Statistics 210B Lecture 4 Notes

Daniel Raban

January 27, 2022

1 Bernstein's Inequality, the Johnson-Lindenstass Lemma, and More Concentration Inequalities

1.1 Bernstein condition for sub-exponentiality

A bounded random variable is sub-Gaussian and hence is sub-expoenntial, but we can get a tighter quantitative sub-exponential bound.

Proposition 1.1. *Suppose X has a mean μ and variance σ^2. Suppose that $\mathbb{E}[(X-\mu)^k] \leq \frac{1}{2}k!\sigma^2 b^{k-2}$ for all $k \geq 2$. Then X is $(\sqrt{2}\sigma, 2b)$-sub-exponential.*

Note that the units in this inequality condition make sense. This condition is called the **Bernstein condition**.

Proof. We just need to show that the moment generating function is bounded: Do a Taylor expansion:

$$\begin{aligned}\mathbb{E}[e^{\lambda(X-\mu)}] &= 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \lambda^k \frac{\mathbb{E}[(X-\mu)^k]}{k!} \\ &\leq 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2}\sum_{k=3}(|\lambda|b)^{k-2}\end{aligned}$$

This is a geometric series, so we can simplify it.

$$\begin{aligned}&\leq 1 + \frac{\lambda\sigma^2/2}{1-b|\lambda|} \\ &\leq e^{(\lambda^2\sigma^2/2)/(1-b|\lambda|)}\end{aligned}$$

When $|\lambda| \leq \frac{1}{2b}$,

$$\leq e^{\lambda^2(\sqrt{2}\sigma)^2/2}. \qquad \square$$

Now let X be a random variable with $\mathrm{Var}(X) = \sigma^2$ and $0 \le X \le b$. Then

$$\begin{aligned}\mathbb{E}[|X-\mu|^k] &\le \mathbb{E}[|X-\mu|^2 \cdot b^{k-2}] \\ &= \sigma^2 b^{k-2} \\ &\le \frac{k!}{2}\sigma^2 b^{k-2},\end{aligned}$$

so X is $(\sqrt{2}\sigma, 2b)$-sub-exponential. Last time, we had that X is b-sub-Gaussian. So the sub-exponential tail bound here is stronger in the region where the sub-exponential and sub-Gaussian tail behaviors are similar.

1.2 Bernstein's inequality

Lemma 1.1 (Bernstein's inequality)**.** *Let $\{X_i\}_{i\in[n]}$ be independent with $\mathbb{E}[X_i] = \mu_i$ and X_i (ν_i, α_i)-sub-exponential. Then $\sum_{i=1}^n (X_i - \mu_i)$ is sub exponential with parameters $\nu_* = \sqrt{\sum_{i=1}^n \nu_i^2}$ and $\alpha_* = \max_i \alpha_i$. Moreover,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_i-\mu_i) \ge t\right) \le \begin{cases} e^{-nt^2/(2\nu_*^2)} & t \le \nu_*^2/\alpha_* \\ e^{-nt/(2\alpha_*)} & t > \nu_*^2/\alpha_* \end{cases}$$

Proof.

$$\begin{aligned}\mathbb{E}[e^{\lambda\sum_{i=1}^n (X_i-\mu_i)}] &= \prod_{i=1}^n \mathbb{E}[e^{\lambda(X_i-\mu_i)}] \\ &\le e^{\lambda^2 \sum_{i=1}^n \nu_i^2/2}.\end{aligned}$$

for all $\lambda \le 1/\max_{i\in[n]} \alpha_i$. □

Let $(X_i)_{i\in[n]} \overset{\text{iid}}{\sim} X$ be (ν, b)-sub-sexponential. Then

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_i-\mu_i) \ge t\right) \le e^{-n\min\{\frac{t^2}{2\nu^2}, \frac{t}{2b}\}}.$$

(a) How do we extract the order of $\frac{1}{n}\sum)i = 1^n X_i - \mu$? Set $\delta = \exp(-n\min\{\frac{t^2}{2\nu^2}, \frac{t}{2b}\})$, and solve for t to get

$$t = \max\left\{\nu\sqrt{\frac{2\log(1/\delta)}{n}}, b\frac{2\log(1/\delta)}{n}\right\}.$$

