CS229M/STATS214: Machine Learning Theory

Lecturer: Tengyu Ma — Lecture # 3
Scribe: Brad Ross and Robbie Jones — Jan 20, 2021

3.1 Review and overview

At the end of last lecture, we were left with the following bound on the excess risk:

$$L(\hat{\theta}) - L(\theta^\star) \le |L(\theta^\star) - \hat{L}(\theta^\star)| + \sup_\theta (L(\theta) - \hat{L}(\theta)). \tag{3.1}$$

Bounding the right-hand side involves bounding each of the two terms separately. In a future lecture we will introduce the concept of *uniform convergence*, which we will employ to bound $\sup_\theta (L(\theta) - \hat{L}(\theta))$. Today's lecture will develop the notion of *concentration inequalities* that we use to bound $|L(\theta^\star) - \hat{L}(\theta^\star)|$.

As it turns out, the material from today's lecture constitutes arguably the most important content in the entire course. No matter what area of machine learning one wants to study, if it involves sample complexity, some kind of concentration result will typically be required. Hence, concentration inequalities are some of the most important tools in modern statistical learning theory.

Assume that we have independent random variables $X_1, \ldots, X_n$. In this lecture, we will develop tools to show results that formalize the intuition for these statements:

1. $X_1 + \ldots + X_n$ concentrates around $\mathbb{E}[X_1 + \ldots + X_n]$.
2. More generally, $f(X_1, \ldots, X_n)$ concentrates around $\mathbb{E}[f(X_1, \ldots, X_n)]$.

Then, in the next few lectures, these results will be used to show that $\hat{L}(\theta) \approx L(\theta)$ and $\sup_\theta (L(\theta) - \hat{L}(\theta)) \approx \mathbb{E}\left[\sup_\theta (L(\theta) - \hat{L}(\theta))\right]$.

3.2 Warm-up: Chebyshev's inequality

As a warm-up, consider an arbitrary random variable Z with finite variance. One of the most famous results characterizing its tail behavior is the following theorem:

Theorem 3.1 (Chebyshev's inequality)**.** *Let Z be a random variable with finite expectation and variance. Then*

$$\Pr[|Z - \mathbb{E}[Z]| \ge t] \le \frac{\mathrm{Var}(Z)}{t^2}, \quad \forall t > 0. \tag{3.2}$$

Intuitively, this means that as we approach the tails of the distribution of Z, the density decreases at a rate of at least $1/t^2$. Moreover, for any $\delta \in (0, 1]$, by plugging in $t = \mathrm{sd}(Z)/\sqrt{\delta}$ to (3.2) we see that

$$\Pr\left[|Z - \mathbb{E}[Z]| \le \frac{\mathrm{sd}(Z)}{\sqrt{\delta}}\right] \ge 1 - \delta. \tag{3.3}$$

Unfortunately, it turns out that Chebyshev's inequality is a rather weak concentration inequality. To illustrate this, assume $Z \sim \mathcal{N}(0,1)$. We can show (using the Gaussian tail bound derived in Problem 3(c) in Homework 0) that

$$\Pr\left[|Z - \mathbb{E}[Z]| \leq \mathrm{sd}(Z)\sqrt{2\log(2/\delta)}\right] \geq 1 - \delta. \tag{3.4}$$

for any $\delta \in (0,1]$. In other words, the density at the tails of the normal distribution is decreasing at an exponential rate, while Chebyshev's inequality only gives a quadratic rate. The discrepancy between (3.3) and (3.4) is made more apparent when we consider inverse-polynomial $\delta = \frac{1}{n^c}$ for some parameter n and degree c (we will see concrete instances of this setup in future lectures). Then the tail bound for the normal distribution (3.4) implies that

$$|Z - \mathbb{E}[Z]| \leq \mathrm{sd}(Z) \cdot \sqrt{\log O\left(n^c\right)} = \mathrm{sd}(Z) \cdot O\left(\sqrt{\log n}\right) \quad w.p.\ 1 - \delta, \tag{3.5}$$

while Chebyshev's inequality gives us the weaker result

$$|Z - \mathbb{E}[Z]| \leq \mathrm{sd}(Z) \cdot \sqrt{O(n^c)} = \mathrm{sd}(Z) \cdot O(n^{c/2}) \quad w.p.\ 1 - \delta. \tag{3.6}$$

