UC Berkeley
Department of Electrical Engineering and Computer Sciences

EECS 126: Probability and Random Processes

Homework 06
Fall 2023

1. **Breaking a Stick**

   I break a stick n times, $n \geq 1$, in the following manner: the ith time I break the stick, I keep a fraction $X_i \sim \text{Uniform}((0,1])$ of the remaining stick. Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. Let $P_n = \prod_{i=1}^{n} X_i$ be the fraction of the original stick that I end up with at time n.

   a. Show that $P_n^{1/n}$ converges almost surely, and find its limit.

   b. Compute $\mathbb{E}(P_n)^{1/n}$.

   c. Now compute $\mathbb{E}(P_n^{1/n})$. Do you find the same answer as in part b? Is the limit of $\mathbb{E}(P_n^{1/n})$ equal to the limit you found in part a?

2. **The CLT Implies the WLLN**

   a. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables. Show that if X_n converges in distribution to a constant c, then X_n converges in probability to c.

   b. Now let $(X_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d. random variables with mean μ and finite variance σ^2. Show that the CLT implies the WLLN: that is,

$$\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(X_i-\mu) \xrightarrow{\mathsf{d}} Z \sim \mathcal{N}(0,1) \implies \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} \mu,$$

   where $\xrightarrow{\mathsf{d}}$ is short for "converges in distribution" and $\xrightarrow{\mathbb{P}}$ for "converges in probability."

3. **Borel–Cantelli and the Strong Law**

In this problem, we walk through a proof of the strong law (assuming finite 4th moments) that relies only on basic probability. In class we covered the *Borel-Cantelli lemma*, which states that for events $(A_n)_{n=1}^{\infty}$, if $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then

$$\mathbb{P}(A_n \text{ i.o.}) = 0,$$

where we define the event $\{A_n \text{ i.o.}\} = \cap_{n\geq 1} \cup_{m\geq n} A_n$ as the event where infinitely many A_n occur.

a. Let $X_1, X_2, \ldots$ be i.i.d. with $\mathbb{E}\, X_i = 0$ and $\mathbb{E}\, X_i^4 < \infty$ (and so we also have finite second and third moments). Let $S_n = X_1 + \cdots + X_n$, and compute $\mathbb{E}[S_n^4]$. Write your answer in terms of the moments $\mathbb{E}[X_i^2], \mathbb{E}[X_i^3], \mathbb{E}[X_i^4]$.

b. Fix an $\varepsilon > 0$, and use Markov's inequality to show that, for any n,

$$\mathbb{P}(|S_n/n| > \varepsilon) \leq O(n^{-2}).$$

c. Finally, use Borel-Cantelli to conclude that $\mathbb{P}(\lim_{n\to\infty} S_n/n = 0) = 1$. This a weaker (the full theorem assumes only finite first moments) form of the *strong law of large numbers*.

4. **Jensen's Inequality and Information Measures**

Note: This problem set is designed to be worked on in the order that the questions appear. You may cite results from previous problems in your solutions.

a. Prove **Jensen's inequality**: if φ is a convex function from $\mathbb{R}$ to $\mathbb{R}$ and Z is a random variable, then $\varphi(\mathbb{E}(Z)) \leq \mathbb{E}(\varphi(Z))$.

Hint: A convex function $\varphi\colon \mathbb{R} \to \mathbb{R}$ is lower bounded by all *tangent lines* ℓ that intersect φ at some point(s) and lie below φ everywhere else.

b. Show that $H(X) \leq \log|\mathcal{X}|$ for any distribution p_X. Conclude that for random variables taking values in $[n] \coloneqq \{1, \ldots, n\}$, the distribution which maximizes $H(X)$ is Uniform($[n]$).

Hint: log is a concave function, for which $\log \mathbb{E}(Z) \geq \mathbb{E}(\log Z)$.

c. For two random variables X, Y, we define their *mutual information* to be

$$I(X;Y) = \sum_x \sum_y p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x)\, p_Y(y)},$$

where the sums are taken over all outcomes of X and Y. Show that $I(X;Y) \geq 0$.

d. The *conditional entropy* of X given Y is defined to be

$$\begin{aligned} H(X \mid Y) &= \sum_y p_Y(y) \cdot H(X \mid Y = y) \\ &= \sum_y p_Y(y) \sum_x p_{X|Y}(x \mid y) \log \frac{1}{p_{X|Y}(x \mid y)}. \end{aligned}$$

Show that $H(X) \geq H(X \mid Y)$. Intuitively, conditioning will only ever reduce or maintain our uncertainty, never increase it. *Hint*: Use part c.

