CIS3333: Mathematics of Machine Learning Fall 2025

Lecture: Concentration II: Hoeffding's Inequality

Date: September 22nd, 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.

Recap and Motivation In the last lecture, we developed a powerful set of tools, culminating in the Gaussian tail bound. This gave us a strong, exponential guarantee on how much a sample mean can deviate from its true mean, but it was limited to Gaussian random variables. However, most variables in machine learning, such as the 0-1 loss for classification, are not Gaussian. The goal of this lecture is to generalize our powerful exponential bounds to a much wider and more practical class of random variables, which will lead us directly to Hoeffding's Inequality.

1 From Gaussians to Bounded Variables

Gaussian Tail Bound. Recall that for a Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ and any $u > 0$,

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

The proof, via the Chernoff method, relied on a single key property: the MGF of the centered variable $X - \mu$ was bounded by $\exp(\frac{1}{2}\sigma^2 t^2)$.

Sub-Gaussian Random Variables This suggests a powerful generalization: any random variable whose MGF is similarly bounded will also have Gaussian-like exponential tails. This motivates the definition of a **sub-Gaussian** random variable.

A random variable X with mean μ is sub-Gaussian if there exists a $\sigma > 0$ such that its MGF is dominated by a Gaussian's MGF:

$$\mathbb{E}[\exp(t(X - \mu))] \leq \exp(\sigma^2 t^2/2) \quad \text{for all } t \in \mathbb{R}.$$

Intuitively, this condition is exactly what the Gaussian proof used, and so any X satisfying it inherits the same exponential tail. Many common distributions are sub-Gaussian:

- Gaussians: if $X \sim \mathcal{N}(\mu, \sigma^2)$, then equality holds with this σ.
- Bernoulli: centered $\{0, 1\}$ variables are sub-Gaussian with a universal constant (e.g., $\sigma^2 \leq 1/4$ for Bernoulli).
- Bounded variables: any $X \in [a, b]$ is sub-Gaussian with $\sigma = (b - a)/2$.

Theorem 4 (Tail Bound for Sub-Gaussian Variables). If X is a σ-sub-Gaussian random variable with mean μ, then for any $u > 0$:

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

Proof. The proof is identical to that of the Gaussian Tail Bound, simply replacing the MGF equality with the inequality from the sub-Gaussian definition. □

Property of Sub-Gaussian Averages. A crucial property is that the average of i.i.d. sub-Gaussian random variables is also sub-Gaussian, with a smaller variance proxy. This scaling of the parameter by $1/\sqrt{N}$ is the sub-Gaussian analogue of the fact that the standard deviation of a sample mean of i.i.d. variables is $\sigma/\sqrt{N}$, which we saw in the last lecture.

If $X_1, \ldots, X_N$ are i.i.d. σ-sub-Gaussian random variables with mean μ, then their sample mean $\hat{\mu}_N$ is $\sigma/\sqrt{N}$-sub-Gaussian. This is because:

$$\begin{aligned}
\mathbb{E}[\exp(t(\hat{\mu}_N - \mu))] &= \mathbb{E}\left[\exp\left(\frac{t}{N}\sum_i (X_i - \mu)\right)\right] \\
&= \prod_i \mathbb{E}\left[\exp\left(\frac{t}{N}(X_i - \mu)\right)\right] \\
&\leq \prod_i \exp\left(\frac{t^2}{N^2}\frac{\sigma^2}{2}\right) = \exp\left(\frac{t^2\sigma^2}{2N}\right)
\end{aligned}$$

This directly implies the two-sided tail bound for the average of sub-Gaussian random variables:

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

Bounded Random Variables are Sub-Gaussian. The sub-Gaussian condition can be hard to check directly from the MGF definition. Fortunately, a vast and practical class of random variables is automatically sub-Gaussian: **bounded random variables**. The intuition is that if a variable physically cannot take values outside a range $[a, b]$, it cannot have "heavy tails", that is, extreme deviations are impossible. This property is enough to guarantee its MGF is well-behaved. We will prove this for the simple case of a Rademacher variable, and then state the more general result formalized by Hoeffding's Lemma.

