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 $i$th 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?

**Solution**:

a. By the Strong Law of Large Numbers,

$$\lim_{n\to\infty} \ln P_n^{1/n} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \ln X_i = \mathbb{E}(\ln X_1) \text{ a.s.}$$

We also find that $\mathbb{E}(\ln X_1) = \int_0^1 \ln x \, \mathrm{d}x = -1$. Thus, we have that

$$\mathbb{P}(\ln P_n^{1/n} \to -1) = \mathbb{P}(P_n^{1/n} \to e^{-1}) = 1,$$

so the almost-sure limit is $e^{-1}$.

b. By independence,

$$\mathbb{E}(P_n)^{1/n} = \mathbb{E}\left(\prod_{i=1}^n X_i\right)^{1/n} = (\mathbb{E}(X_1)^n)^{1/n} = \mathbb{E}(X_1) = \frac{1}{2}.$$

c. We now find that

$$\mathbb{E}(P_n^{1/n}) = \mathbb{E}(X_1^{1/n}) \cdots \mathbb{E}(X_n^{1/n}) = \mathbb{E}(X_1^{1/n})^n = \left(\int_0^1 x^{1/n} \, \mathrm{d}x\right)^n = \left(\frac{n}{n+1}\right)^n.$$

This differs from part b because the expectation of a nonlinear transformation is not necessarily equal to the nonlinear transformation of the expectation. However,

$$\lim_{n\to\infty} \mathbb{E}(P_n^{1/n}) = \lim_{n\to\infty} \left(\frac{n}{n+1}\right)^n = \lim_{n\to\infty} \left(1 - \frac{1}{n+1}\right)^n = e^{-1},$$

which is indeed the almost-sure limit of $P_n^{1/n}$ we 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 $\mu$ and finite variance $\sigma^2$. 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."

**Solution**:

a. Since $X_n$ converges in distribution to $c$, we know that for all $\varepsilon>0$,

$$\lim_{n\to\infty}F_{X_n}(c-\varepsilon)=F_c(c-\varepsilon)=0$$
$$\lim_{n\to\infty}F_{X_n}(c+\tfrac{\varepsilon}{2})=F_c(c+\tfrac{\varepsilon}{2})=1.$$

Using these limits, we have convergence in probability:

$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}(|X_n-c|\geq\varepsilon)&=\lim_{n\to\infty}\mathbb{P}(X_n\leq c-\varepsilon)+\lim_{n\to\infty}\mathbb{P}(X_n\geq c+\varepsilon)\\&\leq\lim_{n\to\infty}\mathbb{P}(X_n\leq c-\varepsilon)+\lim_{n\to\infty}\mathbb{P}(X_n>c+\tfrac{\varepsilon}{2})\\&=\lim_{n\to\infty}F_{X_n}(c-\varepsilon)+\lim_{n\to\infty}1-F_{X_n}(c+\tfrac{\varepsilon}{2})\\&=0+1-1=0.\end{aligned}$$

(The reason we take $c+\frac{\varepsilon}{2}$ instead of $c+\varepsilon$ is because $1-F_{X_n}(x)=\mathbb{P}(X_n>x)$, but we have $\mathbb{P}(X_n\geq c+\varepsilon)$, which is not a strict inequality.)

b. From the CLT, we know that

$$Z_n\coloneqq\frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\right)\text{ converges to }Z\sim\mathcal{N}(0,1)\text{ in distribution.}$$

Additionally, $a_n\coloneqq\frac{\sigma}{\sqrt{n}}\to 0$. Then $Y_n\coloneqq a_nZ_n=\frac{1}{n}\sum_{i=1}^{n}X_i-\mu\to 0$ in distribution. By part a, since $c=0$ is a constant, $Y_n$ also converges to 0 in probability. In other words,

$$\frac{1}{n}\sum_{i=1}^{n}X_i\to\mu\text{ in probability,}$$

which is precisely the Weak Law of Large Numbers.

*Note.* The claim that "if $Z_n\to Z$ in distribution and $a_n\to 0$ as constants, then $a_nZ_n\to 0$ in distribution" requires proof, which we present below.

