CIS3333: Mathematics of Machine Learning Fall 2025

Lecture: Concentration I: Tail Inequalities and the Chernoff Method

Date: September 17th, 2025 *Author: Surbhi Goel*

Attribution. These notes are extremely similar to the beginning lectures of Larry Wasserman's Intermediate Statistics course from CMU (https://www.stat.cmu.edu/~larry/=stat705/), with some slight notation tweaks to match the course.

1 Concentration Basics

In the first lecture, we introduced the empirical risk $\hat{R}(f)$ and the true risk $R(f)$, arguing that for a model to generalize, these values must be close. This lecture provides the mathematical machinery to prove this. Our goal is to show that for a given **error tolerance** $\epsilon > 0$ and **confidence level** $1 - \delta$, the following holds:

$$\mathbb{P}\left(|\hat{R}(f) - R(f)| < \epsilon\right) > 1 - \delta$$

where the two risks are defined as:

$$\hat{R}(f) = \frac{1}{N}\sum_{i=1}^{N} \ell\big(f(x_i), y_i\big) \quad \text{and} \quad R(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell\big(f(x), y\big)\big]$$

In short, we want to be highly confident that the error on our training data is a good approximation of the true error. To prove this, we will develop a set of powerful tools called **tail inequalities**. These inequalities formalize the idea of **concentration**: the phenomenon that an empirical average (like our empirical risk) gets exponentially unlikely to be far from its true mean as we collect more data.

1.1 Coin flips

Instead of risk, let's consider a much simpler example. Each coin toss is a Bernoulli trial, a concept from our last lecture, where we record $x_i \sim \text{Ber}(p)$ with $p = 0.5$. The average of these trials is:

$$\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} x_i$$

It is easy to see that $\mathbb{E}[\hat{\mu}_N] = 1/2$. How far away is $\hat{\mu}_N$ from its expectation? For example, if $x_i = 1$ for all N flips, then $\hat{\mu}_N = 1$ and it is very far.

Concentration of measure phenomenon says that $\hat{\mu}_N$ "concentrates" closer to $\mathbb{E}[\hat{\mu}_N]$, i.e.

The average of N i.i.d. variables concentrates within an interval of length roughly $1/\sqrt{N}$ around the mean.

Intuitively, for the average to be far from the expectation, many independent variables would need to "conspire" to pull it away, which is extremely unlikely.

This phenomenon is formally captured by the Central Limit Theorem (CLT), which states that for large N, the distribution of the sample mean $\hat{\mu}_N$ becomes approximately Normal. This result is a cornerstone of statistics and machine learning, and the tail inequalities we will study provide a way to make this notion of concentration precise for any finite N.

1.2 Tail inequalities

Our goal is to bound the probability that our sample average $\hat{\mu}$ deviates far from the true mean μ. We will now look at a series of inequalities, each one making slightly stronger assumptions about our random variables to give us a much tighter bound. This is a classic story in statistics: the more you know about your distribution, the stronger a guarantee you can prove.

Theorem 1 (Markov's Inequality). For any non-negative random variable X with finite mean $\mathbb{E}[X] = \mu$, and any $t > 0$:

$$P(X \geq t) \leq \frac{\mu}{t}$$

Proof. The proof follows directly from the definition of expectation for a non-negative random variable:

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty xp(x)dx \\
&= \int_0^t xp(x)dx + \int_t^\infty xp(x)dx \\
&\geq \int_t^\infty xp(x)dx \quad (\text{since } xp(x) \geq 0 \text{ for } x \in [0, t]) \\
&\geq \int_t^\infty tp(x)dx \quad (\text{since } x \geq t \text{ in the integral}) \\
&= t\int_t^\infty p(x)dx \\
&= t\mathbb{P}(X \geq t)
\end{aligned}$$

Rearranging the inequality gives the desired result. □

This bound is very crude, as it makes no assumptions about the distribution beyond non-negativity and a finite mean, but it captures the basic intuition that a random variable is unlikely to be much larger than its mean. To get a tighter bound, we need to make more assumptions about our random variable. A natural next step is to assume the variable has a finite variance, which measures its spread. By incorporating this information, we can derive a much stronger, two-sided guarantee.

