Information and Coding Theory **Autumn 2017**

Lecture 3: October 2, 2017

Lecturer: Madhur Tulsiani

1 Shearer's lemma and applications

In the previous lecture, we saw the following statement of Shearer's lemma.

Lemma 1.1 (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*

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

We saw some applications of this lemma in the previous lecture and will see some more in this one. However, let us first prove the lemma.

Proof: The proof of the lemma follows simply from the chain rule for entropy and the fact that conditioning reduces entropy (on average).

$$\begin{aligned}
\mathop{\mathbb{E}}_{S\sim D}[H(X_S)] &= \mathop{\mathbb{E}}_{S\sim D}\left[\sum_{i\in S} H(X_i|X_{S\cap[i-1]})\right] && \text{by Chain rule} \\
&\geq \mathop{\mathbb{E}}_{S\sim D}\left[\sum_{i\in S} H(X_i|X_{[i-1]})\right] && H(X_i|X_A) \geq H(X_i|X_B) \text{ for } A \subset B \\
&= \mathop{\mathbb{E}}_{S\sim D}\left[\sum_{i\in [n]} \mathbb{1}_S(i) \cdot H(X_i|X_{[i-1]})\right] && \mathbb{1}_s \text{ is indicator function for set } S \\
&= \sum_{i\in[n]} \mathop{\mathbb{P}}_{S\sim D}[i \in S] \cdot H(X_i|X_{[i-1]}) \\
&\geq \mu \cdot \sum_{i\in[n]} H(X_i|X_{[i-1]}) = \mu \cdot H(X_1, \ldots, X_m)
\end{aligned}$$

■

1.1 Counting graph homomorphisms

Shearer's lemma can be used to give an estimate of the number of ways of "embedding" a small graph G into a large graph H. For two graphs $G : (V_G, E_G)$ and $H = (V_H, E_H)$, an embedding (also called a homomorphism) of G in H is defined as a function $f : V_G \to V_H$ such that for all $(u, v) \in E_G$, we have $(f(u), f(v)) \in V_H$. Note that the definition does not prevent the image of non-edge pairs in E_G from being edges in E_H.

We will show an upper bound on the maximum number of embeddings for a graph G into any H with at most m edges. For now, let us take G to be the 5-cycle with vertex set $\{1, 2, 3, 4, 5\}$. Consider any graph H with at most m edges and let $F = (F(1), \ldots, F(5))$ be a collection of random variables denoting an embedding of G chosen uniformly from the set of all embeddings. Using Shearer's lemma, we can write

$$2 \cdot H(F(1), \ldots, F(5)) \;\leq\; H(F(1), F(2)) + H(F(2), F(3)) + \cdots + H(F(5), F(1)) \,.$$

Since $\{1, 2\}$ is an edge in G, the pair $(F(1), F(2))$ must correspond to an (ordered) edge in H. Since the number of edges in H is at most m, we get that $H(F(1), F(2)) \leq \log(2m)$. Using the same bound for all terms on the right, we get

$$H(F(1), \ldots, F(5)) \;\leq\; \frac{5}{2} \cdot \log(2m) \,,$$

which gives a bound of $(2m)^{5/2}$ on the number of embeddings.

Exercise 1.2 *Check that the exponent of* $5/2$ *in the above bound is tight.*

The above method can also be used to give a tight estimate for any graph G (of constant size). In general, the exponent depends on a parameter known as the *fractional independent set number* of G. I will divide this proof in a few parts and add this as an extra problem in the homework. The solution to this problem need not be submitted. The proof, along with many other combinatorial applications can also be found in the surveys by Radhakrishnan [Rad03] and [Gal14].

1.2 An application to counting primes

Consider the problem of finding the number of prime numbers less than or equal to a given number n. This quantity, denoted by $\pi(n)$ is the subject of the famous prime number theorem [Gol73] which shows that

$$\pi(n) \;\sim\; \frac{n}{\ln n} \,.$$

Here, $f(n) \sim g(n)$ is used to denote $\lim_{n\to\infty}(f(n)/g(n)) = 1$. A weaker estimate was proved by Chebyshev, who proved that for sufficiently large n,

$$c_1 \cdot \frac{n}{\ln n} \;\leq\; \pi(n) \;\leq\; c_2 \cdot \frac{n}{\ln n}\,.$$

Chebyshev, indeed used (and proved) the following estimate to obtain his bounds:

$$\sum_{p\leq n} \frac{\log p}{p} \;\sim\; \log n\,.$$

An information theoretic proof of the above estimate was obtained by Kontoyiannis [Kon07, Kon08]. We sketch the argument for one side of this estimate below.

Let us first obtain a very simple lower bound on $\pi(n)$. For a given n, let $k = \pi(n)$ and let $p_1, \ldots, p_k$ be the prime numbers in $[n]$. Let R be a uniform random number in $[n]$. Using prime factorization, we can express R as

$$R \;=\; p_1^{X_1} \cdots p_k^{X_k}\,,$$

for random variables $X_1, \ldots, X_k$. Using sub-additivity of entropy, we can write

$$\log n \;=\; H(R) \;=\; H(X_1, \ldots, X_k) \;\leq\; H(X_1) + \cdots + H(X_k)\,.$$