Despite the previous example, Chebyshev's inequality is actually optimal without further assumptions, in the sense that there exist distributions with finite variance for which the bound is tight. With that in mind, we will need to assume more about our random variables if we want to improve upon the Chebyshev inequality's $1/t^2$ rate of tail decay. As an example, recall that when $0 \leq X_i \leq 1$ for $i = 1, \ldots, n$, Hoeffding's inequality is applicable:

$$\Pr[|Z - \mathbb{E}[Z]| \geq t] \leq 2\exp(-2t^2/n). \tag{3.7}$$

This tail probability is exponentially decaying in t instead of polynomially decaying as in Chebyshev's inequality! Certainly requiring boundedness in $[0,1]$ (or $[a,b]$ more generally) is limiting, so it is worth asking what types of distributions permit such an exponential tail bound. The following section will explore such a class of random variables: *sub-Gaussian* random variables.

3.3 Sub-Gaussian random variables

We begin by defining the class of sub-Gaussian random variables by way of a bound on their moment generating functions, after which we will see how this bound guarantees the exponential tail decay we are after.

Definition 3.2 (Sub-Gaussian Random Variables)**.** A random variable X with finite mean μ is *sub-Gaussian* with parameter σ if

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\sigma^2\lambda^2/2}, \quad \forall \lambda \in \mathbb{R}. \tag{3.8}$$

We say that X is σ-sub-Gaussian and say it has *variance proxy* σ^2.

Remark 3.3. As it turns out, (3.8) is quite a strong condition, requiring that infinitely many moments of X exist and do not grow too quickly. To see why, assume without loss of generality that $\mu = 0$ and take a power series expansion of the moment generating function:

$$\mathbb{E}[\exp(\lambda X)] = \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{(\lambda X)^k}{k!}\right] = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}\,\mathbb{E}[X^k]. \tag{3.9}$$

A bound on the moment generating function then is a bound on infinitely many moments of X, i.e. a requirement that the moments of X are all finite and grow slowly enough to allow the power series to converge.

Although (3.8) is not a particularly intuitive definition, it turns out to imply exactly the type of exponential tail bound we want:

Theorem 3.4 (Tail bound for sub-Gaussian random variables)**.** *If a random variable X with finite mean μ is σ-sub-Gaussian, then*

$$\Pr[|X - \mu| \geq t] \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right), \quad \forall t \in \mathbb{R}. \tag{3.10}$$

Proof. Fix $t > 0$. For any $\lambda > 0$,

$$\begin{aligned}
\Pr[X - \mu \geq t] &= \Pr[\exp(\lambda(X - \mu)) \geq \exp(\lambda t)] && && (3.11)\\
&\leq \exp(-\lambda t)\,\mathbb{E}[\exp(\lambda(X - \mu))] && \text{(by Markov's inequality)} && (3.12)\\
&\leq \exp(-\lambda t)\exp(\sigma^2\lambda^2/2) && \text{(by (3.8))} && (3.13)\\
&= \exp(-\lambda t + \sigma^2\lambda^2/2). && && (3.14)
\end{aligned}$$

Because the bound (3.14) holds for any choice of λ and $\exp(\cdot)$ is monotonically increasing, we can optimize the bound (3.14) by finding λ which minimizes the exponent $-\lambda t + \sigma^2\lambda^2/2$. Differentiating and setting the derivative equal to zero, we find that the optimal choice is $\lambda = t/\sigma^2$, yielding the one-sided tail bound

$$\Pr[X - \mu \geq t] \leq \exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{3.15}$$

