Bernstein's Inequality and the Johnson–Lindenstrauss Lemma

Lecturer: Yun Wei *Scribe: Shirou Zeng*

1 A Sharper Concentration Inequality under Bernstein's Condition

Recall from Topic 3 that a random variable X with mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \mathrm{Var}(X)$ is said to satisfy *Bernstein's condition with parameter* $b \geq 0$ if

$$\left|\mathbb{E}\left[(X-\mu)^k\right]\right| \leq \frac{1}{2}\, k!\, \sigma^2 b^{k-2}, \qquad \text{for all } k = 2, 3, \dots. \tag{1}$$

Topic 3 showed that (1) implies that X is $(\sqrt{2}\,\sigma, 2b)$-sub-exponential. We now redo this computation while keeping track of the full range of λ, which yields a sharper bound on the moment generating function and, via the Chernoff bound, a genuinely two-sided *Bernstein-type* concentration inequality.

Proposition 1.1 (Bernstein-type bound)**.** *Suppose X satisfies Bernstein's condition* (1) *with parameter b, mean μ, and variance σ^2. Then*

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq \exp\left(\frac{\lambda^2\sigma^2/2}{1-b\,|\lambda|}\right) \qquad \textit{for all } |\lambda| < \frac{1}{b}, \tag{2}$$

and, moreover, the concentration inequality

$$\mathbb{P}\left(|X-\mu| > t\right) \leq 2\exp\left(-\frac{t^2}{2(\sigma^2+bt)}\right) \qquad \textit{for all } t \geq 0. \tag{3}$$

Proof. **Proof of** (2)**.** Without loss of generality assume $\mu = 0$ (replace X by $X-\mu$). Expanding the exponential in a Taylor series and using $\mathbb{E}[X] = 0$,

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda X}\right] &= \mathbb{E}\left[\sum_{k=0}^{\infty}\frac{\lambda^k X^k}{k!}\right] = 1 + \lambda\mathbb{E}[X] + \sum_{k=2}^{\infty}\frac{\lambda^k\mathbb{E}\left[X^k\right]}{k!} = 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty}\frac{\lambda^k\mathbb{E}\left[X^k\right]}{k!} \\
&\leq 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty}\frac{|\lambda|^k\left|\mathbb{E}\left[X^k\right]\right|}{k!} \leq 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty}\frac{|\lambda|^k\cdot\frac{1}{2}k!\sigma^2 b^{k-2}}{k!} \qquad (4) \\
&= 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\sigma^2}{2}\sum_{k=3}^{\infty}|\lambda|^k b^{k-2} = 1 + \frac{\lambda^2\sigma^2}{2}\left(1 + \sum_{k=3}^{\infty}(|\lambda|\,b)^{k-2}\right) = 1 + \frac{\lambda^2\sigma^2}{2}\sum_{s=0}^{\infty}(|\lambda|\,b)^s, \qquad (5)
\end{aligned}$$

where the interchange of sum and expectation in the first line is justified because (1) guarantees absolute convergence of the series, and the last equality reindexes $s = k-2$. For $|\lambda| < 1/b$ the geometric series in (5) sums to $1/(1-b\,|\lambda|)$, so that

$$\mathbb{E}\left[e^{\lambda X}\right] \leq 1 + \frac{\lambda^2\sigma^2/2}{1-b\,|\lambda|} \leq \exp\left(\frac{\lambda^2\sigma^2/2}{1-b\,|\lambda|}\right), \qquad |\lambda| < \frac{1}{b}, \tag{6}$$

using $1 + x \leq e^x$ in the last step. This proves (2).

Remark 1.2. Restricting further to $|\lambda| < \frac{1}{2b}$ gives $b\,|\lambda| < \frac{1}{2}$, hence $1 - b\,|\lambda| > \frac{1}{2}$ and

$$\mathbb{E}\left[e^{\lambda X}\right] \leq \exp\left(\frac{\lambda^2\sigma^2/2}{1/2}\right) = e^{\lambda^2\sigma^2} = e^{\frac{\lambda^2(\sqrt{2}\sigma)^2}{2}}, \qquad |\lambda| < \frac{1}{2b},$$

which recovers the Topic 3 conclusion that X is $(\sqrt{2}\sigma, 2b)$-sub-exponential. Bound (2) is sharper: it holds on the full range $|\lambda| < 1/b$ instead of only $|\lambda| < 1/(2b)$, and this extra room is exactly what lets us optimize the Chernoff bound below.

