UC Berkeley
Department of Electrical Engineering and Computer Sciences

EECS 126: Probability and Random Processes

**Problem Set 6**
Fall 2021

## 1. The CLT Implies the WLLN

(a) Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables. Show that if $X_n \xrightarrow{\mathsf{d}} c$, where $c$ is a constant, then $X_n \xrightarrow{P} c$.

(b) 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, i.e. if

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

then

$$\frac{1}{n}\sum_{i=1}^{n}X_i \xrightarrow{P} \mu.$$

**Solution:**

(a) Since $X_n \xrightarrow{\mathsf{d}} c$, we can deduce that for any $\epsilon > 0$, we have

$$\lim_{n\to\infty} F_{X_n}(c-\epsilon) = 0,$$
$$\lim_{n\to\infty} F_{X_n}\Big(c+\frac{\epsilon}{2}\Big) = 1.$$

Using this fact we have that

$$\begin{aligned}
\lim_{n\to\infty} P(|X_n-c|\geq\epsilon) &= \lim_{n\to\infty}[P(X_n\leq c-\epsilon)+P(X_n\geq c+\epsilon)] \\
&= \lim_{n\to\infty} P(X_n\leq c-\epsilon) + \lim_{n\to\infty} P(X_n\geq c+\epsilon) \\
&= \lim_{n\to\infty} F_{X_n}(c-\epsilon) + \lim_{n\to\infty} P(X_n\geq c+\epsilon) \\
&\leq 0 + \lim_{n\to\infty} P\Big(X_n > c+\frac{\epsilon}{2}\Big) \\
&= 1 - \lim_{n\to\infty} F_{X_n}\Big(c+\frac{\epsilon}{2}\Big) \\
&= 0.
\end{aligned}$$

Therefore $\lim_{n\to\infty} P(|X_n-c|\geq\epsilon) = 0$, for all $\epsilon > 0$ which means that $X_n \xrightarrow{P} c$.

(b) From the CLT we know that

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

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

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \xrightarrow{\mathsf{d}} 0$$

or stated another way

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathsf{d}} \mu.$$

Finally using Part (a) we can conclude that

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu.$$

**2. 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 $\mu$, and towards this goal we use the sample mean estimator

$$\bar{X}_n \triangleq \frac{X_1 + \cdots + X_n}{n}.$$

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 in order 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:

$$P(|\bar{X}_n - q| \geq \epsilon) \leq 2\epsilon^{-2n\epsilon^2}, \quad \text{for any } \epsilon > 0.$$

Use this inequality in order to show that $\mu$ lies in the interval

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

with probability at least $1-a$.

(c) Show that if $Z \sim \mathcal{N}(0,1)$, then

$$P(|Z| \geq \epsilon) \leq 2\epsilon^{-\frac{\epsilon^2}{2}}, \quad \text{for any } \epsilon > 0.$$

(d) Use the Central Limit Theorem, and Part (c) in order to heuristically argue that $\mu$ lies in the interval

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

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 $\mu$:

$$\begin{aligned}
P\Big\{\mu \in \Big(\bar{X}_n - \frac{\sigma}{\sqrt{n}}\frac{1}{\sqrt{a}}, \bar{X}_n + \frac{\sigma}{\sqrt{n}}\frac{1}{\sqrt{a}}\Big)\Big\} &= P\Big(|\bar{X}_n - \mu| \le \frac{\sigma}{\sqrt{n}}\frac{1}{\sqrt{a}}\Big) \\
&= 1 - P\Big(|\bar{X}_n - \mu| > \frac{\sigma}{\sqrt{n}}\frac{1}{\sqrt{a}}\Big) \\
&\ge 1 - \frac{\operatorname{var} \bar{X}_n}{(\sigma^2/n)(1/a)} = 1 - a,
\end{aligned}$$

because $\operatorname{var} \bar{X}_n = \sigma^2/n$.

