UC Berkeley
Department of Electrical Engineering and Computer Sciences

EECS 126: Probability and Random Processes

Homework 03
Spring 2023

1. **Matrix Sketching**

   Matrix sketching is an important technique in randomized linear algebra for doing large computations efficiently. For example, to compute $\mathbf{A}^T \times \mathbf{B}$ for 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 $(\mathbf{SA})^T(\mathbf{SB}) = \mathbf{A}^T\mathbf{S}^T\mathbf{SB}$ 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 let the dimension of the sketch matrix $\mathbf{S}$ be $d \times n$ (where typically $d \ll n$).

   a. **Gaussian sketch**. Let the sketch matrix be

   $$\mathbf{S} = \frac{1}{\sqrt{d}} \begin{bmatrix} S_{1,1} & \cdots & S_{1,n} \\ \vdots & \ddots & \vdots \\ S_{d,1} & \cdots & S_{d,n} \end{bmatrix},$$

   where the $S_{i,j}$ are chosen i.i.d. from $\mathcal{N}(0,1)$ for all $i \in [1,d]$ and $j \in [1,n]$. Show that the elementwise mean and variance of the matrix $\hat{\mathbf{I}}$, as functions of d, are

   $$\mathbb{E}(\hat{I}_{i,j}) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

   $$\text{var}(\hat{I}_{i,j}) = \begin{cases} \frac{2}{d} & \text{if } i = j \\ \frac{1}{d} & \text{otherwise.} \end{cases}$$

   You can use without proof the fact that $\mathbb{E}(Z^4) = 3$ for $Z \sim \mathcal{N}(0,1)$.

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

   $$S_{i,j} = \begin{cases} 1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2}, \end{cases}$$

   and assign $S_{k,j} = 0$ for all $k \neq i$. An example of a 3×8 count sketch matrix is

   $$\begin{bmatrix} 0 & -1 & 1 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{bmatrix}.$$

Show that the elementwise mean and variance of the matrix $\hat{\mathbf{I}}$ are

$$\mathbb{E}(\hat{I}_{i,j}) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

$$\operatorname{var}(\hat{I}_{i,j}) = \begin{cases} 0 & \text{if } i = j \\ \frac{1}{d} & \text{otherwise.} \end{cases}$$

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

Solution:

a. For the Gaussian sketch matrix $\mathbf{S}$, we have

$$\hat{I}_{i,j} = \frac{1}{d} \sum_{k=1}^{d} S_{k,i} S_{k,j}.$$

By the linearity of expectation, and the $S_{k,i}$ being drawn i.i.d. from $\mathcal{N}(0,1)$, we get

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

Then, by the definition of variance, we have

$$\begin{aligned} d^2 \operatorname{var}(\hat{I}_{i,j}) &= \mathbb{E}[(d\hat{I}_{i,j})^2] - \mathbb{E}[d\hat{I}_{i,j}]^2 \\ &= \mathbb{E}\left[\left(\sum_{k=1}^{d} S_{k,i} S_{k,j}\right)^2\right] - d^2\, \mathbb{1}_{i=j}\,. \end{aligned}$$

Now we consider the two cases of $i = j$ and $i \neq j$, starting with the former:

$$\begin{aligned} d^2 \operatorname{var}(\hat{I}_{i,i}) &= \sum_{k=1}^{d} \mathbb{E}(S_{k,i}^4) + \sum_{k \neq \ell} \mathbb{E}(S_{k,i}^2)\, \mathbb{E}(S_{\ell,i}^2) - d^2 \\ &= 3d + d(d-1) - d^2 = 2d. \end{aligned}$$

For the case of $i \neq j$, we can use the independence of $S_{k,i}$ and $S_{k,j}$:

$$\begin{aligned} d^2 \operatorname{var}(\hat{I}_{i,j}) &= \sum_{k=1}^{d} \mathbb{E}(S_{k,i}^2)\, \mathbb{E}(S_{k,j}^2) + \sum_{k \neq \ell} \mathbb{E}(S_{k,i})\, \mathbb{E}(S_{k,j})\, \mathbb{E}(S_{\ell,i})\, \mathbb{E}(S_{\ell,j}) \\ &= d + 0 = d. \end{aligned}$$

