**UC Berkeley** **Jonathan Shafer**
**CS 294-220, Spring 2021**

**Unit 3**

# Basic Concentration of Measure: Hoeffding's Inequality

## 1 Introduction

When learning in a statistical setting, the more evidence we see, the more confident we become. For example, if we tossed a (possibly biased) coin once and it came out heads, that does not necessarily mean that heads is a more likely outcome than tails. But if we tossed the coin 100 times and it always came out heads, then we can be fairly certain that the coin is strongly biased in favor of heads.

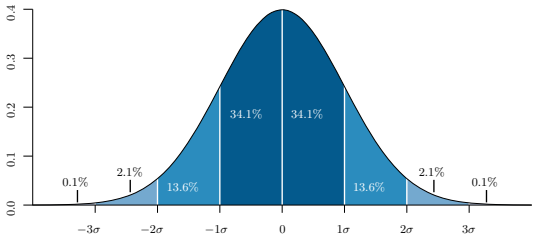


**Figure 1** A Gaussian random variable is concentrated near its mean. Source: Wikipedia.

Formally, this phenomena is captured by *concentration of measure*. In a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, more than 95% of the probability mass is concentrated within two standard deviations of the mean. Namely, $\mathbb{P}_{X\sim\mathcal{N}(\mu,\sigma^2)}\left[|X-\mu| \leq 2\sigma\right] \geq 0.95$. The central limit theorem tells us that this phenomenon is not unique to the Gaussian distribution: for any sequence of i.i.d. random variables $X_1, X_2, \ldots$ with finite mean $\mu$ and variance $\sigma$, the average $\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ converges in distribution to a Gaussian random variable:[1]

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right).$$

In particular, the weak law of large numbers says that for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[\left|\overline{X}_n - \mu\right| > \varepsilon\right] = 0.$$

This implies that if we observe enough i.i.d. random variables then we can obtain an excellent estimate of $\mu$. Hence in the biased coin example above, the more times we toss the coin, the better we can estimate its bias.

The problem is that the central limit theorem and the law of large numbers are asymptotic results that only tell us what happens in the limit. In order to use concentration of measure for learning with a finite sample complexity, we need a quantitative version of these results that can tell us how good our estimates are after seeing a finite number $n$ of examples.

[1] So the Gaussian distribution is not unusual in this respect, it is very normal – hence the name *the normal distribution*.

In this unit we will prove a quantitative theorem of this form called Hoeffding's inequality, which roughly says that

$$\mathbb{P}\left[\left|\overline{X}_n - \mu\right| > \varepsilon\right] \leq e^{-\Omega(n\varepsilon^2)}.$$

This theorem will be very useful for us when we continue to investigate the PAC model of learning.

## 2 Moment Generating Functions

▶ **Definition 1.** *Let $X$ be a real-valued random variable. For any $n \in \mathbb{N}$, the n-th moment of $X$ is $\mathbb{E}[X^n]$ (if the integral exists).*

The first moment of a random variable is its mean, the second moment is its variance (if $\mathbb{E}[X] = 0$), and the higher moments convey further types of "global" information about the random variable.

▶ **Definition 2.** *Let $X$ be a real-valued random variable. The moment generating function (MGF) of $X$ is the function $M_X : \mathbb{R} \to \mathbb{R}$ given by*

$$M_X(t) = \mathbb{E}\left[e^{tX}\right].$$

Note that the MGF might not always exist (namely, $e^{tX}$ is not necessarily integrable). Some facts about the MGF when it exists:

**(a)** As its name suggests, the MGF generates the moments of the random variable, in the following sense.

▷ Claim 3. Let $X$ be a random variable and assume $M_X$ exists and is finite in some neighborhood of 0. For any positive integer $n$, let $M_X^{(n)}$ denote the $n$-th derivative of $M_X$. Then $M_X^{(n)}(0) = \mathbb{E}[X^n]$.

