APC486/ELE486: Transmission and Compression of Information

Bounds on the Expected Length of Code Words

Scribe: Kiran Vodrahalli

September 18, 2014

1 Notations

In these notes,

- $\mathcal{X}$ denotes a finite set, called the source alphabet in this lecture. The elements of $\mathcal{X}$ are called the symbols.
- P is a probability distribution on $\mathcal{X}$.
- $\mathcal{C}$ is a source encoder on $\mathcal{X}$.
- $L_{\mathcal{C}}(x)$ denotes the length of the codeword $\mathcal{C}(x)$ for some element $x \in \mathcal{X}$.
- The expected value of the length of a codeword for some P and $\mathcal{C}$ is given by:

$$\bar{L}(P, \mathcal{C}) = \sum_{x \in \mathcal{X}} P(x) L_{\mathcal{C}}(x) \tag{1}$$

- UD is the set of codes that are uniquely decodeable.

2 Generalizing Kraft's Inequality to all Uniquely Decodeable Codes

In this section, we show that we can generalize Kraft's Inequality to any uniquely decodeable code.

Theorem 1 (McMillian)**.**
For all uniquely decodeable codes $\mathcal{C}$ on $\mathcal{X}$,

$$\sum_{x \in \mathcal{X}} 2^{-L_{\mathcal{C}}(x)} \leq 1 \tag{2}$$

Note that we already have the converse from Kraft's Inequality: If (2) is satisfied, then there exists a uniquely decodeable code (just choose a prefix-free one).

Proof. Let $\mathcal{C}$ be a uniquely decodeable code on $\mathcal{X}$, and let

$$\alpha = \sum_{x \in \mathcal{X}} 2^{-L_C(x)} \tag{3}$$

We only assume $\alpha > 0$. We claim that $\exists$ a constant $\beta > 0$ such that

$$\alpha^k \leq \beta k \tag{4}$$

for all k. This inequality implies that $\alpha \leq 1$, for otherwise, there would exist a k for which $\alpha^k > \beta k$ since we would have increasing exponential growth on the LHS, and linear growth on the RHS. So we will show (4) and conclude the proof.

We have that

$$\alpha^k = \prod_{i=1}^{k} (\sum_{x_i \in \mathcal{X}} 2^{-L_{\mathcal{C}}(x_i)}) = \sum_{x_1, x_2, \ldots, x_k \in \mathcal{X}} 2^{-(L_{\mathcal{C}}(x_1) + \ldots + L_{\mathcal{C}}(x_k))} \tag{5}$$

Now, we know that the minimum value of any $L_{\mathcal{C}}(x)$ is 1, so we let $L_{\min} = 1$, and let $L_{\max} = \max_{x \in \mathcal{X}} L_{\mathcal{C}}(x)$. Then, we have that

$$\sum_{i=1}^{k} L_{\mathcal{C}}(x_i) \in [k, k * L_{\max}] \tag{6}$$

Thus, we can rewrite (5) as

$$\sum_{l=k}^{kL_{\max}} \sum_{x_1, \ldots, x_k \in \mathcal{X} \text{ s.t. } \sum_{i=1}^{k} L_{\mathcal{C}}(x_i) = l} 2^{-l} \tag{7}$$

Then let

$$A(l) = \sum_{x_1, \ldots, x_k \in \mathcal{X} \text{ s.t. } \sum_{i=1}^{k} L_{\mathcal{C}}(x_i) = l} 1 \tag{8}$$

Then we have that

$$\alpha^k = \sum_{l=k}^{kL_{\max}} 2^{-l} A(l) \tag{9}$$

Since $\mathcal{C}$ is uniquely decodeable, there are at most 2^l ways to generate codewords so that the sum of the lengths of the code words is l. The reason is as follows: the max size of a codeword is l (otherwise, we violate the sum condition). Supposing every codeword was length l, there are a maximum of 2^l distinct such codewords – we require them to be distinct, otherwise $\mathcal{C}$ is not uniquely decodeable. Thus, the number of ways to assign unique codewords cannot be more than 2^l. Therefore, we have an upper bound $A(l) \leq 2^l$ and we can say that

$$\alpha^k = \sum_{l=k}^{kL_{\max}} 2^{-l} A(l) \leq \sum_{l=k}^{kL_{\max}} 1 \leq kL_{\max} = k\beta \tag{10}$$

Therefore, we have shown that $\alpha^k \leq k\beta$, as desired, implying the result. $\square$

3 Entropy Lower Bound on $\bar{L}(P, \mathcal{C})$

Definition 1. *Entropy is defined as*

$$H(P) = \sum_{x \in \mathcal{X}} P(x) \log_2 \frac{1}{P(x)} \tag{11}$$

Theorem 2. *$\forall \mathcal{X}, P,$ and $\mathcal{C}$ on $\mathcal{X}$ with $\mathcal{C}$ uniquely decodeable, $H(P) \leq \bar{L}(P, \mathcal{C})$.*

Proof. We will express the minimization problem as follows: Find

$$\min\{\sum_{x \in \mathcal{X}} P(x) L_K(x)\} \leq \bar{L}(P, \mathcal{C}) \tag{12}$$

minimizing over all K that are uniquely decodeable. Therefore $\{L_K(x)\}_{x \in \mathcal{X}}$ satisfies McMillian's Theorem, $\sum_{x \in \mathcal{X}} 2^{-L_K(x)} \leq 1$.

