1. Midterm

Solve again the midterm problems which you got incorrect. Please demonstrate understanding of the questions without simply copying the solutions.

Solution: See midterm solutions.
2. The CLT Implies the WLLN

a. Let \((X_n)_{n\in\mathbb{N}}\) be a sequence of random variables. Show that if \(X_n \overset{d}{\to} c\) for some constant \(c\), then \(X_n \overset{P}{\to} c\).

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,

\[
\frac{1}{\sigma \sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \overset{d}{\to} \mathcal{N}(0,1) \implies \frac{1}{n} \sum_{i=1}^{n} X_i \overset{P}{\to} \mu.
\]

Solution:

a. Since \(X_n \overset{d}{\to} c\), we know that for all \(\varepsilon > 0\),

\[
\lim_{n \to \infty} F_{X_n}(c - \varepsilon) = 0 \\
\lim_{n \to \infty} F_{X_n} \left( c + \frac{\varepsilon}{2} \right) = 1.
\]

Using these limits, we have convergence in probability:

\[
\lim_{n \to \infty} \mathbb{P}(|X_n - c| \geq \varepsilon) = \lim_{n \to \infty} \mathbb{P}(X_n \leq c - \varepsilon) + \mathbb{P}(X_n \geq c + \varepsilon) \\
= \lim_{n \to \infty} F_{X_n}(c - \varepsilon) + \lim_{n \to \infty} \mathbb{P}(X_n \geq c + \varepsilon) \\
\leq 0 + \lim_{n \to \infty} \mathbb{P} \left( X_n > c + \frac{\varepsilon}{2} \right) \\
= 0 + 1 - \lim_{n \to \infty} F_{X_n} \left( c + \frac{\varepsilon}{2} \right) = 0.
\]

b. From the CLT, we know that

\[
\frac{\sqrt{n}}{\sigma} \left( \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right) \overset{d}{\to} Z \sim \mathcal{N}(0,1).
\]

In addition, \(\frac{\sigma}{\sqrt{n}} \to 0\), so

\[
\frac{1}{n} \sum_{i=1}^{n} X_i - \mu \overset{d}{\to} 0, \text{ or, stated another way, } \frac{1}{n} \sum_{i=1}^{n} X_i \overset{d}{\to} \mu.
\]

By part a), we can conclude that

\[
\frac{1}{n} \sum_{i=1}^{n} X_i \overset{P}{\to} \mu.
\]
3. Confidence Intervals: Chebyshev vs Chernoff vs CLT

Let \( X_1, \ldots, X_n \) be i.i.d. Bernoulli(\( q \)) random variables, with common mean \( \mu = \mathbb{E}(X_1) = q \) and variance \( \sigma^2 = \text{var}(X_1) = q(1 - q) \). We want to estimate the mean \( \mu \), so towards this goal we use the sample mean estimator

\[
\bar{X}_n := \frac{X_1 + \cdots + X_n}{n}.
\]

Given some confidence level \( a \in (0, 1) \), we want to construct a confidence interval around \( \bar{X}_n \) such that \( \mu \) lies in this interval with probability at least \( 1 - a \).

a. Use Chebyshev’s inequality to show that \( \mu \) lies in the interval

\[
\left( \bar{X}_n - \frac{\sigma}{\sqrt{n}} \frac{1}{\sqrt{a}}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} \frac{1}{\sqrt{a}} \right)
\]

with probability at least \( 1 - a \).

b. A Chernoff bound for this setting can be computed to be

\[
\mathbb{P}(|\bar{X}_n - q| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}
\]

for any \( \varepsilon > 0 \). Use this inequality in order to show that \( \mu \) lies in the interval

\[
\left( \bar{X}_n - \frac{1}{\sqrt{n}} \sqrt{2 \ln \frac{1}{a}}, \bar{X}_n + \frac{1}{\sqrt{n}} \sqrt{2 \ln \frac{1}{a}} \right)
\]

with probability at least \( 1 - a \).

c. Show that if \( Z \sim \mathcal{N}(0, 1) \), then for any \( \varepsilon > 0 \),

\[
\mathbb{P}(|Z| \geq \varepsilon) \leq 2e^{-\frac{\varepsilon^2}{2}}.
\]

d. Use the Central Limit Theorem and part c) to argue heuristically that \( \mu \) lies in the interval

\[
\left( \bar{X}_n - \frac{\sigma}{\sqrt{n}} \sqrt{2 \ln \frac{1}{a}}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} \sqrt{2 \ln \frac{1}{a}} \right)
\]

with probability at least \( 1 - a \).

e. Compare the three confidence intervals.

