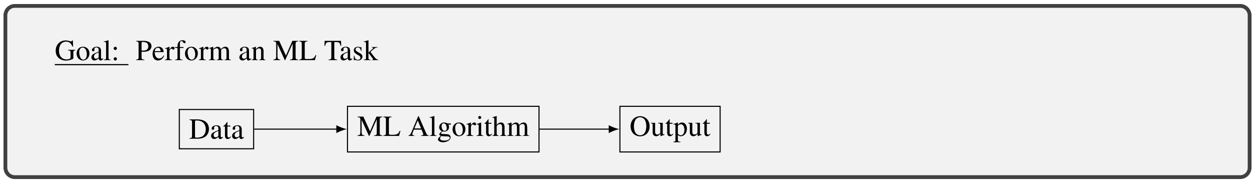
Instructor: Adarsh Barik
Scribed by: Adarsh Barik

Lecture 1, COV878
August 14,2025

Lecture 1: Markov's Inequality and Applications

In this course, we will focus on establishing rigorous performance guarantees for machine learning (ML) algorithms. These guarantees are commonly expressed through generalization bounds, sample complexity estimates, convergence rates, etc. To derive such results, we typically leverage the statistical properties inherent to the problem setting.

A typical ML pipeline can be modeled as follows:



Consider the following instantiation of this setup.

1. *Goal:* We want to estimate the mean of a Gaussian distribution.
2. *Data:* We draw n i.i.d. samples $x_1, \ldots, x_n$ from a one-dimensional Gaussian distribution with (unknown) mean μ and variance σ^2, i.e., $x_i \underset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2),\ \forall i \in \{1, \ldots, n\}$.
3. *ML Algorithm:* We simply take an average of the n samples.
4. *Output:* The final output is given by $\hat{\mu}$, where

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i \ .$$

We would like $\hat{\mu}$ to be a good estimate of μ. To that end, we ask the following precise question:

Question 1 (Estimating Mean of a Gaussian random variable)**.** *Given an $\epsilon > 0$ and $\delta > 0$, how many i.i.d. samples are sufficient to ensure that $|\hat{\mu} - \mu| \leq \epsilon$ with probability at least $1 - \delta$?*

In this lecture, we will see how Markov's inequality, a fundamental result in probability theory, can be used to answer this question.

1 Some Probability Basics

Consider a random variable x with support on $\mathcal{X}$ and probability density function $p(\cdot)$. Let $f : \mathcal{X} \to \mathbb{R}$ be a function of the random variable x. Then,

$$\mathbb{E}\left[f(x)\right] = \int_{x \in \mathcal{X}} f(x) p(x) dx \ .$$

If we choose $f(x) = (x - \mathbb{E}\left[x\right])^2$, we obtain the definition of the variance of x, given by $\mathbb{V}\text{ar}\left[x\right] = \mathbb{E}\left[(x - \mathbb{E}\left[x\right])^2\right]$. Another important example is $f(x) = \mathbb{1}_A(x)$, the indicator function for an event A, which is defined as

$$f(x) = \mathbb{1}_A(x) = \begin{cases} 1, \ x \in A \\ 0, \ \text{otherwise} \end{cases} \quad .$$

Note that $\mathbb{P}\left[A\right] = \int_{x \in A} p(x)dx = \int_{x \in \mathcal{X}} \mathbb{1}_A(x)p(x)dx = \mathbb{E}\left[\mathbb{1}_A(x)\right]$.

2 Markov's Inequality

Next, we will present Markov's inequality for a non-negative random variable x.

Theorem 1 (Markov's inequality)**.** *Let x be a non-negative random variable with probability density function $p(\cdot)$. For any fixed $a > 0$, we have that*

$$\mathbb{P}\left[x \geq a\right] \leq \frac{\mathbb{E}\left[x\right]}{a} .$$

Proof. We begin by the definition of $\mathbb{E}\left[x\right]$.

$$\begin{aligned} \mathbb{E}\left[x\right] &= \int_0^\infty xp(x)dx \\ &= \underbrace{\int_0^a xp(x)dx}_{\geq 0} + \int_a^\infty xp(x)dx \\ &\geq \int_a^\infty xp(x)dx \\ &\geq a\int_a^\infty p(x)dx \\ &= a\mathbb{P}\left[x \geq a\right] \end{aligned}$$

□

We get Chebyshev's inequality as an immediate consequence of Markov's inequality.

