COMS 6998-2: Information Theory in Theoretical CS September 12, 2017

Lecture 2 - Source Coding, Conditional Entropy, Shearer's Lemma

Instructor: *Omri Weinstein* Scribes: *Anand Sundaram, Sandip Sinha*

1 Review of Key Concepts

1.1 Entropy

In the previous (first) lecture, we introduced entropy conceptually as a measure of surprise. The more surprised we would be by a particular observation, the higher its entropy should be. With this motivation, we derived the definition that for any observation $x \sim \mu$ (the event x observed in a context where the underlying probability distribution over events is μ), the entropy of this observation is:

$$H_\mu(x) := \sum_i \mu_i \log \frac{1}{\mu_i}$$

Here, μ_i is the probability of event i occurring under μ.

A special case of this is binary entropy, where μ is a Bernoulli distribution $Ber(p)$ where there are only two possible events occurring with probabilities p and $1-p$ respectively. Without loss of generality, we can consider μ to be a fair coin biased to flip heads with probability p. In this case we have:

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

1.2 Lossless Compression

Changing perspectives, we considered an operational interpretation of the same entropy measure defined above as the expected cost (in bits, or any other standard units of information) per observation of communicating observations from the distribution. We proved that under amortized analysis lossless compression is possible in the limit.

Theorem 1. ***The Fundamental Source Coding Theorem***
Suppose $X^n = (X_1, X_2, \cdots, X_n) \sim \mu^n$ is a random variable composed of n iid draws $X_i \sim \mu$, and $C(X^n)$ is the cost of communicating X^n under the most efficient possible coding scheme. Then:

$$\lim_{n\to\infty} \frac{C(X^n)}{n} = H_\mu(X)$$

We won't formally prove this, but we will provide a quick intuitive argument. Suppose that the sequence $(x_1, x_2, \ldots, x_n)$ is a draw from the joint distribution μ^n, and suppose that $\mu^n(x_1, x_2, \ldots, x_n)$ is the probability of this draw. Then:

Definition 2. ***Typical Sequence***
*A **typical sequence** is a sequence $(x_1, x_2, \ldots, x_n)$ such that:*

$$(\forall \epsilon > 0)(\exists N_\epsilon)(\forall n > N_\epsilon) 2^{-nH_\mu - \epsilon} \leq \mu^n(x_1, x_2, \ldots, x_n) \leq 2^{-nH_\mu + \epsilon}$$

The number of typical sequences is approximately 2^{nH_μ} for n large enough.

In Theorem 1, if we replace $C(X^n)$ by $\mathbb{E}_{\mu^n}(C(X^n))$, then it follows by an application of the bounds for single-shot compression, established below. However, this theorem is true even without the expectation. The reason is that as the n draws are iid, we can use concentration bounds like Chernoff inequalities to argue that the distribution μ^n is supported almost entirely (at least $1 - o(1)$ mass) on typical sequences for the distribution μ, so almost all the codewords can be assumed to be of the same length. Moreover, the distribution is (almost) uniform on these typical sequences.

To understand this notion, focus on the binary case $X \sim Ber(p)$ for some p. We assume wlog $p \leq \frac{1}{2}$ (otherwise we can flip all bits). Then $X = (X_1, X_2, \ldots, X_n) \sim Bin(n, p)$. The typical sequences in this case are all sequences with Hamming weight np. We formally show later that the number of sequences of Hamming weight at most np is at most $2^{nH(p)}$. Further, the number of sequences of Hamming weight exactly np approaches $2^{nH(p)}$ as $n \to \infty$. Clearly, the distribution is uniform on these sequences of Hamming weight np. In this case, as was shown in the previous lecture, it is clearly enough to only consider the lengths of codewords corresponding to these sequences, in the limit. We claim without proof that this is generally true for any distribution μ.

2 Single-Shot Compression

Today we extend the Fundamental Theorem of Source Coding, which provides an operational interpretation of entropy as the expected cost of communication in the limit for large sequences using amortized analysis, to provide a similar interpretation for communicating a single observation. We will show that the expected cost of communicating one observation is bounded by its entropy; for $X \sim \mu$, if $C(X)$ is the length of the code word for X in an optimally efficient code:

$$H_\mu(X) \leq \mathbb{E}_\mu[C(X)] < H_\mu(X) + 1.$$

2.1 Prefix-Free Codes

Definition 3. ***Prefix-Free Codes***
A Prefix-Free (PF) code is a code such that no code word for any event (or input symbol) is a prefix of the code word for any other event.