**Proof sketch.**[2] Recall that the Taylor series of $e^x$ around 0 is

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots.$$

Therefore,

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{(tX)^k}{k!}\right] \overset{(\star)}{=} \sum_{k=0}^{\infty} \mathbb{E}\left[\frac{(tX)^k}{k!}\right] = \sum_{k=0}^{\infty} \frac{t^k}{k!}\mathbb{E}\left[x^k\right].$$

Equality $(\star)$ uses the fact that we can exchange integration and summation.[3] Furthermore, in this case it is possible to exchange differentiation and summation,[4] yielding

$$M_X^{(n)}(t) = \frac{\partial^n}{\partial t^n} \sum_{k=0}^{\infty} \frac{t^k}{k!}\mathbb{E}\left[x^k\right] = \sum_{k=0}^{\infty} \mathbb{E}\left[x^k\right] \frac{\partial^n}{\partial t^n} \frac{t^k}{k!}.$$

[2] A complete proof is available as Theorem 4.8 in [7].
[3] We will not prove this fact, which is a consequence of Fubini's theorem. See also here.
[4] A power series may be differentiated term-by-term within its radius of convergence. See here.

Observe that

$$\left.\frac{\partial^n}{\partial t^n}\frac{t^k}{k!}\right|_{t=0} = \begin{cases} 1 & k = n \\ 0 & k \neq n \end{cases},$$

and therefore $M_X^{(n)}(0) = \mathbb{E}[X^n]$. ◀

**(b)** The MGF completely determines the distribution. Formally,

▶ **Theorem 4** (Uniqueness Theorem). *If there exists $\delta > 0$ such that*

$$\forall t \in (-\delta, \delta): \; M_X(t) = M_Y(t) < \infty$$

*then $X$ and $Y$ are equal in distribution.*

This is a non-trivial fact. We will not prove it, but it is useful to keep in mind.[5]

**(c)** The MGF for the sum of independent random variables is the product of their MGFs. Namely, if $X$ and $Y$ are independent random variables and $M_X(t)$ and $M_Y(t)$ exist for some $t$, then

$$M_{X+Y}(t) = \mathbb{E}\left[e^{t(X+Y)}\right] = \mathbb{E}\left[e^{tX}e^{tY}\right] = \mathbb{E}\left[e^{tX}\right]\mathbb{E}\left[e^{tY}\right] = M_X(t)M_Y(t).$$

## 3 Sub-Gaussian Distributions

As mentioned in the introduction, the Gaussian distribution is concentrated near its mean.

▶ **Definition 5.** *Let $\mu, \sigma \in \mathbb{R}$. The Gaussian (normal) distribution with mean $\mu$ and variance $\sigma^2$ is denoted $\mathcal{N}(\mu, \sigma^2)$ and is defined by the following probability density function:*

$$\forall x \in \mathbb{R}: \; p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

▷ Claim 6. Let $X$ be a random variable with distribution $\mathcal{N}(\mu, \sigma^2)$. Then for any $t > 0$,

$$\mathbb{P}[X - \mu > t] \leq \frac{1}{\sqrt{2\pi}} \frac{e^{-\frac{t^2}{2\sigma^2}}}{t},$$

$$\mathbb{P}[X - \mu < -t] \leq \frac{1}{\sqrt{2\pi}} \frac{e^{-\frac{t^2}{2\sigma^2}}}{t},$$

and therefore

$$\mathbb{P}[|X - \mu| > t] \leq \sqrt{\frac{2}{\pi}} \frac{e^{-\frac{t^2}{2\sigma^2}}}{t}.$$

▶ **Exercise 7.** *Prove Claim 6.*

In other words, a Gaussian random variable satisfies

$$\mathbb{P}[|X - \mu| > t] \leq \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

[5] Proofs are available e.g. in [1, 3].

The following definition captures this concentration property. It will be convenient to express this using the logarithm of the MGF,

$$\psi_X(s) = \ln M_X(s) = \ln \mathbb{E}\left[e^{sX}\right].$$

▶ **Definition 8.** *Let $X$ be a random variable. We say that $X$ is sub-Gaussian with variance factor $v$ if $\mathbb{E}[X] = 0$, $\psi_X(s)$ exists for all $s \in \mathbb{R}$, and*

$$\forall s \in \mathbb{R}: \ \psi_X(s) \leq \frac{s^2 v}{2}.$$

The bound on the tails of a sub-Gaussian distribution can be derived using the Cramér–Chernoff method as follows.

