**Information and Coding Theory** **Winter 2021**

# Lecture 2: January 14, 2021

Lecturer: Madhur Tulsiani

## 1 Source Coding (continued)

We will show that the concept of entropy, defined in the previous lecture, provides a lower bound on the expected length of any prefix free code. In particular, we will now show that *any* prefix-free code for communicating the value of a random variable $X$ must use at least $H(X)$ on average.

**Claim 1.1.** *Let $X$ be a random variable taking values in $\mathcal{X}$ and let $C : \mathcal{X} \to \{0,1\}$ be a prefix-free code. Then the expected number of bits used by $C$ to communicate the value of $X$ is at least $H(X)$.*

**Proof:** The expected number of bits used is $\sum_{x\in\mathcal{X}} p(x) \cdot |C(x)|$. We consider the quantity

$$\begin{aligned} H(X) - \sum_{x\in\mathcal{X}} p(x)\cdot|C(x)| &= \sum_{x\in\mathcal{X}} p(x)\cdot\left(\log\left(\frac{1}{p(x)}\right) - |C(x)|\right) \\ &= \sum_{x\in\mathcal{X}} p(x)\cdot\log\left(\frac{1}{p(x)\cdot 2^{|C(x)|}}\right). \end{aligned}$$

We consider a random variable $Y$ with takes the value $\frac{1}{p(x)\cdot 2^{|C(x)|}}$ with probability $p(x)$. The above expression then becomes $\mathbb{E}[\log(Y)]$. Using Jensen's inequality gives

$$\mathbb{E}[\log(Y)] \;\leq\; \log(\mathbb{E}[Y]) \;=\; \log\left(\sum_{x\in\mathcal{X}} p(x)\cdot\frac{1}{p(x)\cdot 2^{|C(x)|}}\right) \;=\; \log\left(\sum_{x\in\mathcal{X}}\frac{1}{2^{|C(x)|}}\right)$$

which is non-positive since $\sum_{x\in U}\frac{1}{2^{|C(x)|}} \leq 1$ by Kraft's inequality. ∎

**The Shannon code:** We now construct a (prefix-free) code for conveying the value of $X$, using at most $H(X)+1$ bits on average (over the distribution of $X$). For an element $x \in \mathcal{X}$ which occurs with probability $p(x)$, we will use a codeword of length $\lceil\log(1/p(x))\rceil$. By Kraft's inequality, there exists a prefix-free code with these codeword lengths, since

$$\sum_{x\in\mathcal{X}}\frac{1}{2^{|C(x)|}} \;=\; \sum_{x\in\mathcal{X}}\frac{1}{2^{\lceil\log(1/p(x))\rceil}} \;\leq\; \sum_{x\in\mathcal{X}}\frac{1}{2^{\log(1/p(x))}} \;=\; \sum_{x\in\mathcal{X}} p(x) \;=\; 1\,.$$

Also, the expected number of bits used is

$$\sum_{x\in\mathcal{X}} p(x)\cdot\lceil\log(1/p(x))\rceil \;\leq\; \sum_{x\in\mathcal{X}} p(x)\cdot(\log(1/p(x))+1) \;=\; H(X)+1\,.$$

This code is known as the Shannon code.

## 2 Joint Entropy

We have two random variables $X$ and $Y$. The joint distribution of the two random variables $(X,Y)$ takes values $(x,y)$ with probability $p(x,y)$. Merely by using the definition, we can write down the entropy of $Z=(X,Y)$ trivially. However what we are more interested in is seeing how the entropy of $(X,Y)$, the joint entropy, relates to the individual entropies, which we work out below:

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

Denoting $\mathbb{E}_x\left[H(Y|X=x)\right]$ as $H(Y|X)$, this can simply be written as

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

If we were to redo the calculations, we could similarly obtain:

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

This is called the *Chain Rule* for Entropy. Note that in the calculations above, we treat $(Y|X=x)$ as a random variable, with distribution given by $\mathbb{P}\left[Y=y \mid X=x\right] = p(y|x)$. Also note that $H(Y|X)$ is a simply a shorthand for the *expected* entropy of $(Y|X=x)$, with the expectation taken over the values for $X$.

