ISyE 8803 – Special Topics in Modern Mathematical Data Science

Dmitrii M. Ostrovskii

April 13, 2025

Lecture 1: Deviation bounds for random variables

Preliminaries. Let X be a scalar random variable (i.e. a Borel-measurable function on $\mathbb{R}$) with c.d.f. $F_X(x)$. In this lecture, we shall focus on bounding the tail function $\Phi_X(x) := 1 - F_X(x) = \mathbb{P}\{X > x\}$, as well as the two-sided counterpart $\Phi_{|X|}(x) = \mathbb{P}\{|X| > x\}$, under various assumptions on X. Sometimes, when the random variable of interest is clear from the context, we drop the subscript. Note that we can reduce the task of bounding $\Phi_{|X|}(x)$ to that of bounding the right and left tails separately. However, it might be more convenient to study $|X|$ directly, since this random variable is nonnegative – as Markov's inequality requires.

1 Markov's inequality and Chernoff's method

Our first inequality is the most basic one: it only requires that X is nonnegative and has a finite expectation.

Theorem 1.1 (Markov)**.** *Assume $X \geqslant 0$ a.s. and $\mathbb{E}[X] < \infty$. Then for $u > 0$ one has $\mathbb{P}\{X > u\} \leqslant \dfrac{\mathbb{E}[X]}{u}$.*

Proof. Assuming that the distribution of X is absolutely continuous, and denoting its p.d.f. with f_X, we get

$$\begin{aligned}\mathbb{E}[X] = \int_0^\infty x f_X(x)dx &= \int_0^u x f_X(x)dx + \int_{xu}^\infty x f_X(x)dx \\ &\geqslant \int_u^\infty x f_X(x)dx \\ &\geqslant \int_u^\infty u f_X(x)dx = u\Phi_X(u).\end{aligned}$$

Here the first inequality is since X is nonnegative, and the second one is since we integrate over $x \geqslant u$. In the general case, we proceed in the same way but integrating against some reference measure $d\mu(x)$ on $\mathbb{R}_+$. □

Corollary 1.1. *Applying Markov's inequality to r.v. $|X|$, we get $\mathbb{P}\{|X| > u\} \leqslant \dfrac{\mathbb{E}[|X|]}{u}$ provided $\mathbb{E}[|X|] < \infty$.*

Chernoff's bounding method. Let $g : \mathbb{R} \to \mathbb{R}_+$ be an increasing function. Then $g(X)$ is a nonnegative random variable (even when X is not), and $\{X \geqslant u\}$ is the same event as $g(X) \geqslant g(u)$. Thus, we can bound the tail function of X by applying Markov's inequality to $g(X)$, with the bound in terms of $g(X)$ and $g(u)$:

$$\mathbb{P}\{X \geqslant u\} = \mathbb{P}\{g(X) \geqslant g(u)\} \leqslant \frac{\mathbb{E}[g(X)]}{g(u)}.$$

Of course, for this bound not to be vacuous, it must be that $\mathbb{E}[g(X)] < \infty$. This trick is sometimes called "Chernoff's bounding method," though the idea can be traced back at least to Cramér in 1930s. Of course, the question is how to select the mapping $g(\cdot)$: on the one hand, one would like to make the right-hand side as small as possible; on the other hand, we would also like $g(\cdot)$ to "tensorize:" informally, to be well-behaved under convolution. This means that if X, Y are jointly independent r.v.'s, then $g(X+Y)$ must be expressed in terms of $g(X)$ and $g(Y)$. As it turns out, there is a nice way to impose this tensorization requirement with *almost* no loss of tightness. Before elaborating further on this, let us consider some concrete examples.

Chebyshev's inequality and moment bounds

We start with a straightforward generalization of Theorem 1.1. In the next result, the assumptions that $X \geqslant 0$ and $\mathbb{E}[X] < \infty$ replaced with the assumption that $\mathbb{E}[|X|] < \infty$. Note that this assumption implies $\mathbb{E}[X] < \infty$ and the existence of the first absolute central moment $\mathbb{E}[|X - \mathbb{E}[X]|] < \infty$ by the triangle inequality.

Theorem 1.2. *Assume that $\mathbb{E}[|X|] < \infty$. Then for $u > 0$ it holds that $\mathbb{P}\{|X - \mathbb{E}[X]| > u\} \leqslant \dfrac{\mathbb{E}[|X - \mathbb{E}[X]|]}{u}$.*

Exercise 1.1. *Prove Theorem 1.2 via Chernoff's method. Note: g might (and will) depend on the law of X.*

Theorem 1.3 (Chebyshev)**.** *If $X \in \mathbb{R}$ has a finite variance, then $\mathbb{P}\{|X - \mathbb{E}[X]| > u\} \leqslant \dfrac{\operatorname{Var} X}{u^2}$ for $u > 0$.*

Of course, we can run this trick for any absolute moment, even fractional ones, to get the following result.

Theorem 1.4 (Moment bound)**.** *Assume $\mathbb{E}[X] < \infty$ and $\mathbb{E}[|X - \mathbb{E}[X]|^p] < \infty$ for some $p > 0$. Then $\forall u > 0$,*

$$\mathbb{P}\{|X - \mathbb{E}[X]| > u\} \leqslant \frac{\mathbb{E}\left[|X - \mathbb{E}[X]|^p\right]}{u^p}.$$

Remark. *For any $p \geqslant 1$, the quantity $\|Z\|_{L^p} := \mathbb{E}^{1/p}[|Z|^p]$ is a norm over distributions, and a quasinorm—i.e. it does not satisfy the triangle inequality—whenever $0 < p < 1$. Using Hölder's inequality, one may verify that $\|Z\|_{L^p}$ increases in $p > 0$.*[1] *As such, the assumptions of Theorem 1.4 get stronger with increasing $p > 0$.*

2 Moment-generating function and the MGF method

The *moment-generating function* (MGF) of X is defined by $M_X(t) := \mathbb{E}[e^{tX}]$, where the value $+\infty$ is allowed. Note that M_X is an extended-value convex function, as a weighted sum of convex functions. A somewhat less trivial fact, to which we shall return many times later, is the convexity of the *cumulant* $K_X(t) := \log M_X(t)$.

Exercise 2.1. *Show that K_X is convex. Use Young's inquality: for $a, b \in \mathbb{R}^d$ and $p, q \in [1, \infty]$ with $\frac{1}{p} + \frac{1}{q} = 1$,*

$$|a^\top b| \leqslant \|a\|_p \|b\|_q. \tag{1}$$

Note that $M_X(0) = 1$, so the domain of M_X is nonempty. In fact, by Exercise 2.1, the domain of M_X is a convex subset of $\mathbb{R}$, i.e. a "segment" (both $\mathbb{R}$ and $\mathbb{R}_+$ qualify). The name comes from the following result.

Proposition 2.1. *Assume M_X exists in an open interval around 0. Then for any $k \in \mathbb{N}$ with a finite r.h.s.,*

$$M_X^{(k)}(0) = \mathbb{E}[X^k]. \tag{2}$$

Proof. It suffices to differentiate under the integral sign,

$$M_X^{(k)}(t) = \frac{d^k}{dt^k}\left(\int_{\mathbb{R}} e^{tx} f_X(x) dx\right) = \int_{\mathbb{R}} x^k e^{tx} f_X(x) dx = \mathbb{E}[X^k e^{tX}],$$

and take the limit as $t \to \infty$. Of course, one must ensure this is allowed, under the assumption on M_X. □

Revisiting our previous discussion, MGF tensorizes: if $X, Y \in \mathbb{R}$ are jointly independent, then

$$M_{X+Y}(t) = \mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}] = M_X(t) M_Y(t). \tag{3}$$

These transitions are by the property of the exponential $e^{a+b} = e^a e^b$, and since the functions of independent random variables are also independent. This suggests to take e^{tx} in the role of $g = g_t(x)$, then select the best t, i.e. $t > 0$ minimizing the right-hand side (note that e^{tx} is increasing in x whenever $t > 0$). This method, called the Chernoff method "proper" (or exponential Markov), bounds the tails in terms of MGF:

$$\mathbb{P}\{X \geqslant u\} \leqslant \inf_{\lambda > 0} M_X(\lambda) e^{-\lambda u}. \tag{4}$$

Note that the infinum here is formally taken over all $\lambda \in \mathbb{R}$, but can only be attained at λ where $M_X(\lambda) < \infty$.

In the sequel, we shall use the product property (3) to bound the tail probabilities for sums $\sum_{k \in [n]} X_k$ of independent random variables via the MGF method. But before, let us discuss how some matters related to the accuracy of this method, in the context of comparing it with the moment bounds and instantiating it for Gaussian tails. Our first take on this is rather superficial, but later on we shall revisit this discussion at a deeper level, through the prism of large deviations theory.

[1] This is a nice exercise: Young's inequality (1) does the trick in the case $p \geqslant 1$; for $p < 1$ one needs a change of variables.

MGF method vs. moment bounds

A natural question is to compare the best exponential bound (4) with the moment bound: either with the infinum of the r.h.s. in the inequality of Theorem 1.4, or with the simplified moment bound assuming $X > 0$,

$$\mathbb{P}\{X > u\} \leqslant \inf_{p>0} \mathbb{E}\left[X^p\right] u^{-p}.$$

As it turns out, the best moment bounds are generally sharper, even if we only use the integer moments.

Exercise 2.2. *(a) Show that if $X > 0$ a.s., then for any $u > 0$,*

$$\inf_{\lambda>0} M_X(\lambda) e^{-\lambda u} \geqslant \inf_{k\in\mathbb{Z}_+} \mathbb{E}\left[X^k\right] u^{-k}.$$

(b) Show that if X is symmetric (i.e. X and $-X$ have the same distribution), then for any $u > 0$,

$$\inf_{\lambda>0} M_X(\lambda) e^{-\lambda u} \geqslant \frac{1}{2} \inf_{k\in\mathbb{Z}_+} \mathbb{E}\left[X^{2k}\right] u^{-2k}.$$

Gaussian tail bounds

We now apply the MGF method to control the devations of $X \sim \mathcal{N}(\mu, \sigma^2)$ from $\mathbb{E}[X] = \mu$. Here the p.d.f. is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right).$$

Exercise 2.3. *Show that the d-variate Gaussian integral $I_d := \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} e^{-\frac{z_1^2}{2}} \cdots e^{-\frac{z_d^2}{2}} dz_1 \cdots dz_d$ equals $(2\pi)^{d/2}$. To this end, reduce to the general case to $d = 2$ and handle the latter in polar coordinates (the Jacobian is r).*

W.l.o.g. we can consider $Z \sim \mathcal{N}(0,1)$, using that $X = \mu + \sigma Z$. For example, $\mathbb{P}\{X - \mu \geqslant \sigma z\} = \mathbb{P}\{Z \geqslant z\}$.

Lemma 2.1. *For $Z \sim \mathcal{N}(0,1)$, the MGF is $M_Z(\lambda) = \exp\left(\frac{\lambda^2}{2}\right)$.*

Proof. It suffices to complete the square:

$$M_Z(\lambda) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(\lambda z - \frac{z^2}{2}\right) dz = \frac{e^{\frac{\lambda^2}{2}}}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(-\frac{(z-\lambda)^2}{2}\right) dz = e^{\frac{\lambda^2}{2}}. \qquad \square$$

Now, let us apply the MGF method:

$$\mathbb{P}\left\{Z \geq z\right\} \leqslant \inf_{\lambda\geqslant 0} M_Z(\lambda) e^{-\lambda z} \leqslant \inf_{\lambda\geqslant 0} \exp\left(\frac{\lambda^2}{2} - \lambda z\right) = \exp \inf_{\lambda\geqslant 0} \left(\frac{\lambda^2}{2} - \lambda z\right) = \exp\left(-\frac{z^2}{2}\right).$$

In the last two identities, we first used that exp is monotonically increasing; then we minimized the convex quadratic in $\lambda \in \mathbb{R}$ and used that the unconstrained minimum is attained at a positive $\lambda = z$. As the result,

$$\begin{aligned} \mathbb{P}\left\{X - \mu > \sigma z\right\} &\leq \exp\left(-\frac{z^2}{2}\right), \\ \mathbb{P}\left\{|X - \mu| > \sigma z\right\} &\leq 2\exp\left(-\frac{z^2}{2}\right). \end{aligned} \tag{5}$$

Bounds on the tail probabilities can be recast as confidence intervals, by inverting the tail function. In particular, we have just shown that for any fixed $\delta \in (0,1)$, each of the two events

$$X - \mu \leqslant \sigma\sqrt{2\log\left(\frac{1}{\delta}\right)}; \qquad \sigma\sqrt{2\log\left(\frac{2}{\delta}\right)} \leqslant \quad |X - \mu| \leqslant \sigma\sqrt{2\log\left(\frac{2}{\delta}\right)} \tag{6}$$

holds with probability at least $1 - \delta$. This form is convenient when dealing with maxima over a multiple random variables, due to the simplicity of taking the union bound. In particular, (6) implies the result below.