This tells us that

$$\frac{1}{n}\sum_{i=1}^n X_i - \mu \le \max\left\{\nu\sqrt{\frac{2\log(1/\delta))}{n}}, b\frac{2\log(1/\delta)}{m}\right\} \qquad \text{with probability at least } 1-\delta.$$

For small δ, the first term is the dominant term while the second is a *burn-in term.*

(b) How many samples do we need to have $\frac{1}{n}\sum_{i=1}^n X_i - \mu \le t$ with probability $1-\delta$? Set $\delta = \exp(-n\min\{\frac{t^2}{2\nu^2}, \frac{t}{2b}\})$ and solve for n to get

$$n = \max\left\{\frac{2\nu^2}{t^2}\log(1/\delta), \frac{2b}{t}\log(1/\delta)\right\}.$$

When t is small, the first term is dominant, while the second is of smaller order.

Example 1.1. Let X_i be iid with support in $[0,b]$ and $\mathrm{Var}(X_i) \le \nu^2$. We know that X_i is b-sub-Gaussian and (ν, b)-sub-exponential. In order for $|\frac{1}{n}\sum_{i=1}^n X_i - \mu| \le \varepsilon$ with probability $1-\delta$,

$$\mathrm{sG}(1) \implies n \ge \frac{b^2}{\varepsilon^2}\log\left(\frac{1}{\delta}\right),$$

$$\mathrm{sE}(\nu, 1) \implies n \ge \max\left\{\frac{\nu^2}{\varepsilon^2}\log\left(\frac{1}{\delta}\right), \frac{b}{\varepsilon}\log\left(\frac{1}{\delta}\right)\right\}.$$

When $\varepsilon \le b$, $\frac{b}{\varepsilon}\log(\frac{1}{\delta}) \le \frac{b^2}{\varepsilon^2}\log(\frac{1}{\delta})$. So the sub-exponential bound is a stronger bound.

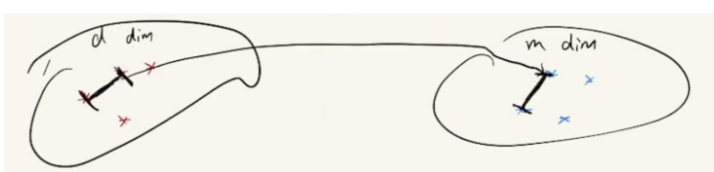
1.3 An application: the Johnson-Lindenstrass Lemma

Let $Y = \sum_{i=1}^n Z_i$ with $Z_i \sim N(0,1)$. Then $Y \sim \chi^2(n)$. Last time, we showed that Z_i^2 is $\mathrm{sE}(2,4)$, so $Y \sim \mathrm{sE}(2\sqrt{n}, 4)$. By Bernstein's inequality,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i^2 - 1\right| \ge t\right) \le 2e^{-nt^2/8} \qquad \forall t \le 1.$$

Here is a problem: Suppose we have $\{u_1, u_2, \dots, u_N\} \subseteq \mathbb{R}^d$ with a high dimension d. Can we find a $F : \mathbb{R}^d \to \mathbb{R}^m$ with some small m such that the distances are preserved? That is, we want

$$1-\delta \le \frac{\|F(u_i) - F(u_j)\|_2^2}{\|u_i - u_j\|_2^2} \le 1+\delta, \qquad \forall i,j \in [N].$$



How small can we make m? The Johnson-Lindenstrass says that we can achieve this by random projection.

Lemma 1.2 (Johnson-Lindenstrass)**.** *Let $X \in \mathbb{R}^{m\times d}$ have entries $X_{i,j} \overset{\text{iid}}{\sim} N(0,1)$, and let $F : \mathbb{R}^d \to \mathbb{R}^m$ be defined as $R(u) = \frac{1}{\sqrt{m}} X \cdot u$. Then for any fixed $\{u_1, \ldots, u_N\} \subseteq \mathbb{R}^d$, as long as $m \gtrsim \frac{1}{\varepsilon^2}\log(\frac{N}{\delta})$, then with probability $1-\delta$, we have*

$$1-\varepsilon \le \frac{\|F(u_i)-F(u_j)\|_2^2}{\|u_i-u_j\|_2^2} \le 1-\varepsilon, \qquad \forall i,j \in [N].$$

Remark 1.1. The dimension that we can reduce to is of order $\log N$, where N is the number of points. So no matter the dimension d, we can always reduce the dimension to order $\log N$.

