CS 7880: Algorithms for Big Data (Fall'22) **Northeastern University**

Lecture 2

September 13, 2022

Instructor: Soheil Behnezhad *Scribe: Erika Melder*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

1 Overview

Most big data algorithms rely on randomization. In this lecture, we will see some basic probabilistic tools that are helpful in the analysis of such randomized algorithms. In particular, we will discuss linearity of expectation, Markov's inequality, Chebyshev's inequality, and the Chernoff bound.

2 Probabilistic Tools

2.1 The Balls and Bins Problem

The following problem is an excellent example to demonstrate the various power levels of probabilistic bounds:

Problem 1 (Balls and Bins)**.**

Input: We have n balls and n bins. Each ball is independently placed into one bin uniformly randomly.

Goal: Define the *max load* to be $L = \max_{b \in \text{Bins}}[\#\text{ balls in } b]$. Find a function $f(n)$ such that, with probability at least $1 - \frac{1}{n}$, $L \leq f(n)$.

From the structure of the problem, it is possible with vanishingly small probability for the max load to be n. To be precise, this probability is $\frac{n}{n^n} = \frac{1}{n^{n-1}}$. This is why we seek a function which upper bounds L with probability $1 - \frac{1}{n}$, which is still very high but eliminates this outlier possibility from contention; it describes what happens the majority of the time. To find such a function, we will introduce various probabilistic tools which bound the probability of error.

2.2 Linearity of Expectation

The following theorem relates the expected values of two random variables:

Theorem 1 (Linearity of Expectation)**.** *For any two random variables X and Y, not necessarily independent,* $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

Proof. We will prove the theorem for discrete random variables. The expected value of $X + Y$ is given by

$$\begin{aligned}
\mathbb{E}[X + Y] &= \sum_{x \in X} \sum_{y \in Y} (x + y) \Pr[X = x, Y = y] \\
&= \sum_{x \in X} \sum_{y \in Y} x \Pr[X = x, Y = y] + \sum_{x \in X} \sum_{y \in Y} y \Pr[X = x, Y = y] \\
&= \sum_{x \in X} x \sum_{y \in Y} \Pr[X = x, Y = y] + \sum_{y \in Y} y \sum_{x \in X} \Pr[X = x, Y = y]
\end{aligned}$$

Observe that

$$\sum_{y \in Y} \Pr[X = x, Y = y] = \Pr[X = x]$$

By symmetry this holds for $\sum_{x \in X} \Pr[X = x, Y = y]$ as well. Then the sum collapses to

$$\mathbb{E}[X + Y] = \sum_{x \in X} x \Pr[X = x] + \sum_{y \in Y} y \Pr[Y = y] = \mathbb{E}[X] + \mathbb{E}[Y]. \quad \square$$

Remark. One may prove by induction that linearity holds for any linear combination of random variables (again, they may not necessarily be independent).

In the Balls and Bins problem, we may apply this to compute the expectation of a certain bin. Without loss of generality, we will focus on bin 1. Define L_1 to be the number of balls in bin 1 after all of them have been placed. Then define indicator functions

$$B_i := \mathbb{1}[\text{Ball } i \text{ is placed into bin 1}]$$

Observe that

$$\mathbb{E}[L_1] = \mathbb{E}\left[\sum_{i=1}^{n} B_i\right] = \sum_{i=1}^{n} \mathbb{E}[B_i]$$

by linearity of expectation. Since each ball is placed into a bin uniformly randomly, it follows that $\mathbb{E}[B_i] = \frac{1}{n}$ for all i. Substitution then yields

$$\mathbb{E}[L_1] = \sum_{i=1}^{n} \frac{1}{n} = 1$$

2.3 Concentration Bounds and Markov's Inequality

Note that the expected value of L_1 does not tell us anything about the probability of L_1 being in a certain range. In order to fully bound the max load of the problem, we must use a *concentration bound*. A concentration bound attempts to place a bound on the probability that a value differs from the expected value μ by more than a specified threshold r. A visual is given in Figure 1.

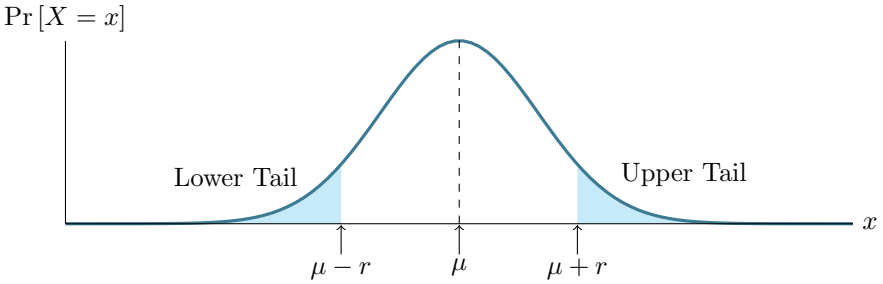

Figure 1: The two tails of a concentration bound, where $\mu = \mathbb{E}[X]$. In the Balls and Bins problem, we seek an upper bound on the size of the upper tail.