Proposition 2.2. *Let $M_n := \max_{j\in[n]}\{X_j - \mu\}$, where $X_j \sim \mathcal{N}(\mu, \sigma^2)$ for $k \in [n]$. With prob. $\geqslant 1-\delta$,*

$$M_n \leqslant \sigma\sqrt{2\log\left(\frac{n}{\delta}\right)}.$$

Two remarks are in order. First, the result does not impose any assumption on the *joint* distribution of the random vector $(X_1, ..., X_n)$, other than $X_j \sim \mathcal{N}(\mu, \sigma^2)$ marginally. Second, since $\log(\frac{n}{\delta}) = \log(n)+\log(\frac{1}{\delta})$ and $\sqrt{a+b} \leqslant \sqrt{a}+\sqrt{b}$ for $a, b \geqslant 0$, we conclude that

$$M_n \leqslant \sigma\sqrt{2\log n} + \sigma\sqrt{2\log\left(\delta^{-1}\right)}.$$

The r.h.s. is the sum of a δ-independent term corresponding (up to a constant factor) to the *expectation* of $M_n = \sigma\max_{j\in[n]} Z_j$, and a δ-dependent term scaled by σ, the standard deviation of an *individual* r.v. Such additive structure is related to the *concentration of measure* phenomenon, to which we shall yet return.

Exercise 2.4. *Show that for any $n \in \mathbb{N}$ and some constant $c > 0$, one has $\mathbb{E}\max_{j\in[n]}|Z_j| \leqslant \sqrt{2\log(2n)}+c$.*

Refined bounds for Gaussian tails

In the Gaussian case, one can refine the tail bounds coming from the MGF method. As in many other situations, the key is to "get more analytical" – here, by using the exact expression for the Gaussian density. In particular, letting ϕ and Φ be, respectively, the p.d.f. and tail function of $\mathcal{N}(0,1)$, one has the following:

$$\left(\frac{1}{u} - \frac{1}{u^3}\right)\phi(u) \leqslant \Phi(u) \leqslant \frac{1}{u}\phi(u) \quad \forall u > 0. \tag{7}$$

Exercise 2.5. *Prove (7). Start with the upper bound, then prove the lower bound using integration by parts.*

Note that for u large enough, the upper bound in (7) is stronger than the one obtained with the MGF method; moreover, the lower bound matches it in the first order when $u \to \infty$; in this sense, (7) is tight for large deviations. Remarkably, the trick you employed to pass from the upper bound to the lower bound can be reiterated: applying it to the lower bound, we get a *new upper bound*.

Exercise 2.6. *Prove the refined upper bound:*

$$\Phi(u) \leqslant \left(\frac{1}{u} - \frac{1}{u^3} + \frac{3}{u^5}\right)\phi(u) \quad \forall u > 0. \tag{8}$$

Applying this method iteratively we express $\Phi(u)/\phi(u)$, so-called *Mills ratio*, as a convergent power series.

Theorem 2.1. *It holds that*

$$\Phi(u) = \phi(u)\sum_{k=0}^{\infty}(-1)^k\frac{(2k-1)!!}{u^{2k+1}}, \quad \forall u > 0. \tag{9}$$

Moreover, stopping this series at any positive or negative term gives an upper or lower bound, respectively.

Let us note that one can also approximate the tail function as a power series in u rather than $1/u$ (i.e., center the Taylor expansion at 0 rather than ∞). Deriving the following result is a nice analytical exercise.

Exercise 2.7. *Show that*

$$\frac{1}{2} - \Phi(u) = \frac{1}{\sqrt{2\pi}}\sum_{k=0}^{\infty}\frac{(-1)^k u^{2k+1}}{2^k k!(2k+1)}.$$

Hint: change variable to remove u from the integration limits, then take derivatives in u under the integral.

For other convergent approximations of $\Phi(u)/\phi(u)$, see the paper [Due10] and the reference book [AS65].

Lecture 2: Subgaussian distributions

Disclaimer. This lecture follows Section 5 of Vershynin's lectures [Ver12], in particular Lemma 5.5 therein.

Definition 1. *A random variable X is called K-subgaussian (or* subgaussian *with parameter K), $K > 0$, if*

$$M_X(\lambda) \leqslant \exp\left(\frac{K^2\lambda^2}{2}\right) \quad \forall \lambda \in \mathbb{R}.$$

Remark. *The subgaussian parameter K describes the spread of X, and its square K^2 is sometimes referred to as* variance proxy. *Note that the MGF argument λ has the units inverse to those of X, and K has the same units as X. (It only makes sense to exponentiate or take logarithm of a unitless quantity – think why!)*

The subgaussian property is homogeneous: the subgaussian parameter of αX, $\alpha > 0$, is α times that of X.

1 Examples

Gaussian case. The name "subgaussian" comes from the fact that in the Gaussian case, the inequality of Definition 1 is tight. In other words (by Lemma 2.1):

Distribution $\mathcal{N}(0, \sigma^2)$ is σ-subgaussian.

For what is to follow, the following calculation is instructive.

Exercise 1.1. *Show that for $Z \sim \mathcal{N}(0,1)$, one has $\mathbb{E}[Z^{2k}] = (2k-1)!!$ for $k \in \mathbb{N}$. Conclude that $\|Z\|_{L^p} \leqslant \sqrt{p}$.*

Bounded case. It is a classical result by Vassily Hoeffding that bounded random variables are subgaussian.

Theorem 1.1 (Hoeffding's lemma)**.** *Let X be zero-mean and supported on $[a, b]$, then X is $\frac{b-a}{2}$-subgaussian:*

$$M_X(\lambda) \leqslant \exp\left(\frac{\lambda^2(b-a)^2}{8}\right). \tag{10}$$

Proof. We can assume $a < 0 < b$, the other cases being trivial. By Jensen, $e^{\lambda x} \leqslant \frac{x-a}{b-a}e^{\lambda a} + \frac{b-x}{b-a}e^{\lambda b}$, whence

$$M_X(\lambda) = \mathbb{E}[e^{\lambda X}] \leqslant \frac{\mathbb{E}[X]-a}{b-a}e^{\lambda a} + \frac{b-\mathbb{E}[X]}{b-a}e^{\lambda b} = \frac{be^{\lambda b} - ae^{\lambda a}}{b-a}$$

Thence one may proceed via some (not completely trivial) calculus, showing that the right-hand side is dominated by that in (10). For simplicity, we only consider the symmetric case $a = -b$. Here, we have to show that, for all $b > 0$,

$$\frac{be^{\lambda b} + be^{-\lambda b}}{2b} \leqslant \exp\left(\frac{\lambda^2 b^2}{2}\right) \quad \forall \lambda \in \mathbb{R},$$

which amounts to showing that $\log\cosh(u) \leqslant \frac{u^2}{2}$ for all $u \in \mathbb{R}$. For $\phi(u) = \frac{u^2}{2} - \log\cosh(u)$, we find that $\phi(0) = \phi'(0) = 0$ and $\phi''(u) \geqslant 0$, whence the desired inequality follows by Jensen. □

Exercise 1.2. *Complete the proof of Theorem 1.1 without assuming $a + b = 0$.*

Exercise 1.3. *Prove that Theorem 1.1 is tight, by exhibiting a distribution for which the equality is attained.*

2 Equivalent subgaussian properties

Recall that for $X \sim \mathcal{N}(0, \sigma^2)$, the L_p-norms grow as $\mathbb{E}^{1/p}[|X|^p] \sim \sigma\sqrt{p}$, the MGF is $M(\lambda) = \exp\left(\frac{\sigma^2\lambda^2}{2}\right)$, and the tails decay as $\mathbb{P}\{|X| \geqslant x\} \leqslant 2\exp\left(-\frac{x^2}{2\sigma^2}\right)$. These properties generalize to subgaussian distributions.

Proposition 2.1 ([Ver12, Lemma 5.5])**.** *Consider the following three properties for a random variable X.*

i. Subgaussian tail decay:

$$\mathbb{P}\{X \geq x\} \leq \exp\left(-\frac{x^2}{2K^2}\right) \qquad \forall x \geqslant 0. \tag{i.a}$$

$$\mathbb{P}\{|X| \geq x\} \leq 2\exp\left(-\frac{x^2}{2K^2}\right) \qquad \forall x \geqslant 0. \tag{i.b}$$

ii. Subgaussian moment growth:

$$\mathbb{E}^{1/p}|X|^p \leq K\sqrt{p} \qquad \forall p \geqslant 1. \tag{ii}$$

iii. Subgaussian MGF:

$$M_X(\lambda) \leqslant \exp\left(\frac{K^2\lambda^2}{2}\right) \qquad \forall \lambda \in \mathbb{R}. \tag{iii}$$

Then (i.b) $\Longrightarrow$ (ii) $\overset{\text{if } \mathbb{E}[X]=0}{\Longrightarrow}$ (iii) $\Longrightarrow$ (i.a)*. All these implications hold with a constant-factor distortion of K.*

Before we prove the theorem, we introduce a simple lemma that generalizes the identity for $X \geqslant 0$:

$$\mathbb{E}[X] = \int_0^{+\infty} \Phi_X(u)du.$$

Lemma 2.1 (Stack-up identity)**.** *If X is absolutely continuous and nonnegative, for any $p \geqslant 1$ one has*

$$\mathbb{E}[X^p] = \int_{\mathbb{R}_+} pu^{p-1}\Phi_X(u)du.$$

Exercise 2.1. *Prove Lemma 2.1 for $p \in \mathbb{N}$. Draw a picture for $p = 1$.*

Proof of Proposition 2.1

1°: (iii) $\Longrightarrow$ (i.a). We have already proved this: see Eq. (5) in the previous lecture. Indeed, to prove those bounds for $\mathcal{N}(0,1)$ we only used the information about $\mathcal{N}(0,1)$ contained in its MGF, and nothing more.

2°: (i.b) $\Longrightarrow$ (ii). By homogeneity, we can assume that (i.b) holds with $K = 1/\sqrt{2}$ and verify that (ii) then follows with some $K = c > 0$. By Lemma 2.1 applied to the random variable $|X|$, (i.b) with $K = 1/\sqrt{2}$ gives

$$\begin{aligned}\mathbb{E}[|X|^p] = \int_0^{+\infty} pu^{p-1}\mathbb{P}\{|X| \geq u\}\,du &\leqslant 2\int_0^{+\infty} pu^{p-1}\exp\left(-u^2\right)du \\ &\leqslant \int_0^{+\infty} pv^{\frac{p}{2}-1}\exp\left(-v\right)dv = p\Gamma(p/2)\end{aligned}$$

where we used the definition of Gamma function. Now, we only consider the case of even p, leaving the general case as an exercise. Here, $\Gamma(\frac{p}{2}) = (\frac{p}{2}-1)!$ and it remains to use that $n! \leqslant n^n$. Indeed, this implies

$$\mathbb{E}^{1/p}[|X|^p] \leqslant p^{1/p}\left(\frac{p}{2}\right)^{\frac{1}{2}} \leqslant \sqrt{p}.$$

3°: (ii) $\Longrightarrow$ (iii) assuming that $\mathbb{E}[X] = 0$. This part is a bit more tedious than the others. On the one hand,

$$\begin{aligned} M_X(t) = \mathbb{E}[e^{tX}] &= 1 + \underbrace{t\mathbb{E}[X]}_{=0} + \sum_{k=2}^{\infty} \frac{t^k}{k!}\mathbb{E}[X^k] \\ &\leqslant 1 + \sum_{k=2}^{\infty} \frac{t^k}{k!}\mathbb{E}[|X|^k] \\ &\leqslant 1 + \sum_{k=2}^{\infty} \frac{t^k k^{k/2}}{k!} \end{aligned}$$

where in the last step we used (ii) with $K = 1$. Now, recall that the trivial bound $n! \leqslant n^n$ admits the companion lower bound (Stirling's approximation):

$$\left(\frac{n}{e}\right)^n \leqslant n! \leqslant n^n.$$

(Actually, the lower bound is way tighter here.) From the lower bound and the previous calculation, we get

$$M_X(t) \leqslant 1 + \sum_{k=2}^{\infty} \left(\frac{et}{\sqrt{k}}\right)^k.$$

On the other hand, by the *upper* bound one has, for any $c, t > 0$

$$\exp\left(c^2t^2\right) = 1 + \sum_{m=1}^{\infty} \frac{(ct)^{2m}}{m!} \geqslant 1 + \sum_{m=1}^{\infty} \left(\frac{ct}{\sqrt{m}}\right)^{2m} \geqslant 1 + \sum_{k=2m, m\in\mathbb{N}} \left(\frac{\sqrt{2}ct}{\sqrt{k}}\right)^k.$$

One can show (do this!) that, for c large enough, the sum over odd k is dominated by that over even k. □

3 Monotonicity of L_p-norms

Recall that the L_p-norm of a univariate distribution $X \in \mathbb{R}$ is $\|X\|_{L_p} := \mathbb{E}^{1/p}[|X|^p]$. (Verify this is a norm!)