Corollary 1 (Chebyshev's inequality)**.** *Let x be a random variable with probability density function $p(\cdot)$. For any fixed $a > 0$, we have that*

$$\mathbb{P}\left[\left|x - \mathbb{E}\left[x\right]\right| \geq a\right] \leq \frac{\mathbb{V}ar\left[x\right]}{a^2} . \tag{1}$$

Proof. The proof is left as an exercise. □

Chebyshev's inequality immediately gives us a way to answer Question 1. Consider

$$\begin{aligned} z &= \frac{1}{n}\sum_{i=1}^n x_i = \hat{\mu} \\ \mathbb{E}\left[z\right] &= \mu \\ \mathbb{V}\text{ar}\left[z\right] &= \frac{\sigma^2}{n} . \end{aligned}$$

Using Chebyshev's inequality (1), it follows that

$$\mathbb{P}\left[|z - \mathbb{E}\left[z\right]| \geq \epsilon\right] = \mathbb{P}\left[|\hat{\mu} - \mu| \geq \epsilon\right] \leq \frac{\sigma^2}{n\epsilon^2}\,.$$

Taking $n \geq \frac{\sigma^2}{\delta\epsilon^2}$ provides an answer to Question 1. While this bound is valid, it does not exploit any specific distributional properties of the x_i's. Next, we will demonstrate that by leveraging the fact that the x_i's are i.i.d. Gaussian random variables, we can obtain significantly improved sample complexity bounds.

3 Concentration Inequalities for the Mean of Gaussian Random Variables

Recall that two random variables x and y with supports $\mathcal{X}$ and $\mathcal{Y}$ and probability density functions $p_x(\cdot)$ and $p_y(\cdot)$ respectively are independent if and only if their joint probability density function $p_{xy}(\cdot,\cdot)$ decomposes as:

$$p_{xy}(x, y) = p_x(x)p_y(y),\ \forall x \in \mathcal{X}, y \in \mathcal{Y}\,. \tag{2}$$

Equation (2) leads to the following result.

Lemma 1. *Let x and y be two independent random variables with supports on $\mathcal{X}$ and $\mathcal{Y}$. For two arbitrary functions $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$, we have that:*

$$\mathbb{E}\left[f(x)g(y)\right] = \mathbb{E}\left[f(x)\right]\mathbb{E}\left[g(y)\right]\ . \tag{3}$$

Proof. The proof is left as an exercise. □

Now we are ready to prove the following stronger result:

Theorem 2. *Let $x_1, \ldots, x_n$ are n independent standard Gaussian random variables, i.e., $x_i \underset{i.i.d.}{\sim} \mathcal{N}(0, 1),\ \forall i \in \{1, \ldots, n\}$. Then, for any fixed $\epsilon > 0$, we have that:*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} x_i \geq \epsilon\right] \leq \exp\left(\frac{-n\epsilon^2}{2}\right)\ .$$

Proof. Pick some arbitrary $t > 0$. We will set an appropriate t later in the proof.

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} x_i \geq \epsilon\right] &= \mathbb{P}\left[\sum_{i=1}^{n} x_i \geq n\epsilon\right] \\
&= \mathbb{P}\left[\exp\left(t\sum_{i=1}^{n} x_i\right) \geq \exp\left(tn\epsilon\right)\right] && (4)\\
&\leq \frac{\mathbb{E}\left[\exp\left(t\sum_{i=1}^{n} x_i\right)\right]}{\exp\left(tn\epsilon\right)} && (5)\\
&= \frac{\mathbb{E}\left[\prod_{i=1}^{n}\exp\left(tx_i\right)\right]}{\exp\left(tn\epsilon\right)} \\
&= \frac{\prod_{i=1}^{n}\mathbb{E}\left[\exp\left(tx_i\right)\right]}{\exp\left(tn\epsilon\right)} && (6)
\end{aligned}$$

where (4) follows due to non-negativity of t and $\exp(\cdot)$, (5) uses Markov's inequality, and (6) holds due to (3). Furthermore, recall that x_i's are independent draws of standard Gaussian random variable (denoted as x below) with probability density function $p(x) = \frac{\exp\left(\frac{-x^2}{2}\right)}{\sqrt{2\pi}}$. Therefore,