Now define $q(x) = 2^{-L_K(x)}$. Note $q(x) > 0 \; \forall x \in \mathcal{X}$, and that we have

$$\sum_{x \in \mathcal{X}} q(x) \leq 1 \tag{13}$$

by McMillian's Theorem. We now rewrite the optimization problem in terms of $q(x)$:

$$\min\{\sum P(x) L_K(x)\} = \min\{\sum P(x) \log_2 \frac{1}{q(x)}\} \tag{14}$$

with $q(x) > 0, \sum q(x) \leq 1$. However, there was also the hidden constraint that $L_K(x) \in \mathbb{Z}^+$. If we only look for $q(x)$ that are positive, we're relaxing a constraint. However, if we denote by M the minimization problem defined over integral $q(x)$, we have that since $\mathbb{Z}^+ \subset \mathbb{R}^+$,

$$\min\{\sum P(x) \log_2 \frac{1}{q(x)}\} \leq M \tag{15}$$

where $q(x) > 0, \sum_{x \in \mathcal{X}} q(x) \leq 1$ and $q(x)$ not necessarily in $\mathbb{Z}^+$. Thus, our goal is now to show that

$$H(P) = \sum_{x \in \mathcal{X}} P(x) \log_2 \frac{1}{P(x)} \leq \min\{\sum_{x \in \mathcal{X}} P(x) \log_2 \frac{1}{q(x)}\} \tag{16}$$

for $\{q(x)\}_{x \in \mathcal{X}}$ such that $q(x) > 0, \sum q(x) \leq 1$. To do this, we show that the LHS - RHS of (16) is ≤ 0. We have that LHS - RHS =

$$\sum_{x \in \mathcal{X}} P(x) \log_2 \frac{q(x)}{P(x)} \tag{17}$$

Note that we can provide a simple upper bound for $\log_2(x)$ with the tangent line at $x = 1$, $f(x) = \frac{1}{\ln(2)}(x - 1)$.

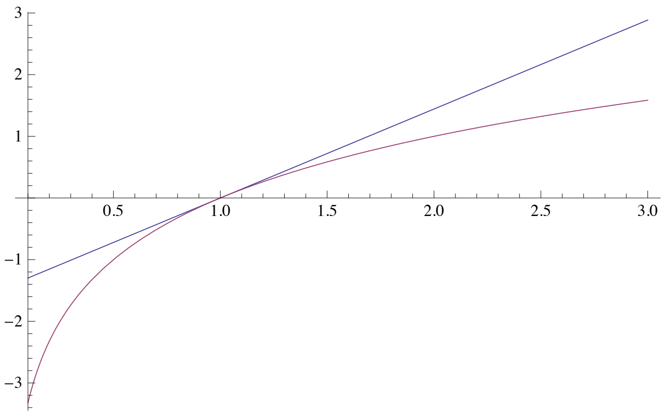

Figure 1: Tangent Upper Bound on $\text{Log}_2(x)$

Then from (17), we have

$$\sum_{x\in\mathcal{X}} P(x)\log_2\frac{q(x)}{P(x)} \le \frac{1}{\ln(2)}\sum_{x\in\mathcal{X}} P(x)(\frac{q(x)}{P(x)}-1) = \frac{1}{\ln(2)}(\sum_{x\in X} q(x) - \sum_{x\in\mathcal{X}} P(x)) \tag{18}$$

Then, since a probability distribution sums to 1, $\sum_{x\in\mathcal{X}} P(x) = 1$. From our assumptions about $q(x)$, we have $\sum_{x\in X} q(x) \le 1$. Therefore,

$$\frac{1}{\ln(2)}(\sum_{x\in X} q(x) - \sum_{x\in\mathcal{X}} P(x)) \le \frac{1}{\ln(2)}(1-1) = 0 \tag{19}$$

and we have

$$\sum_{x\in\mathcal{X}} P(x)\log_2\frac{q(x)}{P(x)} \le 0 \tag{20}$$

as desired. We conclude

$$H(P) \le \bar{L}(P,\mathcal{C}) \tag{21}$$

for all uniquely decodeable $\mathcal{C}$. □

3.1 The Information Inequality: A Brief Digression

Definition 2 (Kullback-Leibler Divergence)**.** *Let p and q be probability distributions on $\mathcal{X}$. Let*

$$D(p||q) = \sum_{x\in\mathcal{X}} p(x)\log_2\frac{p(x)}{q(x)} \tag{22}$$

This quantity is also known as relative entropy.