Going through the same line of reasoning but for $-X$ and $-t$, we can also show that for any $t > 0$,

$$\Pr[X - \mu \leq -t] \leq \exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{3.16}$$

We can then obtain (3.10) by applying the union bound:

$$\Pr[|X - \mu| \geq t] = \Pr[X - \mu \geq t] + \Pr[X - \mu \leq -t] \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{3.17}$$

□

Remark 3.5 (Tail bound implies sub-Gaussianity)*.* In addition to being a necessary condition for sub-Gaussianity (Theorem 3.4), the tail bound (3.10) for sub-Gaussian random variables is also a sufficient condition up to a constant factor. In particular, if a random variable X with finite mean μ satisfies (3.10) for some $\sigma > 0$, then X is $O(\sigma)$-sub-Gaussian. Unfortunately, the proof of this reverse direction is somewhat more involved, so we refer the interested reader to Theorem 2.6 and its proof in Section 2.4 of [Wai19] and Proposition 2.5.2 in [Ver18] for details. While the tail bound is the property we ultimately care about most when studying sub-Gaussian random variables, the definition in (3.8) is a more technically convenient characterization, as we will see in the proof of Theorem (3.7).

Remark 3.6. Note that in light of Remark 3.3, the tail bound (3.4) requires all central moments of X to exist and not grow too quickly. In contrast, Chebyshev's inequality (and more generally any polynomial variant of Markov's inequality $\Pr[|X-\mu| \geq t] = \Pr[|X-\mu|^k \geq t^k] \leq t^{-k}\mathbb{E}[|X-\mu|^k]$) only requires that the second central moment $\mathbb{E}[(X-\mu)^2]$ (more generally, the kth central moment $\mathbb{E}[|X-\mu|^k]$) is finite to yield a tail bound. If infinite moments exist however, it turns out that $\inf_{k\in\mathbb{N}} t^{-k}\mathbb{E}[|X-\mu|^k] \leq \inf_{\lambda>0} \exp(-\lambda t)\,\mathbb{E}[\exp(\lambda(X-\lambda))]$, i.e. the optimal polynomial tail bound is tighter than the optimal exponential tail bound (see Exercise 2.3 in [Wai19]). As we will see shortly though, using exponential functions of random variables allows us to prove results about sums of random variables more conveniently, which is why most researchers use exponential tail bounds in practice.

Having defined and derived exponential tail bounds for sub-Gaussian random variables, we can now accomplish the first of the goals we set out at the beginning of the lecture: to show that under certain conditions, namely independence and sub-Gaussianity of $X_1, \ldots, X_n$, the sum $Z = \sum_{i=1}^n X_i$ concentrates around $\mathbb{E}[Z] = \mathbb{E}[\sum_{i=1}^n X_i]$.

Theorem 3.7 (Sum of sub-Gaussian random variables is sub-Gaussian)**.** *If $X_1, \ldots, X_n$ are independent sub-Gaussian random variables with variance proxies $\sigma_1^2, \ldots, \sigma_n^2$, then $Z = \sum_{i=1}^n X_i$ is sub-Gaussian with variance proxy $\sum_{i=1}^n \sigma_i^2$. As a consequence, we have the tail bound*

$$\Pr[|Z - \mathbb{E}[Z]| \geq t] \leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}\right), \tag{3.18}$$

for all $t \in \mathbb{R}$.

Proof. Using the independence of $X_1, \ldots, X_n$, we have that for any $\lambda \in \mathbb{R}$:

$$\begin{aligned}
\mathbb{E}\left[\exp\left\{\lambda(Z - \mathbb{E}[Z])\right\}\right] &= \mathbb{E}\left[\prod_{i=1}^n \exp\left\{\lambda(X_i - \mathbb{E}[X_i])\right\}\right] && (3.19)\\
&= \prod_{i=1}^n \mathbb{E}\left[\exp\left\{\lambda(X_i - \mathbb{E}[X_i])\right\}\right] && (3.20)\\
&\leq \prod_{i=1}^n \exp\left(\frac{\lambda^2\sigma_i^2}{2}\right) && (3.21)\\
&= \exp\left(\frac{\lambda^2\sum_{i=1}^n \sigma_i^2}{2}\right), && (3.22)
\end{aligned}$$

so Z is sub-Gaussian with variance proxy $\sum_{i=1}^n \sigma_i^2$. The tail bound then follows immediately from (3.10). □

The proof above demonstrates the value of the moment generating functions of sub-Gaussian random variables: they factorize conveniently when dealing with sums of independent random variables.