(b) Use the same idea as the previous part, but using the stronger tail inequality.

$$\begin{aligned}
&P\Big\{\mu \in \Big(\bar{X}_n - \frac{1/2}{\sqrt{n}}\sqrt{2\ln\frac{2}{a}}, \bar{X}_n + \frac{1/2}{\sqrt{n}}\sqrt{2\ln\frac{2}{a}}\Big)\Big\} \\
&\quad = P\Big(|\bar{X}_n - \mu| \le \frac{1}{2\sqrt{n}}\sqrt{2\ln\frac{2}{a}}\Big) \\
&\quad = 1 - P\Big(|\bar{X}_n - \mu| > \frac{1}{2\sqrt{n}}\sqrt{2\ln\frac{2}{a}}\Big) \ge 1 - 2\exp\Big(-\ln\frac{2}{a}\Big) = 1 - a.
\end{aligned}$$

(c) For any $t > 0$ we have that

$$\begin{aligned}
P(Z \ge \epsilon) &= P(tZ \ge t\epsilon) \\
&= P(e^{tZ} \ge e^{t\epsilon}) \\
&\le \frac{\mathbb{E}[e^{tZ}]}{e^{t\epsilon}} \\
&= \epsilon^{\frac{1}{2}t^2 - t\epsilon}.
\end{aligned}$$

Optimizing over $t > 0$, yields

$$P(Z \ge \epsilon) \le e^{-\frac{\epsilon^2}{2}}.$$

The final result follows by a union bound.

(d) From the CLT and the previous part we have that

$$P\Big(\Big|\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)\Big| \ge \epsilon\Big) \approx P(|Z| \ge \epsilon) \le 2\epsilon^{-\frac{\epsilon^2}{2}}.$$

We are going to set $\epsilon$ to be such that $a = 2\epsilon^{-\frac{\epsilon^2}{2}}$, which yields $\epsilon = \sqrt{2\ln\frac{2}{a}}$. Plugging in this value of $\epsilon$ we have that

$$P\Big(|\bar{X}_n - \mu| \ge \frac{\sigma}{\sqrt{n}}\sqrt{2\ln\frac{2}{a}}\Big) \le a,$$

or equivalently

$$P\Big(\bar{X}_n - \frac{\sigma}{\sqrt{n}}\sqrt{2\ln\frac{2}{a}} < \mu < \bar{X}_n + \frac{\sigma}{\sqrt{n}}\sqrt{2\ln\frac{2}{a}}\Big)$$

$$= P\Big(-\frac{\sigma}{\sqrt{n}}\sqrt{2\ln\frac{2}{a}} < \mu - \bar{X}_n < \frac{\sigma}{\sqrt{n}}\sqrt{2\ln\frac{2}{a}}\Big)$$
$$= P\Big(|\bar{X}_n - \mu| < \frac{\sigma}{\sqrt{n}}\sqrt{2\ln\frac{2}{a}}\Big) \gtrapprox 1 - a.$$

.

(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 $1/2$, which is only an upper bound on $\sigma$, since $\sigma^2 = \sigma^2(q) = q(1-q) \leq 1/4$.
Chebyshev's inequality is able to capture the standard deviation term, but on the other hand it has a poor dependence of the form $1/\sqrt{a}$ on the confidence level $a$. Chernoff's inequality and the CLT have a way 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/provable confidence intervals, we can only argue heuristically about the interval derived via the CLT.

**3. Transform Practice**

Consider a random variable $Z$ with transform

$$M_Z(s) = \frac{a - 3s}{s^2 - 6s + 8}, \qquad \text{for } |s| < 2.$$

Calculate the following quantities:

(a) The numerical value of the parameter $a$.

(b) $\mathbb{E}[Z]$.

(c) $\mathrm{var}(Z)$.

