Stats 300b: Theory of Statistics **Winter 2018-2019**

Lecture 19 – March 11

Lecturer: John Duchi *Scribe: John Sholar*

Warning: *these notes may contain factual errors*

Announcements: The take-home final exam is 10 AM - 10 AM, Sunday - Monday or Monday - Tuesday. You may use all available resources (textbooks, problem sets, problem set solutions, online materials), but you may **not** collaborate with any other students.

Reading: HDP 3.1 and 9.1

Outline: Random vectors, concentration of norms, sub-Gaussianity of matrix processes

1 Recap

Theorem 1 (Generic Chaining)**.** *If $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ are processes and Y_t is Gaussian and $\|X_t - X_s\|_{\psi_2} \le \sigma \|Y_t - Y_s\|_{L^2} = \sigma \mathbb{E}[(Y_t - Y_s)^2]^{1/2}$ then*

$$\mathbb{E}\left[\sup_{t\in T} |X_t - X_{t_0}|\right] \le C\sigma \mathbb{E}\left[\sup_{t\in T} |Y_t - Y_{t_0}|\right].$$

Equivalently,

$$\mathbb{P}\left(\sup_{t\in T} |X_t - X_{t_0} \ge C\sigma \mathbb{E}\left[\sup_{t\in T} |Y_t - Y_{t_0}|\right](1+\delta)\right) \le \exp(-\delta^2).$$

Morally, any time you can upper bound the ψ_2-norm of your process by the L^2 norm of a Gaussian process, you can control expected suprema of X by controlling the expected suprema of the Gaussian. Said differently, if X_t has sub-Gaussian increments, it is very well-controlled by a Gaussian process.

Today, we will use the ideas presented in theorem 1 to argue that random design matrices are well-conditioned (e.g. satisfy restricted strong convexity). In particular, we will study the process $\|X\theta\|_2 - \sqrt{n}\|\theta\|_2$

2 Starting point: norms of random vectors

Question: How does the ℓ_2-norm $\|X\|_2$ concentrate for $X \in \mathbb{R}^n$? Remarkably, we will see that this concentration is dimension free (i.e. not dependent on n).
We now state a useful theorem, which we will use throughout this lecture.

Theorem 2. *Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ have indepent[1] σ^2-sub-Gaussian entries. Assume $\mathbb{E}[X_i^2] = 1$, $\|X_i\|_{\psi_2} \le \sigma$. Then*

$$\left\| \|X\|_2 - \sqrt{n} \right\|_{\psi_2} \le C\sigma^2.$$

That is, for all $\lambda \in \mathbb{R}$

$$\mathbb{E}\left[\exp\left(\lambda\left(\|X\|_2 - \mathbb{E}[\|X\|_2]\right)\right)\right] \le \exp\left(\lambda^2\sigma^2 C_1\right)$$

[1] John made a big show of not correcting "indepent" to "independent", so we've faithfully reproduced this here.

Proof Consider $\|X\|_2^2 = \sum_{i=1}^n X_i^2$, which is an IID sum. From the homework we have that $\|X_i^2 - 1\|_{\psi_1} \leq 2\|X_i^2\|_{\psi_1} \leq 2\|X_i\|_{\psi_2}^2 = 2\sigma^2$. From the Bernstein-type inequalities we have

$$\mathbb{P}\left(\left|\sum_{i=1}^n (X_i^2 - 1)\right| \geq t\right) \leq 2\exp\left(-c\min\left\{\frac{t^2}{n\sigma^4}, \frac{t}{\sigma^2}\right\}\right)$$

As $2\sigma^2 \geq \max_i \|X_i^2 - 1\|_{\psi_1}$. Equivalently,

$$\begin{aligned}\mathbb{P}\left(\left|\frac{1}{n}\|X\|_2^2 - 1\right| \geq t\right) &\leq 2\exp\left(-cn\min\left\{\frac{t^2}{\sigma^4}, \frac{t}{\sigma^2}\right\}\right)\\ &\leq 2\exp\left(-\frac{cn}{\sigma^4}\min\left\{t^2, t\right\}\right)\end{aligned}$$

as $\sigma^2 \geq \mathbb{E}[X_i^2] = 1$. It is a True Fact$^{\text{TM}}$ that $|Z - 1| \geq \delta$ implies that $|Z^2 - 1| \geq \max \delta, \delta^2$, because $\sqrt{\cdot}$ is contractive toward 1. Applying this, we have

$$\begin{aligned}\mathbb{P}\left(\left|\frac{1}{\sqrt{n}}\|X\|_2 - 1\right| \geq t\right) &\leq \mathbb{P}\left(\left|\frac{1}{n}\|X\|_2^2 - 1\right| \geq \max\{t, t^2\}\right)\\ &\leq 2\exp\left(-\frac{cn}{\sigma^4}t^2\right)\end{aligned}$$

