IDS.160 – Mathematical Statistics: A Non-Asymptotic Approach

Lecturer: P Rigollet
Scribe: Andy Haupt, David Hughes

Lecture 4
Feb. 13, 2020

In the last lecture, we studied sub-exponential random variables. A random variable is sub-exponential if it is centered ($\mathbb{E}[X] = 0$) and its tail satisfies

$$\mathbb{P}[|X| > t] \leq 2e^{-ct}.$$

We also defined the ψ_1-norm

$$\|X\|_{\psi_1} = \inf\{t > 0 : \exp(|X|/t) \leq 2\}$$

and showed that centered random variables with finite ψ_1-norm are sub-exponential. Next, we proved Bernstein's inequality, which provides a tail bound for sums of independent, sub-exponential random variables. More specifically, we proved that if $X_1, X_2, \ldots, X_n$ are independent and sub-exponential, then

$$\mathbb{P}[|\bar{X}_n| > t] \leq 2\exp\left(-Cn\left(\frac{t^2}{\bar{\sigma}^2} \wedge \frac{t}{\sigma_{\max}}\right)\right) \tag{0.1}$$

where $\bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \|X_i\|_{\psi_1}^2$ and $\sigma_{\max} = \max_{i=1,2,\ldots,n} \|X_i\|_{\psi_1}$. The minimum in the exponential shows a change in the behavior of the tails for small deviations versus larger deviations (see Figure 1).

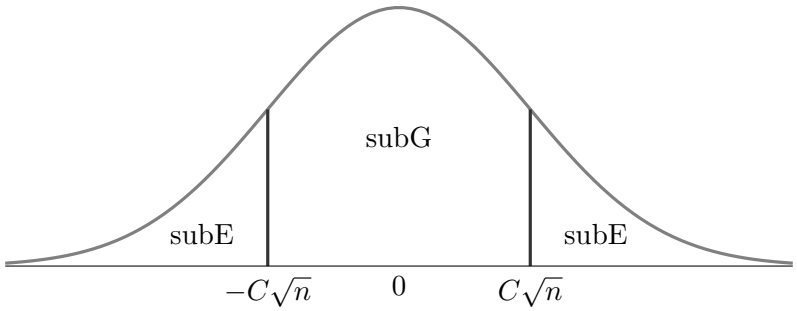

Figure 1: Probability density function of $\sqrt{n}\bar{X}_n$ for sub-exponential random variables $X_1, X_2, \ldots, X_n$. Bounds on the tails are sub-Gaussian for small deviations, but sub-exponential for $t > C\sqrt{n}$.

Alternatively, by setting the bound above equal to some δ and solving for t, Bernstein's inequality tells us that with probability at least $1 - \delta$

$$|\bar{X}_n| \lesssim \frac{\sigma_{\max}}{n}\log\left(\frac{2}{\delta}\right) + \frac{\bar{\sigma}}{\sqrt{n}}\sqrt{\log\left(\frac{2}{\delta}\right)}$$

This uses $\lesssim$ as new notation. For two sequences $(a_n)_{n\in\mathbb{N}}, (b_n)_{n\in\mathbb{N}}$, we denote by $a_n \lesssim b_n$ the fact that $a_n \in O(b_n)$.

Goals: Today, we will first give a more familiar formulation of the Bernstein inequality, and then apply these results to classification and mean estimation.

1. THE BERNSTEIN CONDITION

Definition: A centered random variable X satisfies the *Bernstein condition* with parameter $b > 0$ if its k-th moment satisfies the bound

$$\mathbb{E}[|X|^k] \leq \frac{\mathrm{var}(X)}{2} k! b^{k-2}. \tag{BC(b)}$$

Theorem: If $X_1, \ldots, X_n$ are independent, centered random variables that satisfy BC(b_i), for $i = 1, \ldots, n$, then

$$\mathbb{P}[|\bar{X}_n| > t] \leq 2 \exp\left(-\frac{nt^2}{2(\bar{\sigma}^2 + b_{\max} t)} \right),$$

where $\bar{\sigma}^2 := \frac{1}{n} \sum_{i=1}^n \mathrm{var}(X_i)$ and $b_{\max} := \max_{i=1,\ldots,n}(b_i)$.