Theorem 2 (Chebyshev's Inequality). For a random variable X with finite mean μ and finite variance $\mathbb{V}[X] = \sigma^2$, for any $t > 0$:

$$\mathbb{P}\left(|X - \mu| \geq t\sigma\right) \leq \frac{1}{t^2}$$

Proof. The proof is a clever application of Markov's inequality. Let $Y = (X - \mu)^2$. This is a non-negative random variable with expectation $\mathbb{E}[Y] = \mathbb{E}[(X-\mu)^2] = \sigma^2$. We can apply Markov's inequality to Y with the value $t^2\sigma^2$:

$$\begin{aligned}
\mathbb{P}(|X - \mu| \geq t\sigma) &= \mathbb{P}((X-\mu)^2 \geq t^2\sigma^2) \\
&= \mathbb{P}(Y \geq t^2\sigma^2) \\
&\leq \frac{\mathbb{E}[Y]}{t^2\sigma^2} \quad \text{(by Markov's inequality)} \\
&= \frac{\sigma^2}{t^2\sigma^2} = \frac{1}{t^2}.
\end{aligned}$$

□

This provides a much faster decay rate than Markov's. It bounds the probability of deviating from the mean by t standard deviations.

Weak Law of Large Numbers (almost). Returning to the sample mean, $\hat{\mu}_N = \frac{1}{N}\sum_i X_i$. We assume each X_i is an i.i.d. random variable with mean μ and variance σ^2. Observe that by linearity of expectation,

$$\mathbb{E}[\hat{\mu}_N] = \mathbb{E}\left[\frac{1}{N}\sum_i X_i\right] = \frac{1}{N}\sum_i \mathbb{E}[X_i] = \mu$$

and by the properties of variance for independent variables,

$$\mathbb{V}[\hat{\mu}_N] = \mathbb{V}\left[\frac{1}{N}\sum_i X_i\right] = \sum_i \mathbb{V}\left[\frac{X_i}{N}\right] = \frac{1}{N^2}\sum_i \mathbb{V}[X_i] = \frac{N\sigma^2}{N^2} = \sigma^2/N$$

Applying Chebyshev's inequality to $\hat{\mu}_N$ (which has standard deviation $\sigma/\sqrt{N}$) gives:

$$\mathbb{P}\left(|\hat{\mu}_N - \mu| \geq t\frac{\sigma}{\sqrt{N}}\right) \leq \frac{1}{t^2}$$

This is a form of the Weak Law of Large Numbers. To see how this works, we can set the probability of failure to $1/t^2 = 0.01$ (i.e., $t = 10$). Then with 99% probability, the sample average is within $10\sigma/\sqrt{N}$ of the true mean. The key property is that for any fixed deviation, the probability of that deviation goes to 0 as $N \to \infty$.

The Chernoff Method: A Trick for Exponential Bounds The previous inequalities give bounds that decay polynomially (like $1/t^2$). To get stronger, exponential bounds, we use a powerful technique called the Chernoff method. The core idea is to apply Markov's inequality to the non-negative random variable e^{tX} for some $t > 0$. For any $\epsilon > 0$ and any $t > 0$, we have:

$$\mathbb{P}\left(X \geq \epsilon\right) = \mathbb{P}\left(\exp(tX) \geq \exp(t\epsilon)\right) \leq \frac{\mathbb{E}[\exp(tX)]}{\exp(t\epsilon)} = e^{-t\epsilon}M_X(t)$$

where $M_X(t) = \mathbb{E}[e^{tX}]$ is the Moment Generating Function of X. The exponential function is particularly useful here because for a sum of independent random variables, the MGF of the sum

becomes the product of the individual MGFs, which is easy to work with. Since the inequality holds for any $t > 0$, we can optimize the bound over our choice of t:

$$\mathbb{P}(X \geq \epsilon) \leq \inf_{t>0} e^{-t\epsilon} M_X(t)$$

This is the general form of the Chernoff bound. In practice, we usually want to bound the deviation of a random variable from its mean, so we will apply this result to the centered random variable $X - \mathbb{E}[X]$.

The Moment Generating Function (MGF) The **moment generating function (MGF)**, $M_X(t) = \mathbb{E}[e^{tX}]$, is the key object in the Chernoff bound. It's a deep and useful concept that you can think of as a "blueprint" for a distribution; a key fact from probability theory is that the MGF uniquely defines a distribution. It is so-named because its derivatives with respect to t, evaluated at $t = 0$, can be used to "generate" all the moments of the random variable (e.g., the mean, variance, etc.). For our purposes, its most important role is as the tool that allows us to derive tight, exponential concentration bounds.