Remark. *Note that for a random variable X with $\mathbb{E}[X] = 0$, the subgaussian parameter is* lower-bounded, *up to a constant factor, with its standard deviation $\sigma := \mathbb{E}^{1/2}[X^2]$. Indeed, by Proposition 2.1, one has*

$$K \gtrsim \sup_{p\in\mathbb{N}} \left\{\frac{1}{\sqrt{p}}\|X\|_{L_p}\right\} \geqslant \frac{\sigma}{\sqrt{2}}.$$

Exercise 3.1. *Show that the L_p-norms are nondecreasing in p, i.e. $\|X\|_{L_p} \leqslant \|X\|_{L_q}$ for $1 \leqslant p \leqslant q \leqslant +\infty$.*

Remark. *One might (and should!) get confused here upon remembering that the usual ℓ_p-norms $\|\cdot\|_p$ on $\mathbb{R}^n$ are* nonincreasing *in p, i.e. $\|x\|_p \geqslant \|x\|_q$ whenever $p \leqslant q$. There is no mistake here: by Hölder's inequality,*

$$\|x\|_q \leqslant \|x\|_p \leqslant n^{\frac{1}{p}-\frac{1}{q}}\|x\|_q. \tag{11}$$

for $1 \leqslant p \leqslant q \leqslant +\infty$, so that the power in the right-hand side is nonnegative. This implies the result of Exercise 3.1 for uniform distribution with n outcomes, i.e. on $\{y_1, \ldots, y_n\}$. In general, for $1 \leqslant p \leqslant q \leqslant +\infty$,

$$\left(\int_\Omega d\mu\right)^{1/p} \|f\|_{L_p(\mu)} \leqslant \left(\int_\Omega d\mu\right)^{1/q} \|f\|_{L_q(\mu)}$$

for any measurable space $(\mathcal{X}, \mathcal{F}, \mu)$ and μ-measurable function f, with μ not necessarily a probability *measure. The case of counting measure gives us (11), and that of a probability measure gives Exercise 3.1. This also shows that $\|\cdot\|_{L_p(\nu)}$ norms are generally incomparable when $\nu(\Omega) = \infty$, e.g. when ν is a Lebesgue measure.*

4 Interlude: Hamburger's moment problem

In the exercise below, you shall explore the problem of recovering a distribution from its moment sequence. In the case $X \in \mathbb{R}_+$, this is called Hamburger's moment problem, and for $X \in \mathbb{R}$ it's the Stieltjes problem.

Exercise 4.1. *Show that one can recover a distribution X on $\mathbb{R}_+$ from the sequence $m_k = \mathbb{E}[X^k]$, $k \in \mathbb{N}$ of its moments. In other words, for any sequence $m_1, m_2, ... \in \mathbb{R}_+$, there is a unique continuously differentiable nonincreasing $\Phi : \mathbb{R}_+ \to [0,1]$ such that $\Phi(0) = 1$, $\lim_{x\to+\infty} \Phi(x) = 0$, and $\int_{\mathbb{R}_+} u^k d\Phi(u) = m_k$ for $k \in \mathbb{N}$.*

1. *First assume the bounded case, i.e. that there exists some B such that $m_k \leqslant B^k$ for all $k \in \mathbb{N}$.*
2. *Approximate the integral with a finite sum with step $1/N$, for N large enough, replacing Φ with its piecewise constant approximation (which can be considered a finite-dimensional vector).*
3. *Write down each condition $m_k = \int_{\mathbb{R}_+} ...$ as a linear equation for this vector, and interpret the resulting sequence of conditions as a linear system.*
4. *The system matrix is a Vandermonde one: $V_{jk} = {x_j}^k$. Use the classical result for $V \in \mathbb{C}^{N\times N}$:*

$$\det(V) = \prod_{1\leqslant i<j\leqslant n} x_i - x_j,$$

In particular V is nonsingular if and only if $x_i \neq x_j$ for $i \neq j$.

5. *Conclude with a boring approximation argument, showing that things will work out as $N \to \infty$.*

Summary. The key takeaway from this lecture: we have provided different equivalent characterizations of subgaussian distributions on $\mathbb{R}$. In the next lectures, we shall (a) generalize the subgaussian behavior in the framework of Orlicz norms; (b) extend these notions to multivariate disributions (a.k.a. random vectors).

Lecture 3: Briefly on Orlicz norms

Disclaimer. In this brief lecture, we first introduce the $\|\cdot\|_{\psi_2}$-norm (or subgaussian norm) for univariate distributions; this gives a convenient formalism that will accommodate random vectors in the next lecture. Then we shall discuss an extension of $\|\cdot\|_{\psi_2}$-norm to so-called Orlicz norms; this is supplementary material.

1 Subgaussian norm, a.k.a. ψ_2-norm

For what follows, and especially when dealing with the maxima of random processes, it is convenient for us to use the formalism of Orlicz norms. We do not give a general definition here; instead, we only define ψ_2-norm and later ψ_1-norm. For a deeper treatment of the subject, refer to David Pollard's online lecture notes [Pol].

Definition 2 (Subgaussian norm)**.** *Given a distribution $X \in \mathbb{R}$, its ψ_2-norm (or subgaussian norm) is*

$$\|X\|_{\psi_2} := \inf\left\{K \geqslant 0 : \mathbb{E}\left[\exp\left(\frac{|X|^2}{K^2}\right)\right] \leqslant 2\right\}.$$

From this definition and Proposition 2.1 of the previous lecture, we see that the terminology is consistent: up to a universal constant, $\|X\|_{\psi_2}$ is the subgaussian parameter of X. Moreover, $\|\cdot\|_{\psi}$ is indeed a norm on the vector space of random variables *on the same probability space*: one has symmetry $\|-X\|_{\psi_2} = \|X\|_{\psi_2}$, positive homogeneity $\|\lambda X\|_{\psi_2} = |\lambda|\,\|X\|_{\psi_2}$, and subadditivity a.k.a. triangle inequality: $\|X+Y\|_{\psi_2} \leqslant \|X\|_{\psi_2} + \|Y\|_{\psi_2}$.

Exercise 1.1. *Show that $\|X+Y\|_{\psi_2} \leqslant \|X\|_{\psi_2} + \|Y\|_{\psi_2}$ and exemplify $(X,Y) \in \mathbb{R}^2$ that attain the equality.*

2 General Orlicz norms

Here we briefly discuss general Orlicz norms $\|\cdot\|_{\psi_2}$. For more on this, see [Pol90, BK00] and [KC18, Sec. 3].

Definition 3. *Function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is a* Young function *if it is convex, increasing, and such that $\psi(0) = 0$.*

Exercise 2.1. *Verify the following properties of Young functions:*

- *Super-additivity:*
$$\psi(a+b) \geqslant \psi(a) + \psi(b) \quad \forall a, b \in \mathbb{R}_+.$$
- *Super-homogeneity:*
$$\psi(\lambda x) \geqslant \lambda\psi(x) \quad \forall \lambda \geqslant 1.$$
Note that for $\lambda \in \mathbb{N}$ this follows from super-additivity.

You may assume ψ is $C^2(\mathbb{R}_+)$ (by Alexandrov's theorem, ψ is C^2 almost everywhere on $\mathbb{R}_+$).

Definition 4. *Given a Young function ψ, the corresponding Orlicz norm of a random variable $X \in \mathbb{R}$ is*

$$\|X\|_{\psi} := \min\left\{K > 0 : \mathbb{E}\left[\psi\left(\frac{|X|}{K}\right)\right] \leqslant 1\right\}.$$

One may show that any Orlicz norm is, actually, a *seminorm* over the sigma-algebra of random variables on a common probabilty space; moreover, it is a *norm* when ψ is *strictly* convex. Consistently with our previous notation, the subgaussian norm $\|\cdot\|_{\psi_2}$ is actually the Orlicz norm corresponding to $\psi_2(x) := e^{x^2} - 1$, which is a legitimate Young function. Now, recall that the expected maximum of N random variables with subgaussian norm K grows at worst as $K\sqrt{\log(N+1)}$, i.e. as $\psi_2^{-1}(N)$. This generalizes to all Orlicz norms:

Proposition 2.1. *Given an Orlicz norm $\|\cdot\|_{\psi}$, let $X_j \in \mathbb{R}$ be such that $\|X_j\|_{\psi} \leqslant K_j$ for all $j \in [N]$. Then*

$$\mathbb{E}\left[\max_{j\in[N]} |X_j|\right] \leqslant \psi^{-1}(N) \max_{j\in[N]} K_j.$$

Exercise 2.2. *Prove Proposition 2.1. It suffices to show that $\mathbb{E}\left[\max_{j\in[N]} |Z_j|\right] \leqslant \psi^{-1}(N)$ when $\|Z_j\|_{\psi} \leqslant 1$.*

3 Interlude: ψ_α-norms and sub-Weibull distributions

This material is supplementary; it shall also be used in a later-on, in the context of estimating higher-order moment tensors (which is also a supplementary topic). For any $\alpha \geqslant 1$, consider the function $\psi_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$,

$$\psi_\alpha(x) := \exp(x^\alpha) - 1.$$

These are Young (and strictly convex, in fact), and we already know that $\alpha = 2$ gives the subgaussian norm. In Lecture 6: Section 1 we shall consider the so-called *subexponential or* $\|\cdot\|_{\psi_1}$*-norm*, corresponding to $\alpha = 1$. From these definitions, it is clear that $|X^2|_{\psi_1} = \|X\|_{\psi_2}^2$, so $\|\cdot\|_{\psi_1}$ naturally arises in the context of controlling linear combinations of the *squares* of independent subgaussians (a.k.a. χ^2-type statistics). Deferring further discussion to Lecture 6: Section 1, let us only remark here that, naturally, one would hope to use "ψ_α-norm" with $\alpha < 1$ to the $p > 2$ power of a subgaussian, as

$$|X^p|_{\psi_{2/p}} = \|X\|_{\psi_2}^p,$$

where one *could* take $\exp(x^\alpha) - 1$ in the role of $\psi_\alpha(x)$, for any $\alpha > 0$ (as already done for $\alpha \geqslant 1$). Yet, there is a technical issue: for $\alpha \in (0,1)$, the function $\widetilde{\psi}_\alpha(x) := \exp(x^\alpha) - 1$ on $\mathbb{R}_+$ is *nonconvex* (and so *not* a Young function). Fortunately, inspection shows that $\widetilde{\psi}_\alpha''(x) \geqslant 0$ for x large enough. This suggests to adjust $\widetilde{\psi}_\alpha$ via linear interpolation near the origin, and contextualizes the following exercise.

Exercise 3.1. *For $\alpha \in (0,1)$, define the threshold $x_\alpha = \alpha^{-1/\alpha}$ and the following function $\psi_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$:*

$$\psi_\alpha(x) := \exp(x^\alpha)\, \mathbb{1}_{x \geqslant x_\alpha} + \frac{e^{1/\alpha}}{x_\alpha} x\, \mathbb{1}_{x < x_\alpha}. \tag{12}$$

For intuition, plot $\psi_{1/2}$ and $\widetilde{\psi}_{1/2}$ in one plot. Show the following properties of ψ_α for all $\alpha \in (0,1)$.

1. *ψ_α is C_1-smooth and is a Young function.*
2. *$\psi_\alpha(x_\alpha) = \exp(x_\alpha^\alpha)$ and $\psi_\alpha'(x_\alpha) = \exp(x^\alpha)'|_{x=x_\alpha}$. For all $x, t \in \mathbb{R}_+$, one has*

$$\exp(x^\alpha) - 1 \leqslant \psi_\alpha(x) \leqslant \exp(x^\alpha),$$
$$\log^{1/\alpha}(t) \leqslant \psi_\alpha^{-1}(t) \leqslant \log^{1/\alpha}(t+1).$$

   Hint 1: *for the upper bound in the first line, study the monotonicity of $x \exp(-x^\alpha)$ on $x \in (0, x_\alpha)$.*
   Hint 2: *for the lower bound in the first line, study the monotonicity of $\frac{e^{1/\alpha}}{x_\alpha} x - \exp(x^\alpha)$ on $x \in (0, x_\alpha)$.*
   Hint 3: *the second line follows from the first line (draw a plot).*

Exercise 3.2 (Sub-Weibull distributions)**.** *Define X to be (K,α)-sub-Weibull if it holds that*

$$\mathbb{E}\left[\exp\left(\frac{|X|^\alpha}{K^\alpha}\right)\right] \leqslant 1.$$

1. *Using the results of Exercise 3.1, conclude that if X is (K,α)-sub-Weibull, then $\|X\|_{\psi_\alpha} \leqslant K$.*
2. *Using the results of Exercise 3.1 and Proposition 2.1, conclude that if each X_j is (K_j,α)-sub-Weibull,*

$$\mathbb{E}\left[\max_{j\in[N]} |X_j|\right] \leqslant \log^{1/\alpha}(N+1) \max_{j\in[N]} K_j.$$

3. *Clearly, power p of a σ-sub-Gaussian random variable is (K,α)-sub-Weibull with $K = \sigma^p$ and $\alpha = \frac{2}{p}$. Show the near-converse: if $\|X\|_{\psi_\alpha} \leqslant K$ (in particular if X is (K,α)-sub-Weibull), then for any $p \geqslant 0$*

$$\|X^p\|_{\psi_{\alpha/p}} \leqslant (1 + \mathbb{1}_{\alpha<1}) K^p,$$

   In particular, power $\frac{\alpha}{2}$ of an α-sub-Weibull random variable is subgaussian. Hint: *use super-additivity.*

Lecture 4: Subgaussian random vectors and concentration

Disclaimer. This lecture covers [Ver12, Lemma 5.39] but also discusses some tangentially related topics. We shall first observe that jointly independent subgaussian random variables remain subgaussian under linear combinations. This property mimics that of the Gaussian class of distributions. We shall then use these results to introduce—and then investigate—the natural class of subgaussian random *vectors,* for which we shall prove an *approximate rotation invariance* property – which mimicks the rotation invariance of $\mathcal{N}(0,1)$.