3.3.1 Examples of sub-Gaussian random variables

We now provide several examples of classes of random variables that are sub-Gaussian, some of which will appear repeatedly throughout the remainder of the course.

Example 3.8 (Rademacher random variables)**.** A *Rademacher random variable* ϵ takes a value of 1 with probability 1/2 and a value of -1 with probability 1/2. To see that ϵ is 1-sub-Gaussian, we follow Example 2.3 in [Wai19] and upper bound the moment generating function of ϵ by way of a power series expansion of $\exp(\cdot)$:

$$\begin{aligned}
\mathbb{E}[\exp(\lambda\epsilon)] &= \frac{1}{2}\left\{\exp(-\lambda) + \exp(\lambda)\right\} && (3.23)\\
&= \frac{1}{2}\left\{\sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!} + \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\right\} && (3.24)\\
&= \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k)!} && \text{(for odd } k\text{, } (-\lambda)^k + \lambda^k = 0) \quad (3.25)\\
&\leq 1 + \sum_{k=1}^{\infty}\frac{\left(\lambda^2\right)^k}{2^k k!} && (2^k k! \text{ is every other term of } (2k)!) \quad (3.26)\\
&= \exp(\lambda^2/2), && (3.27)
\end{aligned}$$

which is exactly the moment generating function bound (3.8) required for 1-sub-Gaussianity.

Example 3.9 (Random variables with bounded distance to mean)**.** Suppose a random variable X satisfies $|X - \mathbb{E}[X]| \leq M$ almost surely for some constant M. Then X is $O(M)$-sub-Gaussian.

We now provide an even more general class of sub-Gaussian random variables that subsume the random variables in Example 3.9:

Example 3.10 (Bounded random variables)**.** If X is a random variable such that $a \leq X \leq b$ almost surely for some constants $a, b \in \mathbb{R}$, then

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

i.e., X is sub-Gaussian with variance proxy $(b-a)^2/4$. (We will prove this in Question 2(a) of Homework 1.) Note that combining the $(b-a)/2$-sub-Gaussianity of i.i.d. bounded random variables $X_1, \ldots, X_n$ and Theorem 3.7 yields a proof of Hoeffding's inequality introduced in Lecture 2.

Example 3.11 (Gaussian random variables)**.** If X is Gaussian with variance σ^2, then X satisfies (3.8) and (3.10) with equality. In this special case, the variance and the variance proxy are the same.

3.4 Concentrations of functions of random variables

We now introduce some important inequalities related to the second of our two goals, namely showing that for independent $X_1, \ldots, X_n$ and certain functions f, $f(X_1, \ldots, X_n)$ concentrates around $\mathbb{E}[f(X_1, \ldots, X_n)]$.

Theorem 3.12 (McDiarmid's inequality)**.** *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the* bounded difference condition*: there exist constants $c_1, \ldots, c_n \in \mathbb{R}$ such that for all real numbers $x_1, \ldots, x_n$ and x_i',*

$$|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i. \tag{3.28}$$

(Intuitively, (3.28) *states that f is not overly sensitive to arbitrary changes in a single coordinate.) Then, for any independent random variables $X_1, \ldots, X_n$,*

$$\Pr\left[f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \geq t\right] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right). \tag{3.29}$$

Moreover, $f(X_1, \ldots, X_n)$ is $O\left(\sqrt{\sum_{i=1}^n c_i^2}\right)$-sub-Gaussian.