▷ Claim 9. Let $X$ be a sub-Gaussian random variable with variance factor $\sigma^2$. Then for any $t > 0$,

$$\mathbb{P}[X > t] \leq \exp\left(-\frac{t^2}{2\sigma^2}\right), \text{ and}$$
$$\mathbb{P}[X < -t] \leq \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

**Proof.** For the first inequality, let $s > 0$ be some scalar to be chosen later. Then

$$\begin{aligned}
\mathbb{P}[X > t] &= \mathbb{P}[sX > st] = \mathbb{P}\left[e^{sX} > e^{st}\right] && \text{(monotinicity)} \\
&\leq \frac{\mathbb{E}\left[e^{sX}\right]}{e^{st}} && \text{(Markov's inequality)} \\
&= e^{\psi_X(s) - st} \leq e^{\frac{s^2\sigma^2}{2} - st}. && (1)
\end{aligned}$$

To obtain the tightest bound, we choose the value of $s$ that minimize this expression. The exponent function is monotone increasing, and the expression inside the exponent above is a (U-shaped) parabola. Therefore the minimum is obtained at the stationary point of this parabola.

$$\frac{\partial}{\partial s}\left(\frac{s^2\sigma^2}{2} - st\right) = s\sigma^2 - t = 0 \implies s = \frac{t}{\sigma^2}.$$

Plugging this back into Eq. (1) yields

$$\mathbb{P}[X > t] \leq \exp\left(-\frac{t^2}{2\sigma^2}\right),$$

as desired. The proof for the second inequality is similar. ◀

## 4 Hoeffding's Inequality

The following theorem is our main concentration of measure result for sums of independent random variables.

▶ **Theorem 10** (Hoeffding's Inequality [5]). *Let $Z_1, \ldots, Z_m$ be a sequence of real-valued i.i.d. random variables. Assume that there exist $a, b, \mu \in \mathbb{R}$ such that for all $i \in [m]$, $\mathbb{E}[Z_i] = \mu$*

*and* $\mathbb{P}[a \le Z_i \le b] = 1$. *Then for any* $\varepsilon \ge 0$,

$$\mathbb{P}\left[\frac{1}{m}\sum_{i\in[m]} Z_i - \mu > \varepsilon\right] \le \exp\left(-2m\left(\frac{\varepsilon}{b-a}\right)^2\right), \text{ and} \tag{2}$$

$$\mathbb{P}\left[\frac{1}{m}\sum_{i\in[m]} Z_i - \mu < -\varepsilon\right] \le \exp\left(-2m\left(\frac{\varepsilon}{b-a}\right)^2\right). \tag{3}$$

*Therefore,*

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i\in[m]} Z_i - \mu\right| > \varepsilon\right] \le 2\exp\left(-2m\left(\frac{\varepsilon}{b-a}\right)^2\right). \tag{4}$$

The proof of Hoeffding's inequality relies on the following two lemmas.

▷ Claim 11 (Popoviciu's Inequality). Let $a \le b$ be real numbers, let $X$ be a random variable, and assume that $\mathbb{P}[a \le X \le b] = 1$. Then

$$\mathrm{Var}[X] \le \frac{(b-a)^2}{4}.$$

**Proof.** Consider the function $g(t) = \mathbb{E}\left[(X-t)^2\right] = \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\,t + t^2$. This is a U-shaped parabola, so it achieves its minimum at the stationary point.

$$g'(t) = -2\mathbb{E}[X] + 2t = 0 \implies t = \mathbb{E}[X].$$

Let $c = \frac{a+b}{2}$. Then

$$\mathrm{Var}[X] = g(\mathbb{E}[X]) \le g(c) = \mathbb{E}\left[\left(X - \frac{a+b}{2}\right)^2\right] = \frac{1}{4}\mathbb{E}\left[((X-a)+(X-b))^2\right]$$

Seeing as $X - a \ge 0$ and $X - b \le 0$,

$$((X-a)+(X-b))^2 \le ((X-a)-(X-b))^2 = (b-a)^2.$$

Hence

$$\mathrm{Var}[X] \le \frac{1}{4}\mathbb{E}\left[(b-a)^2\right] = \frac{(b-a)^2}{4}.$$ ◀

