COMP 585 **Spring 2026**
Probabilistic Toolkit for Learning and Computing *Instructor: Maryam Aliakbarpour*
January 22, 2026 *Scriber: Maryam Aliakbarpour*

Lecture 4

Concentration of Random Variables

A fundamental phenomenon in probability theory is that while a single random experiment is unpredictable, the average of many independent experiments is remarkably stable. As we gather more data, the empirical average of our observations tends to "concentrate" around the true underlying expectation. This behavior allows us to make rigorous predictions about large-scale systems even when individual components are uncertain.

In this lecture, we explore this phenomena. We introduce the quantitative tools used to measure this concentration, known as *tail bounds*. We will use the problem of estimating a coin's bias as a running example to illustrate how these tools help us determine the amount of data needed to reach a specific level of confidence.

Running Example: Estimating Coin Bias

Suppose we have a coin with an unknown probability $p = \mathbf{Pr}[\text{head}]$. We wish to design an (ϵ, δ)-tester to determine if the coin is fair. Specifically, with probability at least $1 - \delta$:

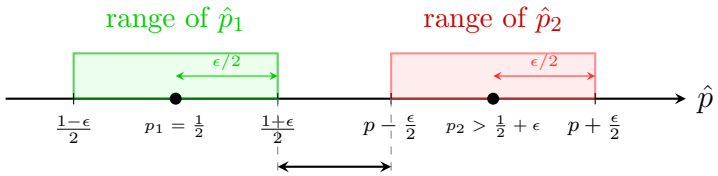
- If $p = \frac{1}{2}$, the algorithm outputs accept.
- If $|p - \frac{1}{2}| > \epsilon$, the algorithm outputs reject.

The algorithm is simple: flip the coin m times, and let X denote the number of observed heads. We compute the empirical mean $\hat{p} = \frac{X}{m}$ and return accept if $\left|\hat{p} - \frac{1}{2}\right| \leq t$, and reject otherwise. The central question is how to choose the parameters m and t so that the algorithm satisfies the desired guarantees for given δ and ϵ.

A reasonable choice for t is $\epsilon/2$. The reason is that if we can estimate p with error at most $|\hat{p} - p| \leq \epsilon/2$, then the two cases are fully separated. When $p_1 = \frac{1}{2}$, the empirical mean $\hat{p}_1$ lies in the interval

$$\left[\tfrac{1-\epsilon}{2},\ \tfrac{1+\epsilon}{2}\right],$$

and deviations within this range can be attributed to sampling noise. In contrast, when $|p_2 - \frac{1}{2}| > \epsilon$, the empirical mean is centered outside this interval. One can distinguish these two cases, by looking at where $\hat{p}$ lands and decide to accept or reject.



Separation exists because $p_2 > \frac{1}{2} + \epsilon$.

This reduces to understanding how well $\frac{X}{m}$ concentrates around p, and how large m must be to ensure this concentration.

Asymptotic Results

Before continuing, it is instructive to contrast concentration inequalities with two classical limit theorems: the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT). These results explain *why* empirical averages stabilize, while concentration inequalities quantify *how fast* this stabilization occurs for a finite number of samples.

Law of Large Numbers (LLN). Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ, and let

$$\overline{X}_m = \frac{1}{m} \sum_{i=1}^{m} X_i.$$

The (weak) Law of Large Numbers states that for every $\epsilon > 0$,

$$\mathbf{Pr}\big[|\overline{X}_m - \mu| > \epsilon\big] \xrightarrow[m\to\infty]{} 0.$$

This theorem establishes *convergence in probability* of the empirical mean to the true mean. However, it does not provide a quantitative bound on how large m must be to achieve a given confidence level δ, nor does it describe the rate at which the probability decays.

The Law of Large Numbers (LLN)

The Law of Large Numbers is the most basic form of concentration phenomenon. It describes the behavior of the sample average of a large number of independent and identically distributed (i.i.d.) random variables. It provides the mathematical justification for the "average" outcome in long-term experiments.