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

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

*Now, let us calculate the following:*

1. $H(X) = H(Y) = \frac{1}{3}\log 3 + \frac{2}{3}\log\frac{3}{2}$
2. $H(Y|X=0) = 0$
3. $H(Y|X=1) = \frac{1}{2}\log\frac{1}{\frac{1}{2}} + \frac{1}{2}\log\frac{1}{\frac{1}{2}} = 1$
4. $H(Y|X) = \frac{1}{3}\cdot 0 + \frac{2}{3}\cdot 1 = \frac{2}{3}$
5. $H(X,Y) = \frac{1}{3}\log 3 + \frac{1}{3}\log 3 + \frac{1}{3}\log 3 = \log 3$

From the above we see that:

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

this is actually *always* true and we prove this fact below.

**Proposition 2.2.** $H(Y) \geq H(Y|X)$

**Proof:** We want to show that $H(Y|X) - H(Y) \leq 0$. Consider the quantity on the left hand side.

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

Now consider a random variable $Z$ that takes value $\frac{p(x)p(y)}{p(x,y)}$ with probability $p(x,y)$. Then we can use Jensen's inequality to get:

$$\sum_{x,y} p(x,y) \left( \log \frac{p(x)p(y)}{p(x,y)} \right) \leq \log \left( \sum_{x,y} \frac{p(x)p(y)}{p(x,y)} p(x,y) \right) = \log(1) = 0\,.$$

■

Note however the fact that conditioning on $X$ reduces the entropy of $Y$ is only true *on average over all fixings of* $X$. In particular, in the above example we have $H(Y|X=1) = 1 > H(Y)$. But $H(Y|X)$, which is an average over all fixings of $X$, is indeed smaller than $H(Y)$. Also, check that above inequality is tight only when $X$ and $Y$ are independent.

**Exercise 2.3.** *Show that $H(Y) = H(Y|X)$ if and only if $X$ and $Y$ are independent.*

Using induction, we can use the chain rule to show that the following also holds for a tuple of random variables $(X_1, \ldots, X_m)$.

$$H(X_1, X_2, \ldots, X_m) \;=\; H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) \ldots H(X_m|X_1, \ldots, X_{m-1})\,.$$

Combining this with the fact that conditioning (on average) reduces the entropy, we get the following inequality which is referred to the sub-additivity property of entropy.

$$H(X_1, X_2, \ldots, X_m) \;\leq\; H(X_1) + H(X_2) + H(X_3) + \cdots + H(X_m)\,.$$

Sub-additivity of entropy is very useful in many applications to combinatorics and counting. However, we first use the chain rule to show that the upper bound on the expected code length of $H(X) + 1$ can be improved if we are communicating many symbols and encode a large block of them at once, rather than sending the code for one symbol at a time.

## 2.1 Source Coding Theorem

We begin by recalling the Shannon Code. We considered a random variable $X$ that took on values $a_1, a_2, \ldots, a_n$ with probabilities $p_1, p_2, \ldots, p_n$. We wanted to encode the values of $X$ such that the expected number of bits needed is small. If $\ell_1, \ell_2, \ldots, \ell_n$ are the number of bits needed to encode $a_1, a_2, \ldots, a_n$, then we saw that a prefix free code exists iff:

$$\sum_{i=1}^{n} 2^{\ell_i} \;\leq\; 1$$

Furthermore, we saw that the expected length of the encoding is lower bounded by $H(X)$ and upper bounded by $H(X) + 1$ (a code as specified as above, the Shannon code may be constructed by setting $\ell_i = \lceil \log(1/p_i) \rceil$).

We will now try to improve this upper bound and we will do so by considering multiple copies of $X$. The idea is that by amortizing the loss over many symbols, we can start to approach an expected length equal to the lower bound i.e. the entropy.

The design may be done as follows: Consider $m$ copies of the random variable $X$, $\{X_1, \ldots, X_m \in U\}$ and a code $C : \mathcal{X}^m \to \{0,1\}^*$. Let $|\mathcal{X}|^m = N$. Now, we know that:

$$H(X_1, \ldots, X_m) \;\leq\; \sum_{i=1}^{N} p_i \left\lceil \log \frac{1}{p_i} \right\rceil \;\leq\; H(X_1, \ldots, X_m) + 1$$

Let us also assume that the $m$ copies of $X$ are drawn i.i.d. Using this assumption we try to work out the quantity $H(X_1, \ldots, X_m)$. Which may be expanded using the chain rule and

independence:

$$\begin{aligned} H(X_1, \ldots, X_m) &= H(X_1) + H(X_2|X_1) + \cdots + H(X_m|X_1, \ldots, X_{m-1}) \\ &= H(X_1) + H(X_2) + \cdots + H(X_m) \\ &= m \cdot H(X) \end{aligned}$$

Therefore, we get

$$\mathbb{E}\left[|C(X_1, \ldots, X_m)|\right] \leq m \cdot H(X) + 1 .$$

Thus, we used $H(X) + \frac{1}{m}$ bits on average per copy of $X$. This leads us to the source coding theorem.