Remark 3.13. Note that McDiarmid's inequality is a generalization of Hoeffding's inequality with $f(x_1, \ldots, x_n) = \sum_{i=1}^n \min\{\max\{x_i, b\}, a\}$.

Proof. See the proof of Corollary 2.21 in [Wai19], which relies on the Azuma-Hoeffding inequality for martingale difference sequences. □

A more general version of McDiarmid's inequality comes from Theorem 3.18 in [vH16]. The setup for this theorem requires defining the *one-sided differences* of a function $f : \mathbb{R}^n \to \mathbb{R}$:

$$D_i^- f(x) = f(x_1, \ldots, x_n) - \inf_z f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n) \tag{3.30}$$

$$D_i^+ f(x) = \sup_z f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_n). \tag{3.31}$$

These two quantities are functions of $x \in \mathbb{R}^n$, and hence can be interpreted as describing the sensitivity of f *at a particular point.* (Contrast this with the bounded difference condition (3.28), which bounds the sensitivity of f universally over all points.) For convenience, define

$$d^+ = \left\| \sum_{i=1}^n |D_i^+ f|^2 \right\|_\infty = \sup_{x_1, \ldots, x_n} \sum_{i=1}^n [|D_i^+ f(x_1, \ldots, x_n)]^2 \tag{3.32}$$

$$d^- = \left\| \sum_{i=1}^n |D_i^- f|^2 \right\|_\infty = \sup_{x_1, \ldots, x_n} \sum_{i=1}^n [D_i^- f(x_1, \ldots, x_n)]^2. \tag{3.33}$$

Theorem 3.14 (Bounded difference inequality)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$, and let $X_1, \ldots, X_n$ be independent random variables. Then, for all $t \geq 0$,*

$$\Pr[f(X_1, \ldots, X_n) \geq \mathbb{E}[f(X_1, \ldots, X_n)] + t] \leq \exp\left(-\frac{t^2}{4d^-}\right) \tag{3.34}$$

$$\Pr[f(X_1, \ldots, X_n) \leq \mathbb{E}[f(X_1, \ldots, X_n)] - t] \leq \exp\left(-\frac{t^2}{4d^+}\right). \tag{3.35}$$

Proof. See Theorem 3.18 in [vH16]. □

3.5 Bounds for Gaussian random variables

Unfortunately, the bounded difference condition (3.28) is often only satisfied by bounded random variables or a bounded function. To get similar concentration inequalities for unbounded random variables, we need some other special conditions. The following inequalities assume that the random variables have the standard normal distribution.

Theorem 3.15 (Gaussian Poincaré inequality, Corollary 2.27 in [vH16]). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be smooth. If $X_1, \ldots, X_n$ are independently sampled from $\mathcal{N}(0,1)$, then*

$$\mathrm{Var}(f(X_1, \ldots, X_n)) \leq \mathbb{E}\left[\|\nabla f(X_1, \ldots, X_n)\|_2^2\right]. \tag{3.36}$$

Before introducing the next theorem, we recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is *L-Lipschitz* with respect to the ℓ_2-norm if there exists a non-negative constant $L \in \mathbb{R}$ such that for all $x, y \in \mathbb{R}^n$,

$$|f(x) - f(y)| \leq L\|x - y\|_2. \tag{3.37}$$

We emphasize that L is universal for all points in $\mathbb{R}^n$.

Theorem 3.16 (Theorem 2.26 in [Wai19]). *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz with respect to Euclidean distance, and let $X = (X_1, \ldots, X_n)$, where $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Then for all $t \in \mathbb{R}$,*

$$\Pr[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2\exp\left(-\frac{t^2}{2L^2}\right). \tag{3.38}$$

In particular, $f(X)$ is sub-Gaussian.

Bibliography

[Ver18] Roman Vershynin, *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge university press, 2018.

[vH16] Ramon van Handel, *Probability in high dimension: Apc 550 lecture notes*, December 2016.

[Wai19] Martin J Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48, Cambridge University Press, 2019.