1 Linear combinations of independent subgaussian random variables

Recall from the last lecture that a distribution $X \in \mathbb{R}$ is *K-sub-Gaussian* if its MGF $M_X(t) := \mathbb{E}e^{tX}$ satisfies

$$M_X(t) \leqslant \exp\left(t^2K^2/2\right) \quad \forall t \in \mathbb{R}.$$

Then, for some $c > 0$ the tail function $\Phi_X(x) = \mathbb{P}\{X \geqslant x\}$ decays as $\Phi_X(x) \leqslant \exp(-cx^2K^{-2})$ on $\mathbb{R}$. Let us now consider the *product-distribution setup*: $X_1, X_2, \ldots, X_n$ are jointly distributed on $\mathbb{R}$ and independent. That is, the random vector $X_{1:n} := (X_1, \ldots, X_n) \in \mathbb{R}^n$ has a product distribution: $f_{1:n}(x_{1:n}) = \prod_{i\in[n]} f_i(x_i)$ where $x_{1:n} := (x_1, \ldots, x_n)$ is a candidate value of $X_{1:n}$, $f_{1:n}$ is the p.d.f. of $X_{1:n}$, and f_i is the p.d.f. of X_i.

Proposition 1.1 (Sums of independent subgaussians are subgaussian)**.** *Let $S_n := \sum_{i=1}^n X_i$, where each X_i is K_i-subgaussian, and $X_1, \ldots, X_n$ are independent. Then S_n is subgaussian with parameter $(\sum_{i=1}^n K_i^2)^{1/2}$.*

Proof. We have already seen this in Lecture 1: Eq. (3) for two random variables; here is the general case:

$$M_{S_n}(t) := \mathbb{E}\left[\exp\left(t\sum_{i=1}^n X_i\right)\right] \overset{(*)}{=} \prod_{i=1}^n \mathbb{E}\left[\exp\left(tX_i\right)\right] \leqslant \prod_{i=1}^n \exp\left(\frac{1}{2}t^2K_i^2\right) = \exp\left(\frac{1}{2}t^2\sum_{i=1}^n K_i^2\right),$$

Here in $(*)$ we used independence of the random variables $\exp(tX_i)$, which are independent since X_i's are. □

A simple check shows that subgaussianity is a positive-homogeneous property: $X \in \mathbb{R}$ is K-subgaussian if and only if λX is λK subgaussian for any $\lambda > 0$. More generally, if $|X|$ is K-subgaussian if and only if $|\lambda X|$ is $|\lambda|K$-subgaussian for any $\lambda \in \mathbb{R}$. Together with the previous result, this implies the following one.

Corollary 1.1. *Let $X := X_{1:n} \in \mathbb{R}^n$ be a random vector with independent entries, $|X_i|$ being K_i-subgaussian. Then, for any fixed vector $a \in \mathbb{R}^n$, the linear combination $\langle a, X\rangle = \sum_{i\in[n]} a_iX_i$ is K-subgaussian, with*

$$K = \left(\sum_{i\in[n]}^n a_i^2K_i^2\right)^{1/2}.$$

We have just seen that all one-dimensional marginals $\langle a, X\rangle$ of a random vector X with *independent* subgaussian entries turn out to be subgaussian as well, with variance proxies depending on a in a quantifiable manner. However, *entrywise* independence is a very strong assumption. In the sequel, we relax it by "*hard-wiring*" the subgaussian behavior of one-dimensional marginals into the *definition* of subgaussianity in $\mathbb{R}^d$.

2 Subgaussian random vectors

Definition 5. *A random vector $X \in \mathbb{R}^d$ is called* K-subgaussian*, denoted $\|X\|_{\psi_2} \leq K$, if its* one-dimensional marginal *$\langle X, u\rangle$, for any unit-norm vector u, is a K-subgaussian random variable. In other words, one has*

$$\sup_{u\in\mathbb{S}^{d-1}} \|\langle X, u\rangle\|_{\psi_2} \leq K$$

with $\mathbb{S}^{d-1} := \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$. Equivalently, X is K-subgaussian *if $\|\langle X, w\rangle\|_{\psi_2} \leq K\|w\|_2$ for all $w \in \mathbb{R}^d$.*

Definition 6. *The ψ_2-norm of $X \in \mathbb{R}^d$ is defined as the largest ψ_2-norm of its one-dimensional projections:*

$$\|X\|_{\psi_2} := \sup_{u \in \mathbb{S}^{d-1}} \|\langle X, u\rangle\|_{\psi_2}.$$

Recall that $X \sim \mathcal{N}(0, I_d)$ is rotationally invariant: for any (deterministic) orthogonal matrix $Q \in \mathbb{R}^{d\times d}$, i.e. such that $QQ^\top = Q^\top Q = I_d$, the vector $QX \sim \mathcal{N}(0, I_d)$ as well; in particular, one has $\langle u, X\rangle \sim \mathcal{N}(0, 1)$ for any $u \in \mathbb{S}^{d-1}$. As the next result shows, this property approximately generalizes to subgaussian vectors.

Proposition 2.1 (Approximate rotational invariance of subgaussian random vectors[2])**.** *For some universal $c > 0$, if the entries of $X \in \mathbb{R}^d$ are independent and K-subgaussian then X is K-subgaussian.*

Proof. Observe that this claim is nothing else but a reformulation of Corollary 1.1. □

Remark (Orlicz norms for vectors)**.** *Of course, one can extend Orlicz norms to multivarite distributions in the same fashion: given a Young function ψ, the corresponding Orlicz norm of a random vector $X \in \mathbb{R}^d$ is*

$$\|X\|_{\psi} := \sup_{u \in \mathbb{S}^{d-1}} \|\langle X, u\rangle\|_{\psi}.$$

We shall use this later, in particular in the lectures devoted to χ^2-type statistics and covariance estimation.

3 Inequalities due to Hoeffding and Khinchine

Applying Proposition 2.1 again to convert an MGF bound to a tail bound, we arrive at *Hoeffding's inequality.*

Proposition 3.1 (Hoeffding's inequality)**.** *Let $\{X_i\}_{i=1}^n$ be a a sequence of independent random variables such that each X_i is sub-Gaussian with parameter K_i. Then, for any $\varepsilon \geqslant 0$,*

$$\mathbb{P}\left\{\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geqslant \varepsilon\right\} \leqslant \exp\left(-\frac{\varepsilon^2}{2\sum_{i=1}^n K_i^2}\right).$$

On the other hand, we also have a bound for the L_p-norms $\|Y\|_{L_p} := \mathbb{E}^{1/p}[|Y|^p]$ of a linear combination.

Proposition 3.2 (Khintchine's inequality)**.** *Let $X_1, X_2, \dots, X_n$ be a sequence of independent, K-subgaussian random variables such that $\mathbb{E}[X_i] = 0$ for all $i \in [n]$. Then, for any $p \geqslant 1$, it holds that*

$$\left\|\sum_{i=1}^n w_i X_i\right\|_{L_p} \lesssim K\sqrt{p}\|w\|_2.$$

Remark. *Assuming $\mathbb{E}[X_i^2] = 1$, the bound is sharp for $p = 2$:* $\|\sum_{i=1}^n w_i X_i\|_{L_2} = \sqrt{\sum_{i=1}^n w_i^2 \mathbb{E}[X_i^2]} = \|w\|_2$.

[2]Our terminology follows Vershynin's [Ver12]: he defined $\|\cdot\|_{\psi_2}$ via moments rather than MGF, so his version of this proposition has distortion of K by a constant factor due to using Proposition 2.1. In our case, "approximate" can be dropped.

Lecture 5: Gaussian maxima and weighted union bounds

Disclaimer. In this lecture, we focus on the classical topic of controlling the uniform norm of a Gaussian vector. The subgaussian case is considered only briefly: the upper bounds extend from the Gaussian case, but not the lower bounds. We also discuss possible refinement of these bounds, leading to an open problem.

Throughout, we assume $\xi \in \mathbb{R}^n$ is standard Gaussian: $\xi \sim \mathcal{N}(0, I_n)$. Then $X = A\xi$, for a fixed $A \in \mathbb{R}^{m\times n}$, is also Gaussian, namely $X \sim \mathcal{N}(0, \Sigma)$ with $\Sigma = AA^\top$. We shall focus on bounding the sup-norm of X, i.e.

$$\|X\|_\infty = \|A\xi\|_\infty = \max_{j\in[m]} |a_j^\top \xi|$$

where $a_j^\top$ are the rows of A, and $X_j = a_j^\top \xi$ are the entries of X.

1 Diagonal case

We start with the case $A = \sigma I_n$. Recall that Lecture 2: Proposition 2.2 implies that, w.p. $\geqslant 1-\delta$, one has

$$\|X\|_\infty \leqslant \sigma\sqrt{2\log(2n\delta^{-1})}. \tag{13}$$

Let us repeat the argument explicitly: $\|X\|_\infty$ is the maximum of n random variables $|X_j|$, $X_j \sim \mathcal{N}(0,\sigma^2)$, so

$$\mathbb{P}\{\|X\|_\infty \geqslant r\} \leqslant \sum_{j\in[n]} \mathbb{P}\{|X_j| \geqslant r\} = 2n\Phi(r/\sigma)$$

by the union bound. Plugging in the Gaussian tail bound $\Phi(r/\sigma) \leqslant \exp(-\frac{r^2}{2\sigma^2})$ of the MGF method, and solving the equation $2n\exp(-\frac{r^2}{2\sigma^2}) = \delta$ for $r = r(\delta)$, we arrive at (13). Note that we have *never* used that X_j's are independent, and the bound does not require this. However, when X_j are highly dependent, it might "overshoot:" indeed, for $X_1 = \ldots = X_n \sim \mathcal{N}(0,\sigma^2)$ we get $\mathbb{P}\{\|X\|_\infty \geqslant r\} = \mathbb{P}\{|X_1| \geqslant r\} = 2\Phi(r/\sigma)$, that is

$$\|X\|_\infty \geqslant \sigma\Phi^{-1}(\tfrac{1}{2}\delta)$$

with probability $1-\delta$. By Lecture 2: Exercise 2.5, this in particular implies that with probability $\geqslant 1-\delta$,

$$\|X\|_\infty \geqslant \sigma\sqrt{2\log(c\delta^{-1})} \quad \text{as long as} \quad \delta \leqslant c_0 < 1.$$

Note that regardless of the interdependence of X_j's, this gives a (trivial) *lower bound* for the $(1-\delta)$-quantile of $\|X_\infty\|$. On the other hand, in the case of independent entries, the *upper bound* (13) turns out to be tight.

Proposition 1.1. *Assume that $X \sim \mathcal{N}(0,\sigma^2 I_n)$, then for some constants $c, c_0, c_1 > 0$, with probability $\geqslant 1-\delta$*

$$\|X\|_\infty \geqslant \sigma\sqrt{2\log(cn\delta^{-1})}\left(1 + \frac{1+\log\log(2n)}{2\log(2n)}\right)^{-1} \tag{14}$$

as long as $n \geqslant 2$ and $\delta^{-1} \geqslant c_0\log(2n)^{1+c_1}$.

The proof of this proposition is technical, was omitted in class, and can be skipped at first (see Section 4).

Remark. *In the above argument, the key idea is to use the "second level" of the inclusion-exclusion formula.*

Executive summary. What we have so far: if $X_j \sim \mathcal{N}(0,\sigma^2)$ marginally for all $j \in [n]$, then the bounds

$$\sigma\sqrt{2\log(2\delta^{-1})} \leqslant \|X\|_\infty \leqslant \sigma\sqrt{2\log(2n\delta^{-1})} \tag{15}$$

hold with probability $\geqslant 1-\delta$ each (albeit with the lower bound restricted to $\delta \leqslant c_0 < 1$). Both these bounds can actually be (nearly) attained: the lower one when X_j's are equal, and the upper one when they are independent. Note also that the upper bound extends to the subgaussian case, since it only used the MGF bound for the tail function. (This cannot be said about the lower bounds; yet, one can furnish natural analogues of the corresponding results, expressed in terms of the quantile function of the entries X_j of X.)

2 General case: vanilla union bound

For $A = [a_1^\top; \cdots; a_m^\top]$, we let $\sigma_j := \|a_j\|_2$ and note that $X_j = a_j^\top \xi \sim \mathcal{N}(0, \sigma_j^2)$. Clearly, (15) generalizes to

Proposition 2.1. *Let $A \in \mathbb{R}^{m\times n}$, with σ_j being the norm of j^{th} row. Each of the following holds w.p. $\geqslant 1-\delta$:*

$$\max_{j\in[m]} \sigma_j \sqrt{2\log(2\delta^{-1})} \leqslant \|X\|_\infty \leqslant \max_{j\in[m]} \sigma_j \sqrt{2\log(2m\delta^{-1})}. \tag{16}$$

Here, the upper bound is valid for any $\delta \in (0,1)$, and the lower bound is restricted to $\delta \in (0, c_0]$ with $c_0 < 1$.

Proof. For the upper bound, proceeding as in the proof of (13) we get

$$\begin{aligned}\mathbb{P}\left\{\|X\|_\infty \geqslant r\right\} \leqslant \sum_{j\in[m]} \mathbb{P}\left\{|X_j| \geqslant r\right\} = 2\sum_{j\in[m]} \Phi(r/\sigma_j) \leqslant 2m \max_{j\in[m]} \Phi(r/\sigma_j) \leqslant 2m \max_{j\in[m]} \exp\left(-\frac{r^2}{2\sigma_j^2}\right) \\ = 2m\exp\left(-\frac{r^2}{2\max_{j\in[m]}\sigma_j^2}\right).\end{aligned}$$

For the lower bound, we proceed as in the diagonal case but choose the "worst" X_j to get $\max_{j\in[m]}\sigma_j$. □

Note that there is an $O(\sqrt{\log m})$ additive gap between the upper and lower quantile bounds in (16). This is rather crucial: the two bounds become of the same order only in the high-confidence regime: $\delta = O(m^{-1})$. Our subsequent discussion is focused on narrowing it down. The first suggestion is a simple and cute trick.