PF codes can be represented by trees where each leaf is labeled with an event (or input symbol), each edge in the tree is labeled with an output symbol, and the code word corresponding to each event is the concatenation from root to leaf of the output symbols found on the path from the root to the leaf labeled with that event.

Let τ be any tree representing a PF code. Our objective is to choose an optimally efficient PF code, which uses the shortest possible expected cost per event encoded:

$$\min_{\tau \in T} \sum \mu_i l_i$$

Note that optimizing the worst-case cost per event is trivial; any tree which can encode n distinct events in alphabet Σ needs at least $\lceil \log_{|\Sigma|}(n) \rceil$ depth in order to have n leaves, so the worst case cost

must be at least $\lceil \log_{|\Sigma|}(n) \rceil$, and this is tight because we can always construct such a tree. Expected cost is more interesting because it is non-trivial.

Lemma 4. ***Key Lemma: Kraft's Inequality***
A PF code for $X \sim \mu$ with code word lengths $(l_1, l_2, \ldots, l_n)$ in alphabet Σ exists if and only if:

$$\sum_{i=1}^{n} |\Sigma|^{-l_i} \leq 1$$

We will focus on binary codes, in which case a PF code exists if and only if:

$$\sum_{i=1}^{n} 2^{-l_i} \leq 1$$

.

Proof. **2.1.1 Lower Bound**

We will show that any binary PF code with lengths $(l_1, l_2, \ldots, l_n)$ must satisfy the condition from 4 (this is the only if direction).

Suppose there exists a binary tree with lengths $(l_1, l_2, \ldots, l_n)$ to its leaves. Take a random walk on the tree, stopping when a leaf is reached or the walk falls off the tree (there is no edge with the label generated by the walk at the current location). Consider the probability of reaching a leaf.

$$\Pr[\text{reach leaf}] = \sum_i \Pr[\text{reach leaf a}_\text{i}]$$

But because we're doing a random walk, all leaves at the same depth d are reached with equal probability $\frac{1}{2^d}$:

$$\Pr[\text{reach leaf a}_\text{i}] = 2^{-l_i}$$

But obviously:

$$\Pr[\text{reach leaf}] \leq 1$$

So:

$$\sum_i 2^{-l_i} \leq 1$$

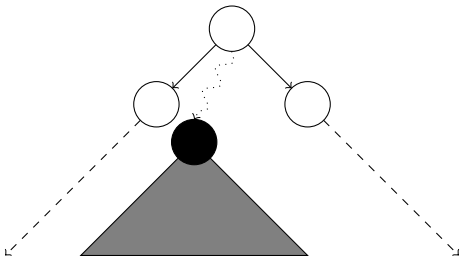
An equivalent lower bound can be proven not just for PF codes but also for uniquely decodable codes, which is a much more general class. Therefore, focusing on PF codes is not an arbitrary restriction; in this case, the analysis for PF codes captures the nature of the problem but is simple to complete. The more general argument is more complicated, so we won't cover it; at a high level, the idea would be to take a random walk over an n-tuple of concatenated code words instead, since the encoding for a sequence of words should be a PF code even if the encoding for each individual word is not.

2.1.2 Upper Bound

Here we will show that for any sequence of code word lengths $(l_1, l_2, \ldots, l_n)$ in alphabet Σ that satisfies the condition from 4, there does exist a PF code with these lengths.

Assume $l_1 \leq l_2 \leq \ldots \leq l_n$. We can do this without loss of generality because we could always reorder and relabel the lengths without changing the nature of the problem.

Suppose we choose an arbitrary leaf from a full binary tree at depth l_1. This prevents us from choosing any leaves in the larger tree that are contained in the subtree rooted at this leaf, since any such choice would violate the Prefix-Free constraint. How many such possible codes are ruled out by the choice of this first leaf (in the picture below, if the chosen node is the black one, how many nodes are ruled out in the shaded gray subtree)?