▶ **Lemma 12** (Hoeffding's Lemma [5]). *Let* $a, b \in \mathbb{R}$, $a \le b$, *and let* $X$ *be a real-valued random variable such that* $\mathbb{P}[a \le X \le b] = 1$ *and* $\mathbb{E}[X] = 0$. *Then* $X$ *is sub-Gaussian with variance factor* $\frac{(b-a)^2}{4}$. *Namely, for all* $t \in \mathbb{R}$,

$$\psi_X(t) \le \frac{t^2(b-a)^2}{8}.$$

**Proof of Lemma 12.** First, for fixed $t \in \mathbb{R}$, we claim there exists a random variable $U$ such that for any integrable function $f : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[f(U)] = \frac{\mathbb{E}\left[f(X)e^{tX}\right]}{\mathbb{E}\left[e^{tX}\right]}. \tag{5}$$

This is called *exponential change of measure*.[6] We can define $U$ via its Radon-Nikodym derivative with respect to $X$:

$$\frac{d\,\mathbb{P}_U}{d\,\mathbb{P}_X} = \frac{e^{tx}}{c},$$

where $\mathbb{P}_U$ and $\mathbb{P}_X$ are the measures of $U$ and $X$ respectively, and $c = \mathbb{E}\left[e^{tX}\right]$. Namely, $U$ is a random variable with probability measure $\mathbb{P}_U$ such that for any event $A$,

$$\mathbb{P}_U(A) = \int_A d\,\mathbb{P}_U = \frac{1}{c}\int_A e^{tx}\,d\,\mathbb{P}_X(x) = \frac{\mathbb{E}\left[\mathbb{1}(X \in A)e^{tX}\right]}{\mathbb{E}\left[e^{tX}\right]}.$$

$\mathbb{P}_U$ is a valid measure function because it returns results in $[0,1]$, $\mathbb{P}_U(\varnothing) = 0$, $\mathbb{P}_U(\mathbb{R}) = 1$, and it satisfies countable additivity. Furthermore, for any integrable function $f : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}\left[f(U)\right] = \int_{\mathbb{R}} f(s)\,d\,\mathbb{P}_U(s) = \frac{1}{c}\int_{\mathbb{R}} f(s)e^{ts}\,d\,\mathbb{P}_X(s) = \frac{\mathbb{E}\left[f(X)e^{tX}\right]}{\mathbb{E}\left[e^{tX}\right]},$$

as desired.

Second, we note that Eq. (5) implies:

**(a)** $\mathbb{E}\left[U\right] = \frac{\mathbb{E}\left[Xe^{tX}\right]}{\mathbb{E}\left[e^{tX}\right]}$ and $\mathbb{E}\left[U^2\right] = \frac{\mathbb{E}\left[X^2e^{tX}\right]}{\mathbb{E}\left[e^{tX}\right]}$.

**(b)** $U \in [a,b]$ with probability 1. This follows from,

$$\mathbb{P}\left[a \le U \le b\right] = \mathbb{E}\left[\mathbb{1}(a \le U \le b)\right] = \frac{\mathbb{E}\left[\mathbb{1}(a \le X \le b)e^{tX}\right]}{\mathbb{E}\left[e^{tX}\right]} = 1,$$

where we used $f(U) = \mathbb{1}(a \le U \le b)$.

Third, we calculate the derivatives of $\psi(t) = \psi_X(t)$.

$$\psi'(t) = \frac{\partial}{\partial t}\ln\left(\mathbb{E}\left[e^{tX}\right]\right) = \frac{\frac{\partial}{\partial t}\mathbb{E}\left[e^{tX}\right]}{\mathbb{E}\left[e^{tX}\right]} \overset{(\star)}{=} \frac{\mathbb{E}\left[\frac{\partial}{\partial t}e^{tX}\right]}{\mathbb{E}\left[e^{tX}\right]} = \frac{\mathbb{E}\left[Xe^{tX}\right]}{\mathbb{E}\left[e^{tX}\right]},$$

where $(\star)$ follows from Leibniz's rule for differentiation under the integral sign. Similarly,