5. **Compression of a Random Source**

Suppose I'm trying to send a text message to a friend. In general, I need $\log_2(26)$ bits for every letter I want to send, as there are 26 letters in the English alphabet, but if I have some information on the distribution of the letters, I can do better. For example, I might give the most common letter 'e' a shorter bit representation. It turns out the number of bits needed on average is precisely the entropy of the distribution: let us see why that is.

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

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

a. Here, we extend the notation $p(\cdot)$ to denote the joint PMF of $(X_1, \ldots, X_n)$, so that $p(x_1, \ldots, x_n) = p(x_1) \cdots p(x_n)$. Show that

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

b. Fix $\varepsilon > 0$ and define $A_\varepsilon^{(n)}$ to be the set of all sequences $(x_1, \ldots, x_n) \in \mathcal{X}^n$ such that

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

Show that for all n sufficiently large,

$$\mathbb{P}((X_1, \ldots, X_n) \in A_\varepsilon^{(n)}) > 1 - \varepsilon.$$

Consequently, $A_\varepsilon^{(n)}$ is called the **typical set**, because the observed sequences lie within $A_\varepsilon^{(n)}$ with high probability.

c. Show that for all n sufficiently large,

$$(1-\varepsilon) 2^{n(H(X_1)-\varepsilon)} \le \left| A_\varepsilon^{(n)} \right| \le 2^{n(H(X_1)+\varepsilon)}.$$

Hint: Use the union bound.

Parts (b) and (c) are called the **asymptotic equipartition property** (AEP), because they state there are $\approx 2^{nH(X_1)}$ possible observed sequences, each with probability $\approx 2^{-nH(X_1)}$. Thus, by discarding the sequences outside of $A_\varepsilon^{(n)}$, we need only keep track of $2^{nH(X_1)}$ sequences, which means that a sequence of length n 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 sets $B_n \subseteq \mathcal{X}^n$ with $|B_n| \le 2^{n(H(X_1)-\delta)}$, $n \ge 1$, we have

$$\mathbb{P}((X_1, \ldots, X_n) \in B_n) \to 0 \text{ as } n \to \infty.$$

In other words, we cannot compress the possible observed sequences of length n into any set smaller than size $2^{nH(X_1)}$; the typical set is in this sense *minimal*.

Hint: Consider the intersection of B_n and $A_\varepsilon^{(n)}$.

e. Finally, we turn towards using the AEP for compression. Recall that encoding a set of size n in binary requires $\lceil \log_2(n) \rceil$ bits, so a naïve encoding of the message sequence requires $\lceil \log_2 |\mathcal{X}| \rceil$ bits per symbol.

From the previous parts, if we use $\log_2 |A_\varepsilon^{(n)}| \approx nH(X_1)$ bits to encode the sequences in the typical set, 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 using $1 + \lceil \log_2 |A_\varepsilon^{(n)}| \rceil$ bits to encode the sequences in the typical set and $1 + n\lceil \log_2 |\mathcal{X}| \rceil$ bits for other sequences, where the first bit is used to indicate whether the sequence belongs in $A_\varepsilon^{(n)}$ or not. Let L_n be the length of the encoding of $(X_1, \ldots, X_n)$ using this error-free code. Show that

$$\lim_{n\to\infty} \frac{\mathbb{E}(L_n)}{n} \le H(X_1) + \varepsilon.$$

In other words, asymptotically, we can compress the message sequence so that the number of bits per symbol is arbitrary close to the entropy.

6. **Crafty Bounds**

   We have an alphabet $\mathcal{X}$ containing n letters $\{x_1, \ldots, x_n\}$, where each letter x_i occurs with probability p_i. We wish to *encode* the alphabet by assigning to each letter x_i a binary string of length ℓ_i. Let $L = \sum_{i=1}^n p_i \ell_i$ be the expected codeword length, and let $H(p)$ be the entropy of the distribution on $\mathcal{X}$.

   a. Prove the lower bound $H(p) \leq L$. You may cite well-known results.

   b. A code is *prefix-free* if no codeword is a prefix of another codeword. For example, `011` is a prefix of `01101`. Show that if we have a prefix-free code where each x_i is mapped to a codeword of length ℓ_i, then

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

   Hint: Consider the codewords as sequences of coin flips that we can feed into a decoder to recover the original letters, and revisit midterm 1 question 2b.

   c. Prove the converse of part b: If $\ell_1, \ell_2, \ldots, \ell_n$ satisfy $\sum_{i=1}^n 2^{-\ell_i} \leq 1$, then there exists a prefix-free code where each x_i is mapped to a codeword of length ℓ_i.

   Hint: Consider induction. Can you assume without loss of generality that $\sum_{i=1}^n 2^{-\ell_i} = 1$?

   d. Show that there exists a prefix-free code with $\ell_i = \lceil -\log_2 p_i \rceil$ for $i = 1, \ldots, n$.

   e. Conclude that there exists a prefix-free code such that $L \leq H(p) + 1$.