Let $L = l_n$ be the depth of the full binary tree we will need to construct to include all the code words. Then there are 2^L total possible code words before we choose any code words. After we choose a code word of length l_1, there are 2^{L-l_1} code words which are no longer valid because they have a prefix which is already a code word (they belong to the subtree of depth $L - l_1$ rooted at the chosen node).

After we repeat the process for each length, the total number of code words we have ruled out is:

$$2^{L-l_1} + 2^{L-l_2} + \ldots + 2^{L-l_n} = \sum_{i=1}^{n} 2^{L-l_i} = 2^L \sum_{i=1}^{n} 2^{-l_i}.$$

But by the assumption that the condition from Kraft's lemma is satisfied:

$$\sum_{i=1}^{n} 2^{-l_i} \leq 1$$

We get

$$2^L \sum_{i=1}^{n} 2^{-l_i} \leq 2^L$$

So the total number of code words ruled out is at most the total number of possible code words in a binary tree of this depth, so we had enough space to fit every code word in the tree and construct a PF code by carrying out this process. □

2.2 Proving single-shot cost

We will now use Kraft's lemma to prove the single-shot analog of the amortized result from the last lecture.

Definition 5. ***Shannon-Fano codes***
A Shannon-Fano code for $X \sim \mu$ is created by choosing $l_i := \left\lceil \log \frac{1}{\mu_i} \right\rceil$ and constructing a PF code with these lengths.

(Throughout these notes, log has base 2 unless otherwise specified).

We will now prove that the choice of lengths in any Shannon-Fano code satisfies Kraft's inequality to show that it is a valid choice for a PF code.
In the derivation below, we use the fact that the exponent is negative, so removing the ceiling makes the result larger. Moreover, μ is a distribution, so $\sum_i \mu_i = 1$.

$$\sum_i 2^{-l_i} = \sum_i 2^{-\left\lceil \log \frac{1}{\mu_i} \right\rceil} \leq \sum_i 2^{-\log \frac{1}{\mu_i}} = \sum_i \mu_i = 1$$

Therefore any Shannon-Fano code satisfies Kraft's Inequality, so it is a valid choice for a PF code.

Claim 6. ***Shannon-Fano expected cost***
Let $C(X)$ be the length of a code word in a Shannon-Fano code, where $X \sim \mu$. Then:

$$H_\mu(X) \leq \mathbb{E}_\mu[C(X)] < H_\mu(X) + 1.$$

Proof. The expected cost is $\mathbb{E}_\mu[C(X)] = \sum_i \mu_i l_i$; we analyze this.

2.2.1 Upper Bound

$$\mathbb{E}_\mu[C(X)] = \sum_i \mu_i l_i = \sum_i \mu_i \left\lceil \log \frac{1}{\mu_i} \right\rceil < \sum_i \mu_i \left(\log \frac{1}{\mu_i} + 1 \right) = \sum_i \mu_i \log \frac{1}{\mu_i} + \sum_i \mu_i = H_\mu(x) + 1$$

So:

$$\mathbb{E}_\mu[C(X)] < H_\mu(X) + 1$$

2.2.2 Lower Bound

$$\mathbb{E}_\mu[C(X)] = \sum_i \mu_i \left\lceil \log \frac{1}{\mu_i} \right\rceil \geq \sum_i \mu_i \log \frac{1}{\mu_i} = H_\mu(X)$$

So:

$$\mathbb{E}_\mu[C(X)] \geq H_\mu(X)$$

□

The upper bound for a Shannon-Fano code is a general upper bound because upper bounds are inherently constructive; by showing that a Shannon-Fano code requires at most $H_\mu(X) + 1$ expected cost, we've shown that there exist coding schemes which are that efficient.

However, the lower bound we just showed is specific to Shannon-Fano codes. It does not prove that any code requires at least $H_\mu(X)$ cost. We will now prove a stronger lower bound.

