = 1/2$, we can save on the space by encoding X_2 so that 0 represents a and all other strings are encoded with a leading 1. With this encoding, the expected amount of space we need to store X_2 is at most

  $$\underbrace{\frac{1}{2}\cdot 1}_{\text{amount of memory if } a \text{ is selected}} + \underbrace{\frac{1}{2}\cdot (n+1)}_{\text{amount of memory otherwise}} = n/2 + 1.$$

  This notion corresponds to Shannon's entropy, which we will define soon.

- We could measure how unpredictable X is in terms of what the probability of success is for an algorithm that tries to guess the value of X before it is sampled. If X_1 is a uniformly random string, the best probability is 2^{-n}. On the other hand, for X_2 is as in the second example, this probability is $1/2$. This corresponds to the notion of "min-entropy". A simple example where it may be important is in assessing the strength of a cryptographic key generation scheme (want the min-entropy to be high).

- We could measure the expected length of the shortest computer program that prints out X (namely the expected Kolmogorov complexity of X). This definition has the advantage of being applicable not only to random variables, but also to individual random strings. For example, it explains why the string "29035472908579286345..." is more random than "4444444444444..." or "31415926535897..." — the former requires a much longer program.

[*]Based on lecture notes by Anup Rao

There are just some of the options we have for measuring the information in X. Each of them sounds reasonable, and there are many other reasonable measures that make sense. Research in pseudorandomness has made gains by asking for analogous measures to those suggested above, under computational restrictions — i.e. what is the shortest encoding of X with an efficient encoding algorithm, or the best probability of predicting X using an efficient predictor algorithm. In this course, we shall focus on a few such measures, and explore applications of these ideas to problems in computer science.

3 Notation

Capital letters like X, Y, Z will be used to denote random variables, letters like S, T, U will denote sets. Calligraphic letters like $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ will be used to denote the supports of random variables (X, Y, Z). Small letters like x, y, z will be used to denote instantiations of random variables, and also elements of sets (s, t, u).

We shall use the shorthand $p(x)$ to denote the probability $\Pr[X = x]$, and $p(x, y)$ to denote $\Pr[X = x, Y = y]$. For conditional probability, we shall often use the notation $p(x|y)$ to denote $\Pr[X = x|Y = y]$.

4 The Entropy Function

The measure of information we shall focus on to start is the *entropy* function, defined as follows:

$$H(X) = \sum_x p(x) \log(1/p(x)),$$

where here we adopt the convention that $0 \log 1/0 = 0$ (which is justified by the fact that $\lim_{x \to 0} x \log(1/x) = 0$). Another way to interpret $H(X)$ is as the expected log of the probability of a sample from X,

$$H(X) = \mathbb{E}\left[\log(1/p(X))\right].$$

The log function here (and throughout most of the course) is the base-2 $\log_2$ function.

For $q \in [0, 1]$, we shall write $H(q)$ to denote the entropy of the Bernoulli random variable B_q for which $\Pr[B_q = 1] = q$ and $\Pr[B_q = 0] = 1 - q$.

Some facts are immediate from the definition:

- $H(X) \geq 0$ (since each term in the sum is non-negative). Moreover, $H(X) = 0$ if and only if $\Pr[X = a] = 1$ for some a, since otherwise, one of the terms in the sum will be strictly positive.
- $H(q) = H(1 - q)$.
- $H(1/2) = 1$.
- $H(0) = H(1) = 0$.

5 Some Examples

- For X_1 as above – a uniformly random n-bit string – $H(X_1) = \sum_x p(x) \log(1/p(x)) = \sum_x 2^{-n} \log(2^n) = n$.
- If X is a uniformly random element of a set S, $H(X) = \sum_{x \in S} (1/|S|) \log(|S|) = \log(|S|)$.
- For X_2 as above, assuming a is not an n-bit string,

$$H(X_2) = \frac{1}{2} \log 2 + \sum_{x \text{ is an } n\text{-bit string}} 2^{-(n+1)} \log(2^{n+1}) = \frac{1}{2} + \frac{n+1}{2} = \frac{n}{2} + 1.$$

- Let X_3, Y_3, Z_3 be uniformly random bits conditioned on their majority being 1, then $H(X_3Y_3Z_3) = 2$, since there are 8 3-bit strings, of which exactly 4 have majority 1. Each of the bits X_3, Y_3, Z_3 is 1 with probability 3/4 and 0 with probability 1/4. Hence

$$H(X_3) = H(Y_3) = H(Z_3) = H(3/4) \approx 0.81.$$

  If we look at the pair of bits X_3Y_3 then their distribution is given by $p(01) = p(10) = 1/4$, $p(11) = 1/2$, and thus by symmetry

$$H(X_3Y_3) = H(X_3Z_3) = H(Y_3Z_3) = \frac{1}{4}\log 4 + \frac{1}{4}\log 4 + \frac{1}{2}\log 2 = 1.5.$$

- Let X_4, Y_4, Z_4 be uniformly random bits conditioned on their parity being 0 (i.e. $X_4 + Y_4 + Z_4 = 0 \mod 2$), then $H(X_4Y_4Z_4) = 2$, since again, the fraction of such strings is 1/2. $H(X_4) = 1$, and $H(X_4Y_4) = 2$. In this case, we see that $H(X_4Y_4) = H(X_4Y_4Z_4)$, indicating that the first two bits already contain all information about the entire string. Indeed this is the case since $Z_4 = X_4 \oplus Y_4$. We shall formalize a way to measure how much information is left over in the last bit soon.