Proof. By the Chernoff bound, we have

$$\mathbb{P}[\bar{X}_n > t] \leq \prod_{i=1}^n \mathbb{E}[e^{sX_i}] e^{-nst}. \tag{1.2}$$

We can bound the moment on the right hand side using using the power series expansion for the exponential:

$$\begin{aligned}
\mathbb{E}[e^{sX_i}] &= \sum_{k=0}^{\infty} \frac{|s|^k \mathbb{E}[|X_i|^k]}{k!} \\
&\leq 1 + \frac{s^2 \mathrm{var}(X_i)}{2} + \sum_{k \geq 3} \frac{|s|^k}{k!} \frac{\mathrm{var}(X_i)}{2} k! b_i^{k-2} \\
&= 1 + \frac{s^2 \mathrm{var}(X_i)}{2} + \frac{s^2 \mathrm{var}(X_i)}{2} \sum_{k \geq 3} (|s| b_i)^{k-2} \\
&= 1 + \frac{s^2 \mathrm{var}(X_i)}{2} \sum_{k \geq 2} (|s| b_i)^{k-2},
\end{aligned}$$

where the inequality in the second line follows from the Bernstein condition. Re-indexing

and evaluating the sum gives

$$\begin{aligned}\mathbb{E}[e^{sX_i}] &\leq 1 + \frac{s^2\mathrm{var}(X_i)}{2}\sum_{k\geq 0}(|s|b_i)^k \\ &= 1 + \left(\frac{1}{1-|s|b_i}\right)\frac{s^2\mathrm{var}(X_i)}{2}, \qquad \text{for } |s| < \frac{1}{b_i} \\ &\leq \exp\left(\frac{s^2\mathrm{var}(X_i)}{2(1-|s|b_i)}\right).\end{aligned}$$

The final line uses the inequality $1 + x \leq e^x$. Combining this result with (1.2), we get

$$\mathbb{P}[\bar{X}_n > t] \leq \exp\left(\frac{s^2 n\bar{\sigma}^2}{2(1-|s|b_{\max})} - nst\right)$$

for any $|s| < \frac{1}{b_{\max}}$. Plugging in $s = \frac{t}{\bar{\sigma}^2 + b_{\max}t}$, which satisfies $|s| < \frac{1}{b_{\max}}$, the argument of the exponential reads

$$\begin{aligned}\frac{t^2}{(\bar{\sigma}^2 + b_{\max}t)^2}\frac{n\bar{\sigma}^2}{2(1-|s|b_{\max})} - \frac{nt^2}{\bar{\sigma}^2 + b_{\max}t} &= \frac{t^2}{\bar{\sigma}^2 + b_{\max}t}\left(\frac{1}{\bar{\sigma}^2 + b_{\max}t}\frac{n\bar{\sigma}^2}{2(1-\frac{tb_{\max}}{\bar{\sigma}^2+b_{\max}t})} - n\right) \\ &= \frac{t^2}{\bar{\sigma}^2 + b_{\max}t}\left(\frac{n\bar{\sigma}^2}{2\bar{\sigma}^2} - n\right) \\ &= -\frac{nt^2}{2(\bar{\sigma}^2 + b_{\max}t)},\end{aligned}$$

so that the bound simplifies to the right-hand side from the theorem. □

Note that in the proof we have shown that the moment-generating function of a random variable satisfying the Bernstein condition with parameter b is bounded by

$$\mathbb{E}[e^{sX}] \leq \exp\left(\frac{s^2\mathrm{var}(X)}{2(1-|s|b)}\right)$$

whenever $|s| < \frac{1}{b}$. This implies that these variable are sub-exponential, with parameter

$$2b \vee \sqrt{\mathrm{var}(X)}$$

by our definition of sub-exponential random variables. In other treatments of sub-exponentiality, this maximum is not necessary, as separate parameters are used to control the variance in the bound on the moment-generating function, and the range over which the bound holds. We do not cover these topics in this course.

Lemma: If $|X| \leq B$ and X is centered, then X satisfies (BC(b)) with parameter $b = B/3$.

Proof. Observe that

$$\mathbb{E}[|X|^k] \leq \mathbb{E}[|X|^2|X|^{k-2}] \leq \mathbb{E}[|X|^2 B^{k-2}] = \mathrm{var}(X)B^{k-2}$$

It remains to check that the Bernstein condition holds, that is

$$B^{k-2} \leq \frac{k!}{2}\left(\frac{B}{3}\right)^{k-2} \iff 3^{k-2} \leq \frac{k!}{2}$$

the above is clearly true for $k = 3$, and since the left-hand side increases by a factor of 3, while the right-hand increases by a factor of k, it is therefore also true for all $k \geq 3$. $\square$

Note that if the random variable is symmetric, i.e. $X \stackrel{\mathrm{d}}{=} -X$, then we can strengthen the statement by replacing the 3 with a $\sqrt{12}$. This is possible since $E[|X|^3] = 0$, so we need only control terms of order $k \geq 4$ (the proof of this is left as an exercise).