Claim 7. ***Entropy Lower Bound***
Let $X \sim \mu$. Fix a codebook for X, and let $C(X)$ be the length of a code word. Then

$$\mathbb{E}_\mu [C_\mu(X)] \geq H_\mu(X).$$

Proof. We will prove this for PF codes, although it is true more generally. We want to show:

$$H_\mu(X) - \mathbb{E}_\mu [C_\mu(X)] = \sum_i \mu_i \log \frac{1}{\mu_i} - \sum_i \mu_i l_i \leq 0$$

But $l_i = \log 2^{l_i}$, so:

$$\begin{aligned}
\sum_i \mu_i \log \frac{1}{\mu_i} - \sum_i \mu_i l_i &= \sum_i \mu_i \log \frac{1}{\mu_i} - \sum_i \mu_i \log 2^{l_i} \\
&= \sum_i \mu_i \log \frac{1}{\mu_i 2^{l_i}} \\
&\leq \log \sum_i \mu_i \frac{1}{\mu_i 2^{l_i}} && \text{(Jensen's inequality; } \log x \text{ is concave in } x\text{)} \\
&= \log \sum_i 2^{-l_i} \\
&\leq 0 && \text{(Kraft's inequality: } \textstyle\sum_i 2^{-l_i} \leq 1 \text{ for PF code)}
\end{aligned}$$

This is what we wanted to show. □

Note that together, the upper and lower bounds allow us to extend the operational interpretation of entropy as a measure of encoding length or communication cost to the single-shot encoding case instead of only applying in the limit over long sequences due to amortized analysis.

2.3 Optimal PF Codes: Huffman Codes

The idea behind Huffman codes is to build the binary tree in a bottom-up manner, by putting the least likely symbols at maximum depth first. We find the two symbols x, y which are least likely to occur, and fix them as sibling leaves with a common parent z (which we introduce). Then we think of z as a composite symbol in place of x and y, and recurse on the rest of the symbols (including z).

Definition 8. ***Huffman Codes***
Suppose we have a distribution μ with a support of size n. If $n \leq 2$ then we assign each element to a leaf in a binary tree of depth 1. Otherwise if $n > 2$, we follow this recursive procedure to construct a Huffman code:

1. *Find* $\arg\min_{x,y}\{\mu(x) + \mu(y)\}$.
2. *Create nodes for x and y (unless they exist already) and make them siblings of each other. Create a node for the common parent z of x and y.*
3. *Define* $\mu(z) \leftarrow \mu(x) + \mu(y)$.

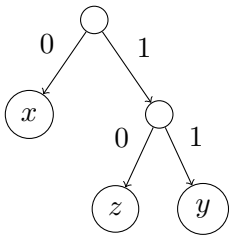
4. This produces a new distribution μ' with a support of size $n-1$ on which we recurse.

Now, we state the optimality of Huffman Codes among all PF codes without proof.

Claim 9. ***Huffman codes are optimal for PF Codes***
Every PF code τ for any distribution μ which is not a Huffman code can be modified in a way that makes it closer to a Huffman code without worsening its expected cost. Formally:

$(\forall \tau)(\exists \tau')$ *such that the least frequent pair (*$\operatorname{argmin}_{X,Y}\{\mu(X)+\mu(Y)\}$*) are siblings at maximum depth in the tree τ', and $\sum_i \mu_i l_i$ is no greater in τ' than in τ.*

Although we will not formally prove this, for some intuition about why this is true, consider the following toy example:



If this is a code for a distribution μ where $\mu(z) > \mu(y) > \mu(x)$ (which would not be a Huffman code), then we can always swap the positions of x and z in the tree to turn it into a Huffman code, and this improves the performance of the PF code because z will be communicated more often and has a shorter code.

3 Noiseless Coding

As a corollary to the single-shot coding analysis, we get Shannon's noiseless coding theorem for sequences.

Proof. **Proof of Theorem 1** Note that the problem of encoding the tuple $X^n = (X_1, \cdots, X_n)$ is equivalent to the problem of encoding the single random variable $X^n \sim \mu^n$ in a single shot. So, we can apply the single-shot bounds derived earlier for X^n. As X^n consists of n iid draws $X_i \sim \mu$, we have $H_{\mu^n}(X^n) = \sum_i H_\mu(X_i) = nH(X)$. Note that we are using the fact that the joint entropy of a collection of independent random variables is the sum of the entropy of the individual random variables, which follows the chain rule (stated later) and the observation that $H(X|Y) = H(X)$ iff X is independent of Y. These facts will be proved in a later lecture.