$$\begin{aligned}
\psi''(t) &= \frac{\mathbb{E}\left[X^2e^{tX}\right]}{\mathbb{E}\left[e^{tX}\right]} - \left(\frac{\mathbb{E}\left[Xe^{tX}\right]}{\mathbb{E}\left[e^{tX}\right]}\right)^2 && \text{(formula for derivative of a fraction)}\\
&= \mathbb{E}\left[U^2\right] - \left(\mathbb{E}\left[U\right]\right)^2 && \text{(from (a))}\\
&= \operatorname{Var}\left[U\right] \le \frac{(b-a)^2}{4}. && \text{(from Claim 11)}
\end{aligned}$$

Lastly, note that $\psi(0) = \ln 1 = 0$ and $\psi'(0) = \mathbb{E}\left[X\right] = 0$. By Taylor's theorem, there exists $\theta \in [0,t]$ such that

$$\psi(t) = \psi(0) + t\psi'(0) + \frac{t^2}{2}\psi''(\theta) \le \frac{t^2(b-a)^2}{8}.$$ ◀

[6] This is also sometimes called an *exponential tilting* (or *twisting*) of $X$, or the *Esscher transform* of $X$.

**Proof of Theorem 10.** First, we prove Eq. (2). It suffices to prove this for the case where $\mu = 0$, because if $\mu \neq 0$ then we can use the result for variables $Z_i' = Z_i - \mu$ (which do have mean 0), and this implies the result for $Z_i$. Denote $S_m = \sum_{i\in[m]} Z_i$. Observe that

$$\begin{aligned}
\psi_{S_m}(t) &= \ln \mathbb{E}\left[e^{tS_m}\right] \\
&= \ln \mathbb{E}\left[e^{t(Z_1+\cdots+Z_m)}\right] \\
&= \sum_{i=1}^{m} \ln \mathbb{E}\left[e^{tZ_i}\right] && (Z_i\text{'s are i.i.d.}) \\
&\leq \sum_{i=1}^{m} \frac{t^2(b-a)^2}{8} && (\text{Lemma 12}) \\
&= \frac{mt^2(b-a)^2}{8} = \frac{t^2\left(\frac{m(b-a)^2}{4}\right)}{2}.
\end{aligned}$$

Hence $S_m$ is sub-Gaussian with variance factor $v = \frac{m(b-a)^2}{4}$. By Claim 9,

$$\mathbb{P}\left[\frac{1}{m}\sum_{i\in[m]} Z_i > \varepsilon\right] = \mathbb{P}\left[S_m > m\varepsilon\right] \leq \exp\left(-\frac{m^2\varepsilon^2}{2v}\right) = \exp\left(-\frac{2m\varepsilon^2}{(b-a)^2}\right).$$

This concludes the proof of Eq. (2). The proof of Eq. (3) is similar. Finally, Eq. (4) follows from Eq. (2) and (3) via a union bound. ◀

## 5 Bibliographic Notes

A very good exposition of the material covered in this unit appears in [2, Chapter 2], as well as in [6, Chapter 1], and [4, Chapter 2].

## References


1 Patrick Billingsley. *Probability and Measure*. John Wiley & Sons, 2008.

2 Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities – A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. doi:10.1093/acprof:oso/9780199535255.001.0001.

3 John H. Curtiss. A note on the theory of moment generating functions. *The Annals of Mathematical Statistics*, 13(4):430–433, 1942.

4 Bruce Hajek and Maxim Raginsky. *ECE 543: Statistical Learning Theory*. University of Illinois at Urbana-Champaign, 2018. URL: https://web.archive.org/web/20210213043003/http://maxim.ece.illinois.edu/teaching/SLT/SLT.pdf.

5 Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi:https://doi.org/10.2307/2282952.

6 Philippe Rigollet. *18.S997: High Dimensional Statistics*. MIT Open-CourseWare, Cambridge, MA, 2015. URL: https://web.archive.org/web/20200729072809/https://ocw.mit.edu/courses/mathematics/18-s997-high-dimensional-statistics-spring-2015/lecture-notes/MIT18_S997S15_CourseNotes.pdf.

7 Thomas A. Severini. *Elements of Distribution Theory*. Cambridge University Press, 2005.