18.656 Notes

Lecturer:

Andrew Liu

Spring 2024

Last updated on Thursday 15th February, 2024.

Contents

1 February 8, 2024

1.1 Recap

Definition 1.1
A random variable X with mean $\mu = \mathbb{E}[X]$ is **σ-sub-Gaussian** if $\mathbb{E}[e^{\lambda(X-\mu)}] \le e^{\lambda^2\sigma^2/2}$ for all $\lambda \in \mathbb{R}$.
Equality if and only if $X \sim \mathcal{N}(\mu, \sigma^2)$.

Example 1.2
Given r.v. $X \in [a, b]$, $\mathbb{E}[X] = \mu$. Then, X is sub-Gaussian with $\sigma = (b-a)$.

It turns out that we can also show $\sigma = (b-a)/2$, but we won't show this during lecture, since the technique used to show the weaker result is more interesting.

Proof. Let $\tilde{X}$ be i.i.d. to X. Then,

$$\begin{aligned} \mathbb{E}_X[e^{\lambda(X-\mu)}] &= \mathbb{E}_X[e^{\lambda(X-\mathbb{E}_{\tilde{X}}[\tilde{X}])}] \\ &\le \mathbb{E}_{X,\tilde{X}}[e^{\lambda(X-\tilde{X})}], \end{aligned}$$

by Jensen's inequality. Since $X, \tilde{X}$ are i.i.d., $(X-\tilde{X})$ has a distribution symmetric around 0. Now, we also have that

$$(X-\tilde{X}) \stackrel{\text{dist}}{=} \varepsilon(X-\tilde{X}),$$

where $\varepsilon \in \{\pm 1\}$ with equal probability (also called a **Rademacher** random variable). Therefore,

$$\begin{aligned} \mathbb{E}_{\tilde{X}}[e^{\lambda(X-\mu)}] &\le \mathbb{E}_{\tilde{X},X}[\mathbb{E}_\varepsilon[e^{\lambda\varepsilon(X-\tilde{X})}]] \\ &\le \mathbb{E}_{X,\tilde{X}}[e^{\lambda^2(X-\tilde{X})^2/2}], \end{aligned}$$

since ε is 1-sub-gaussian. Finally, since X is bounded,

$$\mathbb{E}_{X,\tilde{X}}[e^{\lambda^2(X-\tilde{X})^2/2}] \le e^{\lambda^2(b-a)^2/2}.$$

$\square$

1.2 More on sub-Gaussians

Definition 1.3 (Addition property of Gaussians)
Given $X_1 \sim \mathcal{N}(0, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(0, \sigma_2^2)$,

$$X_1 + X_2 \sim \mathcal{N}(0, \sigma_1^2 + \sigma_2^2).$$

Claim 1.4 (Addition property of sub-Gaussians)
Given $X_i \sim \sigma_i$-sub-Gaussian, $i \in \{1, 2\}$, then $X_1 + X_2$ is $\sqrt{\sigma_1^2 + \sigma_2^2}$-sub-Gaussian.

Proof.

$$\begin{aligned}
\mathbb{E}_{X_1, X_2}[e^{\lambda(X_1 + X_2)}] &= \mathbb{E}_{X_1, X_2}[e^{\lambda X_1} e^{\lambda X_2}] \\
&= \mathbb{E}_{X_1}[e^{\lambda X_1}] \mathbb{E}_{X_2}[e^{\lambda X_2}] \\
&\leq e^{\lambda^2 \sigma_1^2 / 2} e^{\lambda^2 \sigma_2^2 / 2} = e^{\lambda^2 (\sigma_1^2 + \sigma_2^2)/2}.
\end{aligned}$$

□

A consequence of this fact is that given $X_i \sim \sigma$-sub-Gaussian, i.i.d. with zero-mean, then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \sim \sigma\text{-sub-Gaussian},$$

or equivalently

$$\frac{1}{n} \sum_{i=1}^n X_i \sim \frac{\sigma}{\sqrt{n}}\text{-sub-Gaussian}.$$

Example 1.5 (Survey sampling)
Two candidates for an election A and B.

Sample people $i = 1, \ldots, n$, giving responses $X_i = 1$ if A or 0 if B. Let μ^* be the actual fraction of people who will vote A. Let $\hat{\mu}$ be our estimator for μ^*:

$$\hat{\mu} = \sum_{i=1}^n X_i.$$

We can construct a confidence interval $\hat{\mathcal{I}}$ for our estimator, and we would like to know at what point we can say

$$\mathbb{P}[\hat{\mathcal{I}} \ni \mu^*] \geq 1 - \delta.$$

For example, with $\delta = 0.02$, interval width of 0.03, we require $n \approx 10000$ to make this guarantee.

