6.7720/18.619/15.070 Lecture 7
Sub-Gaussian Random Variables

Kuikui Liu

February 24, 2025

Acknowledgements & Disclaimers *In the process of writing these notes, we consulted materials by Phillipe Rigollet, as well as "High-Dimensional Probability" by Roman Vershynin. Please be advised that these notes have not been subjected to the usual scrutiny reserved for formal publications. If you do spot an error, please contact the instructor.*

1 Concentration Beyond Bounded Random Variables

Our aim in this lecture is begin the process of deriving progressively more and more general Chernoff-type bounds. To do this, we define a class of random variables which "behave like Gaussians" in terms of how well they concentrate. These are known as sub-Gaussian random variables, and appear in a wide variety of applications.

Proposition 1.1 (Sub-Gaussian Distributions; see e.g. [Ver18])**.** *We say a random variable X is* sub-Gaussian *if it satisfies any one of the following definitions, which are all equivalent up to rescaling the corresponding factors $K_1, \dots, K_5$ by universal numerical constants:*

1. *The tails of X are upper bounded by Gaussian tails, i.e. there exists $K_1 > 0$ such that*
$$\Pr\left[|X| \geq t\right] \leq 2\exp\left(-t^2/K_1^2\right), \qquad \forall t \geq 0.$$
2. *The (absolute) moments of X are uniformly bounded by the moments of a Gaussian, i.e. there exists $K_2 > 0$ such that*
$$\mathbb{E}\left[|X|^p\right]^{1/p} \leq K_2 \cdot \sqrt{p}.$$
3. *The moment generating function of X^2 is upper bounded by the moment generating function of a Gaussian, i.e. there exists $K_3 > 0$ such that*
$$\mathbb{E}\left(\exp\left(s^2X^2\right)\right) \leq \exp\left(K_3^2 s^2\right), \qquad \forall s \in \left[-\frac{1}{K_3}, \frac{1}{K_3}\right].$$
4. *The moment generating function of X^2 is upper bounded at some point, i.e. there exists $K_4 > 0$ such that $\mathbb{E}\left[\exp\left(X^2/K_4^2\right)\right] \leq 2$.*
5. *The moment generating function of the centered random variable $X - \mathbb{E}[X]$ is upper bounded by the moment generating function of a Gaussian, i.e. there exists $K_5 > 0$ such that*
$$\mathbb{E}\left[\exp\left(s \cdot (X - \mathbb{E}[X])\right)\right] \leq \exp\left(K_5^2 s^2\right), \qquad \forall s \in \mathbb{R}.$$

We omit the proof as it is conceptually straightforward and relies entirely on calculations; see e.g. [Ver18]. The utility of Proposition 1.1 is that it gives us different ways of viewing sub-Gaussianity, some of which may be more convenient to use depending on the application. It also gives us useful ways of quantifying "how sub-Gaussian" a random variable is.

Definition 1 (Sub-Gaussian Norm)**.** *For a sub-Gaussian random variable X, we define its* sub-Gaussian norm *as the quantity*
$$\|X\|_{\psi_2} \stackrel{\text{def}}{=} \inf\left\{K > 0 : \mathbb{E}\left[\exp\left(X^2/K^2\right)\right] \leq 2\right\}.$$

Definition 2 (Variance Proxy)**.** *For a sub-Gaussian random variable X, we define its optimal* variance proxy *to be the quantity*

$$\|X\|_{\mathsf{VP}}^2 \stackrel{\mathsf{def}}{=} \inf\left\{K^2 > 0 : \mathbb{E}\left[\exp\left(s \cdot (X - \mathbb{E}[X])\right)\right] \le \exp\left(K^2 s^2\right), \forall s \in \mathbb{R}\right\}.$$

By the equivalence given by Proposition 1.1, we have $\|X\|_{\psi_2} \asymp \|X\|_{\mathsf{VP}}$. In this lecture, we'll mainly use $\|X\|_{\mathsf{VP}}$ as it tends to lead to cleaner expressions. However, we emphasize that any expression involving the variance proxy can be replaced by the sub-Gaussian norm up to a multiplicative loss by a universal numerical constant. To get a sense of the scale of these quantities, observe that the variance proxy and the sub-Gaussian norm are both bonafide upper bounds on the true variance.