Thus the elementwise variance is

$$\operatorname{var}(\hat{I}_{i,j}) = \begin{cases} \frac{2}{d} & \text{if } i = j \\ \frac{1}{d} & \text{otherwise.} \end{cases}$$

b. For the count sketch matrix $\mathbf{S}$, we have

$$\hat{I}_{i,j} = \sum_{k=1}^{d} S_{k,i} S_{k,j}.$$

By construction of $\mathbf{S}$, the diagonal terms $\hat{I}_{i,i}$ are always 1, so their mean is 1 and their variance is 0, and we only need to worry about the non-diagonal terms.

We also note that in $\mathbf{S}$, entries in a row are independent, but entries in a column are dependent. (There can only be one nonzero entry in one column.) Moreover, for all $i \neq j$,

$$S_{k,i} S_{k,j} = \begin{cases} 1 & \text{with probability } \frac{1}{2d^2} \\ -1 & \text{with probability } \frac{1}{2d^2} \\ 0 & \text{with probability } 1 - \frac{1}{d^2}. \end{cases}$$

Thus the elementwise expectation is

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

Now, for $i \neq j$, using the fact that $\mathbb{E}[\hat{I}_{i,j}]^2 = 0$,

$$\begin{aligned} \operatorname{var}(\hat{I}_{i,j}) &= \mathbb{E}\left[\left(\sum_{k=1}^{d} S_{k,i} S_{k,j}\right)^2\right] \\ &= \sum_{k=1}^{d} \mathbb{E}(S_{k,i}^2)\, \mathbb{E}(S_{k,j}^2) + \sum_{k \neq \ell} \mathbb{E}(S_{k,i} S_{\ell,i})\, \mathbb{E}(S_{k,j} S_{\ell,j}) \\ &= \sum_{k=1}^{d} \frac{1}{d^2} + 0 = \frac{1}{d}. \end{aligned}$$

The term 0 in the last step comes from the fact that in any column j, the product of two elements $S_{k,j} S_{\ell,j} = 0$, since only one can be nonzero. Thus the elementwise variance is

$$\operatorname{var}(\hat{I}_{i,j}) = \begin{cases} 0 & \text{if } i = j \\ \frac{1}{d} & \text{otherwise.} \end{cases}$$

2. **Properties of the CDF**

   The **cumulative distribution function**, or cdf, of a random variable X is the function $F(x) = \mathbb{P}(X \leq x)$.

   a. Using the properties of a probability measure, show that F is nondecreasing: if $x \leq y$, then $F(x) \leq F(y)$.

   b. Show that F is right-continuous: if $x_1, x_2, \ldots$ is a decreasing sequence converging to y, then $F(x_1), F(x_2), \ldots$ converges to $F(y)$.

   c. Show that F is *normalized*: $\lim_{x \to -\infty} F(x) = 0$, and $\lim_{x \to \infty} F(x) = 1$.

   Hint: For parts b and c, it may help to revisit question 1b of discussion 01.

   Solution:

   a. If $x \leq y$, then $\{X \leq x\} \subset \{X \leq y\}$ as events. By the monotonicity of probability, we see that $F(x) = \mathbb{P}(X \leq x) \leq \mathbb{P}(X \leq y) = F(y)$.

   b. Let $x_1, x_2, \ldots$ be a decreasing sequence converging to $y \in \mathbb{R}$. Then $\{X \leq x_1\} \supset \{X \leq x_2\} \supset \cdots \supset \{X \leq y\}$ as events. By the continuity from above of the probability measure $\mathbb{P}$, we see that $F(x_n) = \mathbb{P}(X \leq x_n)$ converges to $\mathbb{P}(X \leq y) = F(y)$.

   c. Let $x_1, x_2, \ldots$ be a sequence decreasing to $-\infty$, and consider the decreasing sequence of events $\{X \leq x_1\} \supset \{X \leq x_2\} \supset \cdots \supset \varnothing$. By continuity from above, $F(x_n) = \mathbb{P}(X \leq x_n)$ converges to $\mathbb{P}(\varnothing) = 0$.