**Theorem 2.4** (Fundamental Source Coding Theorem (Shannon))**.** *For all $\varepsilon > 0$ there exists a $n_0$ such that for all $n \geq n_0$ and given $n$ copies of $X$, $X_1, \ldots, X_n$ sampled i.i.d., it is possible to communicate $(X_1, \ldots, x_n)$ using at most $H(X) + \varepsilon$ bits per copy on average.*

## 2.2 Bounding binomial sums

We use the subadditivity property to obtain an upper bound on the number of subsets of $[n] = \{1, \ldots, m\}$ of size at most $k$ i.e., we need to bound size of the following set

$$S = \{(x_1, \ldots, x_n) \in \{0,1\}^n \mid x_1 + \cdots + x_n \leq k\} .$$

Of course we can write the following expression for the size of $S$

$$|S| = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{k},$$

but how much *is* the value of the above sum? We will estimate it in terms of the binary entropy function defined as

$$H_2(p) := p \cdot \log\left(\frac{1}{p}\right) + (1-p) \cdot \log\left(\frac{1}{1-p}\right)$$

Note that $h(p)$ is the entropy of a random variable $X$, which takes value 1 with probability $p$ and 0 with value $1 - p$ (or vice-versa). This immediately tells us that the maximum possible value of $H_2(p)$ is 1, which is achieved at $p = 1/2$. The function $H_2(p)$ can also easily be shown to be concave. It is also increasing for $p \in (0, 1/2)$ and decreasing for $p \in (1/2, 1)$

**Exercise 2.5.** *Prove that the function $H_2(p)$ is concave, using $H(Y|X) \leq H(Y)$.*

**Exercise 2.6.** *Prove that the function $H_2(p)$ is increasing when $p \in (0, 1/2)$.*

We now return to the estimation problem. Let $(X_1, \ldots, X_n)$ be a uniformly distributed over $S$. Thus, we have that $H(X_1, \ldots, X_n) = \log |S|$. We can also use sub-additivity to say that

$$\log |S| \;=\; H(X_1, \ldots, X_n) \;\leq\; H(X_1) + \cdots + H(X_n) \;=\; n \cdot H(X_1)\,,$$

where the last equality used the symmetry of the variables $X_1, \ldots, X_n$. Now since $X_1$ was an indicator variable, let us say that it takes value 1 with probability $p$ and value 0 with probability $1 - p$. Then $H(X_1) = H_2(p)$. Also, we have that $X_1 + \cdots + X_n \leq k$, which gives by symmetry that $p = \mathbb{E}[x_1] \leq k/n$. Finally, we note that since the function $H_2(p)$ is increasing for $p \leq 1/2$, we get $H(X_1) = H_2(p) \leq H_2(k/n)$. This gives the bound

$$\log |S| \;\leq\; n \cdot H_2(k/n) \quad \Rightarrow \quad |S| \;\leq\; 2^{n \cdot h(k/n)}\,.$$

You can check that the upper bound obtained here is not too bad since the sum is approximately equal to $\frac{2^{n \cdot h(k/n)}}{\sqrt{2\pi \cdot k \cdot (1-k/n)}}$.

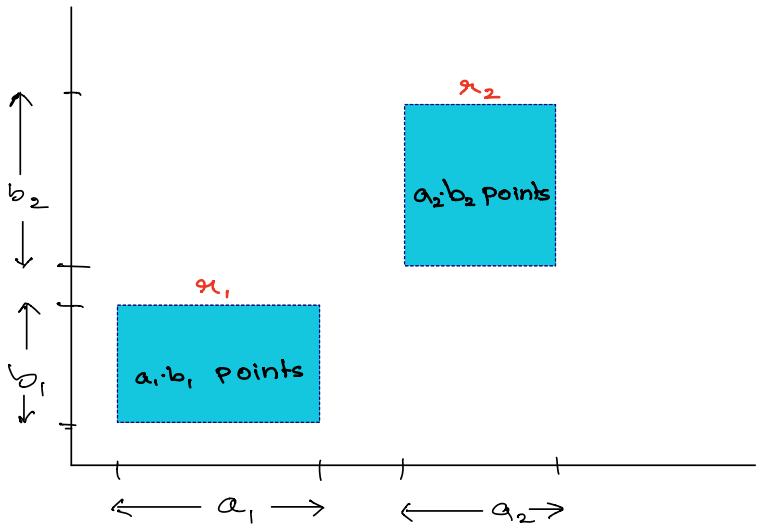
### 2.3 Cauchy-Schwarz and some generalizations

We now give an information-theoretic proof of the Cauchy-Schwarz inequality. While this inequality can of course be proved in many other ways, the method described here was used by Friedgut [Fri04] to prove several interesting generalizations. I highly recommend taking a look at his paper!