6 Conditional Entropy

Let X, Y be two random variables. Then, expanding $H(X, Y)$ gives

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

We denote the second term above $H(Y|X)$. It is the expected entropy that is left in Y after fixing X. In this notation, we have just showed the *chain rule*:

Lemma 1 (Chain Rule)**.** $H(XY) = H(X) + H(Y|X)$.

Repeated applications of this two variable chain rule give:

Lemma 2 (Chain Rule)**.** $H(X_1X_2\ldots X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_2X_1) + \cdots + H(X_n|X_{n-1}\ldots X_1)$.

Revisiting our examples, we see that in the case when $X_4Y_4Z_4$ are three random bits conditioned on the event that their parity is 0, we have that $H(X_4) = 1$, $H(Y_4|X_4) = 1$ and $H(Z_4|X_4Y_4) = 0$, where the last equation states that for every fixing of X_4, Y_4, the value of Z_4 is determined. On the other hand, when $X_3Y_3Z_3$ are random bits whose majority is 1, observe that when $X_3 = 1, Y_3 = 1$, the last bit has some entropy. Therefore $H(Z_3|X_3Y_3) > 0$. Indeed, we can calculate that $H(Z_3|X_3Y_3) = H(X_3Y_3Z_3) - H(X_3Y_3) = 2 - 1.5 = 0.5$ bits.

7 Jensen's Inequality and Subadditivity

Definition 3. *We say that a function $f : (a, b) \to \mathbb{R}$ is* convex *if for every $x, y \in (a, b)$ and every $\lambda \in (0, 1)$,*

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$

Examples of convex functions include x, e^x, x^2 and $\log(1/x)$. If $-f$ is convex, we shall say that f is *concave*.

The following is a useful inequality for dealing with the entropy function and its derivatives:

Lemma 4 (Jensen's Inequality)**.** *If f is a convex function on (a, b) and X is a random variable taking values in (a, b), then*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

Proof We prove the case when X takes on finitely many values. The general case follows by continuity arguments.

We prove the statement by induction on the number of elements in the support of X. If X is supported on 2 elements, the lemma immediately follows from the definition of convexity. In the general case, let us assume X is supported on $x_1, \ldots, x_n$. Then,

$$\begin{aligned}
\mathbb{E}[f(X)] &= p(x_1)f(x_1) + \sum_{i=2}^{n} p(x_i)f(x_i) \\
&= p(x_1)f(x_1) + (1 - p(x_1)) \sum_{i=2}^{n} p(x_i)f(x_i)/(1 - p(x_1)) \\
&\leq p(x_1)f(x_1) + (1 - p(x_1))f\left(\sum_{i=2}^{n} p(x_i)x_i/(1 - p(x_1))\right) \\
&\leq f\left(p(x_1)x_1 + (1 - p(x_1))\left(\sum_{i=2}^{n} p(x_i)x_i/(1 - p(x_1))\right)\right) \\
&= f(\mathbb{E}[X]),
\end{aligned}$$

where the first inequality follows by applying the lemma for the case when there are $n - 1$ elements in the support, and the second inequality is a direct consequence of the definition of convexity. ■

As a first application of this inequality, we show the following lemma:

Lemma 5 (Subadditivity of Entropy)**.** $H(X, Y) \leq H(X) + H(Y)$

Proof

$$\begin{aligned}
H(X, Y) - H(X) - H(Y) &= \sum_{x,y} p(x, y) \log(1/p(x, y)) - \sum_{x} p(x) \log(1/p(x)) - \sum_{y} p(y) \log(1/p(y)) \\
&= \sum_{x,y} p(x, y) \log(1/p(x, y)) - \sum_{x,y} p(x, y) \log(1/p(x)) - \sum_{x,y} p(x, y) \log(1/p(y)) \\
&= \sum_{x,y} p(x, y) \log(p(x)p(y)/p(x, y)) \\
&\leq \log\left(\sum_{x,y} p(x, y)p(x)p(y)/p(x, y)\right) \\
&= \log 1 = 0,
\end{aligned}$$

where the inequality follows from Jensen's inequality applied to the convex function $\log(1/x)$. ■

Note that the above lemma implies in particular that $H(X) + H(Y|X) \leq H(X) + H(Y)$, which means that $H(Y|X) \leq H(Y)$. In other words, conditioning can only reduce the entropy in a random variable on average.

Lemma 6. $H(Y|X) \leq H(Y)$

It is NOT true that $H(Y|X = x)$ is always smaller than $H(Y)$. Indeed, in the example where $X_3Y_3Z_3$ are three uniform bits conditioned on the majority being 1, we see that $H(X_3) = H(3/4) \approx 0.81 < 1$, yet $H(X|Y = 1, Z = 1) = 1$. However, the lemma shows that *average* fixings of Y_3Z_3 do reduce the entropy in X_3. And indeed $H(X_3|Y_3Z_3) = 0.5 < H(X_3)$.