Lemma 1.2. *For a sub-Gaussian random variable X, we have* $\mathrm{Var}(X) \le \|X\|_{\mathsf{VP}}^2$ *and* $\mathrm{Var}(X) \lesssim \|X\|_{\psi_2}^2$.

This lemma is quite intuitive, since $\|X\|_{\psi_2}^2 \le K^2$ roughly means X is "dominated" by a Gaussian with variance K^2. The proof is straightforward and provided in Appendix A. Let us now see a few examples of sub-Gaussian random variables.

Example 1 (Gaussian Random Variables). If $X \sim \mathcal{N}(0, \sigma^2)$, then we have $\|X\|_{\psi_2} \le C\sigma$ for some constant $C > 0$, and $\|X\|_{\mathsf{VP}}^2 = \sigma^2$.

Example 2 (Rademacher Random Variables). If $X \sim \mathsf{Unif}\{\pm 1\}$, then we have $\|X\|_{\psi_2} \le \frac{1}{\sqrt{\ln 2}}$ and $\|X\|_{\mathsf{VP}}^2 \le 1$.

Example 3 (Bounded Random Variables). If X takes values in the bounded interval $[-a, a]$ almost surely, then we have $\|X\|_{\psi_2} \le \frac{a}{\sqrt{\ln 2}}$ and $\|X\|_{\mathsf{VP}}^2 \le a^2$.

2 Concentration for Basic Functionals

The quantities $\|\cdot\|_{\psi_2}, \|\cdot\|_{\mathsf{VP}}$ behave well under natural operations like taking linear combinations of random variables. This is illustrated in the following theorem, which significantly generalizes the Chernoff–Hoeffding concentration inequality we stated in the previous lecture for bounded random variables.

Theorem 2.1 (Generalized Chernoff–Hoeffding Inequality)**.** *Let $X_1, \ldots, X_n$ be independent sub-Gaussian random variables. Then for any $\boldsymbol{v} \in \mathbb{R}^n$,*

$$\left\|\sum_{i=1}^n \boldsymbol{v}_i X_i\right\|_{\mathsf{VP}}^2 \le \sum_{i=1}^n \boldsymbol{v}_i^2 \|X_i\|_{\mathsf{VP}}^2,$$

and the random variable $Y = \sum_{i=1}^n \boldsymbol{v}_i X_i$ satisfies the tail bound

$$\Pr\left[|Y - \mathbb{E}[Y]| \ge t\right] \le 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \boldsymbol{v}_i^2 \|X_i\|_{\mathsf{VP}}^2}\right).$$

Proof. By assumption, we have

$$\mathbb{E}\left[\exp\left(s \cdot (\boldsymbol{v}_i X_i - \mathbb{E}[\boldsymbol{v}_i X_i])\right)\right] \le \exp\left(\boldsymbol{v}_i^2 \|X_i\|_{\mathsf{VP}}^2 s^2\right), \qquad \forall s \in \mathbb{R}.$$

Multiplying both sides across $i = 1, \ldots, n$, and then applying independence to push the product inside the expectation in the left-hand side, we obtain the first inequality. The second claim is then an immediate consequence of the standard application of Markov's Inequality; one may also invoke Proposition 1.1. □

The previous result concerns linear functionals of sub-Gaussian vectors. Now let us turn to the Euclidean norm.

Lemma 2.2. *Let $X_1, \ldots, X_n$ be independent sub-Gaussian random variables with variance proxy $\widehat{\sigma}^2/n$. Then there exist universal numerical constants $C, \alpha^* > 0$ such that the Euclidean norm of the vector $X = (X_1, \ldots, X_n)$ satisfies*

$$\Pr\left[\|X - \mathbb{E}[X]\|_2 \ge \alpha \cdot \widehat{\sigma}\right] \le \exp\left(-C\alpha^2 n\right), \qquad \forall \alpha > \alpha^*.$$

Remark 1. It is actually necessary that α exceed at least some universal constant α^*. Indeed, as we will see in the next lecture, the random variable $\|X\|_2$ actually concentrates around its expectation $\mathbb{E}[\|X\|_2]$, and so an exponentially decaying bound $\Pr[\|X\|_2 \geq \alpha \cdot \widehat{\sigma}]$ cannot hold for arbitrarily small α.

