CS 355: Topics in Cryptography **Spring 2020**

Probability-Theory Cheat Sheet

Common Notation

- For a finite set S, $x \xleftarrow{\mathbb{R}} S$ denotes an element chosen uniformly at random from S. (More formally, x is a random variable taking each value $a \in S$ with probability $1/|S|$.)
- For a positive integer n, the expression $[n]$ denotes the set $\{1, 2, \dots, n\}$.
- When we will want to be explicit about the probability distribution, we will often use either the notation $\mathbb{E}[x^2 :\ x \xleftarrow{\mathbb{R}} \{-1,1\}] = 1$ or $\mathbb{E}_{x \xleftarrow{\mathbb{R}} \{-1,1\}}[x^2] = 1$.
- For a probability event A, the *indicator variable* I_A is a random variable that takes the value 1 when event A occurs, and 0 otherwise.

Linearity of expectation. let X and Y be random variables taking values in $\mathbb{R}$, and let $a, b \in \mathbb{R}$ be constants. Then

$$\mathbb{E}[aX + bY] = a \cdot \mathbb{E}[X] + b\mathbb{E}[Y]\,.$$

Union Bound. For events $E_1, \dots, E_n$,

$$\Pr[E_1 \cup \cdots \cup E_n] \le \Pr[E_1] + \cdots + \Pr[E_n]\,.$$

In many situation, we will be interested in bounding the probability that a random variable deviates from its expectation.

Markov's inequality. Let X be a non-negative random variable and $a > 0$ be a constant. Then

$$\Pr[X \ge a] \le \frac{\mathbb{E}[X]}{a}\,.$$

Chebyshev's inequality. Let X be a random variable and let $\delta > 0$ be a constant. Then

$$\Pr\left[|X - \mathbb{E}[X]| \ge \delta\right] \le \frac{\mathrm{Var}[X]}{\delta^2}\,.$$

Chernoff Bound. Let $X_1, \dots, X_n$ be independent random variables taking values in $[0,1]$, and let $X = \sum_{i \in [n]} X_i$. Then

$$\forall 0 < \epsilon \le 1, \qquad \Pr[X \le (1-\epsilon)\,\mathbb{E}[X]] \le e^{-\frac{\epsilon^2 \mathbb{E}[X]}{2}} \qquad \text{and}$$

$$\forall 0 < \epsilon, \qquad \Pr[X \ge (1+\epsilon)\,\mathbb{E}[X]] \le e^{-\frac{\epsilon^2 \mathbb{E}[X]}{2+\epsilon}}\,.$$

Note that the Chernoff bound holds even if the X_i take real values in the interval $[0,1]$, rather than just integer values in $\{0,1\}$. Moreover, the X_i don't have to be identically distributed, but they *crucially* need to be mutually independent.

Example. Consider throwing n balls independently and uniformly into n bins.

- *What is the expected number of empty bins?*

  For $i \in [n]$, let Z_i be and indicator variable of the event that bin i is empty. It holds

  $$\Pr[Z_i = 1] = \left(1 - \tfrac{1}{n}\right)^n \approx \tfrac{1}{e}.$$

  (More formally, we can use that for $n > 1$, $|x| \le n$, it holds that $e^x \le \left(1 + \frac{x}{n}\right)^n \le e^x \left(1 - \frac{x^2}{n}\right)$).

  We can express the number of empty bins Z as

  $$Z = \sum_{i \in [n]} Z_i,$$

  and, by linearity of expectation, the expected number of empty bins is:

  $$\mathbb{E}[Z] = \mathbb{E}\left[\sum_{i \in [n]} Z_i\right] \approx \tfrac{n}{e}.$$

- *How concentrated is the number of empty bins Z?*

  Ideally, we would want to show that Z is tightly concentrated around its expectation n/e.

  The first attempt could be to use Markov's inequality. However, this only gives us a very weak bound, for example:

  $$\Pr[Z \ge 0.75n] \le \Pr[Z \ge \tfrac{2n}{e}] \le \frac{\mathbb{E}[Z]}{2n/e} \approx 1/2.$$

  Next, we will use Chebyshev's inequality to get a better bound. To this end, let's compute the variance of Z:

  $$\mathrm{Var}[Z] = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2.$$

  To compute the first term,

  $$\mathbb{E}[Z^2] = \mathbb{E}\left[\sum_{i,j \in [n]} Z_i \cdot Z_j\right] = \sum_{i \ne j} E[Z_i Z_j] + \sum_{i \in [n]} E[Z_i^2].$$

  Now note that $Z_i Z_j = 1$ when all n balls miss both bin i and j, which happens with probability $(1 - 2/n)^n$. Since $Z_i \in \{0, 1\}$, it also holds that $E[Z_i^2] = E[Z_i] = (1 - 1/n)^n$. Therefore

  $$\mathbb{E}[Z^2] = n(n-1)(1 - \tfrac{2}{n})^n + n(1 - \tfrac{1}{n})^n, \qquad \text{and} \qquad \mathbb{E}[Z]^2 = n^2 \left(1 - \tfrac{1}{n}\right)^{2n}.$$

  With a little bit of work, one can then show that $\mathrm{Var}[Z] = O(n)$. Therefore, by Chebyshev's inequality,

  $$\Pr[\left|Z - \tfrac{n}{e}\right| \ge \epsilon n] \le \frac{\mathrm{Var}[Z]}{\epsilon^2 n^2} = O(1/n),$$

  which is a much tighter bound that we have obtained before from Markov's inequality.

- *What it the maximum load amongst all bins?*

  We will show that with high probability, no bin has load higher than $O(\ln n)$ (in fact, one can show a tighter bound of $O(\ln n/\ln\ln n)$).

  For $i, j \in [n]$ let X_{ij} be an indicator variable of ball i landing in bin j. Then, we can write the load of bin j as

  $$L_j = \sum_{i\in[n]} X_{ij}\,,$$

  and it holds

  $$\mathbb{E}[L_j] = \sum_{i\in[n]} \mathbb{E}[X_{ij}] = n\cdot \Pr[X_{ij}=1] = n\cdot \tfrac{1}{n} = 1\,.$$

  Now note that, for every j and every $i \neq i'$, the random variables X_{ij} and $X_{i'j}$ are independent random variables taking values in $\{0,1\}$. Therefore, we can use the Chernoff bound, with $\mathbb{E}[L_j] = 1$ and $\epsilon = 2\ln n + 2$, to claim that

  $$\begin{aligned}\Pr[L_j \geq (2\ln n+3)] &= \Pr[L_j \geq (1+(2\ln n+2))\,\mathbb{E}[L_j]] \leq \exp\left(-\frac{(2\ln n+2)^2}{2+2\ln n+2}\right)\\ &\leq \exp\left(-\frac{4\ln^2 n+8\ln n}{2\ln n+4}\right) \leq \exp(-2\ln n) = \frac{1}{n^2}\,.\end{aligned}$$

  Therefore, by the union bound, it holds that

  $$\Pr[\max_{j\in[n]} L_j \geq (2\ln n+3)] \leq \sum_{j\in[n]} \Pr[L_j \geq (2\ln n+3)] \leq n\cdot\frac{1}{n^2} = \frac{1}{n}\,.$$