MGF of a Standard Normal. Let's derive the MGF for $Z \sim \mathcal{N}(0, 1)$, which we will need shortly.

$$\begin{aligned}
M_Z(t) = \mathbb{E}[\exp(tZ)] &= \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2 + tz\right) dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}((z-t)^2 - t^2)\right) dz \quad \text{(completing the square)} \\
&= e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(z-t)^2\right) dz \\
&= e^{t^2/2}
\end{aligned}$$

The last step follows because the integral is the PDF of a Normal distribution $\mathcal{N}(t, 1)$, which integrates to 1. For a general Gaussian $X \sim \mathcal{N}(\mu, \sigma^2)$, we can write $X = \sigma Z + \mu$ where $Z \sim \mathcal{N}(0, 1)$. Its MGF is then:

$$M_X(t) = \mathbb{E}[e^{t(\sigma Z+\mu)}] = e^{t\mu}\mathbb{E}[e^{(t\sigma)Z}] = e^{t\mu} M_Z(t\sigma) = e^{t\mu} e^{(t\sigma)^2/2} = \exp\left(t\mu + \frac{1}{2}t^2\sigma^2\right)$$

Theorem 3 (Gaussian Tail Bound). For $X \sim N(\mu, \sigma^2)$ and any $u > 0$:

$$\mathbb{P}(|X - \mu| \geq u) \leq 2\exp\left(-\frac{u^2}{2\sigma^2}\right)$$

Proof. We will prove the one-sided bound $\mathbb{P}((X - \mu) \geq u)$ using the Chernoff method. Let $X' = X - \mu$. The MGF of this zero-mean Gaussian is $M_{X'}(t) = \exp(\frac{1}{2}t^2\sigma^2)$, as derived from the general

form above by setting $\mu = 0$. Applying the Chernoff bound:

$$\begin{aligned}\mathbb{P}(X - \mu \geq u) &\leq \inf_{t>0} \frac{\mathbb{E}[\exp(t(X-\mu))]}{\exp(tu)} \\ &= \inf_{t>0} \frac{\exp(\frac{1}{2}t^2\sigma^2)}{\exp(tu)} \\ &= \inf_{t>0} \exp\left(\frac{1}{2}t^2\sigma^2 - tu\right)\end{aligned}$$

To find the infimum, we minimize the exponent by taking its derivative with respect to t and setting it to zero: $t\sigma^2 - u = 0 \implies t = u/\sigma^2$. Since this value is positive, we can plug it in:

$$\mathbb{P}(X - \mu \geq u) \leq \exp\left(\frac{u^2}{2\sigma^2} - \frac{u^2}{\sigma^2}\right) = \exp\left(-\frac{u^2}{2\sigma^2}\right)$$

A symmetric argument for $\mathbb{P}(X - \mu \leq -u)$ gives the same bound. Combining them with a union bound (for two events A and B, $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$) gives the two-sided inequality. □

Chebyshev vs. Gaussian Tail Bound Let's compare the bounds for the sample mean $\hat{\mu}_N$, which has a standard deviation of $\sigma/\sqrt{N}$. We are interested in the probability of deviating by t standard deviations from the mean, i.e., $\mathbb{P}(|\hat{\mu}_N - \mu| \geq t\frac{\sigma}{\sqrt{N}})$.

- **Chebyshev's bound** gives: $\mathbb{P}(\dots) \leq \frac{1}{t^2}$
- **Gaussian tail bound** gives: $\mathbb{P}(\dots) \leq 2\exp(-t^2/2)$

The difference is stark. To achieve 99% confidence (a failure probability of $\delta = 0.01$), we can find the required number of standard deviations, t, for each bound:

- For Chebyshev: $0.01 = 1/t^2 \implies t^2 = 100 \implies t = 10$. The deviation is $10\sigma/\sqrt{N}$.
- For the Gaussian bound: $0.01 = 2e^{-t^2/2} \implies t = \sqrt{2\ln(200)} \approx 3.25$. The deviation is $\approx 3.25\sigma/\sqrt{N}$.

The Gaussian bound is far tighter because its probability decays exponentially in t^2, while Chebyshev's only decays polynomially. This holds in general: for a confidence of $1 - \delta$, the required deviation for Chebyshev scales with $1/\sqrt{\delta}$, while for a Gaussian it scales with $\sqrt{\log(1/\delta)}$.

Summary and a Look Ahead In this lecture, we have built up a powerful set of tools. We started with the intuitive but weak polynomial bounds of Markov and Chebyshev and developed the Chernoff method to achieve much stronger, exponential bounds. The Gaussian tail bound is the first major payoff of this machinery.

However, this powerful tool is currently limited to Gaussian random variables. In the next lecture, we will see how to generalize this result to a much broader and more useful class of variables, which will lead us directly to Hoeffding's Inequality, the workhorse of machine learning theory.