**Solution:**

(a) By definition, we know that $M_Z(s) = \mathbb{E}[\epsilon^{sZ}]$. Thus, we know the following must be true:

$$M_Z(0) = \mathbb{E}[\epsilon^{0Z}] = 1 = \frac{a}{8}$$

It follows that $a = 8$.

(b)
$$\mathbb{E}[Z] = \frac{\mathrm{d}}{\mathrm{d}s} M_Z(s)\Big|_{s=0} = \frac{2}{(4-s)^2} + \frac{1}{(2-s)^2}\Big|_{s=0} = \frac{3}{8}.$$

(c) Note that
$$\mathbb{E}[Z^2] = \frac{\mathrm{d}^2}{\mathrm{d}s^2} M_Z(s)\Big|_{s=0} = \frac{4}{(4-s)^3} + \frac{2}{(2-s)^3}\Big|_{s=0} = \frac{5}{16}.$$

Thus,
$$\mathrm{var}(Z) = \frac{11}{64}.$$

4. **Rotationally Invariant Random Variables**

Suppose random variables $X$ and $Y$ are i.i.d., with zero mean, such that their joint density is rotation invariant.

(a) Let $\varphi(t)$ be the characteristic function of $X$. Show that $\varphi(t)^n = \varphi(\sqrt{n}t)$.

(b) Show that $\varphi(t) = \exp(ct^2)$ for some constant $c$, and all $t$ such that $t^2 \in \mathbb{Q}$. *Hint:* Let $t^2 = a/b$, where $a, b$ are positive integers.

(c) Conclude that $X$ and $Y$ must be Gaussians.

**Solution:**

(a) For $t \in \mathbb{R}$, $tX + tY$ has the same distribution as $\sqrt{2}tX$, so $\varphi(t)^2 = \varphi(\sqrt{2}t)$. Likewise, note that $t\sqrt{n-1}X + tY$ has the same distribution as $\sqrt{n}X$, so $\varphi(\sqrt{n-1}t)\varphi(t) = \varphi(\sqrt{n}t)$. Inducting on $n$, this implies $\varphi(t)^n = \varphi(\sqrt{n}t)$ for all $n \in \mathbb{N}$.

(b) For positive integers $a, b$, let $t^2 = a/b$. Using part (a), we can write

$$\varphi(t) = \varphi(\sqrt{a/b}) = \varphi(1/\sqrt{b})^a = (\varphi(1/\sqrt{b})^b)^{a/b} = \varphi(1)^{a/b} = e^{ct^2},$$

where we took $c$ to satisfy $e^c = \varphi(1)$.

(c) We have so far shown that $\varphi(t) = \exp(ct^2)$ for all $t$ such that $t^2 \in \mathbb{Q}_{\geq 0}$. But since $\{t : t^2 \in \mathbb{Q}_{\geq 0}\}$ is a dense subset of $\mathbb{R}$, and characteristic functions are continuous, it follows that $\varphi(t) = \exp(ct^2)$ for all $t \in \mathbb{R}$. Finally, note that this is just the characteristic function of a Gaussian, so we conclude that $X$ (and also $Y$) must be Gaussian distributed.

5. **Matrix Sketching**

Matrix sketching is an important technique in randomized linear algebra to do large computations efficiently. For example, to compute the multiplication $\mathbf{A}^T \times \mathbf{B}$ of two large matrices $\mathbf{A}$ and $\mathbf{B}$, we can use a random sketch matrix $\mathbf{S}$ to compute a "sketch" $\mathbf{SA}$ of $\mathbf{A}$ and a "sketch" $\mathbf{SB}$ of $\mathbf{B}$. Such a sketching matrix has the property that $\mathbf{S}^T\mathbf{S} \approx \mathbf{I}$ so that the approximate multiplication $\mathbf{A}^T\mathbf{S}^T\mathbf{SB}$ is close to $\mathbf{A}^T\mathbf{B}$.