Proof of (3)**.** By the Chernoff bound and (2), for any $t \geq 0$,

$$\log \mathbb{P}\left(X - \mu \geq t\right) \leq \inf_{0 \leq \lambda < 1/b} \left\{\log \mathbb{E}\left[e^{\lambda(X-\mu)}\right] - \lambda t\right\} \leq \inf_{0 \leq \lambda < 1/b} \underbrace{\left\{\frac{\lambda^2\sigma^2}{2(1-b\lambda)} - \lambda t\right\}}_{=:\,\theta(\lambda)}. \tag{7}$$

Rather than optimizing θ exactly, it suffices to evaluate it at the specific point

$$\lambda^* := \frac{t}{\sigma^2 + bt}.$$

Since $t, \sigma^2, b \geq 0$ we have $\lambda^* \geq 0$, and $\lambda^* < 1/b \iff bt < \sigma^2 + bt \iff \sigma^2 > 0$, which holds whenever X is non-degenerate (if $\sigma^2 = 0$ then $X \equiv \mu$ a.s. and (3) is trivial). At $\lambda = \lambda^*$,

$$1 - b\lambda^* = 1 - \frac{bt}{\sigma^2 + bt} = \frac{\sigma^2}{\sigma^2 + bt},$$

so that

$$\theta(\lambda^*) = \frac{\sigma^2(\lambda^*)^2(\sigma^2 + bt)}{2\sigma^2} - \lambda^* t = \frac{(\lambda^*)^2(\sigma^2 + bt)}{2} - \lambda^* t = \frac{t^2}{2(\sigma^2 + bt)} - \frac{t^2}{\sigma^2 + bt} = -\frac{t^2}{2(\sigma^2 + bt)}.$$

Combining with (7) (using that λ^* is feasible),

$$\mathbb{P}\left(X - \mu \geq t\right) \leq \exp\left(-\frac{t^2}{2(\sigma^2 + bt)}\right), \qquad t \geq 0. \tag{8}$$

Finally, $-X$ also satisfies Bernstein's condition with the same (σ^2, b), since $\left|\mathbb{E}[(-(X-\mu))^k]\right| = \left|\mathbb{E}[(X-\mu)^k]\right|$ and $\mathrm{Var}(-X) = \sigma^2$. Applying (8) to $-X$ gives $\mathbb{P}(X - \mu \leq -t) \leq \exp(-t^2/(2(\sigma^2 + bt)))$ as well, and adding the two one-sided bounds yields (3). $\square$

Exercise 1.3. Compare the bound (3) with the bound one would obtain by plugging the (weaker) parameters $(\sqrt{2}\sigma, 2b)$ derived in the Remark above directly into the general sub-exponential tail bound of Topic 3. Show that (3) is never worse, and can be substantially tighter when $bt \ll \sigma^2$.

2 Sums of Independent Sub-exponential Random Variables

Recall (Topic 3) that X with mean μ is *sub-exponential with parameters* (ν, α) if $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\nu^2\lambda^2/2}$ for all $|\lambda| < 1/\alpha$, and that in this case

$$\mathbb{P}\left(|X - \mu| > t\right) \leq \begin{cases} 2e^{-t^2/(2\nu^2)}, & 0 \leq t \leq \nu^2/\alpha, \\ 2e^{-t/(2\alpha)}, & t > \nu^2/\alpha. \end{cases} \tag{9}$$