Proof. Denote $Y_{i,j} = \frac{\|F(u_i)-F(u_j)\|_2^2}{\|u_i-u_j\|_2^2}$. We claim that $Y_{i,j} \sim \chi^2(m)/m$. Then Bernstein's inequality will give

$$\mathbb{P}(|Y_{i,j}-1| \ge t) \le 2e^{-mt^2/\delta} \qquad \forall t \le 1.$$

Using a union bound on all $N(N-1) \le N^2$ pairs $i \ne j$, we get

$$\mathbb{P}\left(\exists i,j \in [N] \text{ s.t.} |Y_{i,j}-1| \ge t\right) \le 2N^2 e^{-mt^2/8} \qquad \forall t \le 1.$$

Setting the right hand side equal to δ, we can solve for m to get

$$m \ge \frac{8}{t^2}\log\left(\frac{2N^2}{\delta}\right) = \frac{C}{t^2}\log\left(\frac{N}{\delta}\right).$$

Now let's verify the claim that $Y_{i,j} = \frac{\|F(u_i)-F(u_j)\|_2^2}{\|u_i-u_j\|_2^2} \sim \chi^2(m)/m$. Note that

$$\frac{1}{\sqrt{m}} X(u_i-u_j) \sim N\left(0, \frac{\|u_i-u_j\|_2^2}{m} I_m\right),$$

which implies that

$$\frac{\|X(u_i-u_j)\|_2^2}{m} \sim \|u_i-u_j\|_2^2 \chi^2(m)/m.$$

This proves the claim. □

Remark 1.2. If we use Markov's inequality instead of Bernstein's inequality, we get a worse bound.

1.4 Equivalent characterizations of sub-exponentiality

Theorem 1.1 (2.13 in HDS, 2.7.1 in HDP[1])**.** *The following statements are equivalent:*

(a)

$$\mathbb{P}(|X| \ge t) \le 2\exp(-t/\kappa_1), \qquad \forall t \ge 0.$$

[1] These two theorems actually say something slightly different.

(b)

$$\|X\|_{L^p} = (\mathbb{E}[|X|^p])^{1/p} \le \kappa_2 p, \qquad \forall p \ge 1.$$

(c)

$$\mathbb{E}[\exp(\lambda|X|)] \le \exp(\kappa_3\lambda) \qquad \forall\lambda \text{ } s.t. \text{ } 0 \le \lambda \le \frac{1}{\kappa_3}.$$

(d)

$$\mathbb{E}[\exp(|X|/\kappa_4)] \le 2.$$

Moreover, if $\mathbb{E}[X] = 0$*, (a)-(d) are equivalent to*

5.

$$\mathbb{E}[\exp(\lambda X)] \le \exp(\lambda^2\kappa_5^2/2) \qquad \forall|\lambda| \le \frac{1}{\kappa_5}.$$

Here, $\kappa_1, \dots, \kappa_5$ *are universal constants.*

We will not give the proof here, but you can check either textbook. Here is an example:

Example 1.2. Let $X_1 \sim \mathrm{sG}(\sigma_1)$ and $X_2 \sim \mathrm{sG}(\sigma_2)$ be not necessarily independent with $\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$. We claim that $X_1X_2 \sim \mathrm{sE}(K\sigma_1\sigma_2, K\sigma_1\sigma_2)$ for some universal K. For this, we can use property (b) above: First rescale X_1 and X_2 for simplicity. Using the Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}\left[\left(\left|\frac{X_1}{\sigma_1}\right|\left|\frac{X_2}{\sigma_2}\right|\right)^p\right] &\le \mathbb{E}\left[\left|\frac{X_1}{\sigma_1}\right|^{2p}\right]^{1/2} \mathbb{E}\left[\left|\frac{X_2}{\sigma_2}\right|^{2p}\right]^{1/2} \\ &= \left\|\frac{X_1}{\sigma_1}\right\|_{L^{2p}}^p \left\|\frac{X_2}{\sigma_2}\right\|_{L^{2p}}^p\end{aligned}$$

By the rescaling, $X_i/\sigma_i \sim \mathrm{sG}(1)$ for $i = 1, 2$. The sub-Gaussian condition says that $\|G\|_{L^{2p}} \le K(2p))^p$ for all p.

$$\begin{aligned}&\le K^p(\sqrt{2p})^p \cdot K^p(\sqrt{2p})^p \\ &= K^{2p}(2p)^p.\end{aligned}$$

This tells us that $\|\frac{X_1}{\sigma_1}\frac{X_2}{\sigma_2}\|_{L^p} \le K^2 2p$ for all p.