For $x<0$ and any $N\geq 1$, we know that $\frac{x}{a_n}\leq -N$ eventually, so

$$\lim_{n\to\infty}\mathbb{P}(a_nZ_n\leq x)=\lim_{n\to\infty}\mathbb{P}(Z_n\leq\tfrac{x}{a_n})\leq\lim_{n\to\infty}\mathbb{P}(Z_n\leq -N)=\mathbb{P}(Z\leq -N).$$

The left-hand side does not depend on $N$, so taking the limit as $N \to \infty$ of both sides, by continuity from above, we find that

$$\lim_{n\to\infty} \mathbb{P}(a_n Z_n \leq x) \leq \mathbb{P}(Z = -\infty) = 0.$$

Similarly, for $x > 0$, we know that $\frac{x}{a_n} \geq N$ eventually for any $N$, so

$$\lim_{n\to\infty} \mathbb{P}(a_n Z_n > x) \leq \lim_{n\to\infty} \mathbb{P}(Z_n > N) = \mathbb{P}(Z > N),$$

and taking the limit as $N \to \infty$, we find that $\lim_{n\to\infty} \mathbb{P}(a_n Z_n > x) \leq 0$, or $\lim_{n\to\infty} \mathbb{P}(a_n Z_n \leq x) = 1$. In other words, we have shown that

$$\lim_{n\to\infty} \mathbb{P}(a_n Z_n \leq x) = \mathbb{1}\{0 \leq x\},$$

i.e. $a_n Z_n$ converges to 0 in distribution. This is a specific case of a more general result called *Slutsky's theorem*.

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*.

**Solution**:

a. We expand:

$$\mathbb{E}\, S_n^4 = \mathbb{E}\left(\sum_{i=1}^{n} X_i\right)^4 = \mathbb{E} \sum_{1\leq i,j,k,l\leq n} X_i X_j X_k X_l.$$

Terms of the form $\mathbb{E}[X_i^3 X_j]$, $\mathbb{E}[X_i^2 X_j X_k]$, and $\mathbb{E}[X_i X_j X_k X_l]$ are just 0 by independence. We are left with

$$\mathbb{E}\left(\sum_{i=1}^{n} X_i^4\right) + \mathbb{E}\left[\sum_{i\neq j} X_i^2 X_j^2\right] = n\,\mathbb{E}[X_1^4] + 3n(n-1)\,\mathbb{E}[X_1^2]\,\mathbb{E}[X_2^2].$$

b. By Markov's inequality and the previous part, we have

$$\mathbb{P}(|S_n/n| > \varepsilon) < \varepsilon^{-4}\,\mathbb{E}(S_n/n)^4 = O(\varepsilon^{-4} n^{-2}).$$

c. Letting $A_n = \{|S_n/n| > \varepsilon\}$, we get from the Borel-Cantelli lemma that $\mathbb{P}(|S_n/n| > \varepsilon \text{ i.o.}) = 0$. Since $\varepsilon$ is arbitrary, this implies almost sure convergence.

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 $\varphi$ 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* $\ell$ that intersect $\varphi$ at some point(s) and lie below $\varphi$ 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.

**Solution**:

a. Per the hint, for every $x \in \mathbb{R}$, $\varphi(x) = \sup\{\ell(x) : \ell \text{ an affine function such that } \ell \leq \varphi\}$. Consider any particular $\ell(x) = ax + b$ such that $\ell \leq \varphi$. We have that

$$\mathbb{E}(\varphi(Z)) \geq \mathbb{E}(\ell(Z)) = a\,\mathbb{E}(Z) + b = \ell(\mathbb{E}(Z)).$$

As this is true for all affine functions $\ell \leq \varphi$, we can take the supremum to find that

$$\mathbb{E}(\varphi(Z)) \geq \sup_{\ell \leq \varphi} \ell(\mathbb{E}(Z)) = \varphi(\mathbb{E}(Z)).$$

b. $Z = 1/p_X(X)$ is a function of $X$ and thus a random variable, taking values in $[1, \infty)$. Since log is a concave function, or $-\log$ is a convex function, by Jensen's inequality,

$$\begin{aligned} H(X) = \mathbb{E}\left(\log \frac{1}{p_X(X)}\right) &\leq \log \mathbb{E}\left(\frac{1}{p_X(X)}\right) \\ &= \log \sum_{x \in \mathcal{X}} p_X(x) \frac{1}{p_X(x)} \\ &= \log \sum_{x \in \mathcal{X}} 1 = \log|\mathcal{X}|. \end{aligned}$$

Then, note that for $X \sim \text{Uniform}([n])$, we have

$$H(X) = \sum_{k=1}^{n} \frac{1}{n} \log \frac{1}{1/n} = \log n = \log|\{1, \ldots, n\}|.$$

Hence the uniform distribution maximizes entropy for the finite set $[n]$.