Equivalently,

$$\mathbb{P}\left(\left|\|X\|_2 - \sqrt{n}\right| \geq t\right) \leq 2\exp\left(-\frac{ct^2}{\sigma^4}\right)$$

Which is what we wanted to show. □

Morally, what we have shown in theorem 2 is that ℓ_2-norms of vectors with independent sub-Gaussian entries have dimension-independent sub-Gaussian tails.

3 Main question

Our main focus for this lecture will be how we can control $\{\|X\theta\|_2 - \sqrt{n}\|\theta\|_2\}_{\theta\in\Theta}$, where $X \in \mathbb{R}^{n\times d}$ has rows $X_1^T, \ldots, X_n^T$.

3.1 Definitions

Definition 3.1. *A mean-zero vector $X_i \in \mathbb{R}^d$ is σ^2-sub-Gaussian if*

$$\|X_i\|_{\psi_2} := \sup_{u:\ \|u\|_2 \leq 1} \|\langle X_i, u\rangle\|_{\psi_2} \leq \sigma$$

Equivalently (up to some numerical constant), for all $v \in \mathbb{R}^d$

$$\mathbb{E}[\exp(\langle X_i, v\rangle)] \leq \exp\left(\frac{\sigma^2\|v\|_2^2}{2}\right)$$

Definition 3.2. *X_i is isotropic if $\mathbb{E}[X_iX_i^T] = I$*

For now, we will consider only isotropic vectors, but our argument extends to non-isotropic vectors with some messy eigenvalue and condition number arguments that we wish to avoid.

3.2 Main theorem

The proof of the main theorem for this lecture (which we actually stated and derived corollaries for, but did not prove, last Thursday) is quite long, so we've broken it into sub-parts.

Theorem 3. *let $X \in \mathbb{R}^{n \times d}$ have independent, isotropic, σ^2-sub-Gaussian rows $(X_1^T, \dots, X_n^T)$. Now, define the process $Z_\theta := \|X\theta\|_2 - \sqrt{n}\theta$. Then, Z_θ is $C\sigma^4$-sub-Gaussian for the ℓ_2-norm, i.e.*

$$\|Z_\theta - Z_t\|_{\psi_2} \leq C\sigma^2 \|\theta - t\|$$

where C is independent of the sample size n and dimension d.

Proof This proof is broken down into three cases, which we enumerate in theorems 4, 5, and 6 (though effectively theorem 6 proves this theorem in the most general case, and leans on theorems 4 and 5). Morally, if I give you a random matrix with independent, isotropic, sub-Gaussian rows, its increments are sub-Gaussian with a constant. From this, things like restricted strong convexity immediately pop out. □

Theorem 4. *Theorem 3 holds for $t = 0$, $\|\theta\|_2 = 1$*

Proof In this case, note that $X\theta$ is the vector $(X_1^T\theta, \dots, X_n^T\theta)$, which has n independent σ^2-sub-Gaussian entries. From theorem 2, we have that

$$\|Z_\theta - Z_t\|_{\psi_2} = \|Z_\theta\|_{\psi_2} = \|\|X\theta\|_2 - \sqrt{n}\|_{\psi_2} \leq C\sigma^2 = C\sigma^2\|\theta - t\|_2$$

□

Theorem 5. *Theorem 3 holds for $\|\theta\|_2 = \|t\|_2 = 1$*

Proof We break the proof of this theorem down into two cases: $\varepsilon \leq \frac{4\sqrt{n}}{3}$ and $\varepsilon \geq \frac{4\sqrt{n}}{3}$. We address case 1 in lemma 7 and case 2 in lemma 8 □

Theorem 6. *Theorem 3 holds for $\|\theta\|_2, \|t\|_2 > 0$*

Proof Because we can normalize by $\min\{\|\theta\|_2, \|t\|_2\} > 0$, we assume WLOG that $\|\theta\|_2 = 1$, $\|t\| \geq 1$.
Now, let $\bar{t} = \frac{t}{\|t\|_2}$. From theorem 5 we have because $\|\bar{t}\|_2 = 1$,

$$\|Z_\theta - Z_{\bar{t}}\|_{\psi_2} \leq c\sigma^2\|\theta - \bar{t}\|_2 \tag{1}$$