Proof. Proposition 1.1 ensures there exists a universal numerical constant $L > 0$ such that

$$\mathbb{E}\left[\exp\left(\frac{n}{L\widehat{\sigma}^2}(X_i - \mathbb{E}[X_i])^2\right)\right] \leq 2, \qquad \forall i = 1, \dots, n.$$

By independence, it follows that

$$\mathbb{E}\left[\exp\left(\frac{n}{L\widehat{\sigma}^2}\|X - \mathbb{E}[X]\|_2^2\right)\right] \leq 2^n.$$

Applying Markov's Inequality, we obtain

$$\Pr[\|X - \mathbb{E}[X]\| \geq \alpha \cdot \widehat{\sigma}] \leq 2^n \exp\left(-\frac{\alpha^2 n}{L}\right).$$

Setting $\alpha^* > \sqrt{L \ln 2}$ yields the claim. □

Finally, let's look at the ℓ_∞-norm, or more generally, the supremum of a sub-Gaussian process.

Theorem 2.3. *Let $X_1, \dots, X_n$ be mean-zero sub-Gaussian random variables satisfying $\|X_i\|_{\mathsf{VP}}^2 \leq \widehat{\sigma}^2$ for all $i = 1, \dots, n$; note that they may be arbitrarily correlated. Then*

$$\Pr\left[\max_{i=1,\dots,n} X_i > t\right] \leq n \exp\left(-\frac{t^2}{2\widehat{\sigma}^2}\right), \qquad \Pr\left[\max_{i=1,\dots,n} |X_i| > t\right] \leq 2n \exp\left(-\frac{t^2}{2\widehat{\sigma}^2}\right), \qquad \forall t > 0,$$

and

$$\mathbb{E}\left[\max_{i=1,\dots,n} X_i\right] \leq \widehat{\sigma} \cdot \sqrt{2 \ln n}, \qquad \mathbb{E}\left[\max_{i=1,\dots,n} |X_i|\right] \leq \widehat{\sigma} \cdot \sqrt{2 \ln(2n)}.$$

Remark 2. Note that these bounds are essentially sharp in the case where all the random variables are independent. In the other extreme, when the random variables are maximally correlated (e.g. $X_1 = \cdots = X_n$), the extra multiplicative factors depending on n are completely unnecessary. There is a beautiful theory of *generic chaining* which aims for more refined bounds based on the correlation structure of the random variables $X_1, \dots, X_n$; see [Tal14].

Proof. First, observe if we define n new random variables by setting $X_{n+i} = -X_i$ for $i = 1, \dots, n$, then $\max_{i=1,\dots,n} |X_i| = \max_{1,\dots,2n} X_i$. Hence, the inequalities regarding $\max_{i=1,\dots,n} |X_i|$ are immediate consequences of the inequalities regarding $\max_{i=1,\dots,n} X_i$. For the first inequality, we combine sub-Gaussianity with the Union Bound to obtain

$$\begin{aligned}
\Pr\left[\max_{i=1,\dots,n} X_i > t\right] &= \Pr\left[\bigcup_{i=1}^n \{X_i > t\}\right] \\
&\leq \sum_{i=1}^n \Pr[X_i > t] && \text{(Union Bound)} \\
&\leq n \exp\left(-\frac{t^2}{2\widehat{\sigma}^2}\right). && \text{(Using } \|X_i\|_{\mathsf{VP}}^2 \leq \widehat{\sigma}^2\text{)}
\end{aligned}$$

