UC Berkeley
Department of Electrical Engineering and Computer Sciences

EECS 126: PROBABILITY AND RANDOM PROCESSES

Problem Set 11
Fall 2021

1. **Midterm**

   Solve all of the problems on the midterm again (including the ones you got correct).

   Solution:

   See midterm solutions.

2. **Compression of a Random Source**

   Suppose I'm trying to send a text message to my friend. In general, I know I need $\log_2(26)$ bits for every letter I want to send because there are 26 letters in the alphabet. However, it turns out if I have some information on the distribution of the letters, I can do better. For example, I might give the letter e a shorter bit representation because I know it's the most common. Actually, it turns out the number of bits I need on average is the entropy, and in this problem, we try to show why this is true in general.

   Let $(X_i)_{i=1}^{\infty} \overset{\text{i.i.d.}}{\sim} p(\cdot)$, where p is a discrete PMF on a finite set $\mathcal{X}$. We know the entropy of a random variable X is

   $$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

   Since entropy is really a function of the distribution, we could write the entropy as $H(p)$.

   (a) Show that

   $$-\frac{1}{n} \log_2 p(X_1, \dots, X_n) \xrightarrow{n \to \infty} H(X_1) \qquad \text{almost surely.}$$

   (Here, we are extending the notation $p(\cdot)$ to denote the joint PMF of $(X_1, \dots, X_n)$: $p(x_1, \dots, x_n) := p(x_1) \cdots p(x_n)$.)

   (b) Fix $\epsilon > 0$ and define $A_\epsilon^{(n)}$ as the set of all sequences $(x_1, \dots, x_n) \in \mathcal{X}^n$ such that:

   $$2^{-n(H(X_1)+\epsilon)} \le p(x_1, \dots, x_n) \le 2^{-n(H(X_1)-\epsilon)}.$$

   Show that $P((X_1, \dots, X_n) \in A_\epsilon^{(n)}) > 1 - \epsilon$ for all n sufficiently large. Consequently, $A_\epsilon^{(n)}$ is called the **typical set** because the observed sequences lie within $A_\epsilon^{(n)}$ with high probability.

   (c) Show that $(1-\epsilon)2^{n(H(X_1)-\epsilon)} \le |A_\epsilon^{(n)}| \le 2^{n(H(X_1)+\epsilon)}$, for n sufficiently large. Use the union bound.

   Parts (b) and (c) are called the **asymptotic equipartition property (AEP)** because they say that there are $\approx 2^{nH(X_1)}$ observed sequences which each have probability $\approx 2^{-nH(X_1)}$. Thus, by discarding the sequences outside of $A_\epsilon^{(n)}$, we need only keep track of $2^{nH(X_1)}$ sequences, which means that a length-n sequence can be compressed into $\approx nH(X_1)$ bits, requiring $H(X_1)$ bits per symbol.

(d) Now show that for any $\delta > 0$ and any positive integer n, if $B_n \subseteq \mathcal{X}^n$ is a set with $|B_n| \le 2^{n(H(X_1)-\delta)}$, then $P((X_1, \dots, X_n) \in B_n) \to 0$ as $n \to \infty$.

This says that we cannot compress the observed sequences of length n into any set smaller than size $2^{nH(X_1)}$.

[*Hint*: Consider the intersection of B_n and $A_\epsilon^{(n)}$.]

(e) Next we turn towards using the AEP for compression. Recall that in order to encode a set of size n in binary, it requires $\lceil \log_2 n \rceil$ bits. Therefore, a naïve encoding requires $\lceil \log_2 |\mathcal{X}| \rceil$ bits per symbol.

From (b) and (d), if we use $\log_2 |A_\epsilon^{(n)}| \approx nH(X_1)$ bits to encode the sequences in $A_\epsilon^{(n)}$, ignoring all other sequences, then the probability of error with this encoding will tend to 0 as $n \to \infty$, and thus an asymptotically error-free encoding can be achieved using $H(X_1)$ bits per symbol.

Alternatively, we can create an error-free code by using $1 + \lceil \log_2 |A_\epsilon^{(n)}| \rceil$ bits to encode the sequences in $A_\epsilon^{(n)}$ and $1 + n\lceil \log_2 |\mathcal{X}| \rceil$ bits to encode other sequences, where the first bit is used to indicate whether the sequence belongs in $A_\epsilon^{(n)}$ or not. Let L_n be the length of the encoding of $X_1, \dots, X_n$ using this code; show that $\lim_{n\to\infty} \mathbb{E}[L_n]/n \le H(X_1) + \epsilon$. In other words, asymptotically, we can compress the sequence so that the number of bits per symbol is arbitrary close to the entropy.