Now, using the fact that $\left\|\frac{u}{\|u\|_2} - u\right\|_2 = (\|u\|_2 - 1)$ for any u, we have from theorem 2 that

$$\begin{aligned} \|Z_{\bar{t}} - Z_t\|_{\psi_2} &= \|\|Xt\|_2 - \sqrt{n}\|t\|_2 - (\|X\bar{t}\|_2 - \sqrt{n}\|\bar{t}\|_2)\|_{\psi_2} \\ &= (\|t\|_2 - 1)\|z_{\bar{t}}\|_{\psi_2} \\ &\leq (\|t\|_2 - 1)c\sigma^2 \end{aligned} \tag{2}$$

Then, for $Z_\theta = \|X\theta\|_2 - \sqrt{n}\|\theta\|_2$, combining equations 1 and 2 with the triangle inequality yields

$$\begin{aligned}\|Z_\theta - Z_t\|_{\psi_2} &= \|Z_\theta - Z_{\bar{t}} + Z_{\bar{t}} - Z_t\|_{\psi_2} \\ &\leq \|Z_\theta - Z_{\bar{t}}\|_{\psi_2} + \|Z_{\bar{t}} - Z_t\|_{\psi_2} \\ &\leq c\sigma^2\|\theta - \bar{t}\|_2 + c\sigma^2(\|t\|_2 - 1) \\ &= c\sigma^2(\|\theta - \bar{t}\|_2 + \|\bar{t} - t\|_2)\end{aligned}$$

But this isn't quite the result we need, because the triangle inequality goes the wrong way. We need a slightly different bound. From lemma 10 we have $\|Z_\theta - Z_t\|_{\psi_2} \leq C\sigma^2\|\theta - t\|_2$. □

3.3 Supporting lemmata for the main theorem

Lemma 7. *For $\varepsilon \leq \frac{4\sqrt{n}}{3}$ we have*

$$\mathbb{P}\left(\left|\frac{\|X\theta\|_2 - \|Xt\|_2}{\|t - \theta\|_2}\right| \geq \varepsilon\right) \leq \exp\left(-\frac{c\varepsilon^2}{\sigma^4}\right)$$

Proof Using the fact that $a^2 - b^2 = (a-b)(a+b)$, we have

$$\mathbb{P}\left(\left|\frac{\|X\theta\|_2 - \|Xt\|_2}{\|t - \theta\|_2}\right| \geq \varepsilon\right) = \mathbb{P}\left(\left|\frac{\|X\theta\|_2^2 - \|Xt\|_2^2}{\|t - \theta\|_2}\right| \geq \varepsilon(\|X\theta\|_2 + \|Xt\|_2)\right) =: \mathbb{P}(A)$$

Note that in order for event A to occur, one of two events has to occur

1. $\|X\theta\|_2 \leq 3\sqrt{n}/4$ (the right hand side is small)
2. $\|X\theta\|_2 \geq 3\sqrt{n}/4$ but $\left|\frac{\|X\theta\|_2^2 - \|Xt\|_2^2}{\|\theta - t\|_2}\right| \geq \frac{3\varepsilon\sqrt{n}}{4}$

From theorem 2 we have that the probability of event (1) is bounded by

$$\mathbb{P}((1) \text{ occurs}) = \mathbb{P}\left(\|X\theta\|_2 \leq 3\sqrt{n}/4\right) = \mathbb{P}\left(\|X\theta\|_2 \leq \sqrt{n} - \frac{\sqrt{n}}{4}\right) \leq \exp\left(-\frac{cn}{\sigma^4}\right) \leq \exp\left(-\frac{c\varepsilon^2}{\sigma^4}\right)$$

From lemma 9 we have that the probability of event (2) is bounded by

$$\mathbb{P}((2) \text{ occurs}) \leq \mathbb{P}\left(\left|\frac{\|X\theta\|_2^2 - \|Xt\|_2^2}{\|\theta - t\|_2}\right| \geq \frac{3\varepsilon\sqrt{n}}{4}\right) \leq \exp\left(-\frac{c\varepsilon^2}{\sigma^4}\right)$$

Thus, the theorem holds in the case that $\varepsilon \leq \frac{4\sqrt{n}}{3}$ □