For the bound on the expectation, one way is to simply integrate the tail, leveraging the bound in the preceding display for the large t regime, and the trivial upper bound of 1 for the small t regime. Another way is to appeal to the moment generating function. Observe that for any $t > 0$,

we have

$$\begin{aligned}
\mathbb{E}\left[\max_{i=1,\dots,n} X_i\right] &= \frac{1}{t}\mathbb{E}\left[\ln \exp\left(t \cdot \max_{i=1,\dots,n} X_i\right)\right] \\
&\le \frac{1}{t}\ln \mathbb{E}\left[\max_{i=1,\dots,n} \exp\left(tX_i\right)\right] && \text{(Jensen's Inequality)} \\
&\le \frac{1}{t}\ln \sum_{i=1}^{n} \mathbb{E}\left[\exp\left(tX_i\right)\right] && (*) \\
&\le \frac{\ln n}{t} + \frac{\widehat{\sigma}^2 t}{2}. && \text{(Sub-Gaussianity)}
\end{aligned}$$

Balancing the two terms by setting $t = \sqrt{\frac{2\ln n}{\widehat{\sigma}^2}}$ yields the desired bound. All that remains is to justify $(*)$, which can be achieved by applying the Union Bound. In particular, using the *layered cake representation*[1] for the expectation of a nonnegative random variable, we have

$$\begin{aligned}
\mathbb{E}\left[\max_{i=1,\dots,n} \exp\left(tX_i\right)\right] &= \int_0^\infty \Pr\left[\max_{i=1,\dots,n} \exp\left(tX_i\right) \ge s\right] ds \\
&\le \sum_{i=1}^{n} \int_0^\infty \Pr\left[\exp\left(tX_i\right) \ge s\right] ds \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\exp\left(tX_i\right)\right].
\end{aligned}$$

□

3 A Brief Venture Into the World of Random Matrices

In this section, we illustrate the utility our concentration inequalities by bounding various important quantities associated to well-known random matrix ensembles. One of the landmark applications of random matrix theory was Eugene Wigner's observation that the eigenvalue spacings of a random matrix can be used to fruitfully model the energy levels of heavy atomic nuclei; mathematical physicists often refer to this insight as *Wigner's surmise*.

> *"Perhaps I am now too courageous when I try to guess the distribution of the distances between successive levels (of energies of heavy nuclei). Theoretically, the situation is quite simple if one attacks the problem in a simpleminded fashion. The question is simply what are the distances of the characteristic values of a symmetric matrix with random coefficients."* — Eugene Wigner 1956

Random matrix theory now plays a central role in modern high-dimensional probability and statistics, with abundant applications throughout science and engineering.

3.1 The Sherrington–Kirkpatrick Spin Glass

We begin by looking at optimizing the quadratic form of a random matrix drawn over the Boolean hypercube $\{\pm 1\}^n$. When the matrix is drawn from the *Gaussian orthogonal ensemble (GOE)*, the optimizers are known as the *ground states* of the *Sherrington–Kirkpatrick model* in statistical physics. This model has received a lot of attention in recent years. Its nonrigorous analysis was one of many reasons Giorgio Parisi won the Nobel Prize in Physics in 2021, and its rigorous analysis was one of many reasons Michel Talagrand won the Abel Prize in 2024.

Definition 3 (Gaussian Orthogonal Ensemble (GOE))**.** *For $n \in \mathbb{N}$, a random symmetric matrix $\boldsymbol{J} \in \mathbb{R}^{n\times n}$ is drawn from the* Gaussian Orthogonal Ensemble $\mathsf{GOE}(n)$ *if it is given by $\boldsymbol{J} = \frac{\boldsymbol{G}+\boldsymbol{G}^\top}{\sqrt{2n}}$, where $\boldsymbol{G} \in \mathbb{R}^{n\times n}$ has i.i.d. $\mathcal{N}(0,1)$ entries.*

[1] For a nonnegative random variable Y, we have $\mathbb{E}[Y] = \int_0^\infty \Pr[Y \ge y]\, dy$.