**Solution:**

a. Rewrite the probability that \( \mu \) lies in the specified interval as the probability that \( \bar{X}_n \) lies in an interval of the same width around \( \mu \):

\[
\mathbb{P} \left( \mu \in \left( \bar{X}_n - \frac{\sigma}{\sqrt{n}} \frac{1}{\sqrt{a}}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} \frac{1}{\sqrt{a}} \right) \right) = \mathbb{P} \left( |\bar{X}_n - \mu| \leq \frac{\sigma}{\sqrt{n}} \frac{1}{\sqrt{a}} \right)
\]

\[
= 1 - \mathbb{P} \left( |\bar{X}_n - \mu| > \frac{\sigma}{\sqrt{n}} \frac{1}{\sqrt{a}} \right)
\]

\[
\geq 1 - \frac{\text{var}(\bar{X}_n)}{\left( \sigma^2/n \right) \left( 1/a \right)} = 1 - a,
\]

where \( \text{var}(\bar{X}_n) = \sigma^2/n \).
b. Using the same idea as the previous part, but with the stronger tail inequality,
\[
P \left( \mu \in \left( X_n - \frac{1}{\sqrt{n}} \sqrt{2 \ln \frac{2}{a}}, X_n + \frac{1}{\sqrt{n}} \sqrt{2 \ln \frac{2}{a}} \right) \right) = P \left( |X_n - \mu| \leq \frac{1}{2 \sqrt{n}} \sqrt{2 \ln \frac{2}{a}} \right)
\]
\[
= 1 - P \left( |X_n - \mu| > \frac{1}{2 \sqrt{n}} \sqrt{2 \ln \frac{2}{a}} \right)
\]
\[
\geq 1 - 2 \exp \left( - \ln \frac{2}{a} \right) = 1 - a.
\]

c. For any \( t > 0 \), we have that
\[
P(Z \geq \varepsilon) = P(e^{tZ} \geq e^{t\varepsilon}) \leq \frac{E(e^{tZ})}{e^{t\varepsilon}} = e^{\frac{1}{2}t^2 - t\varepsilon}.
\]
Optimizing over \( t > 0 \) yields \( P(Z \geq \varepsilon) \leq e^{-\frac{\varepsilon^2}{2t}} \), from which the final result follows by a union bound.

d. From the CLT and the previous part, we have that
\[
P \left( \left| \frac{\sqrt{n}}{\sigma} (X_n - \mu) \right| \geq \varepsilon \right) \approx P(|Z| \geq \varepsilon) \leq 2e^{-\frac{\varepsilon^2}{2}}.
\]
Setting \( \varepsilon = \sqrt{2 \ln \frac{2}{a}} \) such that \( a = 2e^{-\frac{\varepsilon^2}{2}} \),
\[
P \left( |X_n - \mu| \geq \frac{\sigma}{\sqrt{n}} \sqrt{2 \ln \frac{2}{a}} \right) \leq a,
\]
or equivalently
\[
P \left( X_n - \frac{\sigma}{\sqrt{n}} \sqrt{2 \ln \frac{2}{a}} < \mu < X_n + \frac{\sigma}{\sqrt{n}} \sqrt{2 \ln \frac{2}{a}} \right)
\]
\[
= P \left( -\frac{\sigma}{\sqrt{n}} \sqrt{2 \ln \frac{2}{a}} < \mu - X_n < \frac{\sigma}{\sqrt{n}} \sqrt{2 \ln \frac{2}{a}} \right)
\]
\[
= P \left( |X_n - \mu| < \frac{\sigma}{\sqrt{n}} \sqrt{2 \ln \frac{2}{a}} \right) \geq 1 - a.
\]

e. We can see that Chebyshev’s inequality and the CLT produce confidence intervals with standard deviation term \( \sigma \) present, while on the other hand using the Chernoff bound the standard deviation is replaced by \( \frac{1}{\sqrt{2}} \), which is only an upper bound on \( \sigma \), as \( \sigma^2 = q(1-q) \leq 1/4 \).