Lemma 8. *For $\varepsilon \geq \frac{4\sqrt{n}}{3}$ we have*

$$\mathbb{P}\left(\left|\frac{\|X\theta\|_2 - \|Xt\|_2}{\|t - \theta\|_2}\right| \geq \varepsilon\right) \leq \exp\left(-\frac{c\varepsilon^2}{\sigma^4}\right)$$

Proof Observe that from the triangle inequality we have

$$|Z_\theta - Z_t| = |\|X\theta\|_2 - \|Xt\|_2| \leq \|X(\theta - t)\|_2$$

By theorem 2 we have

$$\begin{aligned}\mathbb{P}\left(\frac{|Z_\theta - Z_t|}{\|\theta - t\|_2} \geq \varepsilon\right) &\leq \mathbb{P}\left(\frac{\|X(\theta - t)\|_2}{\|\theta - t\|_2} \geq \varepsilon\right)\\ &= \mathbb{P}\left(\frac{\|X(\theta - t)\|_2}{\|\theta - t\|_2} \geq \sqrt{n} + (\varepsilon - \sqrt{n})\right)\\ &\leq 2\exp\left(-\frac{c(\varepsilon - \sqrt{n})^2}{\sigma^4}\right)\\ &\leq 2\exp\left(-\frac{c\varepsilon^2}{\sigma^4}\right)\end{aligned}$$

Where the last line follows from the fact that $(eps - \sqrt{n}) \geq \varepsilon/3$. □

Lemma 9. *For $\varepsilon \geq 3\sqrt{n}/4$ we have*

$$\mathbb{P}\left(\left|\frac{\|X\theta\|_2^2 - \|Xt\|_2^2}{\|t - \theta\|_2}\right| \geq \delta\right) \leq 2\exp\left(-\frac{c\varepsilon^2}{\sigma^4}\right)$$

Proof Note that

$$\frac{\|X\theta\|_2^2 - \|Xt\|_2^2}{\|t - \theta\|_2} = \frac{\langle X(\theta - t), X(\theta + t)\rangle}{\|\theta - t\|_2} = \langle Xu, Xv\rangle = \sum_{i=1}^n \langle X_i, u\rangle \langle X_i, v\rangle$$

where

$$u = \frac{\theta - t}{\|\theta - t\|_2}, \quad v = \theta + 2, \quad \|u\|_2 1, \quad \|v\|_2 \leq 2.$$

We have from exercise 6.7 (a) that $\| \langle X_i, u\rangle \langle X_i, v\rangle \|_{\psi_1} \leq \| \langle X_i, u\rangle \|_{\psi_2} \| \langle X_i, v\rangle \|_{\psi_2} \leq 2\sigma^2$. We also have from the assumption that the rows of X are isotropic that

$$\mathbb{E}\left[\langle X_i, u\rangle \langle X_i, v\rangle\right] = \frac{1}{\|\theta - t\|_2}\mathbb{E}\left[(X_i^T\theta)^2 - (X_i^T t)^2\right] = 0$$

That is

$$\mathbb{E}\left[\frac{\|X\theta\|_2^2 - \|Xt\|_2^2}{\|t - \theta\|_2}\right] = 0$$

That is, this is a sum of mean-zero sub-exponential random variables. By the Bernstein-type bounds and subsequently taking $\delta \mapsto \sqrt{n}$ we have

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{i=1}^n \langle X_i, u\rangle \langle X_i, v\rangle\right| \geq \delta\right) &\leq 2\exp\left(-c\min\left\{\frac{\delta^2}{n\sigma^4}, \frac{\delta}{\sigma^2}\right\}\right)\\ &\leq 2\exp\left(-c\min\left\{\frac{1}{\sigma^4}, \frac{\sqrt{n}}{\sigma^2}\right\}\right)\\ &\leq 2\exp\left(-\frac{c\varepsilon^2}{\sigma^4}\right) \quad \text{if } \varepsilon \leq \sqrt{n}\sigma^2\end{aligned}$$

