

## Lecture 2

Jan 15, 2026

*Instructor: Soheil Behnezhad**Scribe: Soheil Behnezhad, Erika Melder*

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

## 1 Overview

Most modern big-data algorithms rely on randomization. This lecture presents several fundamental probabilistic tools used in the analysis of randomized algorithms, including linearity of expectation, Markov's inequality, Chebyshev's inequality, and the Chernoff bound.

## 2 Probabilistic Tools

### 2.1 The Balls and Bins Problem

Consider the following balls and bins problem.

**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, the maximum load can equal  $n$ , but only with vanishingly small probability. Specifically, this probability is  $\frac{n}{n^n} = \frac{1}{n^{n-1}}$ . Consequently, we seek a function that upper-bounds  $L$  with probability  $1 - \frac{1}{n}$ . This guarantee remains very strong while excluding this extreme outlier from consideration; it characterizes the behavior that occurs with very high probability. To obtain such a bound, we will introduce several probabilistic tools for controlling the probability of error.

### 2.2 Linearity of Expectation

The following theorem relates the expected values of two random variables (recall that we used it in our first session too without proof):

**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

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

This matches our intuition: there are  $n$  balls distributed among the  $n$  bins uniformly. Due to symmetry, we don't expect any specific bin to receive more balls than the others on average, hence this average must be the same for all. Moreover, since sum of these average numbers for all  $n$  bins should equal the number of balls  $n$ , each bin must receive 1 ball in expectation.

## 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  $\mu$  by more than a specified threshold  $r$ . A visual is given in Figure 1.

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  $t > 0$ ,*

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

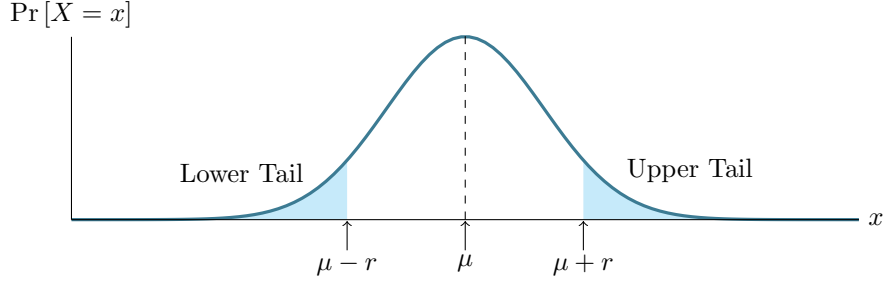


Figure 1: The two tails of a concentration bound, where  $\mu = \mathbb{E}[X]$ . In the Balls and Bins problem, we seek an upper bound on the size of the upper tail.

*Proof.* Let us assume, again for simplicity, that  $X$  takes non-negative integer values. We have

$$\begin{aligned}
 \mathbb{E}[X] &= \sum_{i=0}^{\infty} i \cdot \Pr[X = i] \\
 &\geq \sum_{i=t}^{\infty} i \cdot \Pr[X = i] && \text{(By non-negativity of } X.) \\
 &\geq \sum_{i=t}^{\infty} t \cdot \Pr[X = i] \\
 &= t \sum_{i=t}^{\infty} \Pr[X = i] \\
 &= t \cdot \Pr[X \geq t].
 \end{aligned}$$

Moving the terms implies 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

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

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  $X_1, \dots, X_n$ , 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,

$$\text{Var}[X] = \text{Var}[X_1] + \dots + \text{Var}[X_n].$$

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.**  $\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ .

*Proof.* We have

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

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  $t > 0$ ,*

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

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

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

Thus by Markov's inequality,

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

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

$$\text{Var}[L_1] = \text{Var}[B_1 + \dots + B_n]$$

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

$$\text{Var}[L_1] = \sum_{i=1}^n \text{Var}[B_i]$$

Using the simplified expression for variance, note that

$$\text{Var}[B_i] = \mathbb{E}[B_i^2] - \mathbb{E}[B_i]^2 \leq \mathbb{E}[B_i^2]$$

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

$$\text{Var}[B_i] \leq \mathbb{E}[B_i] = \frac{1}{n}$$

This then yields

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

Finally, we may apply Chebyshev's inequality to obtain

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

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  $X_1, \dots, X_n$  be independent random variables in  $[0, 1]$ . Let  $X = \sum_{i=1}^n X_i$ , and let  $\mu = \mathbb{E}[X]$ . Then:*

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

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

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

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  $\delta$  which achieves our desired bound. In this case, we seek to bound the upper tail by  $\frac{1}{n}$ , giving us

$$\begin{aligned} \Pr[L_1 \geq (1 + \delta) \cdot 1] &\leq \exp\left(-\frac{\delta^2}{2 + \delta}\right) \leq \frac{1}{n} \\ \Rightarrow \frac{\delta^2}{2 + \delta} &\geq \ln n \end{aligned}$$

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

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

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.  $\square$

**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.