Lemma 1 (Rademacher is 1-Sub-Gaussian)**.** *Let X be a Rademacher random variable, i.e., $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2$. Then X is 1-sub-Gaussian. That is, for any $t \in \mathbb{R}$:*

$$\mathbb{E}[e^{tX}] \leq \exp\left(\frac{t^2}{2}\right)$$

Proof. First, we compute the moment generating function of X directly:

$$\mathbb{E}[e^{tX}] = \frac{1}{2}e^t + \frac{1}{2}e^{-t}$$

Now, we use the Taylor series for the exponential function, $e^u = \sum_{k=0}^{\infty} \frac{u^k}{k!}$:

$$\mathbb{E}[e^{tX}] = \frac{1}{2}\left(\sum_{k=0}^{\infty} \frac{t^k}{k!} + \sum_{k=0}^{\infty} \frac{(-t)^k}{k!}\right)$$
$$= \frac{1}{2}\sum_{k=0}^{\infty} \frac{t^k + (-t)^k}{k!}$$

When k is odd, $t^k + (-t)^k = 0$. When k is even, let $k = 2j$, then $t^{2j} + (-t)^{2j} = 2t^{2j}$. This simplifies the sum to only the even terms:

$$\mathbb{E}[e^{tX}] = \frac{1}{2}\sum_{j=0}^{\infty} \frac{2t^{2j}}{(2j)!} = \sum_{j=0}^{\infty} \frac{t^{2j}}{(2j)!}$$

Finally, we compare this to the series for $e^{t^2/2} = \sum_{j=0}^{\infty} \frac{(t^2/2)^j}{j!} = \sum_{j=0}^{\infty} \frac{t^{2j}}{2^j j!}$. The inequality $\mathbb{E}[e^{tX}] \leq e^{t^2/2}$ holds because for every term, $(2j)! \geq 2^j j!$. This is true because $(2j)!$ is the product of all integers up to $2j$, while $2^j j!$ is the product of only the even integers up to $2j$. $\square$

This useful result can be generalized to any bounded random variable.

Lemma 2 (Hoeffding's Lemma)**.** *Let X be a random variable with $\mathbb{E}[X] = 0$ and $a \leq X \leq b$. Then X is $\frac{b-a}{2}$-sub-Gaussian. That is, for any $t \in \mathbb{R}$:*

$$\mathbb{E}[e^{tX}] \leq \exp\left(\frac{t^2(b-a)^2}{8}\right)$$

The proof of this lemma is more involved, but follows a similar Taylor-expansion argument as the Rademacher case. We omit the full proof, but provide an alternative proof technique below for interested students.

Optional: *Proof of a Weaker Hoeffding's Lemma via Symmetrization.* Let X be a zero-mean random variable on $[a, b]$. The proof uses a clever technique called **symmetrization**. We introduce an independent copy of our variable, X', which has the same distribution as X. Since $\mathbb{E}[X'] = 0$, we can write $\mathbb{E}[e^{tX}] = \mathbb{E}_X[e^{t(X - \mathbb{E}_{X'}[X'])}]$.

Now, let's focus on the inner expectation for a fixed value of X. Define a function $g(y) = e^{t(X-y)}$. As a function of y, this is **convex**. Geometrically, a function is convex if the line segment connecting any two points on its graph lies on or above the graph. This geometric property leads to a powerful probabilistic result (known as Jensen's inequality): the function of an average is less than or equal to the average of the function. Applying this to the random variable X', we get:

$$g(\mathbb{E}[X']) \leq \mathbb{E}[g(X')] \implies e^{t(X - \mathbb{E}[X'])} \leq \mathbb{E}_{X'}\left[e^{t(X-X')}\right]$$

This inequality holds for any fixed X. Now, we can take the expectation over X on both sides:

$$\mathbb{E}_X\left[e^{t(X-\mathbb{E}[X'])}\right] \leq \mathbb{E}_X\left[\mathbb{E}_{X'}\left[e^{t(X-X')}\right]\right]$$