Lemma 2.1. *Let $X_1, \dots, X_n$ be independent random variables, with X_i sub-exponential with parameters (ν_i, d_i). Then $S_n := \sum_{i=1}^n X_i$ is sub-exponential with parameters*

$$\left(\sqrt{\sum_{i=1}^n \nu_i^2},\ \max_{1 \le i \le n} d_i \right).$$

Proof. Write $d := \max_{1 \le i \le n} d_i$ and $V^2 := \sum_{i=1}^n \nu_i^2$. For any $|\lambda| < 1/d = \min_{1 \le i \le n}(1/d_i)$, we have $|\lambda| < 1/d_i$ for every i, so by independence,

$$\mathbb{E}\left[e^{\lambda(S_n - \mathbb{E}[S_n])}\right] = \mathbb{E}\left[e^{\lambda \sum_{i=1}^n (X_i - \mathbb{E}[X_i])}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{\lambda(X_i - \mathbb{E}[X_i])}\right] \le \prod_{i=1}^n e^{\nu_i^2 \lambda^2/2} = e^{\lambda^2 V^2/2}.$$

Hence S_n is (V, d)-sub-exponential. □

By (9) applied with $(\nu, \alpha) = (V, d)$, we obtain the tail bound

$$\mathbb{P}\left(|S_n - \mathbb{E}[S_n]| > t\right) \le \begin{cases} 2e^{-t^2/(2V^2)}, & 0 \le t \le V^2/d, \\ 2e^{-t/(2d)}, & t > V^2/d, \end{cases} \qquad V^2 = \sum_{i=1}^n \nu_i^2, \quad d = \max_{1 \le i \le n} d_i. \tag{10}$$

Example 2.2 (Concentration of χ^2 variables)**.** Let $Z_1, \dots, Z_n$i.i.d. $N(0,1)$ and let $Y := \sum_{i=1}^n Z_i^2 \sim \chi_n^2$, so that $\mathbb{E}[Y] = n$. Recall from Topic 3 that each Z_i^2 has mean 1 and is $(2,4)$-sub-exponential, i.e. $\mathbb{E}[e^{\lambda(Z_i^2 - 1)}] \le e^{2\lambda^2}$ for all $|\lambda| < 1/4$. Applying Lemma 2.1 with $\nu_i = 2$, $d_i = 4$ for every i, Y is $(\sqrt{4n}, 4) = (2\sqrt{n}, 4)$-sub-exponential; here $V^2 = 4n$ and $d = 4$, so the threshold in (10) is $V^2/d = n$. Thus

$$\mathbb{P}\left(|Y - n| \ge s\right) \le \begin{cases} 2e^{-s^2/(8n)}, & 0 \le s \le n, \\ 2e^{-s/8}, & s > n. \end{cases}$$

Substituting $s = nt$ (so $s \le n \iff t \le 1$) gives

$$\mathbb{P}\left(\left|\frac{Y}{n} - 1\right| \ge t\right) \le \begin{cases} 2e^{-nt^2/8}, & 0 \le t \le 1, \\ 2e^{-nt/8}, & t > 1. \end{cases} \tag{11}$$

3 Application: The Johnson–Lindenstrauss Lemma

Fix distinct points $u^1, \dots, u^N \in \mathbb{R}^d$, where d is large. We would like a map $F : \mathbb{R}^d \to \mathbb{R}^m$ with $m \ll d$, so that we only need to store the compressed points $\left\{F(u^1), \dots, F(u^N)\right\}$, while approximately preserving all pairwise (squared) distances: given a distortion parameter $\delta \in (0,1)$, we want

$$1 - \delta \ \le\ \frac{\left\|F(u^i) - F(u^j)\right\|_2^2}{\|u^i - u^j\|_2^2} \ \le\ 1 + \delta, \qquad \text{for all } i \ne j. \tag{12}$$

Construction. Let $X \in \mathbb{R}^{m \times d}$ have i.i.d. $N(0,1)$ entries, and write its rows as $X_1, \dots, X_m \in \mathbb{R}^d$, so that $X_1, \dots, X_m$i.i.d. $N(0, I_d)$. Define the (random) linear map

$$F(u) := \frac{Xu}{\sqrt{m}}, \qquad u \in \mathbb{R}^d.$$

Step 1: distribution of $\|Xu\|_2^2$. Fix $u \in \mathbb{R}^d$, $u \ne 0$. Writing $Xu = (\langle X_1, u\rangle, \dots, \langle X_m, u\rangle)^\top$, for each i and any unit vector $v = u/\|u\|_2$,