3 Refinement via weighted union bound

Perhaps somewhat surprisingly, it turns out that there is a computationally cheap way to *universally* improve the upper bound in (16). Even more surprising is the fact that this trick seems to be inpublished, and even some experts in the field are unaware of it, as learned from personal interaction with them. Here is the idea:

> *When taking the union bound, one can distribute the "budget" δ over the probabilities p_j of the violation events* nonuniformly, *so as to minimize the resulting upper bound on the quantile.*

To make this precise, note that each of the events

$$E_j := \left\{|X_j| > \sigma_j\sqrt{2\log\left(\frac{2}{\delta_j}\right)}\right\}$$

holds with probability at least $1-p_j$. Hence, for any selection of δ_j's such that $\sum_j \delta_j = \delta$, one has

$$\|X\|_\infty \leqslant \max_{j\in[m]} \sigma_j\sqrt{2\log\left(\frac{2}{\delta_j}\right)}$$

holds w.p. $\geqslant 1-\delta$. If we define $p_j := \delta_j/\delta$, computing the tightest of these upper bounds amounts to solving

$$Q_\delta := \min_{p\in\Delta_m} \underbrace{\max_{j\in[m]} \sigma_j\sqrt{2\log\left(\frac{2}{\delta p_j}\right)}}_{q_\delta(p)}.$$

Minimization is on the standard simplex $\Delta_m := \{p \in \mathbb{R}^d_+ : p^\top \mathbb{1}_m = 1\}$, and $q_\delta(p)$ is convex on $\mathbb{R}^d_+$ due to the convexity of $-\log(\cdot)$ on $\mathbb{R}_+$, so this is a convex optimization problem . This improves over Proposition 2.1:

$$Q_\delta \leqslant q_\delta(\tfrac{1}{m}\mathbb{1}_m) = \max_{j\in[m]} \sigma_j\sqrt{2\log(2m\delta^{-1})}. \tag{17}$$

Computationally, the improvement comes for free: as the next exercise shows, one can compute Q_δ explicitly.

Exercise 3.1. *Let $S_m := \sum_{j\in[m]} \sigma_j^2$ and $M_m = \max_{j\in[m]} \sigma_j^2$. Show the following.*

1. *It holds that*
$$\tfrac{1}{2}Q_\delta^2 \leqslant \overline{\tfrac{1}{2}Q_\delta^2} := \max_{j\in[m]} \sigma_j^2 \log\left(\frac{2S_m}{\delta\sigma_j^2}\right).$$
Hint: *choose $p \in \Delta_m$ appropriately.*

2. *"Softmax inequality" states that, for any $x \in \mathbb{R}^m$ and $\beta \in \mathbb{R}$,*
$$\max_{j\in[m]} x_j \leqslant \frac{1}{\beta}\log\sum_{j\in m}\exp(\beta x_j) \leqslant \log m + \max_{j\in[m]} x_j.$$
The lower bound is trivial; the upper is Jensen's. Use it to verify that $\overline{\frac{1}{2}Q_\delta^2} \leqslant M_m \log(\frac{2m}{\delta})$.

3. *Show that for $\delta \leqslant 2e^{-1}$, one has*
$$\overline{\tfrac{1}{2}Q_\delta^2} = M_m \log\left(\frac{2S_m}{\delta M_m}\right).$$

Exercise 3.2 (Further refinement)**.**

1. *Find $\beta \in \mathbb{R}$ that minimizes the right-hand side in the (obviously correct) inequality*
$$\tfrac{1}{2}Q_\delta^2 \leqslant \max_{j\in[m]} \sigma_j^2 \log\left(\frac{2\sum_{k\in[m]} e^{-\beta\sigma_k^2}}{\delta e^{-\beta\sigma_j^2}}\right).$$

2. *Find $\frac{1}{2}Q_\delta^2$ explicitly.* Hint: *use that* $\max_{j\in[m]} f_j = \max_{\lambda\in\Delta_m}\sum_{j\in[m]}\lambda_j f_j$ *to express $\frac{1}{2}Q_\delta^2$ as the value of a convex-concave min-max problem; then use Sion's minimax theorem [Kin05] to switch* min *and* max.

4 Interlude: proof of Proposition 1.1

Note that $|X_1|, \dots, |X_n|$ are jointly independent, so for any fixed level $r > 0$, the "violation events"
$$E_j := \{|X_j| \geqslant r\}$$
are independent as well, and
$$\mathbb{P}\{\|X\|_\infty \geqslant r\} = \mathbb{P}\left(\bigcup_{j\in[n]} E_j\right) \geqslant \sum_{j\in[n]}\mathbb{P}(E_j) - \sum_{1\leqslant j<k\leqslant n}\mathbb{P}(E_j\cap E_k) = \sum_{j\in[n]} p_j - \sum_{1\leqslant j<k\leqslant n} p_j p_k$$
where $p_j := \mathbb{P}(E_j)$ is the probability of j^{th} violation event; in the inequality we used the "second level" of the inclusion-exclusion formula, and the last identity is by the independence of E_j and E_k. Whence we get
$$\begin{aligned}\mathbb{P}\{\|X\|_\infty \geqslant r\} &\geqslant \sum_{j\in[n]} p_j\left(1 - \frac{1}{2}\sum_{k\neq j} p_k\right) \geqslant \left(\sum_{j\in[n]} p_j\right)\left(1 - \frac{1}{2}\sum_{j\in[n]} p_j\right)\\ &= 2n\Phi(r/\sigma)\left(1 - n\Phi(r/\sigma)\right).\end{aligned}$$
For $r \geqslant \sigma\sqrt{2\log(2n)}$, one has $\Phi(r/\sigma) \leqslant \exp(-\frac{r^2}{2\sigma^2}) \leqslant \frac{1}{2n}$, and therefore
$$\mathbb{P}\{\|X\|_\infty \geqslant r\} \geqslant n\Phi(r/\sigma) \geqslant \frac{\sigma n}{r\sqrt{2\pi}}\left(1 - \frac{\sigma^2}{r^2}\right)\exp\left(-\frac{r^2}{2\sigma^2}\right) \geqslant \frac{c\sigma n}{r}\exp\left(-\frac{r^2}{2\sigma^2}\right)$$

for a universal constant $c > 0$. Assigning the left-hand side to δ and taking the logarithm, this translates to

$$\frac{r^2}{2\sigma^2} + \log\left(\frac{r}{\sigma}\right) \geqslant \log(cn\delta^{-1}),$$

where $r = r(\delta)$ is the $(1-\delta)$-quantile of $\|X\|_\infty$. Now, to handle the correction term, note: if $r \geqslant \sigma\sqrt{2\log(2n)}$,

$$\frac{r^2}{2\sigma^2} + \log\left(\frac{r}{\sigma}\right) = \frac{r^2}{2\sigma^2}\left(1 + \frac{2\sigma^2\log(r/\sigma)}{r^2}\right) \leqslant \frac{r^2}{2\sigma^2}\left(1 + \frac{\log(\sqrt{2\log(2n)})}{\log(2n)}\right) \leqslant \frac{r^2}{2\sigma^2}\left(1 + \frac{1+\log\log(2n)}{2\log(2n)}\right).$$

This gives (14), so it only remains to verify that our standing assumption $r \geqslant \sigma\sqrt{2\log(2n)}$ is fulfilled whenever $\delta^{-1} \geqslant c_0\log^2(2n)$. To this end, squaring (14) and using the assumption on δ, we get

$$\begin{aligned}
\frac{r^2}{\sigma^2} \geqslant 2\log(cn\delta^{-1})\left(1 + \frac{\log(2\log(2n))}{2\log(2n)}\right)^{-2} &\geqslant 2\log(2n)\left(1 + \frac{\log(\frac{c}{2}\delta^{-1})}{\log(2n)}\right)\left(1 + \frac{\log(2\log(2n))}{2\log(2n)}\right)^{-2} \\
&\geqslant 2\log(2n)\left(1 + \frac{\log(\frac{c}{2}\delta^{-1})}{\log(2n)}\right)\left(1 + \frac{(1+c_1)\log(2\log(2n))}{\log(2n)}\right)^{-1} \\
&\geqslant 2\log(2n) \quad \text{once} \quad \log\left(\tfrac{c}{2}\delta^{-1}\right) \geqslant (1+c_1)\log(2\log(2n)). \quad \square
\end{aligned}$$

5 Challenge: tractable approximation of Gaussian volume (and quantiles)

For general matrix A, we still have a gap between upper and lower bounds, namely

$$\|\vec{\sigma}^2\|_\infty \log(2\delta^{-1}) \leqslant \tfrac{1}{2}Q_\delta^2 \leqslant \|\vec{\sigma}^2\|_\infty \log\left(2\delta^{-1}\mathtt{r}(A)\right),$$

where Q_δ is the $1-\delta$ quantile of $\|X\|_\infty$, and

$$\mathtt{r}(A) = \frac{\|\vec{\sigma}^2\|_1}{\|\vec{\sigma}^2\|_\infty} = \frac{\|A\|_{2,2}^2}{\|A\|_{2,\infty}^2}$$

can be thought of as (some version of) the effective rank of A. In general, the best we can say is that $\mathtt{r}(A) \leqslant m$. Which leads us to the following problem:

> **Tractable quantiles for Gaussian maxima**. Given $\delta \in (0,1)$, $A \in \mathbb{R}^{m\times n}$, and $\varepsilon \in (0,1)$, approximate $Q_\delta(A)$, the $(1-\delta)$-quantile of $\|X\|_\infty$ with relative error ε, i.e. find $\hat{Q}_\delta(A)$ such that
>
> $$(1-\varepsilon)\hat{Q}_\delta(A) \leqslant Q_\delta(A) \leqslant (1+\varepsilon)\hat{Q}_\delta(A).$$
>
> This has to be done via a tractable method, with complexity polynomial in $m, n, 1/\delta$ and $\log(1/\varepsilon)$.

Equivalent formulation in terms of Gaussian volume. Note that the inverse map for $r_A(1-\delta) := Q_\delta(A)$ is the probability that $\|A\xi\|_\infty \leqslant r$, i.e., the *Gaussian volume* $\gamma_n(P) := \mathbb{P}\{\xi \in P\}$ of the polytope

$$P = P_r(A) := \{u \in \mathbb{R}^n : |a_j^\top u| \leqslant r \quad \forall j \in [m]\},$$

as a function of its "radius" r. As such, we reduce the above problem to the problem of approximating the Gaussian volume of a symmetric convex polytope in $\mathbb{R}^n$ with $2m$ facets, for which the normal vectors are $\pm a_j$. [3] In fact, there are general algorithms to approximate $\gamma_n(P(A))$ via sampling; the fastest one

[3] As it stands, the above problem asks to compute $r > 0$ such that, for the scaled polytopes $P_{r-\varepsilon r}(A)$ and $P_{r+\varepsilon r}A$, one has

$$\gamma_n(P_{r-\varepsilon r}(A)) \leqslant 1-\delta \leqslant \gamma_n(P_{r+\varepsilon r}(A)).$$

However, if we can approximate $\gamma_n(P_r(A))$ up to error Δ in $\mathsf{poly}(1/\Delta)$—as is the case e.g. for sampling algorithms [CV14]—then the standing problem can be solved via binary search, with $\log(1/\varepsilon)$ overhead. This motivates shooting for $\log(1/\varepsilon)$ complexity.

running in $O(n^3)$ [CV14] in $\tilde{O}(n^3)$ in the regime $m \geqslant n$. On the other hand, *exact* computation of *ordinary* volume $\int \mathbb{1}_A(u)du$ is nothing else but the computation of $\det(A)$, and so cannot be done faster than in the matrix inversion time n^ω. Nevertheless, it is an interesting problem to get a *deterministic* method for efficiently approximating $\gamma_n(P(A))$. The only known result in this avenue seems to be [BR24], which gives an approximation up to $O(1)^n$ factor. Even an $O(1)$-approximation algorithm, or more precisely with any $\varepsilon \ll \log(m)$, would be desirable. Below we give some preliminary ideas of one such promising approach.

Approach via the inclusion-exclusion formula Let $E_j := \{|X_j| > r_j\}$ be the violation events, then

$$\mathbb{P}\left(\bigcup_{j\in[m]} E_j\right) = \sum_{k\in[m]} (-1)^{k-1} P_k$$

where

$$P_k := \sum_{1\leqslant j_1<\cdots<j_k\leqslant m} \mathbb{P}\left(E_{j_1}\cap\cdots\cap E_{j_k}\right)$$

is the sum over k-tuples, and we have inequalities cutting the decomposition at any level k. In particular,

$$\mathbb{P}\left\{\|X\|_\infty > \max_{j\in[m]} r_j\right\} \leqslant \mathbb{P}\left(\bigcup_{j\in[m]} E_j\right) \leqslant P_1 - P_2 + P_3$$

where the first inequality follows from the fact that $\frac{\max_j a_j}{\max_j b_j} \leqslant \max_j \frac{a_j}{b_j}$ for any $a_j, b_j > 0$. (Check this!) Now,

$$P_1 = \sum_{j\in[m]} \mathbb{P}(E_j) := \delta \sum_{j\in[m]} p_j,$$

where we *defined* $p_j := \delta_j^{-1}\mathbb{P}(E_j)$, with $\delta \in (0,1)$ being a parameter. Moreover, we can upper-bound the next term $-P_2$ in terms of p_j's via the Gaussian correlation inequality.