We can note a few things about the KL-divergence. First, we've already proved that $D(p||q) \geq 0 \ \forall p, q$. (As we saw, it's enough that $\sum q(x) \leq 1$ – we're imposing a stronger condition when we require q to be a probability distribution.) We have that $D(p||q) = 0$ if and only if $p = q$. However, since $D(p||q)$ is not symmetric and the triangle inequality does not hold, it is not a complete distance metric.

In practice, the KL-divergence acts as something like a norm-squared. If p and q are close, we have

$$D(p||q) \approx \sum_x (p(x) - q(x))^2 p(x) = ||p - q||_p^{\ 2} \tag{23}$$

An analogy for this behavior is comparing the way the triangle inequality works to the general Pythagorean theorem. Here, a, b, c will be the sides of a triangle in Euclidean space. For the triangle inequality, we have that $a + b \geq c$ over all permutations of the sides. For the Pythagorean theorem, we have that $a^2 + b^2 = c^2$, or $a^2 + b^2 < c^2$, or $a^2 + b^2 > c^2$, all of which imply different things about the angles of the triangle that a, b, c make. We can do some work to define a notion of angle for probability distributions, but we will not go into that here. For more information about this sort of thing, look up **Information Geometry**.

4 To What Extent is $H(P)$ a Lower Bound?

We might now have some questions regarding the tightness of the entropy lower bound on expected codeword length.

- Can we achieve $H(P)$? When and how often?
- If you can't achieve $H(P)$, how close can you get?
- How do you construct the code of optimal expected length for a given distribution?

4.1 When and With What Frequency Can We Achieve $H(P)$?

If $P(x)$ is a negative power of two (i.e., $P(x) \in 2^{-\mathbb{Z}^+} \ \forall x \in \mathcal{X}$), then $L(x) = \log_2 \frac{1}{P(x)} \geq 0$ and $\sum_{x \in \mathcal{X}} 2^{-L(x)} = \sum_{x \in \mathcal{X}} P(x) = 1$, and we can therefore achieve the entropy bound with a uniquely decodeable code after applying the existence part of Kraft's Inequality.

Example 1. *Suppose our probability distribution is* $P = \{P(x_1) = \frac{1}{2}, P(x_2) = \frac{1}{4}, P(x_3) = \frac{1}{8}, P(x_4) = \frac{1}{8}\}$. *Then we define* $C = \{C(x_1) = 0, C(x_2) = 10, C(x_3) = 110, C(x_4) = 111\}$ *and we can calcuate* $H(P) = \frac{1}{2} * 1 + \frac{1}{4} * 2 + \frac{1}{8} * 6 = 1.75$. *We also calculate* $\bar{L}(P, \mathcal{C}) = \frac{1}{2} * 1 + \frac{1}{4} * 2 + \frac{1}{8} * 6 = 1.75$*, and we see we have equality.*

In fact, the entropy can be achieved if and only if $P(x) \in 2^{-\mathbb{Z}^+} \ \forall x \in \mathcal{X}$, and furthermore, we can achieve it with a prefix-free code.

4.2 How Close to $H(P)$ Can We Get?

Theorem 3. $H(P) \leq \bar{L}^*(P) \leq H(P) + 1$*, where* $\bar{L}^*(P) = \min_{\mathcal{C} \in UD} \{\bar{L}(P, \mathcal{C})\}$.

Proof. If we have that $P(x) \notin 2^{-\mathbb{Z}^+}$ for some $x \in \mathcal{X}$, then $\log_2 \frac{1}{P(x)} \notin \mathbb{Z}$. Let $l(x) = \lceil \log_2 \frac{1}{P(x)} \rceil$. Then,

$$\sum_{x \in \mathcal{X}} l(x)P(x) = \sum_{x \in \mathcal{X}} \lceil \log_2 \frac{1}{P(x)} \rceil P(x) \leq \sum_{x \in \mathcal{X}} (\log_2 (\frac{1}{P(x)}) + 1)P(x) \tag{24}$$

$$= H(P) + \sum_{x \in \mathcal{X}} P(x) = H(P) + 1 \tag{25}$$

□

Next time, we will show that

$$H(P) \leq \frac{\bar{L}^*(X_1 X_2 ... X_n)}{n} \leq \frac{nH(P)+1}{n} = H(P) + \frac{1}{n} \tag{26}$$

where we are considering the average minimum length after n trials X_1 through X_n. As n grows really large, the bound for the average minimum length shrinks more and more tightly just slightly above $H(P)$.

4.3 Construction of Optimal Codes

The short answer is Huffman Codes. We'll go into more detail on this next time.