One simple form of concentration bound is given by Markov's inequality:

Theorem 2 (Markov's Inequality)**.** *For any nonnegative random variable X, for all $t > 0$,*

$$\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$$

Proof. Note that all outcomes satisfy exactly one of $X \geq t$ and $X < t$. As a result,

$$\mathbb{E}[X] = \mathbb{E}[X \mid X \geq t] \Pr[X \geq t] + \mathbb{E}[X \mid X < t] \Pr[X < t]$$

We know $\mathbb{E}[X \mid X \geq t] \geq t$. Also, because X is nonnegative, we have that $\mathbb{E}[X \mid X < t] > 0$. Then

$$\mathbb{E}[X \mid X \geq t] \Pr[X \geq t] + \mathbb{E}[X \mid X < t] \Pr[X < t] \geq t \Pr[X \geq t] + 0 \cdot \Pr[X < t] = t \Pr[X \geq t]$$

Solving for $\Pr[X \geq t]$ yields Markov's inequality. □

We may apply this to the Balls and Bins problem using the expected value we computed. This yields

$$\Pr[L_1 \geq n] \leq \frac{1}{n}$$

which, while true, is not a tight bound and not particularly useful.

2.4 Chebyshev's Inequality

Another statistical measure of a distribution aside from its expected value is its *variance*.

Definition 3. Let X be a random variable. The *variance* of X is given by

$$\mathrm{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

or the average of the squared distance of X from its own expected value.

Remark. Unlike expected value, the variance is *not* generally linear. However, it is linear for certain cases. For example if X is a sum of n *pairwise independent* binary random variables $X_1, \ldots, X_n$, meaning that for all $i, j \in [n]$,

$$i \neq j \Rightarrow \Pr[X_i \cap X_j] = \Pr[X_i] \Pr[X_j],$$

then the variance is linear. That is,

$$\mathrm{Var}[X] = \mathrm{Var}[X_1] + \ldots + \mathrm{Var}[X_n].$$

Note that this is a weaker condition than mutual independence, which requires that the above hold for any subset of X_i's rather than just any pair.

An alternative expression is often used to compute the variance more easily.

Proposition 4. $\mathrm{Var}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$.

Proof. We have

$$\begin{aligned}\mathrm{Var}[X] &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2 - 2X\,\mathbb{E}[X] + \mathbb{E}[X]^2\right] = \mathbb{E}\left[X^2\right] - 2\,\mathbb{E}[X\,\mathbb{E}[X]] + \mathbb{E}\left[\mathbb{E}[X]^2\right] \\ &= \mathbb{E}\left[X^2\right] - 2\,\mathbb{E}[X]^2 + \mathbb{E}[X]^2 = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2. \quad \square\end{aligned}$$

Chebyshev's inequality bounds the probability of deviating from the expected value as a function of the variance.

Theorem 5 (Chebyshev's Inequality)**.** *For any random variable X, for all $t > 0$,*

$$\Pr\left[|X - \mathbb{E}\left[X\right]| \geq t\right] \leq \frac{\mathrm{Var}\left[X\right]}{t^2}$$

Proof. Define $Y = X - \mathbb{E}\left[X\right]$. Observe that

$$|X - \mathbb{E}\left[X\right]| \geq t \;\Leftrightarrow\; Y^2 \geq t^2.$$

Thus by Markov's inequality,

$$\Pr\left[Y^2 \geq t^2\right] \leq \frac{\mathbb{E}\left[Y^2\right]}{t^2} = \frac{\mathrm{Var}\left[X\right]}{t^2}. \quad \square$$

We may apply this to Balls and Bins. First, note that

$$\mathrm{Var}\left[L_1\right] = \mathrm{Var}\left[B_1 + \cdots + B_n\right]$$

Because the B_i are pairwise independent, we may treat the variance as linear and obtain

$$\mathrm{Var}\left[L_1\right] = \sum_{i=1}^{n} \mathrm{Var}\left[B_i\right]$$

Using the simplified expression for variance, note that

$$\mathrm{Var}\left[B_i\right] = \mathbb{E}\left[B_i^2\right] - \mathbb{E}\left[B_i\right]^2 \leq \mathbb{E}\left[B_i^2\right]$$