Another way to describe $\boldsymbol{J}$ is to say that its strictly upper triangular entries are drawn i.i.d. from $\mathcal{N}(0, 1/n)$, its diagonal entries are drawn i.i.d. from $\mathcal{N}(0, 2/n)$, and its strictly lower triangular entries are set to ensure symmetry.

Theorem 3.1. *For $n \in \mathbb{N}$, $\boldsymbol{J} \sim \mathsf{GOE}(n)$, and $H(\varphi) \stackrel{\mathsf{def}}{=} \frac{1}{2}\varphi^\top \boldsymbol{J}\varphi$ we have*

$$\mathbb{E}\left[\max_{\varphi \in \{\pm 1\}^n} H(\varphi)\right] \leq n \cdot \sqrt{\ln 2}.$$

Moreover,

$$\Pr\left[\max_{\varphi \in \{\pm 1\}^n} H(\varphi) \geq \gamma n\right] \leq 2^n \cdot \exp\left(-\gamma^2 n\right), \qquad \forall \gamma > 0.$$

Remark 3. If we let $Y = \max_{\varphi \in \{\pm 1\}^n} H(\varphi)$, then we can actually obtain the significantly more useful bound of

$$\Pr\left[|Y - \mathbb{E}[Y]| \geq \gamma n\right] \leq 2\exp\left(-\frac{\gamma^2 n}{2}\right)$$

by using concentration for Lipschitz functions of Gaussian random vectors.

Proof. For each $\varphi \in \{\pm 1\}^n$, observe that the random variable $H(\varphi)$ is distributed as a mean-zero Gaussian with variance

$$\mathrm{Var}\left(H(\varphi)\right) = \frac{1}{2n}\mathrm{Var}\left(\varphi^\top \boldsymbol{G}\varphi\right) = \frac{1}{2n}\sum_{i,j=1}^n \varphi_i^2\varphi_j^2 = \frac{n}{2},$$

where in the first step, we used that $\boldsymbol{J} = \frac{\boldsymbol{G}+\boldsymbol{G}^\top}{\sqrt{2n}}$, and in the second step we used that $\boldsymbol{G}$ has i.i.d. $\mathcal{N}(0,1)$ entries. It follows that $\|H(\varphi)\|_{\mathsf{VP}}^2 = \frac{n}{2}$, and so we may apply Theorem 2.3 to the collection of random variables $\{H(\varphi)\}_{\varphi \in \{\pm 1\}^n}$ to deduce the desired inequalities. □

3.2 Bounding the Operator Norm

Now let us turn to bounding the operator norm of $\mathsf{GOE}(n)$. Since the vectors of $\{\pm 1\}^n$ have squared Euclidean norm n, Theorem 3.1 already suggests an $O(1)$ upper bound on $\boldsymbol{\lambda}_{\max}(\boldsymbol{J})$ by expressing this quantity using Rayleigh quotients: $\boldsymbol{\lambda}_{\max}(\boldsymbol{J}) = \sup_{\varphi \neq \boldsymbol{0}} \frac{\varphi^\top \boldsymbol{J}\varphi}{\varphi^\top\varphi}$. Hence, we might expect the operator norm of $\mathsf{GOE}(n)$ to be $O(1)$ as well. As we will show, this turns out to be correct. In fact, it is known that $\|\boldsymbol{J}\|_{\mathsf{op}}$ concentrates around 2 for $\boldsymbol{J} \sim \mathsf{GOE}(n)$.

Using a somewhat bare-handed (but still generically useful) approach, we will instead show an $O(1)$ upper bound on $\|\boldsymbol{J}\|_{\mathsf{op}}$ in the more general setting of matrices with independent sub-Gaussian entries. The key challenge we will need to overcome, unlike the case of the discrete hypercube $\{\pm 1\}^n$, is bounding $\|\boldsymbol{J}\varphi\|_2$ over *all* vectors in the unit sphere S^{n-1}, which is an uncountably infinite set.