Theorem 5.1 (e.g. [LM17]). *For any convex, compact, and symmetric sets $K, L \subseteq \mathbb{R}^n$, one has*

$$\gamma_n(K\cap L) \geqslant \gamma_n(K)\gamma_n(L).$$

This bound is sharp, proved in the general case by Thomas Royen a few years ago [Roy14]. In our case, we can invoke it with K and L being symmetric slabs (this has been done in 1960s [Šid67]), to conclude that

$$-P_2 = -\sum_{1\leqslant j_1<j_2\in m} \mathbb{P}\left(E_{j_1}\cap E_{j_2}\right) \leqslant -\delta^2 \sum_{1\leqslant j_1<j_2\leqslant m} p_{j_1}p_{j_2}$$

That is, $P_1 - P_2 \leqslant \delta e_1(p_1,\dots,p_m) - \delta^2 e_2(p_1,\dots,p_m)$ where e_k is the k^{th} elementary symmetric polynomial;

$$\mathbb{P}\left\{\|X\|_\infty > \max_{j\in[m]} r_j\right\} \leqslant \delta e_1(p_1,\dots,p_m) - \delta^2 e_2(p_1,\dots,p_m) + \delta^3 T_A(r_1,\cdots,r_m),$$

and $T_A(r_1,\dots,r_m) = \delta^{-3}P_3$ as the function of $r_1,\dots,r_m$. As before, one has $\delta p_j = 2\Phi(r_j/\sigma_j) \leqslant 2\exp\left(-\frac{r_j^2}{2\sigma_j^2}\right)$, with a nearly matching lower bound when $\delta \leqslant c < 1$.

$$\tfrac{1}{2}Q_\delta^2 \leqslant \min_{p\in\hat{\Delta}_m(\delta)} \left\{\tfrac{1}{2}q_\delta^2(p) := \max_{j\in[m]} \sigma_j^2 \log\left(\frac{2}{\delta p_j}\right)\right\} \tag{18}$$

where the set

$$\hat{\Delta}_m(\delta) := \{p\in\mathbb{R}_+^m : 1 = e_1(p_1,\dots,p_m) - \delta e_2(p_1,\dots,p_m) + \delta^2 T_A(\sigma_1\Phi^{-1}(\tfrac{\delta}{2}p_1),\dots,\sigma_m\Phi^{-1}(\tfrac{\delta}{2}p_m))\}, \tag{19}$$

can be thought of an approximation of the simplex; in particular, $\hat{\Delta}_m(0) = \Delta_m$. Here are some questions:

1. Can one compute $T_A(r_1, \dots, r_m)$ efficiently? Note that

$$\delta^3 T_A(r_1, \dots, r_m) = \sum_{1 \leqslant j_1 < j_2 < j_3 \leqslant m} \mathbb{P}(E_{j_1} \cap E_{j_2} \cap E_{j_3}),$$

where each term $\mathbb{P}(E_{j_1} \cap E_{j_2} \cap E_{j_3})$ depends only on some triple $a_{j_1}, a_{j_2}, a_{j_3}$ and r_1, r_2, r_3. Explicitly,

$$\mathbb{P}(E_1 \cap E_2 \cap E_3) = \mathbb{P}\{|X_1| \leqslant r_1, |X_2| \leqslant r_2, |X_3| \leqslant r_3\} \quad \text{where} \quad (X_1, X_2, X_3) \sim \mathcal{N}(0, \Sigma_{\{1,2,3\}})$$

and $\Sigma_J = \Pi_J A A^\top \Pi_J$ is a $|J| \times |J|$ submatrix of $\Sigma = AA^\top$. $\mathbb{P}(E_1 \cap E_2 \cap E_3)$ can be computed in $O(1)$.

2. What can be said about the geometry of $\hat{\Delta}_m(\delta)$? Note that for $\psi_\delta(p) := e_1(p_1, \dots, p_m) - \delta e_2(p_1, \dots, p_m)$,

$$\psi_\delta(p) = \sum_{j \in [m]} p_j \left(1 - \frac{\delta}{2} \sum_{k \neq j} p_k\right),$$

$$\frac{\partial}{\partial p_j} \psi_\delta(p) = 1 - \delta \sum_{k \neq j} p_k,$$

$$\frac{\partial^2}{\partial p_j \partial p_k} \psi_\delta(p) = -\delta \mathbb{1}_{j \neq k}.$$

It might be helpful to change the parametrization back to $r_1, \dots, r_m$.

3. How much tighter is (18) compared to (??)?

4. Finally, it seems that the complexity of solving (18) is $\sim m^3$, whereas the sampling approaches are $\sim n^3$. Can we somehow "sketch" the problem to reduce its complexity?

Lecture 6: Bernstein's inequality and covariance estimation

Disclaimer. This is our first "data science" lecture: we dicsuss a qualitatively sharp concentration inequality for the sample covariance matrix of a subgaussian vector (and complete its proof in the next lecture). This material includes to [Ver12, Theorem 5.39], with some additional discussions. Subsequently, this result will be used in a more applied statistical context, in the analysuis of random-design linear regression and finite-sample results for generalized linear models.

1 Bernstein's inequality, a.k.a. the deviations of a χ^2-type statistic

We start with a basic result, useful in its own right. $Y \in \mathbb{R}$ is called K_1-*subexponential* if $\|Y\|_{\psi_1} \leqslant K_1$ where

$$\|Y\|_{\psi_1} := \inf\left\{K \geqslant 0 : \mathbb{E}\left[\exp\left(\frac{|Y|}{K}\right)\right] \leqslant 2\right\}.$$

For example, both χ^2_1 and $\chi^2_2 = \mathrm{Exp}(1)$ are $O(1)$-subexponential (with different but *universal* constants). From this definition it is immediately clear that $\|X^2\|_{\psi_1} = \|X\|^2_{\psi_2}$, i.e. X^2 is K_2^2-subexponential if and only if X is K_2-subgaussian. By Lecture 2: Proposition 2.1, $\|Y\|_{\psi_1} \leqslant K_1$ implies *subexponential moment growth*,

$$\|Y\|_{L_p} \lesssim K_1 p \quad \forall p \in \mathbb{N},$$

when Y is nonnegative (so that $Y = |Y| = X^2$ for some X). In fact, this implication remains valid in the general case of K_1-subexponential Y, by an argument similar to the one in Lecture 2: Proposition 2.1; and the reverse implication holds assuming $\mathbb{E}[Y] = 0$.

Similarly to the subgaussian case, a natural question is how the sum of independent subexponential random variables deviates from its expectation. It is answered by the following result.

Proposition 1.1 (Bernstein's inequality)**.** *Let $W_1, W_2, \dots$ be independent, centered, and $\|W_j\|_{\psi_1} \leqslant K_1$. Then for any deterministic sequence $a_1, a_2, \dots$, with probability $\geqslant 1 - \delta$*

$$\left|\textstyle\sum_j a_j W_j\right| \lesssim K_1 \|a\|_2 \sqrt{\log\left(\delta^{-1}\right)} + K_1 \|a\|_\infty \log\left(\delta^{-1}\right).$$

In particular, for any $n \in \mathbb{N}$ one has w.p. $\geqslant 1 - \delta$:

$$\left|\textstyle\sum_{j\in[n]} W_j\right| \lesssim K_1 \sqrt{n \log\left(\delta^{-1}\right)} + K_1 \log\left(\delta^{-1}\right).$$

In homework, you proved a version of this result for $W_j = Y_j - 1$ with $\sim \chi^2_1$, cf. [LM00], namely for $Y \sim \chi^2_n$,

$$|Y - n| \lesssim \sqrt{n \log\left(\delta^{-1}\right)} + \log\left(\delta^{-1}\right). \tag{20}$$

Generalization to subexponential distributions is analogous, since the proof relied on the MGF method. Note that Proposition 1.1 covers the non-uniform case $\|Z_j\|_{\psi_1} \leqslant K_j$ as well: by homogeneity of $\|\cdot\|_{\psi_1}$, this is equivalent to $\|a_j Z_j\|_{\psi_1} \leqslant K$ with $a_j = \frac{K}{K_j}$. Note also that, since $\|\cdot\|_{\psi_1}$ is a norm on the Borel sigma-algebra,

$$\|Y\|_{\psi_1} \leqslant \|Y - \mathbb{E}[Y]\|_{\psi_1} + |\mathbb{E}[Y]|,$$
$$\|Y - \mathbb{E}[Y]\|_{\psi_1} \leqslant \|Y\|_{\psi_1} + |\mathbb{E}[Y]|.$$

This allows to apply Proposition 1.1 to noncentered random variables, as the following simple result shows.

Proposition 1.2. *Let Z be zero-mean and K-subgaussian, then $W := Z^2 - \mathbb{E}[Z^2]$ is $O(K^2)$-subexponential.*

Proof. $\|W_j\|_{\psi_1} \leqslant \|Z_j\|^2_{\psi_2} + \mathrm{Var}\, Z_j = K^2 + \mathrm{Var}\, Z_j \lesssim K^2$. The last step is by Lecture 2: Proposition 2.1. □

In particular, for the χ^2-type statistic $S_n = \sum_{j\in[n]} Z_j^2$, where Z_j's are independent, K-subgaussian and isotropic (i.e. $\mathbb{E}[Z_j] = 0$ and $\mathbb{E}[Z_j^2] = 1$), one has

$$\left|\tfrac{1}{n} S_n - 1\right| \lesssim K^2 \left(\sqrt{\frac{\log\left(\delta^{-1}\right)}{n}} + \frac{\log\left(\delta^{-1}\right)}{n}\right) \tag{21}$$

with probability $\geqslant 1 - \delta$; here the first term dominates as long as $n \geqslant \log(\delta^{-1})$. We shall reuse this result.

2 Covariance estimation: the result

Recall that the operator (or spectral) norm of $A \in \mathbb{R}^{d\times d}$ is, by definition,

$$\|A\| := \sup_{u\in\mathbb{S}^{d-1}} \|Au\|_2,$$

where the supremum can instead be taken over the unit ball $\|u\|_2 \leqslant 1$. Equivalently, $\|A\|$ is the largest singular value of A, and $\|A\| = \max\{\lambda_{\max}(A), -\lambda_{\min}(A)\}$ if A is symmetric. Next we'll prove the following

Theorem 2.1 (cf. [Ver12, Theorem 5.39])**.** *Let $Z \in \mathbb{R}^d$ be isotropic ($\mathbb{E}[Z] = 0$ and $\mathbb{E}[ZZ^\top] = \boldsymbol{I}$) and K-subgaussian. Consider a sample $Z_1, \dots, Z_n$ of independent copies of Z, and define $\hat{\boldsymbol{J}}_n := \frac{1}{n} Z_j Z_j^\top$, which is an unbiased estimate of the $d\times d$ identity matrix $\boldsymbol{I}$. Then w.p. $\geqslant 1-\delta$ the* operator *norm of $\hat{\boldsymbol{J}}_n - \boldsymbol{I}$ satisfies*

$$\big\|\hat{\boldsymbol{J}}_n - \boldsymbol{I}\big\| \lesssim K^2 \left(\sqrt{\frac{d+\log(\delta^{-1})}{n}} + \frac{d+\log(\delta^{-1})}{n}\right). \tag{22}$$

In particular, there exists $c>0$ such that, for any $\varepsilon \in (0,1)$, one has $\big\|\hat{\boldsymbol{J}}_n - \boldsymbol{I}\big\| \leqslant \varepsilon$ w.p. $\geqslant 1-\delta$, as long as

$$n \geqslant cK^4\varepsilon^{-2}(d+\log(\delta^{-1})). \tag{23}$$

This theorem admits the following equivalent formulation (for estimating *arbitrary* covariance matrices).

Theorem 2.2. *Let $X \in \mathbb{R}^d$ be zero-mean, with covariance matrix $\mathbb{E}[XX^\top] = \boldsymbol{\Sigma} \succ 0$, and satisfying the subgaussian moment comparison assumption*

$$\|\langle X, u\rangle\|_{L_p} \leqslant K\sqrt{p}\|\langle X, u\rangle\|_{L_2} \quad \forall u \in \mathbb{S}^{d-1}. \tag{24}$$

Given $X_1, \dots, X_n \sim_{iid} X$, the sample covariance matrix

$$\hat{\boldsymbol{\Sigma}}_n := \frac{1}{n}\sum_{j\in[n]} X_j X_j^\top$$

with probability at least $1-\delta$ satisfies

$$\big\|\boldsymbol{\Sigma}^{-1/2}(\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1/2}\big\| \lesssim K^2 \left(\sqrt{\frac{d+\log(\delta^{-1})}{n}} + \frac{d+\log(\delta^{-1})}{n}\right). \tag{25}$$

In particular, there is $c>0$ such that, $\forall \varepsilon \in (0,1)$ and when $n \geqslant cK^4\varepsilon^{-2}(d+\log(\delta^{-1}))$, with probability $\geqslant 1-\delta$ one has $\big\|\boldsymbol{\Sigma}^{-1/2}(\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1/2}\big\| \leqslant \varepsilon$, or equivalently

$$(1-\varepsilon)\boldsymbol{\Sigma} \preccurlyeq \hat{\boldsymbol{\Sigma}}_n \preccurlyeq (1+\varepsilon)\boldsymbol{\Sigma}.$$

Exercise 2.1. *Show that for $A \succ 0$ and $B \succcurlyeq 0$, $\|A^{-\frac{1}{2}}(B-A)A^{-\frac{1}{2}}\| \leqslant \varepsilon$ if and only if $(1-\varepsilon)A \preccurlyeq B \preccurlyeq (1+\varepsilon)A$. To this end, verify that for symmetric A, B and nonsingular Q with compatible dimensions, it holds that*

$$A \preccurlyeq B \quad \Longleftrightarrow \quad QAQ^\top \preccurlyeq QBQ^\top.$$

Before we proceed to proving Theorem 2.1, let us verify its equivalence to Theorem 2.2 and discuss the important topic of linear invariance/equivariance.