c. Observe that $Z = p(X)\,p(Y)/p(X,Y)$ is a function of $X, Y$ and thus a random variable. Moreover, by the Law of the Unconscious Statistician, we see that

$$I(X;Y) = \mathbb{E}(\log \tfrac{1}{Z}) = \mathbb{E}(-\log Z).$$

Applying Jensen's inequality, we have

$$\begin{aligned} I(X;Y) &\geq -\log\left(\sum_x \sum_y p(x,y) \frac{p(x)\,p(y)}{p(x,y)}\right) \\ &= -\log\left(\sum_x \sum_y p(x)\,p(y)\right) \\ &= -\log\left(\sum_x p(x) \sum_y p(y)\right) \\ &= -\log(1) = 0. \end{aligned}$$

d. We now observe that $H(X) = \mathbb{E}(-\log p(X))$, and

$$H(X \mid Y) = \sum_x \sum_y p(x,y) \log \frac{1}{p(x \mid y)} = \mathbb{E}(-\log p(X \mid Y)).$$

By part c and the linearity of expectation, we find that

$$\begin{aligned} I(X;Y) &= \mathbb{E}[-\log(p(X)/p(X \mid Y))] \\ &= \mathbb{E}(-\log p(X)) - \mathbb{E}(-\log p(X \mid Y)) \\ &= H(X) - H(X \mid Y) \geq 0. \end{aligned}$$

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

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

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,

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

*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 $|B_n| \leq 2^{n(H(X_1)-\delta)}$, $n \geq 1$, 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

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

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

**Solution**:

a. As $(X_i)_{i=1}^\infty$ is a sequence of i.i.d. random variables, so is $(\log_2 p(X_i))_{i=1}^\infty$. Thus, by the Strong Law of Large Numbers,

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

b. As a consequence of part (a), $-\frac{1}{n} \log_2 p(X_1, \ldots, X_n) \to H(X_1)$ in probability, so

$$\mathbb{P}\left( \left| -\frac{1}{n} \log_2 p(X_1, \ldots, X_n) - H(X_1) \right| < \varepsilon \right) \to 1 \qquad \text{as } n \to \infty.$$

In other words, for all $n$ sufficiently large, the left hand side is $> 1 - \varepsilon$.

c. Taking the hint, we have

$$1 = \sum_{x \in \mathcal{X}^n} p(x) \geq \sum_{x \in A_\varepsilon^{(n)}} p(x) \geq \sum_{x \in A_\varepsilon^{(n)}} 2^{-n(H(X_1)+\varepsilon)} = |A_\varepsilon^{(n)}| \cdot 2^{-n(H(X_1)+\varepsilon)}.$$

This shows that $|A_\varepsilon^{(n)}| \leq 2^{n(H(X_1)+\varepsilon)}$. Now, for all $n$ sufficiently large,

$$1 - \varepsilon < \mathbb{P}\left((X_1, \ldots, X_n) \in A_\varepsilon^{(n)}\right) \leq \sum_{x \in A_\varepsilon^{(n)}} 2^{-n(H(X_1)-\varepsilon)} = |A_\varepsilon^{(n)}| \cdot 2^{-n(H(X_1)-\varepsilon)},$$

which gives $|A_\varepsilon^{(n)}| \geq (1 - \varepsilon) 2^{n(H(X_1)-\varepsilon)}$.

d. For any $\delta > 0$, we can choose $0 < \varepsilon < \delta$, so that

$$\begin{aligned}
\mathbb{P}((X_1, \ldots, X_n) \in B_n) &\leq \mathbb{P}((X_1, \ldots, X_n) \in A_\varepsilon^{(n)} \cap B_n) + \mathbb{P}((X_1, \ldots, X_n) \notin A_\varepsilon^{(n)}) \\
&\leq |B_n| \cdot 2^{-n(H(X_1)-\varepsilon)} + \mathbb{P}((X_1, \ldots, X_n) \notin A_\varepsilon^{(n)}) \\
&\leq 2^{-n(\delta-\varepsilon)} + \mathbb{P}((X_1, \ldots, X_n) \notin A_\varepsilon^{(n)}) \to 0
\end{aligned}$$

by part (b), as this inequality holds for all $0 < \varepsilon < \delta$.

e. Separating the sequences in the typical set from other sequences,

$$\frac{\mathbb{E}(L_n)}{n} = \frac{1 + \lceil \log_2 |A_\varepsilon^{(n)}| \rceil}{n} \mathbb{P}((X_1, \ldots, X_n) \in A_\varepsilon^{(n)}) + \frac{1 + n\lceil \log_2 |\mathcal{X}| \rceil}{n} \mathbb{P}((X_1, \ldots, X_n) \notin A_\varepsilon^{(n)})$$

$$\leq \frac{1 + \lceil n(H(X_1) + \varepsilon) \rceil}{n} + (1 + \lceil \log_2 |\mathcal{X}| \rceil) \, \mathbb{P}((X_1, \ldots, X_n) \notin A_\varepsilon^{(n)})$$
$$\to H(X_1) + \varepsilon.$$

Note that the probability of being outside of the typical set tends to 0 asymptotically, so only the first term matters as $n \to \infty$.