In this problem, we will discuss two popular sketching schemes and understand how they help in approximate computation. Let $\hat{\mathbf{I}} = \mathbf{S}^T\mathbf{S}$ and the dimension of sketch matrix $\mathbf{S}$ be $d \times n$ (typically $d \ll n$).

(a) (**Gaussian-sketch**) Define

$$\mathbf{S} = \frac{1}{\sqrt{d}} \begin{bmatrix} S_{11} & \dots & \dots & S_{1n} \\ \vdots & \ddots & & \vdots \\ S_{d1} & \dots & \dots & S_{dn} \end{bmatrix}$$

such that $S_{ij}$'s are chosen i.i.d. from $\mathcal{N}(0,1)$ for all $i \in [1,d]$ and $j \in [1,n]$. Find the element-wise mean and variance (as a function of $d$) of the matrix $\hat{\mathbf{I}} = \mathbf{S}^T\mathbf{S}$, that is, find $\mathbb{E}[\hat{I}_{ij}]$ and $\mathrm{Var}[\hat{I}_{ij}]$ for all $i \in [1,n]$ and $j \in [1,n]$.

(b) (**Count-sketch**) For each column $j \in [1,n]$ of $\mathbf{S}$, choose a row $i$ uniformly randomly from $[1,d]$ such that

$$S_{ij} = \begin{cases} 1, & \text{with probability } 0.5 \\ -1, & \text{with probability } 0.5 \end{cases}$$

and assign $S_{kj} = 0$ for all $k \neq i$. An example of a $3 \times 8$ count-sketch is

$$\mathbf{S} = \begin{bmatrix} 0 & -1 & 1 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{bmatrix}$$

Again, find the element-wise mean and variance (as a function of $d$) of the matrix $\hat{\mathbf{I}} = \mathbf{S}^T\mathbf{S}$.

Note that for sufficiently large $d$, the matrix $\hat{\mathbf{I}}$ is close to the identity matrix for both cases. We will use this fact in the lab to do an approximate matrix multiplication.

**Note:** You can use the fact that the fourth moment of a standard Gaussian is 3 without proof.

**Solution:** Let $\hat{\mathbf{I}} = \mathbf{S}^T\mathbf{S}$.

(a) For the Gaussian-sketch $\hat{I}_{ij} = \frac{1}{d}\sum_{k=1}^{d} S_{ki}S_{kj}$. Thus, by using linearity of expectation and the fact that $S_{ki}$'s are drawn i.i.d. from $\mathcal{N}(0,1)$, we get

$$\mathbb{E}[\hat{I}_{ij}] = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases}$$

By the definition of variance, we have

$$\begin{aligned} \mathrm{Var}[\hat{I}_{ij}] &= \frac{1}{d^2}\left(\mathbb{E}\left[\left(\sum_{k=1}^{d} S_{ki}S_{kj}\right)^2\right] - \left(\mathbb{E}\left[\sum_{k=1}^{d} S_{ki}S_{kj}\right]\right)^2\right) \\ &= \frac{1}{d^2}\left(\mathbb{E}\left[\left(\sum_{k=1}^{d} S_{ki}S_{kj}\right)^2\right] - \left(\sum_{k=1}^{d} \mathbb{E}[S_{ki}S_{kj}]\right)^2\right) \end{aligned}$$

Next, we consider two cases when $i = j$ and when $i \neq j$. When $i = j$

$$\begin{aligned} \mathrm{Var}[\hat{I}_{ii}] &= \frac{1}{d^2}\left(\mathbb{E}\left[\left(\sum_{k=1}^{d} S_{ki}^2\right)^2\right] - \left(\sum_{k=1}^{d} \mathbb{E}[S_{ki}^2]\right)^2\right) \\ &= \frac{1}{d^2}\left(\sum_{k=1}^{d} \mathbb{E}[S_{ki}^4] + \sum_{\substack{k=1,l=1 \\ k\neq l}}^{d} \mathbb{E}[S_{ki}^2]\mathbb{E}[S_{ki}^2] - d^2\right) \\ &= \frac{1}{d^2}\left(\sum_{k=1}^{d} \mathbb{E}[S_{ki}^4] + d(d-1) - d^2\right) \\ &= \frac{1}{d^2}\left(3d + d(d-1) - d^2\right) = \frac{2}{d}. \end{aligned}$$

where we use the fact that the fourth moment of a standard Gaussian random variable is 3.