3 Discussion: linear invariance and equivariance

At first glance, the latter theorem seems more general; in particular, it clearly implies the former one, simply by instantiating Theorem 2.2 for X which *is* isotropic. But Theorem 2.1, in turn, implies Theorem 2.2. Indeed, assume X satisfies the premise of Theorem 2.2, i.e. one has $\mathbb{E}[X] = 0$, $\mathbb{E}[XX^\top] = \mathbf{\Sigma}$, and (24); note that $\|\langle X, u\rangle\|_{L_2} = \|u\|_{\mathbf{\Sigma}}$. Now, consider

$$Z = \mathbf{\Sigma}^{-1/2} X.$$

Clearly, Z is isotropic (so this is also called "to put X in isotropic position") and K-subgaussian: indeed, in (24) one can replace "$\forall u \in \mathbb{S}^{d-1}$" with "$\forall u \in \mathbb{R}^d$" by homogeneity of L_p-norms. But for any deterministic vector $u \in \mathbb{R}^d$, one has $\langle X, u\rangle = \langle Z, v\rangle$ where the transformation $u \mapsto \mathbf{\Sigma}^{1/2} u$ is one-to-one. As the result,

$$\begin{aligned}
(24) \quad &\iff \quad \|\langle X, u\rangle\|_{L_p} \leqslant K\sqrt{p}\|\langle X, u\rangle\|_{L_2} \quad \forall u \in \mathbb{R}^d \\
&\iff \quad \|\langle Z, v\rangle\|_{L_p} \leqslant K\sqrt{p}\|\langle Z, v\rangle\|_{L_2} \quad \forall v \in \mathbb{R}^d \\
&\iff \quad \|\langle Z, v\rangle\|_{L_p} \leqslant K\sqrt{p} \quad \forall v \in \mathbb{S}^{d-1} \quad \iff \quad Z \text{ is } K\text{-subgaussian},
\end{aligned}$$

where the last equivalence is up to a constant-factor distortion of K (by Lecture 2: Proposition 2.1). That is, X has K-subgaussian moment growth, in the sense of (24), if and only if its isotropic position is K-subgaussian – and, in fact, if and only if $\boldsymbol{A}X$ has K-subgaussian moment growth for any nonsingular matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$. In other words, assumption (24) is *invariant* w.r.t. nonsingular linear transformations, a.k.a. with respect to the action of the generalized linear group $\mathsf{GL}(d, \mathbb{R})$. On the other hand, it holds that

$$\hat{\boldsymbol{J}}_n = \frac{1}{n}\sum_{j\in[n]} Z_j Z_j^\top = \frac{1}{n}\sum_{j\in[n]} \mathbf{\Sigma}^{-1/2} X_j X_j^\top \mathbf{\Sigma}^{-1/2} = \mathbf{\Sigma}^{-1/2}\hat{\mathbf{\Sigma}}_n \mathbf{\Sigma}^{-1/2},$$

so (25) is (22). As such, Theorem 2.1 has the same premise and also the same implications as Theorem 2.2.

4 Left-hand side of (22) as the supremum of a subexponential stochastic process

If $A \in \mathbb{R}^{d\times d}$ is symmetric, i.e. $A^\top = A$, then one can rewrite its operator norm as follows (verify this fact):

$$\|A\| = \sup_{u\in\mathbb{S}^{d-1}} |u^\top A u|.$$

(The absolute value can be removed if $A \succcurlyeq 0$.) So, $\|\hat{\boldsymbol{J}}_n - \boldsymbol{I}\|$ is the supremum of a stochastic process on $\mathbb{S}^{d-1}$,

$$\|\hat{\boldsymbol{J}}_n - \boldsymbol{I}\| = \sup_{u\in\mathbb{S}^{d-1}} |u^\top(\hat{\boldsymbol{J}}_n - \boldsymbol{I})u|,$$

where the random values $u^\top(\hat{\boldsymbol{J}}_n - \boldsymbol{I})u$ are "of χ^2-type:" indeed, for any $u \in \mathbb{S}^{d-1}$, the random variable

$$u^\top(\hat{\boldsymbol{J}}_n - \boldsymbol{I})u = \left(\sum_{j\in[n]} (Z_j^\top u)^2\right) - 1$$

is the centered χ^2-type statistic if $Z_j \sim \mathcal{N}(0, \boldsymbol{I})$. More generally, under the premise of Theorem 2.1 one has

$$|u^\top(\hat{\boldsymbol{J}}_n - \boldsymbol{I})u| \leqslant cK^2\left(\sqrt{\frac{\log(\delta^{-1})}{n}} + \frac{\log(\delta^{-1})}{n}\right) \tag{26}$$

with probability $\geqslant 1 - \delta$ for any *fixed* $u \in \mathbb{S}^{d-1}$; cf. (21). Now, let's get some intuition for what is to follow.

- Intuitively, to control $\|\hat{\boldsymbol{J}}_n - \boldsymbol{I}\| = \sup_{u\in\mathbb{S}^{d-1}} |u^\top(\hat{\boldsymbol{J}}_n - \boldsymbol{I})u|$ we would want to somehow "take the union bound" over the unit sphere. Yet, it is unclear how to do this: $\mathbb{S}^{d-1}$ contains a continuum of points.
- Thus, a reasonable idea seems to replace $\sup_{u\in\mathbb{S}^{d-1}} |u^\top(\hat{\boldsymbol{J}}_n - \boldsymbol{I})u|$ with the supremum over some finite subset of $\mathbb{S}^{d-1}$. While this would decrease the supremum, one might hope the approximation to work for the purpose at hand. In fact, comparing (22) with (26) we expect the approximating set to be of cardinality N, where $\log N \sim d$, i.e. $N \sim \exp(d)$. We shall now construct such an approximating set.

Lecture 7: Covariance estimation and covering numbers

Disclaimer. In this lecture, we prove the theorem on the concentration of sample covariance matrices discussed in the previous lecture. For convenience, let us repeat the formulation here.

Theorem 0.1 (cf. [Ver12, Theorem 5.39])**.** *Let $Z \in \mathbb{R}^d$ be isotropic ($\mathbb{E}[Z] = 0$ and $\mathbb{E}[ZZ^\top] = \boldsymbol{I}$) and K-subgaussian. Consider a sample $Z_1, \dots, Z_n$ of independent copies of Z, and define $\hat{\boldsymbol{J}}_n := \frac{1}{n} Z_j Z_j^\top$, which is an unbiased estimate of the $d \times d$ identity matrix $\boldsymbol{I}$. Then w.p. $\geqslant 1 - \delta$ the* operator *norm of $\hat{\boldsymbol{J}}_n - \boldsymbol{I}$ satisfies*

$$\left\|\hat{\boldsymbol{J}}_n - \boldsymbol{I}\right\| \lesssim K^2 \left(\sqrt{\frac{d + \log(\delta^{-1})}{n}} + \frac{d + \log(\delta^{-1})}{n} \right). \tag{27}$$

In particular, there exists $c > 0$ such that, for any $\varepsilon \in (0, 1)$, one has $\left\|\hat{\boldsymbol{J}}_n - \boldsymbol{I}\right\| \leqslant \varepsilon$ w.p. $\geqslant 1 - \delta$, as long as

$$n \geqslant cK^4 \varepsilon^{-2} (d + \log(\delta^{-1})). \tag{28}$$

1 Covering numbers

Let $\rho(\cdot, \cdot)$ be a metric on $\mathbb{R}^d$. In our situation, $\rho(u, v) = \|u - v\|_2$ but the construction does not require this.

Definition 7 (Epsilon-covering)**.** *Given a metric ρ on $\mathbb{R}^d$ and $\varepsilon \geqslant 0$, we say that $S_\varepsilon \subset \mathbb{R}^d$ is an* ε-covering *of $S \subset \mathbb{R}^d$ (with respect to ρ) if any $u \in S$ is ε-close to some $\hat{u} \in S_\varepsilon$; in other words,*

$$\forall u \in S \quad \exists \hat{u} \in S_\varepsilon : \quad \rho(u, \hat{u}) \leqslant \varepsilon. \tag{29}$$

Some trivialities have to be spelled out, e.g.: (a) there are many ε-coverings of a given set; (b) S itself is its own ε-covering for any $\varepsilon \geqslant 0$; (c) any ε-covering is also an ε'-covering for any $\varepsilon' \geqslant \varepsilon$, and so on. Note also that, a priori, S_ε is not required to be a subset of S, though in applications one might want to ensure this.

"Economic" covering is formalized by the following notion.

Definition 8 (Epsilon-net)**.** *Fix a metric ρ and $\varepsilon \geqslant 0$. A set $\mathcal{N}_\varepsilon(S, \rho) \subset \mathbb{R}^d$ is an* ε-net *for $S \subset \mathbb{R}^d$ (with respect to ρ) if $\mathcal{N}_\varepsilon(S)$ is an ε-covering of S of the smallest cardinality. In other words, $|S_\varepsilon| \geqslant |\mathcal{N}_\varepsilon(S, \rho)|$ for any ε-covering S_ε of S (w.r.t. ρ). Moreover, $N_\varepsilon(S, \rho) = |\mathcal{N}_\varepsilon(S, \rho)|$ is the* ε-covering number *of S (w.r.t. ρ).*

For the remainder of the lecture, we let $\rho(u, v) = \|u - v\|_2$ and supress ρ from the notation for $\mathcal{N}_\varepsilon$ and N_ε.

Theorem 1.1. *Let $\mathbb{B}^d$ be the unit ℓ_2-ball in $\mathbb{R}^d$. For all $d > 1$ and $\varepsilon \in (0, 1)$, one has*

$$\left(\frac{1}{\varepsilon}\right)^d \leqslant N_\varepsilon(\mathbb{B}^d) \leqslant \left(\frac{3}{\varepsilon}\right)^d,$$
$$\left(\frac{1}{\varepsilon}\right)^d \leqslant N_\varepsilon(\mathbb{S}^{d-1}) \leqslant \left(\frac{3}{\varepsilon}\right)^d.$$

In either case, the covering set can be chosen as a subset of $\mathbb{B}^d$ or $\mathbb{S}^{d-1}$, respectively.

Proof sketch. For the lower bound in the case of $\mathbb{B}^d$, let $\mathcal{N}_\varepsilon \subset \mathbb{B}^d$ be an ε-net, and draw N_ε balls of radius ε with centers at the net. Their cumulative volume is $N_\varepsilon V_d \varepsilon^d$ where V_d is the volume of $\mathbb{B}^d$. On the other hand, the union of these balls contains $\mathbb{B}^d$, so $N_\varepsilon V_d \varepsilon^d \geqslant V_d$. Rearranging, we get the first lower bound. The other bounds are left as exercises; see e.g. [RH23, Lemma 1.18] for the upper bound with $\mathbb{B}^d$ and [Ver12, Lemma 5.2] for the upper bound with $\mathbb{S}^{d-1}$. Intuitively, note that the case of $\mathbb{S}^{d-1}$ is qualitatively not that different from $\mathbb{B}^d$, simply because the surface area A_d of $\mathbb{S}^{d-1}$ satisfies

$$V_d = A_d \int_0^1 r dr = \frac{1}{2} A_d,$$

i.e. A_d and V_d are of the same order. Also, upper and lower bounds have the same behavior in ε because of the duality between covering and packing; see [RH23, Problem 1.7]. □

Remark. *The first line of inequalities of Theorem 1.1 actually works for any norm $\nu(\cdot)$ on $\mathbb{R}^d$, namely*

$$\left(\frac{1}{\varepsilon}\right)^d \leqslant N_\varepsilon(\mathbb{B}_\nu^d) \leqslant \left(\frac{3}{\varepsilon}\right)^d$$

where $\mathbb{B}_\nu^d := \|u \in \mathbb{R}^d : \nu(u) \leqslant 1\|$ is the unit ball of $\nu(\cdot)$, and $N_\varepsilon(\mathbb{B}_\nu^d)$ is the ε-covering number of B_ν^d with respect to the metric $\rho(u,v) = \nu(u-v)$. This is because of the volume relation $\operatorname{Vol}(\varepsilon\mathbb{B}_\nu^d) = \varepsilon^d \operatorname{Vol}(\mathbb{B}_\nu^d)$.

Covering numbers for a pair of norms. *Computing the covering numbers of $\mathbb{B}_\nu^d$ w.r.t.* another *norm $\nu' \neq \nu$ might be a nontrivial task. Indeed: on the one hand, one can rather easily extend Theorem 1.1 as follows.*

Theorem 1.2. *For any $\varepsilon \in (0,1)$, the ε-covering number $N_\varepsilon(B,\nu')$ of the unit ball B of ν w.r.t. ν' satisfies*

$$\frac{\operatorname{Vol}(B)}{\varepsilon^d \operatorname{Vol}(B')} \leqslant N_\varepsilon(B,\nu') \leqslant \frac{\operatorname{Vol}(B + \frac{\varepsilon}{2}B')}{(\frac{\varepsilon}{2})^d \operatorname{Vol}(B')}$$

where B' is the unit ball of ν', and $K + L = \{u + v : u \in K, v \in L\}$ is the Minkowski sum of two sets K, L.