We can combine the last lemma with Bernstein's inequality to provide a tail bound for sums of bounded random variables of the form

$$\mathbb{P}\left(|\bar{X}_n| > t\right) \leq 2\exp\left(-\frac{nt^2}{2(\sigma^2 + \frac{1}{3}Bt)}\right), \quad \text{if } \mathrm{var}(X_i) \leq \sigma^2$$

The above tail bound implies that, with probability at least $1 - \delta$,

$$|\bar{X}_n| \lesssim \frac{B}{n}\log\left(\frac{1}{\delta}\right) + \frac{\sigma}{\sqrt{n}}\sqrt{\log\left(\frac{1}{\delta}\right)}. \tag{1.3}$$

The term $\frac{\sigma}{\sqrt{n}}$ usually dominates on the right-hand side; however, when σ is very small, the $\frac{B}{n}$-term will dominate.

2. ORLICZ NORM, BOUNDEDNESS AND VARIANCE

We now consider the relationship between random variables being a.s. bounded, boundedness of its ψ_1-norm and the variance of a random variable.

Boundedness implies finite Orlicz norm First, if $|X| \leq B$ holds almost surely, then

$$\mathbb{E}[e^{\frac{|X|}{t}}] \leq e^{\frac{B}{t}} \leq 2$$

holds if and only if $t \geq \frac{B}{\log(2)}$. Hence $\|X\|_{\psi_1} \leq cB$, so that bounded random variables have bounded ψ_1-norm. (This is not surprising given that we know bounded random variables are sub-Gaussian.)

Variance does not control Orlicz norm On the other hand, in general, we cannot guarantee existence of $C < \infty$ such that $\|X\|_{\psi_1}^2 \leq C\mathrm{var}(X)$. Take as an example

$$X_n = \begin{cases} \pm 1 & \text{with probability } \frac{1}{n} \\ 0 & \text{with probability } 1 - \frac{2}{n}. \end{cases}$$

Then, $\mathrm{var}(X_n) = \mathbb{E}[X_n^2] = \frac{2}{n}$. Furthermore, as

$$\mathbb{E}[e^{\frac{|X|}{t}}] = 1 - \frac{2}{n} + \frac{2}{n}e^{\frac{1}{t}} \leq 2 \iff \frac{n}{2} + 1 \geq e^{\frac{1}{t}} \iff t \geq \frac{1}{\log(\frac{n}{2}+1)} = \|X\|_{\psi_1},$$

we get that the ratio of the two sides diverges:

$$\frac{\|X_n\|_{\psi_1}^2}{\mathrm{var}(X_n)} = \frac{n}{2\log(\frac{n+2}{2})} \xrightarrow[n\to\infty]{} \infty.$$

Hence, the ψ_1-norm does not capture variance and it could be that the variance of a random variable becomes very small, even as the ψ_1-norm remains large.

3. APPLICATIONS OF BERNSTEIN'S INEQUALITY

3.1 Classification

Consider the problem of evaluating the performance of the *classifier*

$$f\colon \mathcal{X} \to \{-1, 1\},$$

which predicts labels $Y \in \{-1, 1\}$ given a set of features $X \in \mathcal{X}$. We evaluate the classifier on a set of labelled test examples $(X_1, Y_1), \ldots, (X_n, Y_n)$ using the *indicator loss* function $Z_i = \mathbb{1}(f(X_i) \neq Y_i)$. Observe that $Z_i \sim \mathsf{Ber}(p)$, where $p = \mathbb{E}[f(X) \neq Y]$ is the expected error rate, or *classification error*.

In this case, Hoeffding's lemma applied to the random variable $Z_i - p$ (which is mean zero and is bounded in the interval $[-p, 1-p]$) implies that $(Z_i - p) \sim \mathsf{subG}\left(\frac{1}{4}\right)$. Hoeffding's inequality then gives

$$\mathbb{P}[|\bar{Z}_n - p| > t] \leq 2\exp\left(-2nt^2\right)$$

so that

$$|\bar{Z}_n - p| = \left|\frac{1}{n}\sum_{i=1}^{n}(Z_i - p)\right| \leq \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}$$

with probability $1 - \delta$.

Now imagine that $\bar{Z}_n = 0$, i.e. the classifier produces no error when tested against our sample. In this case, how can we strengthen our bound on the classification error rate p?

Since Z_i is Bernoulli, $\mathrm{var}(Z_i) = p(1-p) \leq p$, and $|Z_i| \leq B$ for $B = \max\{p, 1-p\} \leq 1$. Using Bernstein's inequality for bounded random variables (1.3) and applying $\bar{Z}_n = 0$, we see that the bound on the classification error:

$$p \lesssim \frac{\log(\frac{1}{\delta})}{n} + \sqrt{\frac{p\log(\frac{1}{\delta})}{n}}$$

holds with probability $1 - \delta$. Plugging in $\delta = \frac{1}{100}$ (any other constant would work as well), we get with 99% accuracy that

$$p \lesssim \sqrt{\frac{p}{n}} + \frac{1}{n}. \tag{3.4}$$

We can strengthen this result using a recursive argument: If $p \lesssim \frac{1}{n}$, then (3.4) can be strengthened to $p \lesssim \frac{1}{n}$. Otherwise, for large enough n, we get $p \geq \frac{c}{n}$ for a $c \gg 0$ to be chosen later. The latter is equivalent to $\frac{1}{n} \leq \frac{p}{c}$. Substituting this into (3.4), we get with high probability that $p \leq C_1 p + \frac{C_2}{n}$, where we can choose c large enough to have $C_1 < 1$. Rearranging yields also in this case $p \lesssim \frac{1}{n}$. So when $\bar{Z}_n = 0$ we only need on the order of $\frac{1}{n}$ observations to test the classification error rate.

