CS7880: Rigorous Approaches to Data Privacy, Spring 2017
Probability Review

Instructor: Jonathan Ullman

Optional.

This review sheet is to make sure that you are familiar with the kinds of probability theory that we're going to use in this class. It is for your benefit, and you are not required to write down answers and submit. These problem are not a "prerequisite" for the course—I will cover some of the finer points of this material when we need it in class—but if these questions look extremely daunting please let me know so I can help you catch up.

Problem 1 (Basic Probability Facts)**.**

(a) Let E_1, E_2 be two events. Prove the "union bound," which states $\mathbb{P}[E_1 \cup E_2] \leq \mathbb{P}[E_1] + \mathbb{P}[E_2]$. Prove that if $E_1, \ldots, E_n$ are n events, then $\mathbb{P}\left[\bigcup_{i=1}^n E_i\right] \leq \sum_{i=1}^n \mathbb{P}[E_i]$.

(b) Let X be a random variable and $\mathrm{Var}[X]$ be its variance. Prove that $\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$.

(c) Prove that if X_1, X_2 are independent, then $\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1]\mathbb{E}[X_2]$. Give an example of X_1, X_2 that are *not* independent for which this identity is false.

(d) Let $X_1, \ldots, X_n$ be independent $\{0,1\}$-valued random variables where $\mathbb{P}[X_i = 1] = p$ (i.e. coin tosses where the probability of heads is p). Let $X = \sum_{i=1}^n X_i$. What are $\mathbb{E}[X]$, $\mathbb{E}\left[X^2\right]$, and $\mathrm{Var}[X]$?

Problem 2 (Tail Bounds)**.**

(a) Prove "Markov's Inequality:" If X is a non-negative random variable, then

$$\forall t > 0 \quad \mathbb{P}[X > t] \leq \frac{\mathbb{E}[X]}{t}.$$

(b) Prove "Chebyshev's Inequality:" If X is any random variable, then

$$\forall t > 0 \quad \mathbb{P}[|X - \mathbb{E}[X]| > t] \leq \frac{\mathrm{Var}[X]}{t^2}.$$

(c) Prove the "Chernoff Bound:"[1] If $X_1, \ldots, X_n$ are independent $[0,1]$-valued random variables, and $X = \sum_{i=1}^n X_i$, then

$$\forall t > 0 \quad \mathbb{P}[X - \mathbb{E}[X] > t] \leq \exp\left(\frac{-t^2}{4n}\right).$$

Hint: Start by bounding $\mathbb{E}\left[e^{\lambda X}\right]$ *for arbitrary* $0 \leq \lambda \leq 1/2$ *and apply Markov's Inequality.*

Problem 3 (Useful Distributions)**.**

(a) Let $X_1, \ldots, X_n$ be independent coin tosses with $\mathbb{P}[X_i = 1] = p$. Let $X = \frac{1}{n}\sum_{i=1}^n X_i$. Prove that for every $\varepsilon, \delta > 0$, if $n \geq c \log(1/\delta)/\varepsilon^2$ (for some suitable $c > 0$), then

$$\mathbb{P}[|X - p| \leq \varepsilon] \geq 1 - \delta.$$

In other words, X is a good estimate of its mean p.

(b) Suppose we do the same experiment m times in parallel. That is, we have m independent random variables $Y^1, \ldots, Y^m$ where for every j, Y^j is an average of n independent coin tosses with probability p^j. Prove that for every $\varepsilon, \delta > 0$, if $n \geq c \log(m/\delta)/\varepsilon^2$ then

$$\mathbb{P}\left[\forall j = 1, \ldots, m \quad |Y^j - p^j| \leq \varepsilon\right] \geq 1 - \delta.$$

In other words, the Y^j's are all good estimates of their means p^j.

[1] This is a misnomer. It's really a case of Bernstein's Inequality from about 40 years prior. This has been a public service announcement.

(c) In the setup above, prove that if $p = 1/2$, and $n = o(1/\varepsilon^2)$, then

$$\mathbb{P}[|X - 1/2| > \varepsilon] \geq \Omega(1).$$

In other words, X is not *too* good of an estimate of the mean p. *Hint: One way to prove this is with Stirling's Approximation. There are probably others.*

(d) Consider the Gaussian distribution $N(0, \sigma^2)$ with probability density function

$$\rho(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-x^2}{2\sigma^2}\right).$$

Prove that if we sample m independent random variables $g^1, \ldots, g^m$ from this distribution, then for some $c > 0$,

$$\mathbb{P}\left[\forall j = 1, \ldots, m \quad |g^j| \leq c\sigma\sqrt{\log(m/\delta)}\right] \leq \delta.$$

(e) Consider the Laplace distribution $\mathrm{Lap}(\sigma)$ with probability density function

$$\rho(x) = \frac{1}{2\sigma} \exp\left(\frac{-|x|}{\sigma}\right).$$

Prove that if we sample m independent random variables $\ell^1, \ldots, \ell^m$ from this distribution, then for some $c > 0$,

$$\mathbb{P}\left[\forall j = 1, \ldots, m \quad |\ell^j| \leq c\sigma\log(m/\delta)\right] \leq \delta.$$