Privacy in Statistics and Machine Learning — Spring 2025
In-class Exercises for Lecture 1 (Intro and Randomized Response)
January 21, 2023

Adam Smith (based on materials developed with Jonathan Ullman)

Problems with marked with an asterisk (*) are more challenging or open-ended.

1. Suppose $X_1, \ldots, X_n$ are independent random valriables, each with mean $\mathbb{E}(X_i) = \mu$ and standard deviation $\sigma = \sqrt{\text{Var}(X_i)}$ (for all i).

   What are the expectation and variance of the average $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$?

   (a) $\mathbb{E}(\bar{X}) = \mu n$ and $\sqrt{\text{Var}(\bar{X})} = n\sigma$

   (b) $\mathbb{E}(\bar{X}) = \mu$ and $\sqrt{\text{Var}(\bar{X})} = \sigma$

   (c) $\mathbb{E}(\bar{X}) = \mu$ and $\sqrt{\text{Var}(\bar{X})} = \sigma/\sqrt{n}$

   (d) $\mathbb{E}(\bar{X}) = \mu$ and $\sqrt{\text{Var}(\bar{X})} = \sigma/n$

   (e) $\mathbb{E}(\bar{X}) = \mu/n$ and $\sqrt{\text{Var}(\bar{X})} = \sigma/n$

2. Recall the randomized response mechanism discussed in class. For each input bit x_i, it generates

$$Y_i = \begin{cases} x_i & \text{w.p. } 2/3\,, \\ 1 - x_i & \text{w.p. } 1/3\,. \end{cases}$$

   Give a procedure that, given the outputs $Y_1, \ldots, Y_n$ from randomized response on input $x_1, \ldots, x_n$, returns an estiamte A such that

$$\sqrt{\mathbb{E}\left(\left(A - \sum_{i=1}^{n} x_i\right)^2\right)} = O(\sqrt{n})$$

   Hint: Find a function f that rescales the Y_i so that $\mathbb{E}(f(Y_i)) = x_i$.

3. (*) Suppose now that for each respondent we have some "public" information (that is, known to the analyst), together with the private bit x_i. We might be interested in solving some task that involves both the public and private information, such as finding a model to predict x_i given the public features.

   Consider the following super-simplified version of this: suppose the public information for person i is a real number $a_i \in \mathbb{R}$. Given the Y_i output by randomized response, how can we get an unbiased estimate of $\sum_i a_i x_i$? What is its variance (as a function of the list of a_i's)?

4. (*) Consider the second randomized response mechanism described in class, in which $Y_i = \begin{cases} x_{)} & \text{w.p. } \frac{e^\varepsilon}{e^\varepsilon+1}, \\ 1 - x_i & \text{w.p. } \frac{1}{e^\varepsilon+1}. \end{cases}$ Give a procedure that, given the outputs $Y_1, ..., Y_n$ from randomized response on input $x_1, ..., x_n$, returns an estimate A such that $\sqrt{\mathbb{E}\left(\left(A - \sum_{i=1}^n x_i\right)^2\right)} = \frac{e^{\varepsilon/2}}{e^\varepsilon - 1}\sqrt{n}$.

   Hint: Find a function f that rescales the Y_i so that $\mathbb{E}(f(Y_i)) = x_i$.

Reminders on sums of random variables A good reference on the probability material needed for this class is the book of Mitzenmacher and Upfal [MU17]. We include here a few reminders that will be useful in today's lecture.

- Expectations are linear: If X, Y are random variables (it does *not* matter if they are independent), then for any constants $a, b \in \mathbb{R}$, we have

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

  By induction, linearity extends to finite sums (so $\mathbb{E}(a_1X_1 + \cdots + a_kX_k) = a_1\mathbb{E}(X_1) + \cdots + a_k\mathbb{E}(X_k)$.

- Variances add when random variables are independent: For any *independent* random variables X, Y, and for any constants $a, b \in R$, we have

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y).$$

  Again, by induction, if $X_1, ..., X_k$ are independent, then $\text{Var}\left(\sum_{i=1}^k a_iX_i\right) = \sum_i = 1^k a_i^2 \text{Var}(X_i)$. Note that variances do not necessarily add for *dependent* random variables. For example, if $Y = -X$, what is the variance of $X + Y$?

- Chebyshev's inequality: For any random variable X with finite mean and variance, for every $t > 0$, we have

$$\mathbb{P}\left(|X - \mathbb{E}(X)| \geq t\sqrt{\text{Var}(X)}\right) \leq 1/t^2.$$

- "Chernoff bounds" are a family of concentration inequalities for sums of independent random variables. A useful example is the following:

  Lemma 0.1. *Let $X_1, ..., X_n$ be i.i.d. random variables taking values in $[0, 1]$. Let X denote their sum and let $\mu = \mathbb{E}(X_i)$ (so that $\mathbb{E}(X) = \mu n$. Then,*

  - *For every $\delta \geq 0$, $\mathbb{P}(X > (1 + \delta)\mu n) \leq e^{-\delta^2 \mu n/3}$*
  - *For every $\delta \in [0, 1]$, $\mathbb{P}(X < (1 - \delta)\mu n) \leq e^{-\delta^2 \mu n/2}$.*

  In particular, for every $t > 0$, the probability that $|X - \mu n| \geq t\sqrt{n}$ is at most $2\exp(-t^2/3)$.

References

[MU17] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis.* Cambridge University Press, 2nd edition, 2017.