Solution:

(a) Since $(X_i)_{i=1}^\infty$ is an i.i.d. sequence, so is $(\log_2 p(X_i))_{i=1}^\infty$. Thus:

$$\begin{aligned}-\frac{1}{n}\log_2 p(X_1, \dots, X_n) &= -\frac{1}{n}\sum_{i=1}^n \log_2 p(X_i) \xrightarrow[\text{a.s.}]{n\to\infty} -\mathbb{E}[\log_2 p(X_1)] \\ &= H(X_1)\end{aligned}$$

by the Strong Law of Large Numbers.

(b) As a consequence of (a), $n^{-1}\log_2 p(X_1, \dots, X_n) \to H(X_1)$ in probability as $n \to \infty$, so

$$P\Big(\Big|-\frac{1}{n}\log_2 p(X_1, \dots, X_n) - H(X_1)\Big| < \epsilon\Big) \to 1 \qquad \text{as } n \to \infty.$$

For n sufficiently large, the LHS is $> 1 - \epsilon$.

(c) We have:

$$1 = \sum_{x\in\mathcal{X}^n} p(x) \ge \sum_{x\in A_\epsilon^{(n)}} p(x) \ge \sum_{x\in A_\epsilon^{(n)}} 2^{-n(H(X_1)+\epsilon)} = |A_\epsilon^{(n)}| 2^{-n(H(X_1)+\epsilon)}$$

This shows that $|A_\epsilon^{(n)}| \le 2^{n(H(X_1)+\epsilon)}$. Now, we have, for n sufficiently large:

$$\begin{aligned}1 - \epsilon < P\big((X_1, \dots, X_n) \in A_\epsilon^{(n)}\big) &\le \sum_{x\in A_\epsilon^{(n)}} 2^{-n(H(X_1)-\epsilon)} \\ &= 2^{-n(H(X_1)-\epsilon)}|A_\epsilon^{(n)}|\end{aligned}$$

Thus, $|A_\epsilon^{(n)}| \ge (1-\epsilon)2^{n(H(X_1)-\epsilon)}$.

(d) Pick $\epsilon \in (0, \delta)$. We can write

$$\begin{aligned}
&P\big((X_1, \ldots, X_n) \in B_n\big) \\
&\quad \le P\big((X_1, \ldots, X_n) \in A_\epsilon^{(n)} \cap B_n\big) + P\big((X_1, \ldots, X_n) \notin A_\epsilon^{(n)}\big) \\
&\quad \le \sum_{x \in A_\epsilon^{(n)} \cap B_n} p(x) + P\big((X_1, \ldots, X_n) \notin A_\epsilon^{(n)}\big) \\
&\quad \le |B_n| 2^{-n(H(X_1) - \epsilon)} + P\big((X_1, \ldots, X_n) \notin A_\epsilon^{(n)}\big) \\
&\quad \le 2^{-n(\delta - \epsilon)} + P\big((X_1, \ldots, X_n) \notin A_\epsilon^{(n)}\big) \to 0
\end{aligned}$$

since $\delta > \epsilon$ and by (b).

(e) Separating out the sequences in the typical set from the sequences which are not in the typical set,

$$\begin{aligned}
\frac{\mathbb{E}[L_n]}{n} &= \frac{1 + \lceil \log_2 |A_\epsilon^{(n)}| \rceil}{n} P\big((X_1, \ldots, X_n) \in A_\epsilon^{(n)}\big) \\
&\quad + \frac{1 + n \lceil \log_2 |\mathcal{X}| \rceil}{n} P\big((X_1, \ldots, X_n) \notin A_\epsilon^{(n)}\big) \\
&\le \frac{1 + \lceil n[H(X_1) + \epsilon] \rceil}{n} + (1 + \lceil \log_2 |\mathcal{X}| \rceil) P\big((X_1, \ldots, X_n) \notin A_\epsilon^{(n)}\big).
\end{aligned}$$

Since $P((X_1, \ldots, X_n) \in A_\epsilon^{(n)}) \to 1$ and $P((X_1, \ldots, X_n) \notin A_\epsilon^{(n)}) \to 0$, then the second term $\to 0$. Asymptotically, only the first term matters, and by taking $n \to \infty$ we get $\lim_{n \to \infty} \mathbb{E}[L_n]/n \le H(X_1) + \epsilon$.