3.2 Mean Estimation

We first consider the isotropic case $X_1, X_2, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2 I_d)$. Then, by standard properties of the multivariate normal distribution, we get for $Z \sim \mathcal{N}(0, I_d)$

$$\frac{n\|\bar{X}_n - \mu\|_2^2}{\sigma^2} \overset{\text{d}}{=} \|Z\|_2^2$$

which is χ^2_d-distributed. This implies that

$$\mathbb{E}[\|\bar{X}_n - \mu\|^2] = \frac{d\sigma^2}{n}.$$

Furthermore, $\|Z_i^2\|_{\psi_1} = \|Z_i\|_{\psi_2}^2 \leq C$, as Z_i is sub-Gaussian. Using Bernstein's inequality, we get the sub-exponential tail bound

$$\mathbb{P}\left[\left|\frac{1}{d}\sum_{i=1}^{d}(Z_i^2 - 1)\right| > t\right] \leq 2\exp(-cd(t^2 \wedge t)) =: \delta,$$

where $c > 0$ is another constant. Hence, with probability $1 - \delta$,

$$\left|\frac{1}{d}\sum_{i=1}^{d}(Z_i^2 - 1)\right| \lesssim \sqrt{\frac{\log(\frac{2}{\delta})}{d}} + \frac{\log(\frac{2}{\delta})}{d}.$$

Re-transforming to the X_i, we get

$$\left|\|\bar{X}_n - \mu\|^2 - \frac{\sigma^2 d}{n}\right| \lesssim \frac{\sigma^2}{n}\log\left(\frac{2}{\delta}\right) + \frac{\sigma^2}{n}\sqrt{d\log\left(\frac{2}{\delta}\right)}.$$

As a second application, consider the non-isotropic case $X_1, \dots, X_n \sim \mathcal{N}(\mu, \Sigma)$, where we assume for simplicity that $\mu = 0$. Then we have for a spectral decomposition $\Sigma = U\Lambda U^\top$ that

$$n\|\bar{X}_n\|^2 \overset{\text{d}}{=} n\|\Sigma^{\frac{1}{2}}\bar{Z}_n\|^2 \overset{\text{d}}{=} \|U\Lambda^{\frac{1}{2}}U^\top Z\|^2 \overset{\text{d}}{=} \|\Lambda^{\frac{1}{2}}Z\|^2 \overset{\text{d}}{=} \sum_{i=1}^{d}\lambda_i Z_i^2$$

From the definition of the ψ_1-norm, $\|\lambda_i(Z_i^2 - 1)\|_{\psi_1} = \lambda_i\|(Z_i^2 - 1)\|_{\psi_1} \leq C\lambda_i$ since $(Z_i^2 - 1)$ is sub-exponential and hence has finite norm.

Applying the version of Bernstein's inequality introduced last lecture, we get

$$\mathbb{P}\left[\frac{1}{d}\left|\sum_{i=1}^{d}\lambda_i(Z_i^2 - 1)\right| > t\right] \leq 2\exp\left(-Cd\left(\frac{t^2}{\frac{1}{d}\sum_{i=1}^{d}\lambda_i^2} \wedge \frac{t}{\lambda_{\max}}\right)\right)$$

Re-transforming to X_i, we get

$$\left|\|\bar{X}_n - \mu\|_2^2 - \frac{\operatorname{Tr}(\Sigma)}{n}\right| \lesssim \frac{\|\Sigma\|_{\text{op}}}{n}\log\left(\frac{2}{\delta}\right) + \frac{\|\Sigma\|_{\text{F}}}{n}\sqrt{\log\left(\frac{2}{\delta}\right)}.$$

where we recall that $\|\Sigma\|_{\text{op}} = \lambda_{\max}$ (operator norm) and $\|\Sigma\|_{\text{F}}^2 = \sum_{i=1}^{d}\lambda_i^2$ (Frobenius norm), and hence, the bound depends non-trivially on the spectrum of the covariance matrix.

Summary: In this lecture, we gave a different formulation of Bernstein's inequality based on the so-called *Bernstein condition*. In addition, we discovered applications of Hoeffding's and Bernstein's inequality to classification and mean estimation.