Now, notice that each random variable X_i takes value at most $\log n$ since

$$2^{X_i} \;\leq\; p_i^{X_i} \;\leq\; n\,.$$

Thus, we have $H(X_i) \leq \log(\log n + 1)$ for each X_i. This gives

$$\pi(n) \;=\; k \;\geq\; \frac{\log n}{\log(\log n + 1)}\,.$$

We can refine this estimate significantly by noticing that we actually used a terrible bound on $H(X_i)$ for each i. This bound would hold only if X_i was uniformly distributed in $\{0, \ldots, \log n\}$, which is quite far from it's actual distribution. Each X_i is approximately distributed as a geometric random variable with $\mathbb{P}[X_i \geq t] \approx (1/p_i)^t$. Plugging in the entropy of a geometric random variable gives

$$H(X_i) \;\approx\; \frac{\log p}{p-1} - \log\left(1 - \frac{1}{p}\right) \;\approx\; \frac{\log p}{p}\,.$$

Substituting this bound gives

$$\log n \;\leq\; \sum_{p\leq n} \frac{\log p}{p} \;+\; o(\log n)\,.$$

2 Mutual Information

The mutual information is a quantity which measures the amount of dependence between two random variables. Unlike correlation, which defines the random variables to take values in the same space, the mutual information can be defined for any two random variables. The mutual information between two random variables X and Y is defined by the formula

$$I(X;Y) \;=\; H(X) - H(X|Y)$$

Using the chain rule for entropy, we can see that

$$I(X;Y) \;=\; H(X) - H(X|Y) \;=\; H(Y) - H(Y|X) \;=\; H(X) + H(Y) - H(X,Y)\,.$$

We can use the first two expressions to observe that $I(X;Y) \geq 0$ and the last one to observe that $I(X;Y) = I(Y;X)$.

Example 2.1 *Consider the random variable (X,Y) with $X \vee Y = 1$, $X \in \{0,1\}$ and $Y \in \{0,1\}$ such that:*

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

Then, we can calculate the entropy and mutual information as follows:

$$\begin{aligned} H(X) \;=\; H(Y) \;&=\; \frac{1}{3}\log 3 + \frac{2}{3}\log\frac{3}{2} \;=\; \log 3 - \frac{2}{3} \\ H(X,Y) \;&=\; \log 3 \\ I(X;Y) \;=\; H(X) + H(Y) - H(X,Y) \;&=\; \log 3 - \frac{4}{3} \end{aligned}$$

Conditioning on a third random variable Z, we can also define the conditional mutual information $I(X;Y|Z)$ as

$$\begin{aligned} I(X;Y|Z) \;&:=\; \mathop{\mathbb{E}}_{z}\left[I(X|Z=z;Y|Z=z)\right] \\ &=\; \mathop{\mathbb{E}}_{z}\left[H(X|Z=z) - H(X|Y,Z=z)\right] \\ &=\; H(X|Z) - H(X|Y,Z)\,. \end{aligned}$$

Consider the following example of three random variables.

Example 2.2 *Consider the random variable (X,Y,Z), $X \in \{0,1\}$, $Y \in \{0,1\}$ and $Z = X \oplus Y$ such that:*

$$(X,Y,Z) = \begin{cases} 000 & \textit{w.p 1/4} \\ 011 & \textit{w.p 1/4} \\ 101 & \textit{w.p 1/4} \\ 110 & \textit{w.p 1/4} \end{cases}$$

We can check that in this case, X, Y are independent and thus $I(X;Y) = 0$. However,

$$\begin{aligned} I(X:Y|Z) &= \mathop{\mathbb{E}}_{z}\left[I(X|Z=z;Y|Z=z)\right] \\ &= \frac{1}{2}I(X|Z=0;Y|Z=0) + \frac{1}{2}I(X|Z=1;Y|Z=1) \\ &= \frac{1}{2}\log 2 + \frac{1}{2}\log 2 \;=\; 1 \end{aligned}$$

The above example illustrates that unlike entropy, it is not true that conditioning (on average) decreases the mutual information. In the above example, while $I(X;Y) = 0$, we have $I(X;Y|Z) = 1$ which is in fact the maximum possible. However, as in the case of entropy, mutual information does obey a chain rule.

Lemma 2.3 $I((X_1,\ldots,X_m);Y) = \sum_{i=1}^{m} I(X_i;Y|X_1,\ldots,X_{i-1})$

Proof: The chain rule for mutual information is a simple consequence of the chain rule for entropy. We have

$$\begin{aligned} I((X_1,\ldots,X_m);Y) &= H(X_1,\ldots,X_m) - H(X_1,\ldots,X_m|Y) \\ &= \sum_{i=1}^{m} H(X_i|X_1,\ldots,X_{i-1}) - \sum_{i=1}^{m} H(X_i|Y,X_1,\ldots,X_{i-1}) \\ &= \sum_{i=1}^{m} \left[H(X_i|X_1,\ldots,X_{i-1}) - H(X_i|Y,X_1,\ldots,X_{i-1})\right] \\ &= \sum_{i=1}^{m} I(X_i;Y|X_1,\ldots,X_{i-1}) \end{aligned}$$

■

References

[Gal14] David Galvin, *Three tutorial lectures on entropy and counting*, arXiv preprint arXiv:1406.7872 (2014). 2

[Gol73] Larry J Goldstein, *A history of the prime number theorem*, The American Mathematical Monthly **80** (1973), no. 6, 599–615. 2

[Kon07] Ioannis Kontoyiannis, *Some information-theoretic computations related to the distribution of prime numbers*, arXiv preprint arXiv:0710.4076 (2007). 3

[Kon08] ———, *Counting the primes using entropy*, IEEE Information Theory Society Newsletter (2008), 6–9. 3

[Rad03] Jaikumar Radhakrishnan, *Entropy and counting*, Computational mathematics, modelling and algorithms **146** (2003). 2