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} x_i \geq \epsilon\right] &\leq \frac{\left(\mathbb{E}\left[\exp\left(tx\right)\right]\right)^n}{\exp\left(tn\epsilon\right)} \\
&= \frac{\left(\int_{-\infty}^{\infty} \exp\left(tx\right) \frac{\exp\left(\frac{-x^2}{2}\right)}{\sqrt{2\pi}} dx\right)^n}{\exp\left(tn\epsilon\right)} \\
&= \frac{\left(\exp\left(\frac{t^2}{2}\right)\right)^n}{\exp\left(tn\epsilon\right)} \\
&= \exp\left(\frac{nt^2}{2} - nt\epsilon\right) \qquad (7)
\end{aligned}$$

We observe that right hand side is minimized by choosing $t = \epsilon$. We complete the proof by plugging it back in (7). □

Theorem 2 provides a one-sided bound on the mean of standard Gaussian random variables. However, to fully address Question 1, a two-sided bound is needed. We apply a union bound technique to derive the desired two-sided result.

Lemma 2 (Union bound)**.** *Let A and B be two events that depend on the random variable x. Then*

$$\mathbb{P}\left[A(x) \text{ or } B(x)\right] \leq \mathbb{P}\left[A(x)\right] + \mathbb{P}\left[B(x)\right] \ .$$

Proof. The proof is left as an exercise. □

Lemma 2 along with Theorem 2 leads to the following result.

Theorem 3. *Let $x_1, \ldots, x_n$ be n independent standard Gaussian random variables. For any choice of $\epsilon > 0$ and $\delta > 0$, let $n \geq \frac{2\log\frac{2}{\delta}}{\epsilon^2}$. Then,*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} x_i\right| \leq \epsilon\right] \geq 1 - \delta \ .$$

Proof.

$$\begin{aligned}
\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} x_i\right| \leq \epsilon\right] &= \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} x_i \leq \epsilon \text{ and } \frac{1}{n}\sum_{i=1}^{n} x_i \geq -\epsilon\right] \\
&= 1 - \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} x_i < \epsilon \text{ or } \frac{1}{n}\sum_{i=1}^{n} x_i > -\epsilon\right]
\end{aligned}$$

We observe that x is a standard Gaussian random variable and so is $-x$. Moreover, using Lemma 2 and Theorem 2, we have:

$$\begin{aligned}\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} x_i < \epsilon \text{ or } \frac{1}{n}\sum_{i=1}^{n} x_i > -\epsilon\right] &\leq \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} x_i \leq \epsilon \text{ or } \frac{1}{n}\sum_{i=1}^{n} x_i \geq -\epsilon\right] \\ &\leq \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} x_i \leq \epsilon\right] + \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} x_i \geq -\epsilon\right] \\ &= \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} x_i \leq \epsilon\right] + \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} -x_i \leq \epsilon\right] \\ &\leq 2\exp\left(\frac{-n\epsilon^2}{2}\right) .\end{aligned}$$

We finish the proof by picking $\delta \geq 2\exp\left(\frac{-n\epsilon^2}{2}\right)$. □

Extending results for the standard Gaussian random variables to the general Gaussian distribution random variables is straightforward. Specifically, the following result holds:

Corollary 2. *Let $x_1, \dots, x_n$ be n independent Gaussian random variables with mean μ and variance σ^2. For any choice of $\epsilon > 0$ and $\delta > 0$, let $n \geq \frac{2\sigma^2 \log \frac{2}{\delta}}{\epsilon^2}$. Then,*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right| \leq \epsilon\right] \geq 1 - \delta .$$

Proof. The proof is left as an exercise. □

Disclaimer: These notes have not been scrutinized with the level of rigor usually applied to formal publications. Readers should verify the results before use.