Theorem 3.2. *There exist universal numerical constants $C, \gamma^* > 0$ such that the following holds for all $\gamma > \gamma^*$ and $n \in \mathbb{N}$: For any random matrix $\boldsymbol{J}$ whose entries are independent, zero-mean, and sub-Gaussian with variance proxy $\widehat{\sigma}^2/n$,*

$$\Pr\left[\|\boldsymbol{J}\|_{\mathsf{op}} \geq \gamma \cdot \widehat{\sigma}\right] \leq 2\exp\left(-C\gamma^2 n\right).$$

The overarching strategy will be the same as in the proof of Theorem 3.1. However, since we cannot Union Bound over all uncountably many unit vectors, we will instead Union Bound over a suitable finite subset $\mathcal{N} \subseteq S^{n-1}$. On the one hand, we will want our set $\mathcal{N}$ to be sufficiently dense so that it is a "good approximation" to the entire unit sphere S^{n-1}. On the other hand, we will want the cardinality of $\mathcal{N}$ to be at most exponential in $O(n)$ so that the concentration bound we have for each individual $\|\boldsymbol{J}\varphi\|_2$ isn't overwhelmed when we invoke the Union Bound.

Note that the reason this strategy has any hope of working in the first place is that we expect the function $\varphi \mapsto \|\boldsymbol{J}\varphi\|_2$ to be Lipschitz. If this were the case, then for any point $\psi \in S^{n-1}$, if we find another point $\varphi \in \mathcal{N}$ such that $\|\varphi - \psi\|_2$ is small, then we can bound $\|\boldsymbol{J}\psi\|_2$ using a bound

on $\|\boldsymbol{J}\varphi\|_2$ plus a bound on $\|\boldsymbol{J}\|_{\mathsf{Lip}} \cdot \|\varphi - \psi\|_2$. One potential issue is that $\|\boldsymbol{J}\|_{\mathsf{Lip}} \asymp \|\boldsymbol{J}\|_{\mathsf{op}}$, and so if we could already bound $\|\boldsymbol{J}\|_{\mathsf{Lip}}$, then we'd already be done and there would be no need to execute our strategy. One must be slightly careful to circumvent this circular reasoning, but fortunately for us, this will not be difficult.

Let us now begin formalizing our strategy.

Definition 4 (δ-Covering/δ-Net)**.** *Fix a metric space $(\mathcal{X}, d)$. For $\delta > 0$, a* δ-covering *(or* δ-net*) of $\mathcal{X}$ is a subset $\mathcal{C} \subseteq \mathcal{X}$ such that for every $x \in \mathcal{X}$, there exists $y \in \mathcal{C}$ such that $d(x, y) \le \delta$. In other words, $\mathcal{X} \subseteq \bigcup_{y \in \mathcal{C}} \mathcal{B}_d(y, \delta)$, where $\mathcal{B}_d(y, \delta) = \{x \in \mathcal{X} : d(x, y) \le \delta\}$ denotes the closed ball of radius δ around y with respect to the metric d.*

There is a *dual* concept called δ-*packing*, which we will leverage in a moment. We will need two lemmas about δ-coverings of the unit sphere, whose proofs are straightforward and provided at the end of the section.

Lemma 3.3. *For $0 < \delta < 1$, let $\mathcal{N} \subseteq S^{n-1}$ be a δ-covering of S^{n-1} with respect to Euclidean distance. Then for any matrix $\boldsymbol{J}$, we have*

$$\max_{\varphi \in \mathcal{N}} \|\boldsymbol{J}\varphi\|_2 \le \|\boldsymbol{J}\|_{\mathsf{op}} \le \frac{1}{1-\delta} \cdot \max_{\varphi \in \mathcal{N}} \|\boldsymbol{J}\varphi\|_2 .$$

Lemma 3.4 (Covering Number of the Sphere)**.** *For any $n \in \mathbb{N}$ and $\delta > 0$, there exists a δ-covering of the unit sphere S^{n-1} in $\mathbb{R}^n$ with at most $\left(1 + \frac{2}{\delta}\right)^n$ points.*

