EE/Stats 376A: Information theory	Winter 2017

Lecture 5 — Jan 24

Lecturer: David Tse *Scribe: Michael X, Nima H, Geng Z, Anton J, Vivek B.*

5.1 Outline

- Markov chains and stationary distributions
- Prefix codes

5.1.1 Reading

- CT: 5.6, 5.7, 5.8, 13.4.

5.2 Recap

For a sequence of random variables $X_1, \cdots, X_n \sim p$, we characterized the convergence properties of $-\frac{1}{n}\log\big(p(X_1,\cdots,X_n)\big)$. In particular, for i.i.d. random variables we used the *law of large numbers* to prove that

$$-\frac{1}{n}\log\big(p(X_1,\cdots,X_n)\xrightarrow{p} H(X_1). \qquad [AEP]$$

Whereas for a time invariant Markov chain we proved a weaker result

$$-\frac{1}{n}\mathbb{E}\big[\log\big(p(X_1,\cdots,X_n)\big)\big] \longrightarrow H(X_2|X_1).$$

5.3 Entropy Rate

Definition 1. For a sequence of random variables $X_1,\ldots,X_n \sim p$ generated from a 'source', the *entropy rate of the source* is defined as

$$H \triangleq \lim_{n\to\infty}\frac{1}{n}H(X_1,X_2,\cdots X_n).$$

For an i.i.d. sequence, the entropy rate is $H(X_1)$, and for a sequence generated from a Markov chain the entropy rate is $H\big(X_2|X_1\big)\ \big(\le H(X_1)\big)$. The entropy rate of a source is asymptotically equal to the expected number of bits per symbol required to compress $X_1,\cdots,X_n$; we shall prove the statement in this set of lectures and the next.

5.3.1 L.L.N. for Markov chains

For a Markov chain $X_1, \cdots, X_n \sim p$, we have

$$\begin{aligned}
\frac{1}{n}\log\frac{1}{p(X_1,\cdots,X_n)} &= \frac{1}{n}\log\frac{1}{p(X_1)p(X_2 \mid X_1)\cdots p(X_n \mid X_{n-1})} \\
&= \frac{1}{n}\left[\log\frac{1}{p(X_1)} + \log\frac{1}{p(X_2 \mid X_1)} + \cdots + \log\frac{1}{p(X_n \mid X_{n-1})}\right] \\
&\approx \frac{1}{n}\left[\log\frac{1}{p(X_2 \mid X_1)} + \frac{1}{p(X_3 \mid X_2)} + \cdots + \log\frac{1}{p(X_n \mid X_{n-1})}\right]. \quad (5.1)
\end{aligned}$$

Expression (5.1) is a sum of $n-1$ **identical** random variables of the form $\log\frac{1}{p(X_i|X_{i-1})}$. However, these random variables are **not independent** because X_i and X_j are dependent via Markov chain.

Therefore, the glaring question is - does the expression (5.1) converge to its expectation in spite of the dependencies? The answer is yes!

Theorem 1. (Weak L.L.N. for weak dependency) If a sequence of identical random variables $Y_1, Y_2, \cdots, Y_n \sim p$ satisfy

$$\frac{1}{n^2}\sum_{i,j=1}^{n} Cov(Y_i, Y_j) \longrightarrow 0 \quad (5.2)$$

then

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{p} \mathbb{E}[Y_1].$$

Proof. The proof directly follows from Q. 1, Chapter 3 of the textbook [1]. □

The random variables $Y_i = \log\big(p(X_i|X_{i-1})\big)$ in expression (5.1) satisfy condition 5.2 of Theorem 1 and this can be verified for Markov chain similar to Q. $5d$ in HW 2. Hence we have the following corollary:

Corollary 1. For a time invariant Markov chain $X_1, \cdots, X_n \sim p$

$$\frac{1}{n}\log\frac{1}{p(X_1,\cdots,X_n)} \xrightarrow{p} H(X_2|X_1).$$

5.4 Codes

In previous sets of lectures, we used AEP to obtain a coding scheme that **asymptotically** required $H(X_1)$ bits per symbol to compresses the sequence $X_1, \cdots, X_n \overset{i.i.d}{\sim} p$. In this lecture and the next, we will devise '**optimal**' **coding schemes** to compress $X_1, \cdots, X_n \overset{i.i.d}{\sim} p$ for **finite** values of n. At the same time, we will compare the performance of these coding schemes to the entropy rate H. First, we derive coding schemes for a single random variable $X \sim p$ and then generalize it to a sequence of random variables. For $\mathcal{X} = \{a, b, c\cdots\}$, let

$$X \sim p; \quad X \in \mathcal{X}; \quad p_a \geq p_b \geq p_c \geq \cdots$$

Definition 2. A ***code*** $C : \mathcal{X} \to \{0,1\}^l$ is an injective function which maps letters in $\mathcal{X}$ to (binary) code words. Let $C(x)$ and $\ell_C(x)$ denote the code word and its length for an alphabet $x \in \mathcal{X}$ respectively.