$$\langle X_i, v\rangle = \sum_{k=1}^d X_{i,k} v_k \sim N\left(0, \sum_{k=1}^d v_k^2\right) = N(0,1) \qquad (\text{check: } \|v\|_2 = 1), \tag{13}$$

since $X_i \sim N(0, I_d)$ and any linear combination of independent standard normals with coefficient vector of unit ℓ_2-norm is again standard normal. Moreover $\langle X_1, v\rangle, \ldots, \langle X_m, v\rangle$ are independent because $X_1, \ldots, X_m$ are independent. Hence

$$Y := \frac{\|Xu\|_2^2}{\|u\|_2^2} = \sum_{i=1}^{m} \frac{\langle X_i, u\rangle^2}{\|u\|_2^2} = \sum_{i=1}^{m} \langle X_i, v\rangle^2 \sim \chi_m^2. \tag{14}$$

Step 2: concentration for a single vector. Since $\|F(u)\|_2^2 / \|u\|_2^2 = \|Xu\|_2^2 / (m \|u\|_2^2) = Y/m$, applying (11) (with n replaced by m) gives, for any $\delta \in (0, 1)$,

$$\mathbb{P}\left(\frac{\|F(u)\|_2^2}{\|u\|_2^2} \notin (1-\delta, 1+\delta) \right) = \mathbb{P}\left(\left| \frac{Y}{m} - 1 \right| \geq \delta \right) \leq 2e^{-m\delta^2/8}. \tag{15}$$

Step 3: union bound over all pairs. Apply (15) with $u = u^i - u^j$ for each of the $\binom{N}{2}$ unordered pairs $i \neq j$ (assuming $u^i \neq u^j$):

$$\mathbb{P}\left(\frac{\left\|F(u^i) - F(u^j)\right\|_2^2}{\|u^i - u^j\|_2^2} \notin (1-\delta, 1+\delta) \right) \leq 2e^{-m\delta^2/8}.$$

Let

$$I := \bigcup_{i<j} \left\{ \frac{\left\|F(u^i) - F(u^j)\right\|_2^2}{\|u^i - u^j\|_2^2} \notin (1-\delta, 1+\delta) \right\}$$

be the event that (12) fails for at least one pair. By the union bound,

$$\mathbb{P}(I) \leq \sum_{i<j} \mathbb{P}\left(\frac{\left\|F(u^i) - F(u^j)\right\|_2^2}{\|u^i - u^j\|_2^2} \notin (1-\delta, 1+\delta) \right) \leq \binom{N}{2} \cdot 2e^{-m\delta^2/8} = N(N-1)\, e^{-m\delta^2/8}. \tag{16}$$

Step 4: choosing m. Given a target failure probability $\varepsilon \in (0, 1)$, we want $\mathbb{P}(I) < \varepsilon$. By (16), it suffices that

$$N(N-1)e^{-m\delta^2/8} < \varepsilon \quad \iff \quad m > \frac{8}{\delta^2} \log \frac{N(N-1)}{\varepsilon}.$$

Since $N(N-1) \leq N^2$, a cleaner sufficient condition is

$$m > \frac{16}{\delta^2} \log \frac{N}{\varepsilon}. \tag{17}$$

Theorem 3.1 (Johnson–Lindenstrauss Lemma)**.** *Let distinct $u^1, \ldots, u^N \in \mathbb{R}^d$ and let $\delta, \varepsilon \in (0, 1)$. If*

$$m > \frac{16}{\delta^2} \log \frac{N}{\varepsilon},$$

then, with $X \in \mathbb{R}^{m \times d}$ having i.i.d. $N(0, 1)$ entries and $F(u) := Xu/\sqrt{m}$, with probability at least $1 - \varepsilon$ we have

$$1 - \delta \leq \frac{\left\|F(u^i) - F(u^j)\right\|_2^2}{\|u^i - u^j\|_2^2} \leq 1 + \delta \qquad \textit{simultaneously for all } i \neq j.$$

Proof. Immediate from (16) and (17): the choice of m forces $\mathbb{P}(I) < \varepsilon$, i.e. (12) holds for every pair on the complementary event, which has probability at least $1 - \varepsilon$. □

References

[1] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge University Press, 2019.