1.5 Bennett's inequality

Here is a stronger bound for bounded random variables. Here, we don't require boundedness from below.

Lemma 1.3 (Bennett's inequality)**.** *Let $(X_i)_{i\in[n]}$ be independent, where $X_i - \mathbb{E}[X_i] \le b$ a.s., and $\nu_i^2 := \mathrm{Var}(X_i)$ for all $i \in [n]$. Then*

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \ge t\right) \le \exp\left(-\frac{\sum_{i=1}^n \nu_i^2}{b^2} h\left(\frac{bt}{\sum_{i=1}^n \nu_i^2}\right)\right),$$

where $h(u) = (1+u)\log(1+u) - u$.

Remark 1.3. This has a stronger assumption than Bernstein's inequality and provides a stronger bound than Bernstein's inequality for bounded random variables. However, it doesn't often improve much over Bernstein's inequality.

1.6 Maximal inequality

Lemma 1.4. *Let $(X_i)_{i\in[n]}$ be a sequence of random variables. For any convex, strictly increasing $\psi : \mathbb{R} \to \mathbb{R}_{\ge 0}$, we have*

$$\mathbb{E}\left[\max_{i\in[n]} X_i\right] \le \psi^{-1}\left(\sum_{i=1}^n \mathbb{E}[\psi(X_i)]\right),$$

$$\mathbb{P}\left(\max_{i\in[n]} X_i \ge t\right) \le \sum_{i=1}^n \frac{\mathbb{E}[\psi(X_i)}{\psi(t)}.$$

Proof.

$$\mathbb{E}\left[\max_{i\in[n]} X_i\right] = \mathbb{E}\left[\psi^{-1}\left(\max_{i\in[n]} \psi(X_i)\right)\right]$$

Using Jensen's inequality,

$$\le \psi^{-1}\left(\mathbb{E}\left[\max_{i\in[n]} \psi(X_i)\right]\right)$$

Upper bounding the maximum by the sum,

$$= \psi^{-1}\left(\sum_{i=1}^n \mathbb{E}[\psi(X_i)]\right).$$

□

Example 1.3. For $X_i \sim \mathrm{sG}(\sigma)$, take $\psi(u) = e^{\lambda u}$. Optimizing over λ, we get

$$\mathbb{E}\left[\max_{i\in[n]} X_i\right] \le \sigma\sqrt{2\log(n)}.$$

This gives an important intuition:: n sub-Gausian random variables have maximum of order $\sqrt{\log(n)}$.

1.7 Truncation argument

Here is a very useful technique in research for getting concentration inequalities for random variables which are not sub-Gaussian nor sub-exponential.

Example 1.4. Let $X_i = G_i^4$, where $(G_i)_{i\in[n]} \overset{\text{iid}}{\sim} N(0,1)$. Then $\mathbb{E}[X_i] = \mathbb{E}[G_i^4] = 3$, but $\mathbb{E}[e^{\lambda X_i}]$ doesn't exist. However, we still want to upper bound $\frac{1}{n}\sum_{i=1}^n X_i - 3$.

Here is the technique:

Step 1: Find b_n such that

$$\mathbb{P}\left(\max_{i\in[n]} X_i \geq b_n\right) \leq \frac{\delta}{2}$$

and ε_n such that

$$\mathbb{E}[X_i \mathbb{1}_{\{X_i \geq b_n\}}] \leq \varepsilon_n.$$

Step 2: Apply Hoeffding/Bernstein and get

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_i \mathbb{1}_{\{X_i \leq b_n\}} - \mathbb{E}[X_i \mathbb{1}_{\{X_i \leq b_n\}}]) \leq t_n\right) \geq 1 - \frac{\delta}{2}.$$

Step 3: Combining Steps 1 and 2 implies that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}[X_i] \leq t_n + \varepsilon_n\right) \geq 1 - \delta.$$

As an exercise, figure out b_n, t_n, ε_n as a function of n and δ. The requirement is that $t_n + \varepsilon \sim \widetilde{O}(\frac{1}{\sqrt{n}})$.