$$\begin{aligned} H_{\mu^n}(X^n) \leq C(X^n) &\leq H_{\mu^n}(X^n) + 1 \\ nH(X) \leq C(X^n) &\leq nH(X) + 1 \\ H(X) \leq \frac{C(X^n)}{n} &\leq H(X) + \frac{1}{n} \\ \text{So, } \lim_{n\to\infty} \frac{C(X^n)}{n} &= H_\mu(X). \end{aligned}$$

□

4 Joint and Conditional Entropy

4.1 Entropy for Joint Distributions

For a composite random variable $(X, Y) \sim \mu$, we call the entropy of the joint distribution $H_\mu(X, Y)$ as the joint entropy of (X, Y). We would like to relate this quantity to the individual entropies $H(X)$ and $H(Y)$ of the marginal distributions of X and Y respectively.

Definition 10. ***Conditional Entropy***
Let $(X, Y) \sim \mu$. Let μ_X denote the marginal distribution of X. The conditional entropy of Y given $X = x$ for some fixed x is the entropy of the conditional distribution of Y given $X = x$, and is denoted by $H(Y|X = x)$. The conditional entropy of Y given X is defined to be the expected conditional entropy of Y given $X = x$, where the expectation is taken w.r.t μ_X. It is denoted by $H(Y|X)$. Thus,

$$H_\mu(Y|X) := \mathop{\mathbb{E}}_{X\sim\mu} H_{\mu_{Y|X}}(Y \mid X) = \sum_x \mu_X(x) \cdot H(Y|X = x)$$

We will denote $\mu_{Y|x}[Y = y|X = x]$ by $\mu[y|x]$ and use similar simplifying notation later.

Claim 11. ***Joint Entropy in terms of Conditional Entropy***
Fix a joint distribution $(X, Y) \sim \mu$. Then

$$H_\mu(X, Y) = H_\mu(X) + H_\mu(Y \mid X).$$

Proof.

$$\begin{aligned}
H_\mu(X, Y) &= \sum_{x,y} \mu(x, y) \log \frac{1}{\mu(x, y)} \\
&= \sum_{x,y} \mu(x, y) \log \frac{1}{\mu(x)\mu(y \mid x)} \\
&= \sum_x \mu(x) \log \frac{1}{\mu(x)} \sum_y \mu(y \mid x) + \sum_{x,y} \mu(x)\mu(y \mid x) \log \frac{1}{\mu(y \mid x)} \\
&= \sum_x \mu(x) \log \frac{1}{\mu(x)} + \sum_x \mu(x) \left(\sum_y \mu(y \mid x) \log \frac{1}{\mu(y \mid x)} \right) \\
&= H_\mu(X) + \mathop{\mathbb{E}}_{X\sim\mu} H_{\mu_{Y|X}}(Y \mid X) \\
&= H_\mu(X) + H_\mu(Y \mid X)
\end{aligned}$$

□

Note that the role of X and Y can be interchanged, so

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

4.2 Example

Let $X \in_R \{0,1,2,3\}$ and $Y := X \mod 2$. Then $H(X) = 2, H(Y) = 1$. Further, $H(X|Y) = 1$ because for each value of Y, X is uniform over 2 values of the same parity. Finally, $H(Y|X) = 0$ because if we know X then we know the parity of X. More generally, if f is a function of X, then $H(f(X)|X) = 0$ by the same argument.

4.3 Entropy Chain Rule

Claim 12. ***Entropy Chain Rule***
Let $(X_1, X_2, \ldots, X_n)$ be a draw from the joint distribution μ. For $i \in [n]$, let $X_{<i} = \{X_j \mid j < i\}$. Then:

$$H_\mu(X_1, X_2, \ldots, X_n) = \sum_i H(X_i \mid X_{<i})$$

Proof. For $n = 2$, we have already proved the result in Claim 11. For $n > 2$, we can think of $(X_1, X_2, \ldots, X_n)$ as (X, Y) where $X = (X_1, \cdots, X_{n-1})$ and $Y = X_n$. Then we apply 11 to (X, Y) and inductively apply 12 to X (which has $n-1$ components) to prove the claim. □

4.4 Interpretations of Conditional Entropy

We ask if there is an operational interpretation of the conditional interpretation $H(Y \mid X)$? The answer is yes, at least in the limit.