Since B_i is a zero-one indicator function, $B_i^2 = B_i$ always. Then

$$\mathrm{Var}\left[B_i\right] \leq \mathbb{E}\left[B_i\right] = \frac{1}{n}$$

This then yields

$$\mathrm{Var}\left[L_1\right] \leq \sum_{i=1}^{n} \frac{1}{n} = 1$$

Finally, we may apply Chebyshev's inequality to obtain

$$\Pr\left[|L_1 - 1| \geq \sqrt{n}\right] \leq \frac{1}{n}$$

meaning that with probability at least $1 - \frac{1}{n}$, $L_1 \leq \sqrt{n} + 1$.

2.5 Chernoff Bound

The Chernoff bound independently produces bounds on each tail of the distribution, provided that the random variable can be decomposed into independent random variables.

Theorem 6 (Chernoff Bound)**.** *Let $X_1, \ldots, X_n$ be independent random variables in $[0, 1]$. Let $X = \sum_{i=1}^{n} X_i$, and let $\mu = \mathbb{E}\left[X\right]$. Then:*

- **Upper tail:** $\Pr\left[X \geq (1+\delta)\mu\right] \leq \exp\left(-\frac{\delta^2}{2+\delta}\mu\right), \forall \delta \geq 0.$
- **Lower tail:** $\Pr\left[X \leq (1-\delta)\mu\right] \leq \exp\left(-\frac{\delta^2}{2}\mu\right), \forall \delta \in [0, 1].$

The two tail bounds of Theorem 6 can be combined to imply the following (slightly weaker) two-sided tail bound:

- **Two-sided bound:** $\Pr[|X-\mu| \geq \delta\mu] \leq 2\exp\left(-\frac{\delta^2\mu}{3}\right), \forall \delta \in [0,1]$.

Moreover, sometimes it is more convenient to work with the additive form of the Chernoff bound. Namely, let $t = \delta\mu$ with $t \leq \mu$ (to ensure $\delta \leq 1$). We have

- **Additive Chernoff bound:** $\Pr[|X-\mu| \geq t] \leq 2\exp\left(-\frac{t^2}{3\mu}\right), \forall t \in [0,\mu]$.

In the Balls and Bins problem, we may directly apply the Chernoff bound and solve for a value of δ which achieves our desired bound. In this case, we seek to bound the upper tail by $\frac{1}{n}$, giving us

$$\Pr[L_1 \geq (1+\delta)\cdot 1] \leq \exp(-\frac{\delta^2}{2+\delta}) \leq \frac{1}{n}$$

$$\Rightarrow \frac{\delta^2}{2+\delta} \geq \ln n$$

Choose a suitable value of δ, such as $\delta = 10\ln n$. Then

$$\Pr[L_1 \geq (1+10\ln n)] \leq \exp(\frac{-100\ln^2 n}{2+10\ln n}) \leq \exp(-\frac{100\ln^2 n}{20\ln^2 n}) \leq \exp(-5\ln n) = n^{-5}$$

This means $L_1 \leq 10\ln n + 1 = O(\log n)$ with probability at least $1 - \frac{1}{n^5}$.

2.6 Union Bound

Theorem 7 (Union Bound)**.** *For any two events E_1, E_2,*

$$\Pr[E_1 \cup E_2] \leq \Pr[E_1] + \Pr[E_2]$$

Proof. Note that $\Pr[E_1 \cup E_2] = \Pr[E_1] + \Pr[E_2] - \Pr[E_1 \cap E_2]$. The result follows. □

> **Remark.** One may prove the union bound for arbitrarily large finite collections of events using induction.

Typically, we use the union bound to upper bound the probabilities of undesirable events. We may apply it in this context to the Balls and Bins problem. Define events

$$E_i := \text{ Bin } i \text{ has more than } 10\ln n + 1 \text{ balls in it}$$

Observe that $\bigcup_{i=1}^n E_i$ is the event that some bin obtains more than $10\ln n + 1$ balls. Then by the union bound,

$$\Pr\left[\bigcup_{i=1}^n E_i\right] \leq \sum_{i=1}^n \Pr[E_i] \leq \sum_{i=1}^n \frac{1}{n^5} = \frac{1}{n^4}$$

Our final result is that the max load is $O(\lg n)$ with probability at least $1 - \frac{1}{n^4}$.

Exercise: The optimal bound on the max load is $O(\frac{\lg n}{\lg\lg n})$. Prove that this bound is correct.