Proof of Theorem 3.2. Let $0 < \delta < 1$ be a parameter to be determined later. By Lemma 3.4, there exists a δ-covering $\mathcal{N}$ of S^{n-1} with at most $\left(1 + \frac{2}{\delta}\right)^n$ points. Observe that

$$\begin{aligned}
\Pr\left[\|\boldsymbol{J}\|_{\mathsf{op}} \ge \gamma \cdot \widehat{\sigma}\right] &\le \Pr\left[\bigcup_{\varphi \in \mathcal{N}} \left\{ \|\boldsymbol{J}\varphi\|_2 \ge (1-\delta) \cdot \gamma \cdot \widehat{\sigma} \right\}\right] && \text{(Lemma 3.3)} \\
&\le \sum_{\varphi \in \mathcal{N}} \Pr\left[\|\boldsymbol{J}\varphi\|_2 \ge (1-\delta) \cdot \gamma \cdot \widehat{\sigma} \right]. && \text{(Union Bound)}
\end{aligned}$$

Now, let us bound each term in the sum. Observe that for any fixed vector $\varphi \in S^{n-1}$, each entry of $X = \boldsymbol{J}\varphi$ is sub-Gaussian with variance proxy $\widehat{\sigma}^2/n$ by Theorem 2.1. Since the rows of $\boldsymbol{J}$ are independent and have zero mean, the entries of X are independent and have zero mean. Applying Lemma 2.2, it follows that

$$\Pr\left[\|\boldsymbol{J}\varphi\|_2 \ge (1-\delta) \cdot \gamma \cdot \widehat{\sigma}\right] \le \exp\left(-C(1-\delta)^2\gamma^2 n\right), \qquad \forall \gamma > \gamma^*.$$

Plugging in $t = \gamma \cdot \widehat{\sigma}$ and combining with the above Union Bound, we obtain

$$\Pr\left[\|\boldsymbol{J}\|_{\mathsf{op}} \ge \gamma \cdot \widehat{\sigma}\right] \le 2\left(1 + \frac{2}{\delta}\right)^n \cdot \exp\left(-C(1-\delta)^2\gamma^2 n\right).$$

By setting δ to be some constant, e.g. $1/2$, and then setting γ^* sufficiently large, we obtain the desired result. □

3.3 On δ-Coverings of the Unit Sphere

Proof of Lemma 3.3. The first inequality is immediate by the definition of the operator norm and the fact that $\mathcal{N} \subseteq S^{n-1}$. For the second inequality, let $\psi \in S^{n-1}$ be such that $\|\boldsymbol{J}\psi\|_2 = \|\boldsymbol{J}\|_{\mathsf{op}}$, which exists by continuity and compactness. Since $\mathcal{N}$ is a δ-covering, there exists $\varphi \in \mathcal{N}$ such that $\|\varphi - \psi\|_2 \le \delta$. Hence, we have

$$\begin{aligned}
\|\boldsymbol{J}\|_{\mathsf{op}} &= \|\boldsymbol{J}\psi\|_2 && \text{(Definition of } \psi\text{)} \\
&\le \|\boldsymbol{J}(\varphi - \psi)\|_2 + \|\boldsymbol{J}\varphi\|_2 && \text{(Triangle Inequality)} \\
&\le \delta \cdot \|\boldsymbol{J}\|_{\mathsf{op}} + \|\boldsymbol{J}\varphi\|_2 . && \text{(Definition of } \varphi\text{)}
\end{aligned}$$

Rearranging yields the claim. □

Proof of Lemma 3.4. Let us choose as an inclusion-wise maximal δ-*packing*: We seek a subset $\mathcal{N} \subseteq S^{n-1}$ such that for all pairs of distinct points $\varphi, \psi \in \mathcal{N}$, we have $\|\varphi - \psi\|_2 > \delta$, and moreover, no strict superset has the same property. Such a set can be constructed greedily, for instance. We will argue two points:

1. First, any *maximal* δ-packing is also a δ-covering.
2. Second, any δ-packing has cardinality upper bounded by $\left(1 + \frac{2}{\delta}\right)^n$. In particular, the greedy procedure for constructing $\mathcal{N}$ must terminate after $\left(1 + \frac{2}{\delta}\right)^n$ steps.