Question: What is the minimum number of bits required by Alice to transmit to Bob a sequence $X^n \sim \mu^n$, given that Bob knows Y^n (which is hidden from Alice) and both know $(X^n, Y^n) \sim \mu^n$ is composed on n iid samples from μ? The answer is given by the Slepian-Wolf Theorem.

Theorem 13. ***Slepian-Wolf Theorem [1973]***

$$\lim_{n\to\infty} \mathbb{E}_\mu \left[\frac{C_\mu(X^n \mid Y^n)}{n} \right] = H_\mu(X \mid Y).$$

We will not prove the theorem, but the idea is to build a set of all possible x_i and y_i and use random hash functions. For any given y (unknown to Alice), this y induces a distribution of typical sequences (recall 2) of X conditioned on y (the number of typical sequences in the support of this distribution is $\approx 2^{nH(X|Y=y)}$). Alice doesn't know this distribution because she doesn't know y, but if she knows the size of the neighborhood and sends a similar number of random hashes, this is the key trick to achieve the above result. The full proof is non-trivial. This general approach is called a "random-bin" argument, which is similar to (but not exactly the same as) perfect hashing.

From this result, we have an operational interpretation of conditional entropy as the cheapest amortized cost of this specific communication complexity problem. Is there an analogous single-shot operational interpretation of conditional entropy? If Alice and Bob can interact, then yes, in expectation $H(X \mid Y) + O_\epsilon(1)$ bits is enough to guarantee decoding the correct result with at least $1 - \epsilon$ probability.

Claim 14. ***Conditional Entropy is upper-bounded by Unconditional Entropy***
Let X, Y be random variables. Then

$$H(X \mid Y) \leq H(X).$$

Equality holds if and only if X and Y are independent.

We will formally prove this in the next lecture, but it is so intuitively obvious that this should be true that we should feel comfortable using this result without formal proof for now.

Combining the entropy chain rule 12 with this upper bound on conditional entropy 14, we get the following:

4.5 Subadditivity of Entropy

Claim 15. *Let $X = (X_1, \cdots, X_n)$ be a sequence of random variables. Then*

$$H(X_1, \ldots, X_n) \leq \sum_i H(X_i).$$

5 Application to counting

In this section, we will show that for $n \in \mathbb{N}$ and $0 \leq p \leq \frac{1}{2}$, the number of subsets of $[n]$ with at most pn elements (alternatively, the number of binary sequences of length n with Hamming weight at most pn) is at most $2^{nH(p)}$, where $H(p) := p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$ is the binary entropy, or entropy of a $Ber(p)$ distribution.

Lemma 16. *Fix $n \in \mathbb{N}, 0 \leq p \leq \frac{1}{2}$. Then*

$$\sum_{i=0}^{pn} \binom{n}{i} \leq 2^{nH(p)}.$$

Proof. Let $\binom{[n]}{\leq pn}$ denote the set of subsets of $[n]$ with at most pn elements, and let $\binom{n}{\leq pn} = \sum_{i=0}^{pn} \binom{n}{pn}$ be the number of such subsets. Let $S \in_R \binom{[n]}{\leq pn}$ be drawn uniformly at random from this set. For $i \in [n]$, define $S_i := \mathbb{1}[i \in S]$ to be the indicator random variable denoting whether i is in S. Then it is easy to see that each $S_i \sim Ber(q)$ for some $q \leq p$. The entropy of a uniform distribution supported on m elements is exactly $\log m$. Given either S or the tuple $(S_1, \cdots, S_n)$, we can reconstruct the other, so they must have the same entropy. We have

$$\begin{aligned}
\log \binom{n}{\leq pn} &= H(S) = H(S_1, \cdots, S_n) \\
&= \sum_{i=1}^{n} H(S_i|S_{<i}) && \text{(Chain Rule)} \\
&\leq \sum_{i=1}^{n} H(S_i) && (H(X|Y) \leq H(X)) \\
&= \sum_{i=1}^{n} h(q) \\
&\leq \sum_{i=1}^{n} H(p) && (q \leq p, \text{ and } H(p) \text{ increases with } p \text{ for } p \leq \tfrac{1}{2}) \\
&= nH(p)
\end{aligned}$$

□

Now, we prove the claim that the number of sequences of length n with Hamming weight at most pn is not only upper-bounded by $2^{nH(p)}$ but approaches this bound as $n \to \infty$.