Recall that the (finite version of) Cauchy-Schwarz inequality states that for real numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$, we have that

$$\left(\sum_{i=1}^{n} a_i \cdot b_i\right) \;\leq\; \left(\sum_{i=1}^{n} a_i^2\right) \cdot \left(\sum_{i=1}^{n} b_i^2\right)\,.$$

Note that we can assume numbers are non-negative, since the LHS can only decrease otherwise, while the RHS remains unchanged. Also, by continuity of the expressions on both sides (in the numbers $a_i, b_i$) it suffices to prove the inequality for rational rational number. Finally, since we can scale both sides by the same number, it suffices to only consider *natural numbers* $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$.




Consider disjoint subsets $A_1, \ldots, A_n \subseteq \mathbb{N}$, where $|A_i| = a_i$, and similarly, disjoint subsets $B_1, \ldots, B_n$ with $|B_i| = b_i$. Let $r_i$ denote the rectangle $A_i \times B_i$ with $a_i \cdot b_i$ points. We think of the rectangles being in the $x, y$ plane with $x$ coordinates is the sets $A_i$ and $y$ coordinates in the sets $B_i$. We pick two points random $(X_1, Y_1)$ and $(X_2, Y_2)$ as follows:

- Pick a rectangle $R$ with probability proportional to its area i.e.,

$$\mathbb{P}[R = r_i] = \frac{a_i \cdot b_i}{\sum_j a_j \cdot b_j}$$

- Pick two points $(X_1, Y_1)$ and $(X_2, Y_2)$ independently from $R$.

Note that for the random point $(X_1, Y_1)$, the distribution is uniform over the set of all points (and similarly for $(X_2, Y_2)$). Also, since the sets are disjoint, specifying any of the variables $X_1, Y_1, X_2, Y_2$ reveals $R$, which means that

$$H(X_1, Y_1, R) = H(X_2, Y_2, R) = H(X_1, Y_1) = H(X_2, Y_2) = \log\left(\sum_i a_i \cdot b_i\right)$$

Finally, we use the fact that *given* the choice of the rectangle $R$, all four random variables

$X_1, Y_1, X_2, Y_2$ are independent. We get

$$\begin{aligned} H(X_1, Y_1, R) + H(X_2, Y_2, R) &= 2H(R) + H(X_1, Y_1|R) + H(X_2, Y_2|R) \\ &= 2H(R) + H(X_1|R) + H(Y_1|R) + H(X_2|R) + H(Y_2|R) \\ &= 2H(R) + H(X_1, X_2|R) + H(Y_1, Y_2|R) \\ &= H(X_1, X_2, R) + H(Y_1, Y_2, R) \\ &= H(X_1, X_2) + H(Y_1, Y_2) \\ &\leq \log\left(\sum_i a_i^2\right) + \log\left(\sum_i b_i^2\right) . \end{aligned}$$

Note that the last inequality used the fact that the number of choices for $(X_1, X_2)$ is $\sum_i a_i^2$ and that for $(Y_1, Y_2)$ is $\sum_i b_i^2$. Combining the above, we get

$$2\log\left(\sum_i a_i \cdot b_i\right) \leq \log\left(\sum_i a_i^2\right) + \log\left(\sum_i b_i^2\right) ,$$

which yields the Cauchy-Schwarz inequality. This method can also be used to prove interesting generalizations such as

$$\left(\sum_{i,j,k} a_{ij} \cdot b_{jk} \cdot c_{ki}\right)^2 = \left(\sum_{i,j} a_{ij}^2\right) \cdot \left(\sum_{j,k} b_{jk}^2\right) \cdot \left(\sum_{k,i} c_{ki}^2\right) .$$

Take a look at [Fri04] for details.

## References

[Fri04] Ehud Friedgut, *Hypergraphs, entropy, and inequalities*, The American Mathematical Monthly **111** (2004), no. 9, 749–760. 6, 8