The left side is $\mathbb{E}[e^{tX}]$ and the right side is the expectation over both variables, $\mathbb{E}_{X,X'}[e^{t(X-X')}]$. This gives us the key symmetrization inequality:

$$\mathbb{E}\left[e^{tX}\right] \leq \mathbb{E}_{X,X'}\left[e^{t(X-X')}\right]$$

The distribution of the difference, $X - X'$, is symmetric around 0. This means we can multiply it by an independent Rademacher random variable, ϵ, without changing the expectation. Conditioning on X, X':

$$\mathbb{E}_{X,X'}\left[e^{t(X-X')}\right] = \mathbb{E}_{X,X'}\left[\mathbb{E}_{\epsilon}\left[e^{t\epsilon(X-X')}\right] \mid X, X'\right]$$

The inner term is the MGF of a Rademacher variable, scaled by a factor of $s = t(X - X')$. By Lemma 1, we know that $\mathbb{E}_{\epsilon}[e^{\epsilon s}] \leq e^{s^2/2}$. Plugging this in gives:

$$\mathbb{E}_{X,X'}\left[\mathbb{E}_{\epsilon}\left[e^{t\epsilon(X-X')}\right] \mid X, X'\right] \leq \mathbb{E}_{X,X'}\left[e^{t^2(X-X')^2/2}\right]$$

Finally, since X, X' are both in $[a, b]$, their difference is at most $b - a$. So, $(X - X')^2 \leq (b - a)^2$. This gives the final bound:

$$\mathbb{E}\left[e^{tX}\right] \leq \mathbb{E}_{X,X'}\left[e^{t^2(b-a)^2/2}\right] = e^{t^2(b-a)^2/2}$$

This shows that X is $(b - a)$-sub-Gaussian. Note that this gives a slightly weaker bound than the main lemma (a denominator of 2 instead of 8), but demonstrates a powerful proof technique. □

With the main lemma, we can now state and prove Hoeffding's Inequality.

Theorem 5 (Hoeffding's Inequality). Let $X_1, \ldots, X_N$ be i.i.d. random variables such that $X_i \in [a, b]$ for all i. Let $\hat{\mu}_N = \frac{1}{N}\sum_i X_i$ be the sample mean. Then for any $u > 0$:

$$\mathbb{P}(|\hat{\mu}_N - \mathbb{E}[\hat{\mu}_N]| \geq u) \leq 2\exp\left(-\frac{2Nu^2}{(b-a)^2}\right)$$

Proof. Let $\mu = \mathbb{E}[X_i]$. Define a new set of random variables $Y_i = X_i - \mu$. Each Y_i has zero mean and is bounded in the interval $[a - \mu, b - \mu]$. The length of this interval is $(b - \mu) - (a - \mu) = b - a$. By Hoeffding's Lemma (Lemma 2), each Y_i is $\frac{b-a}{2}$-sub-Gaussian.

The average of these new variables is $\hat{\mu}_Y = \frac{1}{N}\sum_i Y_i = \hat{\mu}_N - \mu$. Since the Y_i are independent, their average is sub-Gaussian with parameter $\frac{(b-a)/2}{\sqrt{N}} = \frac{b-a}{2\sqrt{N}}$. Let's call this new sub-Gaussian parameter $\sigma' = \frac{b-a}{2\sqrt{N}}$. We can now apply the tail bound for sub-Gaussian averages to $\hat{\mu}_Y$:

$$\mathbb{P}(|\hat{\mu}_N - \mu| \geq u) = \mathbb{P}(|\hat{\mu}_Y| \geq u) \leq 2\exp\left(-\frac{u^2}{2(\sigma')^2}\right) = 2\exp\left(-\frac{u^2}{2(\frac{b-a}{2\sqrt{N}})^2}\right) = 2\exp\left(-\frac{2Nu^2}{(b-a)^2}\right).$$

□

This powerful result connects the deviation u to the number of samples N and the range of the data $(b - a)$, without needing to know any other details about the distribution.