The *Weak Law of Large Numbers* states that the sample average converges in **probability** to the expected value. Let $X_1, X_2, \ldots, X_m$ be i.i.d. random variables with $\mu := \mathbf{E}[X_i]$. Let $\bar{X}_m$ denote the empirical mean of the samples $\frac{1}{m}\sum_{i=1}^{m} X_i$. For any $\epsilon > 0$:

$$\lim_{m\to\infty} \mathbf{Pr}\big[|\bar{X}_m - \mu| > \epsilon\big] = 0\,.$$

That is, for a sufficiently large number of samples m, the probability that the average $\bar{X}_m$

is far from μ is arbitrarily small.

The *Strong Law of Large Numbers* is a more powerful statement, asserting that the sample average converges **almost surely** to the expected value. Under the same conditions as above:

$$\mathbf{Pr}\Big[\lim_{m\to\infty} \bar{X}_m = \mu\Big] = 1\,.$$

This means that as the number of samples m goes to infinity, the sequence of sample averages will converge to μ with probability 1.

The Weak Law of Large Numbers guarantees that for any small $\epsilon, \delta > 0$ we can pick a specific m_0, for which the probability that your average of $m \geq m_0$ samples is ϵ-far from the mean is at most δ; however, it stays silent on whether that average might wander away again if you were to keep sampling. In contrast, the Strong Law of Large Numbers makes a much more powerful claim about the entire history of the process, asserting that with probability 1, the running average will eventually enter a small window around the mean and stay there for all eternity. While the Weak Law says that "failures" are unlikely at any given large m, the Strong Law ensures that the total number of times the average deviates significantly from the mean is finite, meaning that in the infinite limit, the long-run behavior of the average becomes locked near the mean.

Central Limit Theorem (CLT). Under mild conditions (e.g. finite variance $\sigma^2 = \mathbf{Var}[X]$), the CLT refines this picture by describing the *distributional* behavior of the error. CLT states that if you take sufficiently large random samples from any population (regardless of its distribution), the distribution of the sample means will follow a normal distribution. More formally, we have

$$\sqrt{m}\,(\overline{X}_m - \mu) \xrightarrow{\text{in distribution}} \mathcal{N}(0, \sigma^2).$$

Equivalently, we have

$$\lim_{m\to\infty} \mathbf{Pr}\big[\sqrt{m}\,(\overline{X}_m - \mu) \leq t\big] = \Phi\left(\frac{t}{\sigma}\right), \qquad \forall\, t \in \mathbb{R}_{\geq 0}.$$

where $\Phi(z)$ is the standard normal cdf evaluated at z.

While the Gaussian approximation provided by the central limit theorem is often accurate in practice (for example, when $m \geq 50$), it is fundamentally *asymptotic*. In particular, it does not yield exact finite-sample guarantees, and it provides no explicit control over how the quality of the approximation depends on m.

As an illustration, in high-dimensional settings the quantity $\left\|\overline{X}_m - \mu\right\|_2$ need not converge in the manner suggested by the CLT when the sample size m is too small compared to the dimension. These phenomenon, however, cannot be deduced from the CLT alone.

Non-Asymptotic Bounds

Concentration inequalities address this limitation by providing *non-asymptotic* bounds that hold for every finite sample size m, rather than describing limiting behavior as $m \to \infty$. Specifically, they yield bounds of the form

$$\mathbf{Pr}\big[|\overline{X}_m - \mu| > \epsilon\big] \leq f(m, \epsilon),$$

where $f(m, \epsilon)$ decays with m. Such bounds offer explicit sample complexity guarantees. Although they typically require stronger assumptions (e.g., boundedness or sub-Gaussian tails), they provide uniform high-probability control that is absent from asymptotic results.

To measure how much our empirical average $\overline{X}_m$ deviates from the mean μ, we start with basic bounds that require very few assumptions.

Markov's Inequality: For any non-negative random variable X and $a > 0$:

$$\mathbf{Pr}[X \geq a] \leq \frac{\mathbf{E}[X]}{a}$$

Proof. Let X be a non-negative continuous random variable with density f_X, and fix $a > 0$. Then

$$\begin{aligned}\mathbf{E}[X] \;&=\; \int_0^\infty x\, f_X(x)\, dx \;=\; \int_0^a x\, f_X(x)\, dx \;+\; \int_a^\infty x\, f_X(x)\, dx \\ &\geq\; 0 + \int_a^\infty a\, f_X(x)\, dx \;=\; a\, \mathbf{Pr}[X \geq a]\,.\end{aligned}$$

Dividing both sides by $a > 0$ yields the statement. □

Applying this to our coin example, if $p \leq 0.01$, then $\mathbf{Pr}\big[\frac{X}{m} > 0.1\big] \leq \frac{\mathbf{E}[X/m]}{0.1} \leq 0.1$. While very general, Markov's inequality is not very meaningful when p is close to one because it is a relatively loose bound that only uses the first moment, the expected value.