Definition 3. For a random variable $X \sim p$, the ***expected length*** L of a coding scheme C is defined as

$$L \triangleq \sum_x \ell_C(x)p(x).$$

Our task is to devise a coding scheme C from a class of prefix codes (defined later) that **minimizes** L.

5.4.1 Prefix Codes

We consider two examples to understand Definitions 2, 3, and motivate the idea behind prefix codes.

Example 1. Let X be a random variable on alphabet $\mathcal{X} = \{a, b\}$ with probability distribution $p(a) = p(b) = 1/2$.

For the code $C(a) = 0$ and $C(b) = 1$, the expected length L is 1 and the entropy $H(X)$ is also equal to 1!

Example 2. Let X be a random variable on alphabet $\mathcal{X} = \{a, b, c\}$ with probability distribution $p(a) = 1/2$, $p(b) = p(c) = 1/4$.

For the code $C(a) = 0$, $C(b) = 10$ and $C(c) = 11$, the expected length L is 1.5 and the entropy $H(X)$ is also equal to 1.5! We encourage the reader to workout the expected code length and entropy of the above two examples.

Definition 4. A *prefix code* is a code (typically of variable length) distinguished by its possession of the 'prefix property', which requires that there is no whole code word that is a prefix (i.e. an initial segment) of any other code word.

The codes from above Examples 1 and 2 are prefix codes.

Remark 1. Using 'non-prefix' codes creates **ambiguity** in decoding. In Example 2, the code $C(a) = 0$, $C(b) = 1$ and $C(c) = 11$ satisfies Definition 2 and has an expected code length equal to 1.25 (< 1.5). However, if we use the same code to encode **sequence** of letters, then both '*bb*' and '*c*' would map to 11, leading to ambiguity in decoding.

In both Examples 1 and 2, the expected length of the code word is equal to the entropy of the random variable. Is this just a mere coincidence? No!

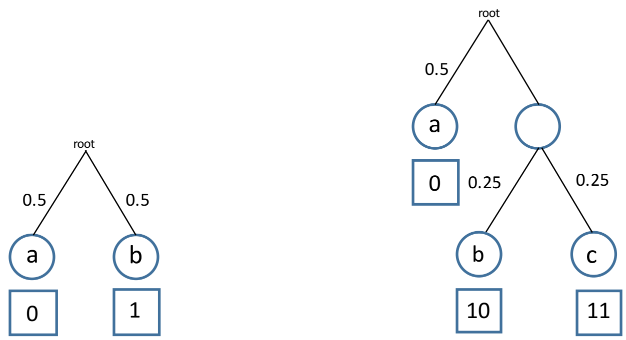

Figure 5.1: Codes for Examples 1 and 2 viewed as Tree codes. The circles contain the letters and the squares contain the corresponding code word.

Claim 2. Consider a random variable $X \sim p$, where p is of the form $p(x) = \frac{1}{2^{k_x}}$, $k_x \in \mathbb{Z}$. Then there exists a prefix code C such that $\ell_C(x) = k_x \; \forall x$. As a result,

$$\begin{aligned} L &= \sum_x p(x)\ell_C(x) \\ &= \sum_x p(x) \log\left(2^{\ell_C(x)}\right) \\ &= \sum_x p(x) \log \frac{1}{p(x)} \\ &= H(X). \end{aligned}$$

We will prove a stronger statement in the next set of lectures, and for this reason we will not prove Claim 2 here.

5.4.2 Tree Codes

For a given tree T with a fixed root, the ***code word*** $C(n)$ corresponding to node n is $C(\text{parent}[n]) + 0$ if n is a left child and $C(\text{parent}[n]) + 1$ if n is a right child. The code word of the root is `null`.

Definition 5. For a random variable $X \in \mathcal{X}$ and tree T with a fixed root, the ***tree code*** C is a map from letters in $\mathcal{X}$ to code word corresponding to the **leaves** of the T. Refer to Figure 5.1 for examples.

The prefix codes in Examples 1 and 2 can be modeled as tree codes as shown in Figure 5.1. All tree codes are prefix codes and vice-versa. The proof is left as an exercise for the reader.

In the next lecture, we will design an optimal code for $X \sim p$ whose expected length is 'close' to $H(X)$. Then, we will generalize it to design optimal codes for a sequence of random variables $X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} p$.

Bibliography

[1] Cover, Thomas M., and Joy A. Thomas. Elements of information theory. John Wiley & Sons, 2012.