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 $\ell_i$. 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 $\ell_i$, 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 $\ell_i$.
*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 $L \leq H(p) + 1$.

**Solution**:

a. This bound follows from Shannon's source coding theorem, namely that the entropy gives a lower bound on the average number of bits required to encode each letter.

b. Consider a sequence of i.i.d. Bernoulli($\frac{1}{2}$) random bits, and let $A_i$ be the event that the first $\ell_i$ bits in the sequence decode to the letter $x_i$. Then $A_1, \ldots, A_n$ are disjoint because the code is prefix-free, and we have that

$$\sum_{i=1}^{n} 2^{-\ell_i} = \sum_{i=1}^{n} \mathbb{P}(A_i) = \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq 1.$$

c. Assume without loss of generality that $\sum_{i=1}^{n} 2^{-\ell_i} = 1$, which we can always achieve by reducing the lengths $\ell_i$. If a prefix-free code exists for the reduced $\ell_i$, then we can simply extend those codewords until we have the desired lengths.

- The base case can be taken to be $n = 1$ (degenerate) or $n = 2$, where $\ell_1 = \ell_2 = 1$ and a prefix-free code is given by `0` and `1`.
- Now, suppose that the proposition holds for $n = k$. Given $\ell_1, \ldots, \ell_{k+1}$ such that $\sum_{i=1}^{k+1} 2^{-\ell_i} = 1$, consider the two longest lengths, without loss of generality $\ell_k$ and $\ell_{k+1}$. Because equality is achieved, we must actually have $\ell_k = \ell_{k+1}$.
  By the inductive hypothesis, there exists a prefix-free code whose codeword lengths are $\ell_1, \ldots, \ell_{k-1}, (\ell_k - 1)$. We can replace the codeword `s` of length $\ell_k - 1$ with two codewords `s0` and `s1`, which have lengths $\ell_k = \ell_{k+1}$, and this is the desired code for $n = k + 1$. This finishes the inductive step and the proof.

*Remark.* Parts b and c are known as the *Kraft–McMillan inequality*.

**Alternate solution**. Suppose without loss of generality that $\ell_1 \le \ell_2 \le \cdots \le \ell_n$, and let us assign codewords one-by-one. In step $k$, given that we have prefix-free codewords of lengths $\ell_1, \ldots, \ell_{k-1}$, there exists a valid codeword of length $\ell_k$ iff

$$2^{\ell_k} \ge 1 + \sum_{i=1}^{k-1} 2^{\ell_k - \ell_i}.$$

The right-hand sum counts the number of bitstrings of length $\ell_k$ that *do* share a prefix with any of the previous $k-1$ codewords. Now, dividing on both sides, this says

$$1 \ge 2^{-\ell_k} + \sum_{i=1}^{k-1} 2^{-\ell_i} = \sum_{i=1}^{k} 2^{-\ell_i}.$$

There exists a prefix-free code with codeword lengths $\ell_1, \ldots, \ell_n$ if and only if the inequality above holds at every step $k = 1, \ldots, n$. But this is precisely equivalent to $\sum_{i=1}^{n} 2^{-\ell_i} \le 1$.

d. For $\ell_i = \lceil -\log_2 p_i \rceil$, we observe that

$$\sum_{i=1}^{n} 2^{-\lceil -\log_2 p_i \rceil} \le \sum_{i=1}^{n} 2^{-(-\log_2 p_i)} = \sum_{i=1}^{n} p_i = 1.$$

By part c, the desired prefix-free code indeed exists.

e. Considering the code identified in part d, we have that

$$L = \sum_{i=1}^{n} p_i \lceil -\log_2 p_i \rceil \le \sum_{i=1}^{n} p_i(-\log_2 p_i + 1) = H(p) + 1.$$

*Remark.* The *Huffman code* is optimal among all prefix-free codes that assign codewords letter-by-letter, so its expected codeword length satisfies the bounds $H(p) \le L \le H(p)+1$.