We can model $X_i \sim \textsc{Bern}(\mu^*)$. Since $X_i \in [0,1]$, our earlier result shows that X_i is 1/2-sub-Gaussian. Using additivity and i.i.d., our sample mean $\hat{\mu}$ is $1/(2\sqrt{n})$-sub-Gaussian. Thus,

$$\mathbb{E}[e^{\lambda(\hat{\mu}-\mu^*)}] \leq e^{\lambda^2/2 \cdot 1/(4n)} = e^{\lambda^2/(8n)}.$$

Using Chernoff,

$$\mathbb{P}[|\hat{\mu} - \mu^*| \geq s] \leq 2e^{2ns^2},$$

for some $s > 0$. So for some fixed δ, we can make a guarantee about interval width $s = \sqrt{\log(2/\delta)/(2n)}$.

Theorem 1.6 (Hoeffding bound)
Let $X_i \sim \sigma_i$-sub-Gaussian, with $\mathbb{E}[X_i] = \mu_i$ and all independent. Then

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right] \leq e^{-t^2/(2\sum_{i=1}^{n}\sigma_i^2)}.$$

Example 1.7
$X \sim \mathcal{N}(0,1)$ is 1-sub-Gaussian. Let $Y = X^2$ is not sub-Gaussian.

$$\mathbb{E}[e^{\lambda X^2}] = \frac{1}{\sqrt{1-\lambda}},$$

$\lambda \in (0,1)$.

Despite not being sub-Gaussian, it is close. Consider an n-dimensional Gaussian, $X = (X_1, \ldots, X_n)$, where each $X_i \sim \mathcal{N}(0,1)$. Then $\mathbb{E}[\|X\|_2^2/n] = \mathbb{E}[\sum X_i^2/n] = 1$, and we can further show that

$$\mathbb{P}\left[\left|\frac{\|X\|^2}{n^2} - 1\right| \geq \delta\right] \leq 2e^{-cn\delta^2},$$

for all $\delta \in (0,1)$. This looks very similar to the sub-gaussian tail bound from earlier, but will only hold for small delta. For larger delta, the tail bound becomes linear in δ (because X is not truly sub-Gaussian).

2 February 15, 2024

2.1 Randomized dimension reduction

Example 2.1

Dataset is a set of N points $\{u^1, \ldots, u^N\}$ where $u^j \in \mathbb{R}^d$.

Storing the entire dataset gets very expensive very quickly when N, d are large. Is there a lower-dimensional representation of this dataset that is still useful?

Things that might be useful to preservere:

- pairwise distances

  $$\|u_i - u_j\|_2^2 \quad \forall i \neq j$$

  useful for estimating clustering algorithms, densities (computing neighborhoods of points)

-

Dimension reduction, formally:

$$\mathcal{F} : \mathbb{R}^d \to \mathbb{R}^m$$

We call m the **sketch dimension**, or the **embedding dimension**. The goal is for us to find a "useful" representation where $m \ll d$, which, using the distance metric as our notion of usefulness, we can bound w.r.t. a new parameter ε:

$$1 - \varepsilon \leq \frac{\|\mathcal{F}(u_i) - \mathcal{F}(u_j)\|_2^2}{\|u^i - u^j\|_2^2} \leq 1 + \varepsilon.$$

We'll solve this from the perspective of a fixed ϵ, so that our goal is to minimize m while preserving some notion of distance. We'll also introduce another parameter δ, so that this equation holds w.p. $1-\delta$.

Example 2.2 (Motivating Bernstein's Condition)
Given $X_i \sim \textsc{Bern}(p)$ for $i \in \{1, \dots, n\}$. We have $\mathbb{E}[X_i] = p$, $\mathrm{Var}[X_i] = p(1-p)$, and $|X_i| \leq b := 1$.

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_i X_i - p\right| \geq t\right) \leq 2e^{-nt^2/(2p(1-p)+2t)}$$

If t is small, then we get a sub-Gaussian tail bound. If t is large, then our bound converges to $e^{-nt/2}$, which are sub-Exponential.

Definition 2.3 (Bernstein's Condition)
Given random variable with parameters $\mu = \mathbb{E}[X]$, $\sigma^2 = \mathrm{Var}[X]$, and b satisfying

$$|\mathbb{E}[(X-\mu)^k]| \leq \frac{1}{2}k!\sigma^2 b^{k-2}$$

for $k = 2, 3, \dots$. Then,

Example 2.4
Bounded random variables satisfy the Bernstein condition. Let's say $|X-\mu| \leq b$. We can show that this satisfies the Condition in a pretty strong sense; we won't need the extra factor of $k!$ on the RHS.