Chebyshev’s inequality is able to capture the standard deviation term, but on the other hand it has a poor dependence of the form \( \frac{1}{\sqrt{a}} \) on the confidence level \( a \). Chernoff’s inequality and the CLT have a much better dependence on \( a \) of the form \( \sqrt{\ln \frac{2}{a}} \).

Finally, while the confidence intervals derived via Chebyshev’s and Chernoff’s inequality are true or provable confidence intervals, we can only argue heuristically about the interval derived via the CLT.
4. Introduction to Information Theory

Recall that the entropy of a discrete random variable $X$ is defined as

$$H(X) \triangleq -\sum_x p(x) \log p(x) = -\mathbb{E}(\log p(X)),$$

where $p(\cdot)$ is the PMF of $X$. Here, the logarithm is taken in base 2, and entropy is measured in the unit of bits.

a. Prove that $H(X) \geq 0$.

b. Entropy is often described as the average information content of a random variable. If $H(X) = 0$, then no new information is given by observing $X$. On the other hand, if $H(X) = m$, then observing the value of $X$ gives you $m$ bits of information on average.

Let $X$ be a Bernoulli($p$) random variable. Would you expect $H(X)$ to be greater when $p = \frac{1}{2}$ or when $p = \frac{1}{3}$? Calculate $H(X)$ in both of these cases and verify your answer.

c. We now consider a binary erasure channel (BEC).

![Channel Model](image)

Figure 1: The channel model for the BEC showing a mapping from channel input $X$ to channel output $Y$. The probability of erasure is $p_e$.

The input $X$ is a Bernoulli random variable with $\mathbb{P}(X = 0) = \mathbb{P}(X = 1) = \frac{1}{2}$. Each time that we use the channel, the input $X$ is either erased with probability $p_e$ or transmitted correctly with probability $1 - p_e$. Using the character ‘?’ to denote erasures, the output $Y$ of the channel can be written as

$$Y = \begin{cases} X & \text{with probability } 1 - p_e \\ ? & \text{with probability } p_e. \end{cases}$$

Compute $H(Y)$.

d. We defined the entropy of a single random variable as a measure of the uncertainty inherent in the distribution of the random variable. We now extend this definition for a pair of random variables $(X, Y)$, but there is really nothing new in this definition: the pair $(X, Y)$ can be considered as a single vector-valued random variable. Define the joint entropy of $(X, Y)$ to be

$$H(X, Y) \triangleq -\mathbb{E}(\log p(X, Y)),$$

where $p(\cdot, \cdot)$ is the joint PMF, and the expectation is taken over the joint distribution of $X$ and $Y$. Compute $H(X, Y)$, for the BEC.

Solution:
a. This follows from \( \log p(x) \leq 0 \) for \( p(x) \leq 1 \).

b. The closer \( p \) is to 0 or 1, the less information you gain from observing \( X \). As an extreme example, when \( p = 1 \), you already know that \( X \) will be 1, so observing \( X \) gives you no new information. Therefore, we expect that the entropy will be greatest when \( p = \frac{1}{2} \).