Claim 17.

$$\lim_{n\to\infty} \frac{\log \binom{n}{\leq pn}}{n} = H(p).$$

First, we explain why we intuitively expect this to hold. In the proof of Lemma 16, the only step which can make the bound loose is when we upper bound $H(S_i|S_{<i})$ by $H(S_i)$. As $n \to \infty$, $pn \to \infty$ as p is a constant. So, for most $i \in [n]$, this conditioning is weak and we have $H(S_i|S_{<i}) \lesssim H(S_i)$.

Proof. We will use Stirling's approximation: for $n \in \mathbb{N}$, $\ln(n!) = n \ln n - n + O(\ln n)$.

$$\begin{aligned}
\log_2 \binom{n}{pn} &= c[\ln n! - \ln(pn)! - \ln((1-p)n)!] && (c = \log_2 e) \\
&= c[n \ln n - n - (pn) \ln(pn) + pn - (n(1-p)) \ln(n(1-p)) + (n - pn) + O(\lg n)] \\
&= c[-pn \ln p - n(1-p) \ln(1-p) + O(\ln n)] \\
&= n \left[p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p} + O\left(\frac{\log n}{n}\right) \right] \\
&= n \left(H(p) + o(1) \right)
\end{aligned}$$

So, $\lim_{n\to\infty} \dfrac{\log \binom{n}{pn}}{n} = H(p)$

The proof actually shows a stronger claim: even if we restrict the sequences to have Hamming weight exactly pn, the number of such sequences approaches $2^{nH(p)}$ as $n \to \infty$, so we get that the number of sequences with Hamming weight at most pn must converge to the same limit, by combining this result with the upper bound. □

6 Shearer's Lemma

To motivate Shearer's Lemma, consider the following setting. Let $\mathbf{X} = (X_1, \cdots, X_n)$ be a set of random variables with joint entropy $H(\mathbf{X})$. Suppose Alice wants to send $\mathbf{X}$ to Bob, but sending the entire random variable $\mathbf{X}$ is too expensive. She would still like to spend some partial information about $\mathbf{X}$ by sending a small index set $S \subset [n]$ randomly and sending $X_i : i \in S$. Shearer's Lemma gives a guarantee that under certain conditions, this set will still have a non-trivial fraction of $H(\mathbf{X})$ in expectation.

Consider the special case when we choose an index $I \in_R [n]$ u.a.r. Then $\Pr[i = I] = \frac{1}{n}$ for all $i \in [n]$. The expected entropy of X_I satisfies

$$\mathbb{E}_{I \in_R n}\left[H(X_I)\right] = \frac{1}{n}\sum_{i=1}^{n} H(X_i) \geq \frac{1}{n} H(\mathbf{X})$$

Shearer's Lemma generalizes this to arbitrary sets $S \subset [n]$.

Lemma 18. *Let $\mathbf{X} = (X_1, \cdots, X_n)$ be a random variable. Suppose $S \subset [n]$ is sampled according to some distribution such that for all $i \in [n]$, $\Pr[i \in S] \geq \alpha$. For $s \subset [n]$, let $X_s = (X_i : i \in s)$. Then*

$$\mathbb{E}_S[H(X_S)] \geq \alpha H(\mathbf{X}).$$

Before proving the lemma, we note that the lemma implies the special case above for $\alpha = \frac{1}{n}$.