      Similarly, let $x_1, x_2, \ldots \uparrow \infty$, and consider the increasing sequence $\{X \leq x_1\} \subset \{X \leq x_2\} \subset \cdots \subset \Omega$. By continuity from below, $F(x_n)$ converges to $\mathbb{P}(\Omega) = 1$.

3. **Change of Variables**

Let X be a continuous random variable with cdf F_X and pdf $f_X > 0$ everywhere, and let $Y = g(X)$, where g is a differentiable function.

a. Suppose that g is also invertible. Find the pdf of Y, f_Y, in terms of g and f_X.
b. Let $U \sim \text{Uniform}([0,1])$. Using the conclusion from part a, show that $F_X^{-1}(U)$ has the same distribution as X. (This allows us to generate a given random variable given only a uniform random number generator.)
c. Now suppose that $g(x) = x^2$. Find the pdf of Y in terms of the pdf of X. Also find the pdf of Y when X is a standard normal random variable in particular.
(Note that this g is not invertible, unlike in part a.)

Solution:

a. g is a continuous invertible function from $\mathbb{R}$ to $\mathbb{R}$, so g must be monotonic, i.e. strictly increasing or strictly decreasing. Let us first find the cdf of Y:

$$F_Y(y) = \mathbb{P}(g(X) \le y) = \begin{cases} \mathbb{P}(X \le g^{-1}(y)) = F_X(g^{-1}(y)) & \text{if } g \text{ is increasing} \\ \mathbb{P}(X \ge g^{-1}(y)) = 1 - F_X(g^{-1}(y)) & \text{if } g \text{ is decreasing.} \end{cases}$$

Then, by the chain rule of differentiation, we find the pdf of Y as

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = \left| \frac{\mathrm{d}}{\mathrm{d}y} F_X(g^{-1}(y)) \right| = f_X(g^{-1}(y)) \cdot \left| \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) \right|.$$

Using the inverse function rule, we can further simplify to

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \frac{1}{|g'(g^{-1}(y))|}.$$

b. Let $Y = F_X^{-1}(U)$. F_X is differentiable because X is a continuous random variable, and strictly increasing because $f_X > 0$ everywhere, so its inverse F_X^{-1} is also differentiable and monotonically increasing. Using the conclusion of part a with $g = F_X^{-1}$,

$$F_Y(y) = F_U(g^{-1}(y)) = F_U(F_X(y)) = F_X(y),$$

which shows that Y has the same distribution as X. Note that $F_U(u) = \mathbb{P}(U \le u) = u$ for $U \sim \text{Uniform}([0,1])$.

c. The cdf of $Y = X^2$ is

$$\mathbb{P}(X^2 \le y) = \mathbb{P}(-\sqrt{y} \le X \le \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} f_X(x) \, \mathrm{d}x.$$

By the fundamental theorem of calculus, the pdf of Y is

$$f_Y(y) = \frac{1}{2\sqrt{y}}(f_X(-\sqrt{y}) + f_X(\sqrt{y})).$$

For $X \sim \mathcal{N}(0,1)$, the pdf of X^2 evaluates to

$$f_Y(y) = \frac{1}{2\sqrt{y}} \left(\frac{1}{\sqrt{2\pi}} e^{-y/2} + \frac{1}{\sqrt{2\pi}} e^{-y/2} \right) = \frac{1}{\sqrt{2\pi y}} e^{-y/2}.$$

(This is known as the **chi-squared** distribution with 1 degree of freedom.)

4. **Gaussian Confidence Interval**

A $C\%$ **confidence interval** for a parameter θ is the interval containing θ of smallest length, such that θ falls in the interval with probability at least $C\%$.

Suppose that a given population has Gaussian distribution with unknown mean μ and variance σ^2. We draw n independent samples; let the average of the samples be $\bar{\mu}$.

a. Find a 95% confidence interval for μ.

b. Suppose $\sigma^2 = 1$. How many independent samples at minimum do we need to construct a 99% confidence interval for μ with length at most 1?

Solution:

a. A 95% confidence interval is given by $\bar{\mu} \pm 1.96 \cdot \sigma/\sqrt{n}$.

b. For $\sigma^2 = 1$, a 99% confidence interval is given by $\bar{\mu} \pm 2.58/\sqrt{n}$. We want $2.58/\sqrt{n} \leq 0.5$, or $n \geq 4 \cdot 2.58^2 \approx 26.6$, so the minimum number of samples we need is 27.