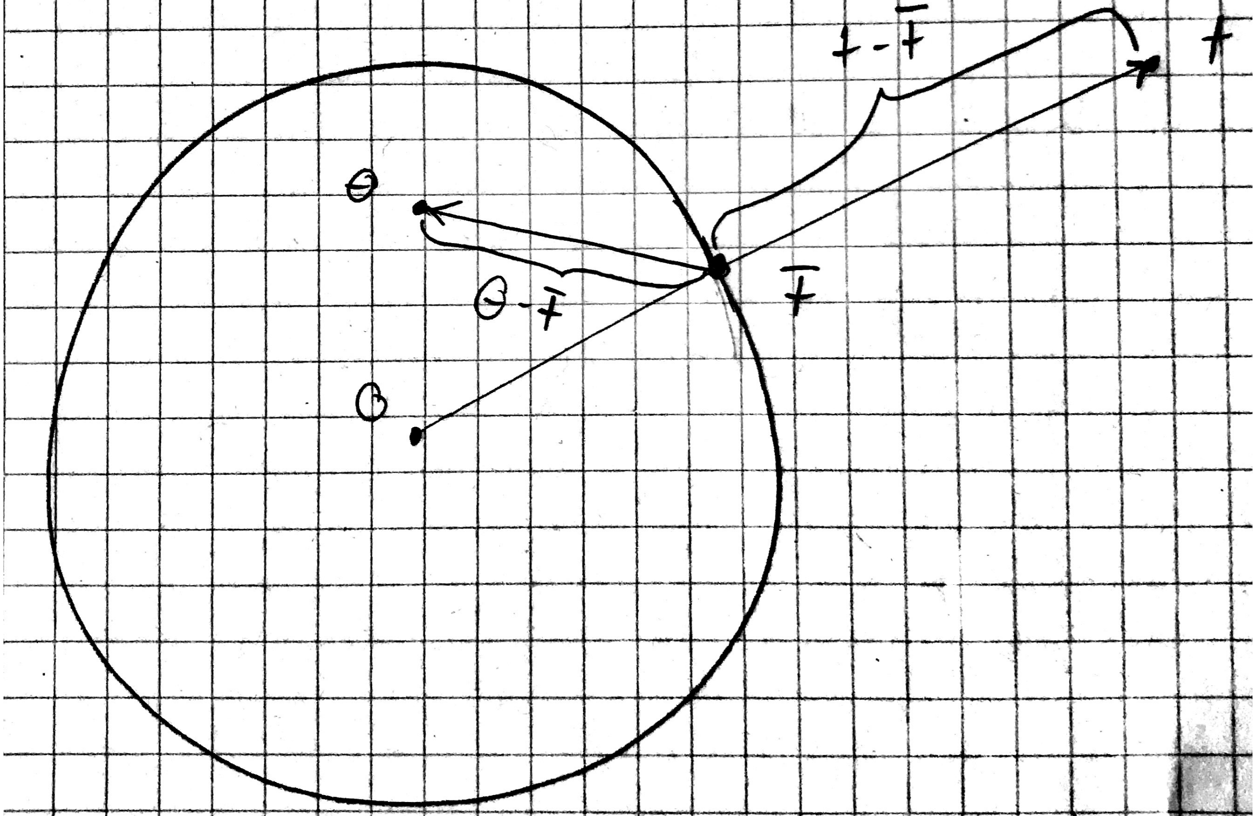
Equivalently, for $\varepsilon \leq \sqrt{n}\sigma^2$, or sufficiently $\varepsilon \leq \sqrt{n}$, we have

$$\mathbb{P}\left(\left|\frac{\|X\theta\|_2^2 - \|Xt\|_2^2}{\|t-\theta\|_2}\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{c\varepsilon^2}{\sigma^4}\right)$$

which was to be shown. □

Lemma 10. *For $\theta, t, \bar{t}$ as defined in theorem 6 we have $\|\theta - \bar{t}\|_2 + \|t - \bar{t}\|_2 \leq 2\|\theta - t\|_2$*

Proof As $\bar{t}$ is a projection of t onto the unit ball, we have $\left\langle \theta - \bar{t}, t - \bar{t}\right\rangle \leq 0$. We prove this by picture in figure 3.3. Equivalently we have $(1/2)\|\bar{t} - t\|_2^2 + (1/2)\|\theta - \bar{t}\|_2^2 - (1/2)\|\theta - t\|_2^2 \leq 0$. So $\|\theta - \bar{t}\|_2^2 + \|t - \bar{t}\|_2^2 \leq \|\theta - t\|_2^2$. Using $(a+b) \leq 2\sqrt{a^2+b^2}$ we finally have $\|\theta - \bar{t}\|_2 + \|t - \bar{t}\|_2 \leq 2\|\theta - t\|_2$. □



$f - \bar{f}$
f
θ
$\theta - \bar{f}$
$\bar{f}$
G