Proof.

$$\begin{aligned}|\mathbb{E}[(X-\mu)^k]| &\leq \mathbb{E}|X-\mu|^2|X-\mu|^{k-2} \\ &\leq \sigma^2 b^{k-2} \ll \frac{1}{2}k!\sigma^2 b^{k-2}.\end{aligned}$$

□

Example 2.5

The Bernstein Condition also implies a nice bound on MGFs. Prof emphasizes that the fact that bounds on the polynomial moments of a r.v. can imply bounds on the MGF is quite a deep idea.

$$\mathbb{E}[e^{\lambda(X-\mu)}] \le e^{(\lambda^2\sigma^2/2)/(1-b|\lambda|)},$$

for all $|\lambda|b < 1$.

Proof.

$$\begin{aligned}
\mathbb{E}[e^{\lambda(X-\mu)}] &= 1 + \lambda\mathbb{E}[X-\mu] + \frac{\lambda^2}{2} + \sum_{k\ge 3} \frac{\lambda^k \mathbb{E}[(X-\mu)^k]}{k!} \\
&\le 1 + \frac{\lambda^2}{2}\sigma^2 + \sum_{k\ge 3} \frac{|\lambda|^k k!/2 \cdot \sigma^2 b^{k-2}}{k!} \\
&\le 1 + \frac{\lambda^2}{2}\sigma^2 + \frac{\lambda^2}{2}\sigma^2 \sum_{k\ge 3} |\lambda|^{k-2} b^{k-2} \\
&\le 1 + \frac{\lambda^2}{2}\sigma^2 \left(\sum_{k\ge 0} |\lambda|^k b^k \right) \\
&= 1 + \frac{\lambda^2\sigma^2/2}{1-b|\lambda|},
\end{aligned}$$

as long as $|\lambda|b < 1$. Now we are done by the fact that $1 + a \le e^a$. □

We will leave as an exercise :skull: that this result can be used to show that

$$\mathbb{P}\left[\left| \frac{1}{n}\sum_{i=1}^n X_i - \mu \right| \ge t \right] \le 2e^{-nt^2/(2\sigma^2 + 2bt)}.$$

Definition 2.6

X is (ν, α)-sub-exponential if $\mathbb{E}[e^{\lambda(x-\mu)}] \le e^{\nu^2\lambda^2/2}$ for all $|\lambda| < 1/\alpha$.

This is a relaxation on sub-Gaussianness. For example, $(\nu, 0)$-sub-exponentials are sub-Gaussian.

Example 2.7

If the previous Example holds, then $|\lambda| < 1/(2b)$ implies $1 - |\lambda| b > 1/2$ implies $\mathbb{E}[e^{\lambda(x-\mu)}] \le e^{\lambda^2(\sqrt{2}\sigma^2)^2/2}$, which is $(\sqrt{2}\sigma, 2b)$-sub-exponential.

Example 2.8

$X \sim \mathcal{N}(0,1)$ and $Z = X^2$.

Then $\mathbb{E}[Z] = 1$, and we can also show that

$$\mathbb{E}[e^{\lambda(X^2-1)}] = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}}$$

as long as $|\lambda| < 1/2$. It can further be shown that

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \le e^{2\lambda^2} = e^{\frac{\lambda^2(2)^2}{2}}$$

for all $|\lambda| < 1/4$. So, this r.v. is $(2,4)$-sub-exponential.

Proposition 2.9

If X is (ν, α)-sub-exponential, then

$$\mathbb{P}[X - \mu \ge t] \le \begin{cases} e^{-t^2/(2\nu^2)} & t \in [0, \nu^2/\alpha] \\ e^{-t/(2\alpha)} & t > \nu^2/\alpha. \end{cases}$$

A more compact way of writing this:

$$\mathbb{P}[X - \mu \ge t] \le e^{\frac{-t^2}{2\nu^2 + 2\alpha t}}.$$

Proof. Using Markov,

$$\begin{aligned} \mathbb{P}[X - \mu \ge t] &\le \mathbb{E}[e^{\lambda(X-\mu)}]e^{-\lambda t} \\ &\le e^{\lambda^2\nu^2/2 - \lambda t} \quad \forall 0 < \lambda < \frac{1}{\alpha}. \end{aligned}$$

Let $g(\lambda)$ be the exponent. $g'(\lambda) = \lambda\nu^2 - t$, so the unconstrained optimum occurs at $\lambda^* = t/\nu^2$. This unconstrained optimum is achievable precisely when $t \in$

$[0, \nu^2/\alpha]$; plugging in this value gives the first bound. When $t > \nu^2/\alpha$, we can choose the boundary point closest to this unconstrained optimum, giving the second bound. $\square$