5. **Binomial with Random Parameter**

Let $U \sim \text{Uniform}([0,1])$, and suppose that X has distribution $\text{Binomial}(n, p)$ given that $U = p$. Find $\mathbb{E}(U^2 X)$ and $\mathbb{E}(U^2 X^2)$.

Hint: Rather than working directly with the definition of expectation, consider the properties of conditional expectation.

Solution: By the tower property,

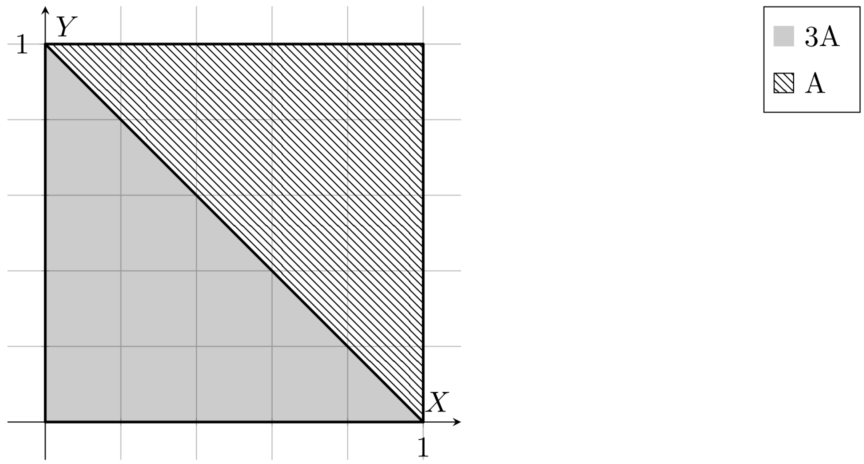
$$\mathbb{E}(U^2 X) = \mathbb{E}(U^2\,\mathbb{E}(X \mid U)) = \mathbb{E}(U^2 \cdot nU) = n\,\mathbb{E}(U^3) = \frac{n}{4}.$$

And, by the tower property again,

$$\begin{aligned}
\mathbb{E}(U^2 X^2) &= \mathbb{E}(U^2\,\mathbb{E}(X^2 \mid U)) \\
&= \mathbb{E}(U^2 \cdot (\text{var}(X \mid U) + \mathbb{E}(X \mid U)^2)) \\
&= \mathbb{E}(U^2 \cdot (nU(1-U) + n^2U^2)) \\
&= n\,\mathbb{E}(U^3) - n\,\mathbb{E}(U^4) + n^2\,\mathbb{E}(U^4) \\
&= \frac{4n^2 + n}{20}.
\end{aligned}$$

6. **Graphical Density**

The following figure depicts the joint density $f_{X,Y}$ of X and Y.



a. Are X and Y independent? Remember to justify your answer.
b. What is the value of A?
c. Compute $f_X(x)$.
d. Compute $\mathbb{E}(Y \mid X = x)$. You may leave your answer as a fraction of terms containing x, but you may not have an integral.
e. What is $\mathbb{E}(X - Y \mid X + Y)$?

Solution:

a. X and Y are not independent. For example, when $X = 0$, the expected value of Y is $\frac{1}{2}$, but when $X = \frac{1}{2}$, the expected value of Y is less than $\frac{1}{2}$, since there is more probability mass in the bottom triangle.
b. The total probability density must integrate to 1, so $\frac{1}{2}(3A) + \frac{1}{2}A = 1$ implies that $A = \frac{1}{2}$.
c. The vertical line at $X = x$ breaks up into two pieces in each triangle:

$$
\begin{aligned}
f_X(x) = \int_0^1 f_{X,Y}(x, y)\,\mathrm{d}y &= \int_0^{1-x} \frac{3}{2}\,\mathrm{d}y + \int_{1-x}^1 \frac{1}{2}\,\mathrm{d}y \\
&= \frac{3}{2}(1 - x) + \frac{1}{2}x = \frac{3}{2} - x.
\end{aligned}
$$

d. As $f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$, we can use the definition of conditional expectation directly:

$$
\begin{aligned}
\mathbb{E}(Y \mid X = x) &= \int_0^{1-x} y \frac{3A}{\frac{3}{2} - x}\,\mathrm{d}y + \int_{1-x}^1 y \frac{A}{\frac{3}{2} - x}\,\mathrm{d}y \\
&= \frac{3A(1 - x)^2}{3 - 2x} + \frac{A(1 - (1 - x)^2)}{3 - 2x} \\
&= \frac{3 - 4x + 2x^2}{2(3 - 2x)}.
\end{aligned}
$$