For the case when $i \neq j$, we use the fact that $S_{ki}$ and $S_{kj}$ are independent and get

$$\mathrm{Var}[\hat{I}_{ij}] = \frac{1}{d^2}\left(\mathbb{E}\left[\left(\sum_{k=1}^{d} S_{ki}S_{kj}\right)^2\right] - \left(\sum_{k=1}^{d} \mathbb{E}[S_{ki}]\mathbb{E}[S_{kj}]\right)^2\right)$$

$$= \frac{1}{d^2}\Bigg(\sum_{k=1}^{d}\mathbb{E}[S_{ki}^2]\mathbb{E}[S_{kj}^2]\Bigg) + \sum_{\substack{k=1,l=1\\k\neq l}}^{d}\mathbb{E}[S_{ki}]\mathbb{E}[S_{kj}]\mathbb{E}[S_{li}]\mathbb{E}[S_{lj}] - 0\Bigg)$$

$$= \frac{1}{d^2}\big(d+0\big) = \frac{1}{d}.$$

Thus, we have

$$\mathrm{Var}[\hat{I}_{ij}] = \begin{cases} 2/d, & \text{if } i=j \\ 1/d, & \text{otherwise.} \end{cases}$$

(b) Note that for Count-sketch, we have $\hat{I}_{ij} = \sum_{k=1}^{d} S_{ki}S_{kj}$. By construction of $\mathbf{S}$, the diagonal terms $\hat{I}_{ii}$ are always one. Thus, we only need to worry about the non-diagonal terms. It is also important to note that in $\mathbf{S}$, entries in a row are independent but the entries in a column are dependent (there can only be one non-zero entry in one column, as shown in the example). Also,

$$S_{ki}S_{kj} = \begin{cases} 1, & \text{with probability } 1/2d \\ -1, & \text{with probability } 1/2d \\ 0, & \text{with probability } 1-1/d. \end{cases} \quad \forall\, i \neq j.$$

Thus,

$$\mathbb{E}[\hat{I}_{ij}] = \begin{cases} 1, & \text{if } i=j \\ 0, & \text{otherwise.} \end{cases}$$

The diagonal terms in $\hat{\mathbf{I}}$ are exactly one, and hence, their variance is zero. For the non-diagonal terms, i.e. when $i \neq j$, we have

$$\mathrm{Var}[\hat{I}_{ij}] = \mathbb{E}\Big[\Big(\sum_{k=1}^{d} S_{ki}S_{kj}\Big)^2\Big] - \Big(\mathbb{E}\Big[\sum_{k=1}^{d} S_{ki}S_{kj}\Big]\Big)^2$$

$$= \sum_{k=1}^{d}\mathbb{E}[S_{ki}^2]\mathbb{E}[S_{kj}^2] + \sum_{\substack{k=1,l=1\\k\neq l}}^{d}\mathbb{E}[S_{ki}S_{li}]\mathbb{E}[S_{kj}S_{lj}] - 0$$

$$= \sum_{k=1}^{d}\frac{1}{d^2} + 0 = \frac{1}{d}.$$

where the 0 in the last step comes from the fact at in any column $j$, the product of two elements $S_{kj}, S_{lj}$ is 0 since only one can be non-zero. Hence, the element-wise variance is

$$\mathrm{Var}[\hat{I}_{ij}] = \begin{cases} 0, & \text{if } i=j \\ 1/d, & \text{otherwise.} \end{cases}$$