COS597D: Information Theory in Computer Science	September 28, 2011
Lecture 4	
Lecturer: Mark Braverman	*Scribe: Abhishek Bhowmick*[*]

1 Convexity/Concavity of Mutual Information

In the previous lecture, we saw that mutual information is concave in p. To be more precise, let (X,Y) have a joint probability distribution $p(x, y) = p(x)p(y|x)$. Write $\alpha = \alpha(x) = p(x)$ and $\pi = \pi(x, y) = p(y|x)$. Then the pair (α, π) specifies the distribution $p(x, y)$.

Lemma 1 (Mutual information is concave in p)**.**
Let I_1 *be* $I(X;Y)$ *where* $(X, Y) \sim (\alpha_1, \pi)$*,*
let I_2 *be* $I(X;Y)$ *where* $(X, Y) \sim (\alpha_2, \pi)$*,*
let I *be* $I(X;Y)$ *where* $(X, Y) \sim (\lambda\alpha_1 + (1 - \lambda)\alpha_2, \pi)$*, for some* $0 \le \lambda \le 1$*.*
then $I \ge \lambda I_1 + (1 - \lambda)I_2$*.*

Now, we prove that mutual information is convex in $p(y|x)$. More formally, we have the following. Let (X, Y) have a joint probability distribution $p(x, y) = p(x)p(y|x)$. Write $\alpha = \alpha(x) = p(x)$ and $\pi = \pi(x, y) = p(y|x)$. Then the pair (α, π) specifies the distribution $p(x, y)$.

Lemma 2 (Mutual information is convex in π)**.** *Let* I_1 *be* $I(X;Y)$ *where* $(X, Y) \sim (\alpha, \pi_1)$*,*
let I_2 *be* $I(X;Y)$ *where* $(X, Y) \sim (\alpha, \pi_2)$*,*
let I *be* $I(X;Y)$ *where* $(X, Y) \sim (\alpha, \lambda\pi_1 + (1 - \lambda)\pi_2)$*, for some* $0 \le \lambda \le 1$*.*
then $I \le \lambda I_1 + (1 - \lambda)I_2$*.*

Proof Let us draw X first according to α. Let S be a B_λ random variable such that S is 1 with probability λ and 0 with probability $1 - \lambda$. If $S = 1$ we select Y using π_1, and otherwise we select Y using π_2. Note that $I(X;Y) = I$.

$$I(SY;X) = I(Y;X) + I(S;X|Y) \ge I(Y;X) = I$$

Also, we have

$$\begin{aligned}
& I(SY;X) \\
= \;& I(S;X) + I(Y;X|S) \\
= \;& 0 + I(Y;X|S) \\
= \;& \lambda I(Y;X|S = 1) + (1 - \lambda)I(Y;X|S = 0) \\
= \;& \lambda I_1 + (1 - \lambda)I_2
\end{aligned}$$

Thus, we have $I \le \lambda I_1 + (1 - \lambda)I_2$.
■

2 Some more inequalities involving mutual information

Lemma 3. *If* $I(B;D|A, C) = 0$*, then* $I(A;B|C) \ge I(A;B|C, D)$*.*

[*]Based on lecture notes by Anup Rao and Punyashloka Biswal

Proof We write $I(A, D; B|C)$ in two different ways.

$$I(A, D; B|C) = I(A; B|C) + I(D; B|A, C) = I(A; B|C)$$

Also,

$$I(A, D; B|C) = I(D; B|C) + I(A; B|D, C) \geq I(A; B|C, D)$$

Therefore, $I(A; B|C) \geq I(A; B|C, D)$ ■

Lemma 4. *If $I(B; D|C) = 0$, then $I(A; B|C) \leq I(A; B|C, D)$.*

Proof We write $I(A, D; B|C)$ in two different ways.

$$I(A, D; B|C) = I(A; B|C) + I(D; B|A, C) \geq I(A; B|C)$$

Also,

$$I(A, D; B|C) = I(D; B|C) + I(A; B|D, C) = I(A; B|C, D)$$

Therefore, $I(A; B|C) \leq I(A; B|C, D)$ ■

Lemma 5. *If $X \to Y \to Z$ form a Markov chain, then $I(X, Z) \leq I(X, Y)$.*

Proof We write $I(X; YZ)$ in two different ways.

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

since Z is independent of X given Y by the Markov chain property. Also,

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

Thus, $I(X; Z) \leq I(X, Y)$. ■

3 Graph Entropy

Now, we shall study a quantity called *graph entropy*. The original motivation for this quantity was to characterize how much information can be communicated in a setting where pairs of symbols may be confused, though we shall see that it is very useful in a variety of settings.

A subset S of the vertices V of an undirected graph $G = (V, E)$ is *independent* if no edge in the graph has both endpoints in S. Given a graph G, define the graph entropy of G

$$H(G) = \min_{X,Y} I(X; Y),$$

where the minimum is taken over all pairs of random variables X, Y such that

- X is a uniformly random vertex in G.
- Y is an independent set containing X.