Alternate solution. We can split this expectation into the cases where Y falls in the $3A$ region (when it falls below $1-x$) and where Y falls in the A region. In the first case, its expectation will be $\frac{1-x}{2}$; in the second case, its expectation will be $\frac{1+1-x}{2} = \frac{2-x}{2}$. It remains to figure out the probability that Y falls below $1-x$. Let B be the (constant) density $f_{Y|X}(y \mid x)$ for y above $1-x$, so that $3B$ is the density of Y below $1-x$. In order to integrate to 1, we must have

$$3B(1-x) + Bx = 1,$$

which implies that $B = \frac{1}{3-2x}$. Then we have

$$\mathbb{E}(Y \mid X = x) = \frac{1-x}{2} \cdot \frac{3(1-x)}{3-2x} + \frac{2-x}{2} \cdot \frac{x}{3-2x},$$

which yields the same result after simplifying.

e. We see that given $X + Y = c$, which is a line parallel to the diagonal line in the graph, the values of $X - Y$ (which is perpendicular to $X + Y$) are uniformly distributed, centered around 0. So $\mathbb{E}(X - Y \mid X + Y) = 0$. Another way to see this is that $\mathbb{E}(X \mid X + Y) = \mathbb{E}(Y \mid X + Y) = \frac{X+Y}{2}$, so by linearity of expectation, $\mathbb{E}(X - Y \mid X + Y) = 0$.

7. **Joint Density for Exponential Distribution**

   a. If $X \sim \text{Exponential}(\lambda)$ and $Y \sim \text{Exponential}(\mu)$ are independent, compute $\mathbb{P}(X < Y)$.

   b. If $X_1, \ldots, X_n$ are independent and Exponentially distributed with parameters $\lambda_1, \ldots, \lambda_n$, show that $\min_{1 \le k \le n} X_k \sim \text{Exponential}(\sum_{j=1}^n \lambda_j)$.

   c. Deduce that
$$\mathbb{P}\left(X_i = \min_{1 \le k \le n} X_k\right) = \frac{\lambda_i}{\sum_{j=1}^n \lambda_j}.$$

Solution:

a. By the law of total probability,
$$\mathbb{P}(X < Y) = \int_0^\infty \mathbb{P}(X < y \mid Y = y) \cdot f_Y(y)\, \mathrm{d}y.$$
Since X and Y are independent, $\mathbb{P}(X < y \mid Y = y) = \mathbb{P}(X < y)$. Plugging in the known $\mathbb{P}(X < y) = 1 - e^{-\lambda y}$ and $f_Y(y) = \mu e^{-\mu y}$, we get
$$\mathbb{P}(X < Y) = \frac{\lambda}{\lambda + \mu}.$$

b. The ccdf of $X := \min_{1 \le k \le n} X_k$ is precisely the ccdf of an $\text{Exponential}(\sum_{j=1}^n \lambda_j)$:
$$\mathbb{P}(X \ge x) = \mathbb{P}(X_1 \ge x, \ldots, X_n \ge x) = \prod_{k=1}^n \mathbb{P}(X_k \ge x) = \prod_{k=1}^n e^{-\lambda_k x} = e^{-x \sum_{k=1}^n \lambda_k}.$$

c. Now, we observe that
$$\mathbb{P}\left(X_i = \min_{1 \le k \le n} X_k\right) = \mathbb{P}\left(X_i \le \min_{k \ne i} X_k\right).$$
By part b, $\min_{k \ne i} X_k \sim \text{Exponential}(\sum_{j \ne i} \lambda_j)$. Then, by part a, the claim follows.