For the first point, observe that if $\mathcal{N}$ weren't a δ-covering, then there would exist some point $\psi \in S^{n-1}$ such that $\|\varphi - \psi\|_2 > \delta$ for all $\varphi \in \mathcal{N}$. But then we could have added ψ to the set $\mathcal{N}$ and preserved the fact that it is a δ-covering, thus contradicting maximality of $\mathcal{N}$.

For the second point, we use a standard *volume argument*. Each point φ added to $\mathcal{N}$ carves out a closed radius-$\frac{\delta}{2}$ Euclidean ball $\mathcal{B}_2\left(\varphi, \frac{\delta}{2}\right)$ of points such that $\mathcal{B}_2\left(\varphi, \frac{\delta}{2}\right) \cap \mathcal{B}_2\left(\psi, \frac{\delta}{2}\right) = \emptyset$ for all distinct pairs $\varphi, \psi \in \mathcal{N}$; this pairwise disjointness is by the Triangle Inequality and the δ-separation between the points of $\mathcal{N}$. Since $\mathcal{B}_2\left(\varphi, \frac{\delta}{2}\right) \subseteq \mathcal{B}_2\left(\mathbf{0}, 1 + \frac{\delta}{2}\right)$ for all $\varphi \in S^{n-1}$ by the Triangle Inequality, it follows that

$$\begin{aligned}
|\mathcal{N}| \cdot \mathrm{Vol}\left(\mathcal{B}_2\left(\mathbf{0}, \frac{\delta}{2}\right)\right) &= \mathrm{Vol}\left(\bigcup_{\varphi \in \mathcal{N}} \mathcal{B}_2\left(\varphi, \frac{\delta}{2}\right)\right) && \text{(Pairwise disjointness)} \\
&\leq \mathrm{Vol}\left(\mathcal{B}_2\left(\mathbf{0}, 1 + \frac{\delta}{2}\right)\right). && \left(\text{Using } \mathcal{B}_2\left(\varphi, \tfrac{\delta}{2}\right) \subseteq \mathcal{B}_2\left(\mathbf{0}, 1 + \tfrac{\delta}{2}\right)\right)
\end{aligned}$$

Rearranging then yields

$$|\mathcal{N}| \leq \frac{\mathrm{Vol}\left(\mathcal{B}_2\left(\mathbf{0}, 1 + \frac{\delta}{2}\right)\right)}{\mathrm{Vol}\left(\mathcal{B}_2\left(\mathbf{0}, \frac{\delta}{2}\right)\right)} = \frac{\left(1 + \frac{\delta}{2}\right)^n}{(\delta/2)^n} = \left(1 + \frac{2}{\delta}\right)^n.$$

□

References

[Tal14] Michel Talagrand. *Upper and Lower Bounds for Stochastic Processes. Modern Methods and Classical Problems*. Springer Berlin, Heidelberg, 2014 (cit. on p. 3).

[Ver18] Roman Vershynin. *High-Dimensional Probability. An Introduction with Applications in Data Science*. Cambridge University Press, 2018 (cit. on p. 1).

A Unfinished Proofs

Proof Sketch of Lemma 1.2. For convenience, we only prove the inequality $\mathrm{Var}(X) \lesssim \|X\|_{\psi_2}^2$. Without loss of generality, we may assume $\mathbb{E}[X] = 0$. In this case, if we write $K = \|X\|_{\psi_2}$, then

$$\begin{aligned}
\mathbb{E}\left[X^2\right] &= \int_0^\infty \Pr[X^2 \geq s]\, ds && \text{(Layered cake representation of an expectation)} \\
&= \int_0^\infty \Pr\left[\exp\left(X^2/K^2\right) \geq \exp\left(s/K^2\right)\right] ds \\
&\leq \mathbb{E}\left[\exp\left(X^2/K^2\right)\right] \cdot \int_0^\infty \exp\left(-\frac{s}{K^2}\right) ds && \text{(Markov's Inequality)} \\
&\leq 2K^2. && \text{(Calculation and definition of } K)
\end{aligned}$$

□