Proof. Fix a subset $s = (s_{i_1}, s_{i_2}, \cdots, s_{i_K}) \subset n$, where $|s| = K$. Assume wlog $s_{i_1} < s_{i_2} < \cdots < s_{i_K}$. For $j \in [n]$, let $X_{<j} = (X_k : 1 \leq k < j)$.

$$\begin{aligned}
H(X_s) &= \sum_{j=1}^{|s|} H\left(X_{i_j} | X_l \ \forall\ l < j, l \in s\right) \\
&\geq \sum_{j=1}^{|s|} H\left(X_{i_j} | X_l \ \forall\ l < j\right) && \text{(E1)} \\
\Rightarrow \mathbb{E}_S[H(X_S)] &= \sum_{s \subset [n]} \Pr[S = s] H(X_s) \\
&\geq \sum_{s} \Pr[S = s] \sum_{j=1}^{|s|} H\left(X_{i_j} | X_l \ \forall\ l < j\right) && \text{(E2)} \\
&= \sum_{j=1}^{n} \Pr[j \in S] H\left(X_j | X_{<j}\right) && \text{(E3)} \\
&\geq \alpha \sum_{j=1}^{n} H\left(X_j | X_{<j}\right) = \alpha H(\mathbf{X}) && \text{(Chain Rule)}
\end{aligned}$$

In (E1), the inequality arises because we condition on more random variables. Fix $j \in [k]$. Then $H\left(X_{i_j} | X_l \ \forall\ l < j, l \in s\right)$ is the entropy of X_{i_j} conditioned on all the random variables that have smaller index in the order *and* are also in S. However, $H\left(X_{i_j} | X_l \ \forall\ l < j\right)$ is the entropy of X_{i_j} conditioned on all the random variables that have smaller index in the order, whether they are in S or not. Since we

are conditioning on a superset, the conditional entropy decreases. Note that after this conditioning, for a fixed $j \in [n]$, the random variable $X_j|X_{<j}$ is independent of $S \subset [n]$. To move from (E2) to (E3), we have used the independence of the random variable on the subset, and interchanged the order of summations. For a fixed $j \in [n]$, the number of times $H(X_j|X_{<j})$ is counted in the sum is precisely the probability that $j \in S$ over the random choice of subset S. □

6.1 Application: Counting Triangles in Graphs

Fix an undirected graph $G = (V, E)$ with l edges. We would like to maximize the number of triangles in G. Formally, an embedding of a triangle into G is a one-one function $f : \{1, 2, 3\} \to V$ such that the 3 edges $(f(1), f(2)), (f(1), f(3)), (f(2), f(3))$ are all in E.

Claim 19. *Fix $l \in \mathbb{N}, l \leq 3$. For $l \in \mathbb{N}$, let $t = t(l)$ be the maximum number of embeddings of a triangle into a graph $G_l = (V, E)$ with l edges. Then $t \leq \frac{(2l)^{3/2}}{6}$.*

Proof. Let $X = (X_1, X_2, X_3)$ be (the vertices of) a uniform random triangle in G_l. Then X is uniform with support size $6t$, because for each fixed triangle, there are $3! = 6$ ways to assign the vertices to X. So, $H(X) = \log(6t)$. Let $T \in_R \{(1, 2), (1, 3), (2, 3)\}$ u.a.r. Then $\Pr[i \in T] = \frac{2}{3}$ for each $i \in [3]$. Let $X_T = (X_i : i \in T)$ denote a random edge in the triangle X. Applying Shearer's Lemma to the random variable $X = (X_1, X_2, X_3)$, we get

$$\mathbb{E}_T[H(X_T)] \geq \frac{2}{3} H(X_1, X_2, X_3)$$

By the probabilistic method, this implies there is a fixed $T = \tilde{T}$ such that

$$H(X_{\tilde{T}}) \geq \frac{2}{3} H(X_1, X_2, X_3).$$

Note that as $\tilde{T}$ is fixed, $X_{\tilde{T}}$ is a random variable corresponding to some distribution over the edges of the graph. As there are l edges and 2 choices of labeling for each edge $e = (u, v)$ (i.e. $X_{\tilde{T}} = (u, v)$ or $X_{\tilde{T}} = (v, u)$), $X_{\tilde{T}}$ is supported on $2l$ elements, so $H(X_{\tilde{T}}) \leq \log(2l)$. Thus we have

$$\log(6t) = H(X_1, X_2, X_3) \leq \frac{3}{2} H\left(X_{\tilde{T}}\right) \leq \frac{3}{2} \log(2l) = \log(2l)^{3/2}$$

Applying the exponential function to this inequality gives the lemma. □

In the next lecture, we will see a generalization of this result to embeddings of arbitrary graph patterns instead of triangles.