Chebyshev's Inequality: For a random variable X with finite mean and variance σ^2:

$$\mathbf{Pr}[|X - \mathbf{E}[X]| \geq k\sigma] \leq \frac{1}{k^2}$$

Proof. Let X be a random variable with finite mean $\mathbf{E}[X]$ and variance $\mathbf{Var}[X] = \sigma^2 < \infty$, and let $t > 0$. Apply Markov's inequality to the non-negative random variable $(X - \mathbf{E}[X])^2$:

$$\mathbf{Pr}[|X - \mathbf{E}[X]| \geq t] \;=\; \mathbf{Pr}\big[(X - \mathbf{E}[X])^2 \geq t^2\big] \;\leq\; \frac{\mathbf{E}[(X - \mathbf{E}[X])^2]}{t^2} \;=\; \frac{\mathbf{Var}[X]}{t^2} \;=\; \frac{\sigma^2}{t^2}.$$

Setting $t = k\sigma$ yields

$$\mathbf{Pr}[|X - \mathbf{E}[X]| \geq k\sigma] \leq \frac{1}{k^2}.$$

□

For our coin example, $\mathbf{E}\left[\frac{X}{m}\right] = p$ and $\mathbf{Var}\left[\frac{X}{m}\right] = \frac{p(1-p)}{m}$. Substituting into Chebyshev's inequality:

$$\mathbf{Pr}\left[|\tfrac{X}{m} - p| > \epsilon\right] \leq \frac{\mathbf{Var}\left[\frac{X}{m}\right]}{\epsilon^2} \leq \frac{1}{4m\epsilon^2}$$

To ensure this error is at most δ, we need $m \approx \frac{1}{\delta\epsilon^2}$. This captures the correct dependency on ϵ, but the dependency on δ is quite poor (linear rather than logarithmic).

We already see that Chebyshev's inequality yields a stronger guarantee than Markov's inequality, as it exploits additional information about the distribution. In particular, it incorporates the variance of the random variable (its second moment) whereas Markov's inequality relies only on the mean. For the coin example, this additional information comes from knowing the variance of the empirical mean.

Chernoff Bound: Let $X_1, \ldots, X_m$ be independent Bernoulli trials. For the empirical mean $\overline{X}_m$, true mean p, and any $\epsilon \in [0, 1]$, we have:

$$\mathbf{Pr}\left[\overline{X}_m - p > \epsilon p\right] \leq e^{-mp\epsilon^2/3} \quad \text{and} \quad \mathbf{Pr}\left[p - \overline{X}_m > \epsilon p\right] \leq e^{-mp\epsilon^2/2}$$

In the proof of this bound, we use the *moment generating function* of the random variable. Such functions provide an even richer description of a random variable, as they encode all of its moments (when they exist) and therefore capture substantially more information about the distribution. This additional structure enables the derivation of much sharper concentration inequalities.

Before proceeding to the proof, we define the moment generating function of a random variable Z as:

$$M_Z(t) = \mathbf{E}\left[e^{tZ}\right]$$

The power of this function lies in its name: it "generates" the moments of the distribution. By taking the n-th derivative with respect to t and evaluating at $t = 0$, we recover the n-th raw moment:

$$\frac{d^n}{dt^n} M_Z(t) = \mathbf{E}\left[Z^n e^{tZ}\right] \implies \mathbf{E}[Z^n] = \frac{d^n}{dt^n} M_Z(t)\bigg|_{t=0}$$

This encodes the entire distribution's profile (mean, variance, skewness, etc.) into a single analytic function, which allows for much sharper tail bounds than those using only the first or second moments.