The entropy of a Bernoulli random variable with bias \( p \) is

\[
H(X) = -p \log p - (1 - p) \log(1 - p).
\]

When \( p = \frac{1}{2} \),

\[
H(X) = -\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2} = \frac{1}{2} + \frac{1}{2} = 1 \text{ bit.}
\]

When \( p = \frac{1}{3} \),

\[
H(X) = -\frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3} \approx 0.918 \text{ bits.}
\]

c. The random variable \( Y \) takes on three values: 0, 1, and ?. The marginal PMF of \( Y \) is

\[
Y = \begin{cases} 
0 & \text{with probability } \frac{1-p_e}{2} \\
1 & \text{with probability } \frac{1-p_e}{2} \\
? & \text{with probability } p_e.
\end{cases}
\]

Therefore the entropy of \( Y \) is

\[
H(Y) = -p_e \log p_e - (1 - p_e) \log \frac{1-p_e}{2}
= 1 - p_e - p_e \log p_e - (1 - p_e) \log(1 - p_e).
\]

d. The joint PMF of \((X, Y)\) can be found as

\[
(X, Y) = \begin{cases} 
(0, 0) & \text{with probability } \frac{1-p_e}{2} \\
(0, ?), & \text{with probability } \frac{p_e}{2} \\
(1, 1) & \text{with probability } \frac{1-p_e}{2} \\
(1, ?), & \text{with probability } \frac{p_e}{2}.
\end{cases}
\]

Therefore the entropy of the pair \((X, Y)\) is

\[
H(X, Y) = -p_e \log \frac{p_e}{2} - (1 - p_e) \log \frac{1-p_e}{2}
= 1 - p_e \log p_e - (1 - p_e) \log(1 - p_e).
\]
5. **Information Theory Bounds**

In this problem we explore some intuitive results which can be formalized using information theory.

(a) **Optional.** Prove Jensen’s inequality: if \( f \) is a convex function and \( X \) a random variable, then \( f(\mathbb{E}(X)) \leq \mathbb{E}(f(X)) \). You may use Jensen’s inequality without proof in later subparts.

*Hint.* Every convex function can be represented as the pointwise supremum of affine functions bounded above by it, i.e.

\[
f(x) = \sup\{\ell(x) = ax + b : \ell(x) \leq f(x) \forall x\}.
\]

(b) It turns out that there is actually a limit to how much “randomness” there is in a random variable \( X \) which takes on \(|X|\) distinct values. Show that \( H(X) \leq \log |X| \) for any distribution \( p_X \). Use this to conclude that if a random variable \( X \) takes values in \([n] := \{1, \ldots, n\}\), then the distribution which maximizes \( H(X) \) is \( X \sim \text{Uniform}([n]) \).

(c) For two random variables \( X, Y \), we define their *mutual information* to be, as in discussion,

\[
I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)},
\]

where the sums are taken over all outcomes of \( X \) and \( Y \). Show that \( I(X;Y) \geq 0 \). You have seen that \( I(X;Y) = H(X) - H(X | Y) \), so the fact that mutual information is nonnegative means intuitively that conditioning will only ever reduce our uncertainty.

**Solution:**

a. Per the hint, taking expectations, we have that for all affine \( \ell \leq f \),

\[
\mathbb{E}(f(X)) \geq \mathbb{E}(aX + b) = a \mathbb{E}(X) + b.
\]

In particular, since this is true for all affine \( \ell \) dominated by \( f \), we have as desired

\[
\mathbb{E}(f(X)) \geq \sup_{\ell \leq f} \ell(\mathbb{E}(X)) = f(\mathbb{E}(X)).
\]

b. Since log is a concave function,

\[
H(X) = \mathbb{E} \left( \log \frac{1}{p_X(X)} \right) \leq \log \mathbb{E} \left( \frac{1}{p_X(X)} \right) = \log \left( \sum_{x \in \mathcal{X}} p_X(x) \frac{1}{p_X(x)} \right) = \log \left( \sum_{x \in \mathcal{X}} 1 \right) = \log |\mathcal{X}|.
\]

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. Applying Jensen’s inequality, we have

\[
I(X;Y) \geq - \log \left( \sum_x \sum_y p(x, y) \frac{p(x)p(y)}{p(x, y)} \right) \\
= - \log \left( \sum_x \sum_y p(x)p(y) \right) \\
= - \log \left( \sum_x p(x) \sum_y p(y) \right) \\
= - \log(1) = 0.
\]