Let us consider some examples:

1. Suppose G has no edges. Then if X is a uniformly random vertex and Y is fixed to be the vertex set V, we get $H(G) \leq I(X; Y) = 0$. But $H(G) \geq 0$, so $H(G)$ must be 0 in this case.

2. Let G be the complete graph on n vertices. Then the only independent set containing a given vertex u is the singleton set $\{u\}$. Thus there is only one available choice for the distribution of X, Y, namely $\Pr[Y = \{X\}] = 1$. $H(G) = H(X) - H(X \mid Y) = \log n - 0$, because X is completely determined by $Y = \{X\}$.

3. Let G be the complete bipartite graph $K_{n,n}$. Call the two parts of the graph A and B. One possible choice of joint distribution for X and Y is to first pick X uniformly at random, and then to choose

$$Y = \begin{cases} A & \text{if } X \in A \\ B & \text{otherwise.} \end{cases}$$

This gives us the upper bound

$$H(G) \leq I(X;Y) = H(X) - H(X \mid Y) = \log(2n) - \log n = 1.$$

On the other hand, we claim that any valid joint distribution must satisfy $H(X \mid Y) \leq \log n$. For if Y is an independent set, then it must be a subset of either A or B. Thus, $H(X|Y) \leq \log |Y| \leq \log n$. This implies that $H(G) \geq \log(2n) - \log n = 1$.

4. Let G be a complete r-partite graph, i.e., $V = [n] \times [r]$ and $E = \{\big((i,j),(k,l)\big) \mid j \neq l\}$. Then we can adapt the proofs from the last two examples to show that $H(G) = \log r$. In fact, we can show further that if G is r-partite with parts $S_1, \ldots, S_r$, the graph entropy of G is the same as $H(Z)$, where $\Pr[Z = i] = \Pr[X \in S_i]$ for uniform vertex X. In particular, $H(G) \leq \log r$ in this case.

5. Let G be the unbalanced complete bipartite graph $K_{m,n}$. We choose X and Y exactly as before and get the bound

$$H(G) \leq \log(m+n) - \frac{m}{m+n}\log m - \frac{n}{m+n}\log n = H\left(\frac{n}{m+n}\right),$$

where $H(\cdot)$ denotes the binary entropy function, or the entropy of a biased coin. As in the previous case, we have that $H(X|Y) \leq \frac{m}{m+n}\log m + \frac{n}{m+n}\log n$, proving that $H(G) = H(\frac{n}{m+n})$.

4 Useful Properties of Graph Entropy

The power of graph entropy comes from the fact that it can be easily controlled even when the underlying graph is manipulated in natural ways.

Proposition 6 (Subadditivity)**.** *Let $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ be graphs on the same vertex set. Then their union $G = (V, E_1 \cup E_2)$ has entropy $H(G) \leq H(G_1) + H(G_2)$.*

Proof Let $p_1(x,y)$ and $p_2(x,y)$ be the distributions that minimize $I(X;Y)$ for G_1 and G_2, respectively, and let us consider the distribution

$$p(x, y_1, y_2) = p(x) \cdot p_1(y_1 \mid x) \cdot p_2(y_2 \mid x).$$

In other words, we pick X uniformly at random, and conditioned on this choice of X we pick Y_1 and Y_2 independently according to each of the conditional distributions. For a given choice of X, observe that $Y_1 \cap Y_2$ contains X and is an independent set in G. Therefore,

$$
\begin{aligned}
& H(G) \\
\leq\; & I(X;(Y_1 \cap Y_2)) \\
\leq\; & I(X;Y_1,Y_2) \\
=\; & H(Y_1,Y_2) - H(Y_1,Y_2 \mid X) \\
=\; & H(Y_1,Y_2) - H(Y_1 \mid X) - H(Y_2 \mid X) \quad \text{(since } Y_1, Y_2 \text{ are independent conditioned on any fixing of } X\text{)} \\
\leq\; & H(Y_1) - H(Y_1 \mid X) + H(Y_2) - H(Y_2 \mid X) \quad \text{(by subadditivity of entropy)} \\
=\; & H(G_1) + H(G_2).
\end{aligned}
$$

■

To give an interesting example where Proposition 6 is tight, consider the representation of the complete graph $G := K_{2^n}$ as a graph on strings $V = \{0,1\}^n$. We've seen that $H(K_{2^n}) = n$. Let

$$E_i := \{(u,w) : \; u_i \neq w_i\},$$

i.e. all pairs of strings that differ in the i-th coordinate. Then $G_i = (V, E_i)$ is a complete balanced bipartite graph, and thus $H(G_i) = 1$. We see that the inequality $H(G) \leq \sum_{i=1}^n H(G_i)$ is tight in this case.