Proof of a simplified version. Here, we prove a simplified version of the upper tail $\mathbf{Pr}\left[\overline{X}_m - p > \epsilon p\right] \leq e^{-mp\epsilon^2/4}$ for $\epsilon < 0.5$ that captures the interesting ideas in the main proof.

We use the *Cramér-Chernoff approach*, which is a standard method for proving concentration bounds in general by relating the tail bound to the moment generating function. Since $\overline{X}_m = \frac{1}{m}\sum X_i$, the inequality $\overline{X}_m - p > \epsilon p$ can be rewritten in terms of the sum as

$\sum X_i > mp(1+\epsilon)$. For any $t > 0$, we multiply by t and take the exponent of both sides:

$$\mathbf{Pr}\big[\overline{X}_m - p > \epsilon p\big] = \mathbf{Pr}\big[e^{t\sum X_i} > e^{tmp(1+\epsilon)}\big]$$

Note that the equality above is due to the fact that e^{tx} is a strictly increasing function. Using Markov's inequality on the non-negative exponential term:

$$\mathbf{Pr}\big[e^{t\sum X_i} > e^{tmp(1+\epsilon)}\big] \leq \frac{\mathbf{E}\big[e^{t\sum X_i}\big]}{e^{tmp(1+\epsilon)}}$$

Let's bound the expected value above. For a binary coin with bias p:

$$\mathbf{E}\big[e^{tX_i}\big] = pe^{t(1)} + (1-p)e^{t(0)} = 1 + p(e^t - 1)$$

Using $1 + x \leq e^x$, we bound this as $\mathbf{E}\big[e^{tX_i}\big] \leq e^{p(e^t-1)}$. By independence, the expectation of the product is the product of expectations. Substituting the above bound back yields:

$$\mathbf{E}\big[e^{t\sum X_i}\big] = \prod_{i=1}^{m} \mathbf{E}\big[e^{tX_i}\big] \leq e^{mp(e^t-1)}$$

Our bound becomes:

$$\mathbf{Pr}\big[\overline{X}_m - p > \epsilon p\big] \leq \frac{e^{mp(e^t-1)}}{e^{tmp(1+\epsilon)}} = e^{mp(e^t-1-t(1+\epsilon))}$$

Note that the above bound is correct for any $t > 0$. To obtain the tightest bound, we aim to minimize the right-hand side for t. Setting $t = \ln(1+\epsilon)$ and applying the Taylor approximation $\ln(1+\epsilon) \geq \epsilon - \epsilon^2/2$ yields:

$$\mathbf{Pr}\big[\overline{X}_m - p > \epsilon p\big] \leq e^{mp\left(\epsilon-(\epsilon-\epsilon^2/2)\cdot(1+\epsilon)\right)} = e^{mp\left(-\epsilon^2/2+\epsilon^3/2\right)} \leq e^{-mp\epsilon^2/4}$$

The last inequality holds for $\epsilon < 1/2$. □

In our coin example, setting these bounds to be at most δ results in a sample complexity of $m = O\left(\frac{\log(1/\delta)}{p\epsilon^2}\right)$. This represents a significant improvement over Chebyshev's inequality, as the required number of samples grows only logarithmically with $1/\delta$. However, the inverse dependence on p is problematic as $p \to 0$. This reflects the inherent difficulty of learning small probabilities: estimating them even up to a constant factor is statistically expensive. For instance, to distinguish an event that happens once in a billion from one that happens twice in a billion, one typically requires roughly a billion samples just to observe the event at all, let alone to estimate its probability accurately.

Hoeffding's Inequality: A related bound for the sum of independent bounded variables provides similar exponential concentration, but with an additive error bound. This is especially useful when p might be too small and hence the Chernoff bound is not useful. Let

$X_1, \ldots, X_m$ be independent Bernoulli trials. For the empirical mean $\overline{X}_m$, true mean p, and any $\epsilon \in [0, 1]$, we have:

$$\mathbf{Pr}\left[|\overline{X}_m - \mu| > \epsilon\right] \leq 2e^{-2m\epsilon^2}$$

We will skip the proof of this bound here. We will prove it in future lectures, where we will dive deeper into tail behavior.

For the coin example, this implies $m \geq \frac{2\ln(2/\delta)}{\epsilon^2}$ is sufficient to guarantee the desired confidence δ, and this is optimal.