Exercise 1.1. *Prove Theorem 1.2. (The proof can be found in [Wai19].)*

On the other hand, it is not guaranteed that the two bounds nearly match in the same sense as before—i.e., are of the same order in ε after extracting the $\frac{1}{d}$-th root. This is because of the volume $\operatorname{Vol}(B + \varepsilon B')$ that might be way larger than $\operatorname{Vol}(B + \varepsilon B')$, and characterizing it might itself be a nontrivial geometric problem.

2 Approximation argument

Key observation: replacing the whole sphere $\mathbb{S}^{d-1}$ with its ε-covering we approximate $\|A\| = \sup_{u \in \mathbb{S}^{d-1}} |u^\top A u|$ up to a factor C_ε depending only on ε; in particular, $\varepsilon = c < 1$ gives a constant approximation factor. Namely:

Proposition 2.1. *For any symmetric $d \times d$ matrix A, any $\varepsilon \in (0, 1/2)$ and ε-covering S_ε of $\mathbb{S}^{d-1}$, one has*

$$\|A\| \leqslant \frac{1}{1 - 2\varepsilon} \sup_{u \in S_\varepsilon} |u^\top A u|.$$

Proof. We use triangle inequality several times to pass from $u^\top A u$ to $\hat{u}^\top A \hat{u}$, where $u, \hat{u}$ are unit vectors:

$$\begin{aligned}
|u^\top A u| &\leqslant |u^\top A \hat{u}| + |u^\top A(\hat{u} - u)| \\
&\leqslant |\hat{u}^\top A \hat{u}| + |u^\top A(\hat{u} - u)| + |\hat{u}^\top A(\hat{u} - u)| \\
&\leqslant |\hat{u}^\top A \hat{u}| + 2\|A(\hat{u} - u)\|_2 \\
&\leqslant |\hat{u}^\top A \hat{u}| + 2\|\hat{u} - u\|_2 \|A\|.
\end{aligned}$$

Choosing $\hat{u} \in S_\varepsilon [\subset \mathbb{S}^{d-1}$ as the closest to u node, and taking the supremum over $u \in \mathbb{S}^{d-1}$, we arrive at

$$\|A\| = \sup_{u \in \mathbb{S}^{d-1}} |u^\top A u| \leqslant 2\varepsilon \|A\| + \sup_{\hat{u} \in S_\varepsilon} |\hat{u}^\top A \hat{u}|,$$

and the claim follows. □

3 Completing the proof of Theorem 0.1

It only remains to combine the pieces:

1. Instantiating Propositon 2.1 for $A = \hat{\boldsymbol{J}}_n - \boldsymbol{I}$ with $\varepsilon = 1/4$, one has (with probability one):

$$\|\hat{\boldsymbol{J}}_n - \boldsymbol{I}\| \leqslant 2 \sup_{u \in S_{1/4}} |u^\top A u|.$$

2. By (26), for any $u \in S_{1/4}$ one has with probability at least $1-\delta$:

$$|u^\top(\hat{\boldsymbol{J}}_n - \boldsymbol{I})u| \leqslant cK^2\left(\sqrt{\frac{\log\left(\delta^{-1}\right)}{n}} + \frac{\log\left(\delta^{-1}\right)}{n}\right).$$

3. Taking the union bound over $N = |S_{1/4}|$ events corresponding to all nodes of the net, we get w.p. $\geqslant 1-\delta$:

$$\|\hat{\boldsymbol{J}}_n - \boldsymbol{I}\| \leqslant 2cK^2\left(\sqrt{\frac{\log\left(N\delta^{-1}\right)}{n}} + \frac{\log\left(N\delta^{-1}\right)}{n}\right)$$

where $N = |S_{1/4}| \leqslant 12^d$, so $\log(N) = O(d)$. The theorem is proved. □

4 Smoothed covariance estimation

At least when X is Gaussian, one can extend Theorems 0.1 (a.k.a. Theorem 2.2 from the previous lecture) to the following result.

Theorem 4.1 ([KL17]). *Let $\lambda \geqslant 0$ be arbitrary, let $\boldsymbol{\Sigma}_\lambda := \boldsymbol{\Sigma} + \lambda \boldsymbol{I}$ and define the degrees-of-freedom number*

$$d_\lambda(\boldsymbol{\Sigma}) := \operatorname{tr}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\lambda^{-1}) = \operatorname{tr}(\boldsymbol{\Sigma}_\lambda^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\lambda^{-1/2}).$$

Assuming that $X \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, with probability $\geqslant 1-\delta$ one has

$$\left\|\boldsymbol{\Sigma}_\lambda^{-1/2}(\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}_\lambda^{-1/2}\right\| \lesssim \sqrt{\frac{d_\lambda(\boldsymbol{\Sigma}) + \log(\delta^{-1})}{n}} + \frac{d_\lambda(\boldsymbol{\Sigma}) + \log(\delta^{-1})}{n}. \tag{30}$$

In one of the subsequent lectures, we shall discuss how to preserve such a generalization for *robust* estimators of covariance under *weak* moment assumptions. To make sone sense of the generality of Theorem 4.1:

Exercise 4.1. *By simple linear algebra, verify the following claims.*

(a) Putting $\lambda \leqslant c\lambda_{\min}(\Sigma)$ in (30), where $c > 0$ is an arbitrary constant, one recovers (25) in the Gaussian case, i.e. the same result as for $\lambda = 0$ up to a constant factor (and under Gaussianity).

(b) Putting $\lambda \geqslant c\|\boldsymbol{\Sigma}\|$ in (30), one recovers the effective rank bound, namely

$$\left\|\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}\right\| \lesssim K^2\|\boldsymbol{\Sigma}\|\left(\sqrt{\frac{r(\boldsymbol{\Sigma}) + \log(\delta^{-1})}{n}} + \frac{r(\boldsymbol{\Sigma}) + \log(\delta^{-1})}{n}\right) \tag{31}$$

with probability $\geqslant 1-\delta$, where

$$r(\boldsymbol{\Sigma}) := \frac{\operatorname{tr}(\boldsymbol{\Sigma})}{\|\boldsymbol{\Sigma}\|}$$

is the ℓ_1-effective rank of $\boldsymbol{\Sigma}$ (called so since $r(\boldsymbol{\Sigma}) = \frac{\|\vec{\lambda}(\boldsymbol{\Sigma})\|_1}{\|\vec{\lambda}(\boldsymbol{\Sigma})\|_\infty}$, where $\vec{\lambda}(\boldsymbol{\Sigma})$ is the vector of eigenvalues).

(c) Show that the case $\lambda = \|\boldsymbol{\Sigma}\|$ is the hardest: applying (31) to the right matrix one gets (30) for smaller λ.

Exercise 4.2. *1. Prove the* min-max theorem *for eigenvalues: if A, B are symmetric $d \times d$ matrices, then*

$$A \succcurlyeq B \quad \Longrightarrow \quad \lambda_k \geqslant \mu_k \ \ \forall k \in [d],$$

where $\lambda_{1:n}, \mu_{1:n}$ are the sorted (in descending order) eigenvalues of A, B respectively. Hint: *use that*

$$\lambda_k = \min_{\mathcal{S} \subset \mathbb{R}^d : \dim(\mathcal{S}) = k} \ \max_{u \in \mathcal{S} : \|u\|_2 \leqslant 1} u^\top A u$$

where the minimum is taken over all subspaces of dimension k.

2. *Use the above result to verify the following implication:*

$$\big\|\boldsymbol{\Sigma}_\lambda^{-1/2}(\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}_\lambda^{-1/2}\big\| \leqslant \varepsilon \quad \Longrightarrow \quad |\hat{\lambda}_j - \lambda_j| \leqslant \varepsilon(\lambda_j + \lambda)$$

where λ_j and $\hat{\lambda}_j$ are the eigenvalues of $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$ respectively (sorted in the descending order). This can be interpreted as follows: $\big\|\boldsymbol{\Sigma}_\lambda^{-1/2}(\hat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}_\lambda^{-1/2}\big\| \leqslant \varepsilon$ implies that eigenvalue λ_j is estimated (with $\hat{\lambda}_j$), up to relative error ε, in its own scale *for $\lambda_j \gtrsim \lambda$, and "in the scale λ" for $\lambda_j \lesssim \lambda$. In particular,* all *eigenvalues are estimated in the scale λ if $\lambda \gtrsim \|\boldsymbol{\Sigma}\|$, and in their own respective scales if $\lambda \lesssim \lambda_{\min}(\boldsymbol{\Sigma})$.*

5 Challenge: proving Theorem 4.1 without generic chaining

The following exercise is not an easy feat, especially if one resists the temptation of looking inside [KL17].

Exercise 5.1 (Covariance estimation with effective rank)**.** *Prove Theorem 4.1 with $\lambda = \|\boldsymbol{\Sigma}\|$. Some ideas:*

1. *Partition the indices of eigenvalues (w.l.o.g. sorted) into the groups*

$$\mathcal{J}_k = \{j \in [d] : \quad 2^{-k}\|\boldsymbol{\Sigma}\| < \lambda_j(\boldsymbol{\Sigma}) \leqslant 2^{1-k}\|\boldsymbol{\Sigma}\|\}, \quad k \in \mathbb{N}.$$

In fact, the last nonempty group is $\mathcal{J}_{k_{\max}}$ with $k_{\max} = \lceil \log_2(\mathrm{cond}(\boldsymbol{\Sigma}))\rceil$, where $\mathrm{cond}(\Sigma) = \frac{\|\boldsymbol{\Sigma}\|}{\lambda_{\min}(\boldsymbol{\Sigma})}$ is the condition number. Observe (and justify) that the sizes of these groups satisfy $|\mathcal{J}_k| \leqslant \min\{2^k r(\boldsymbol{\Sigma}), d\}$.

2. *Consider the orthogonal decomposition of $\mathbb{R}^d$ into subspaces $\mathcal{S}_k = \mathrm{span}(u_j : j \in \mathcal{J}_k)$ for $k \in \{1, ..., k_{\max}\}$; that is, $\mathcal{S}_k$ is the product of the eigenspaces of $\boldsymbol{\Sigma}$ corresponding to all eigenvalues at level k. Note that*

$$d_k := \dim(\mathcal{S}_k) = |\mathcal{J}_k| \leqslant \min\{2^k r(\boldsymbol{\Sigma}), d\}.$$

3. *Observe that for any fixed $u \in \mathbb{R}^d$, one has $\frac{n}{\|u\|_{\boldsymbol{\Sigma}}^2} u^\top(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})u \sim \chi_n^2$, therefore with probability $\geqslant 1 - \delta$,*

$$|u^\top(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})u| \lesssim \|u\|_{\boldsymbol{\Sigma}}^2 \left(\sqrt{\frac{\log(\delta^{-1})}{n}} + \frac{\log(\delta^{-1})}{n}\right) \leqslant \|\boldsymbol{\Sigma}\|\,\|u\|_2^2 \left(\sqrt{\frac{\log(\delta^{-1})}{n}} + \frac{\log(\delta^{-1})}{n}\right).$$

4. *Suggest a discretization set $\mathcal{N}$ for the sphere $\mathbb{S}^{d-1}$ (or for the unit ball, if you prefer) such that*

$$\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\| = \sup_{\|u\|_2 = 1} |u^\top(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})u| \lesssim \sup_{u \in \mathcal{N}} |u^\top(\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})u| \lesssim \|\boldsymbol{\Sigma}\| \left(\sqrt{\frac{r(\boldsymbol{\Sigma}) + \log(\delta^{-1})}{n}} + \frac{r(\boldsymbol{\Sigma}) + \log(\delta^{-1})}{n}\right) \tag{32}$$

where the first inequality holds almost surely (hiding a constant factor), and the second inequality holds w.p. $\geqslant 1 - \delta$ and might hide an additional logarithmic factor. To this end, let $\mathcal{N}_k$ be the ε_k-net (w.r.t. the norm $\|\cdot\|_2$) of the unit ball in $\mathcal{S}_k$, $\mathcal{B}_k = \{u \in \mathcal{S}_k : \|u\| \leqslant 1\}$, with ε_k to be specified; note that

$$N_k = |\mathcal{N}_k| \leqslant \left(\frac{3}{\varepsilon_k}\right)^{d_k}.$$

On the other hand, the product set

$$\mathcal{N}_1 \times \cdots \times \mathcal{N}_{k_{\max}}$$

is an ε-net for the unit ball in $\mathbb{R}^d$, with $\varepsilon^2 = \sum_k \varepsilon_k^2$ by the Pythagorean theorem. Based on these ideas and observations, construct $\mathcal{N}$ and a scheme of discounting the probabilities of the violation events for $u \in \mathcal{N}$ (in the spirit of Lecture 2: Section 2) that verify (32).

Bibliography

[AS65] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*, volume 55. Courier Corporation, 1965.

[BK00] V. Buldygin and Yu. Kozachenko. *Metric characterization of random variables and random processes*, volume 188. American Mathematical Society, 2000.

[BR24] A. Barvinok and M. Rudelson. A quick estimate for the volume of a polyhedron. *Israel Journal of Mathematics*, 262(1):449–473, 2024.

[CV14] B. Cousins and S. Vempala. A cubic algorithm for computing Gaussian volume. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on discrete algorithms*, pages 1215–1228. SIAM, 2014.

[Due10] L. Duembgen. Bounding standard Gaussian tail probabilities. *arXiv:1012.2063*, 2010.

[KC18] A. K. Kuchibhotla and A. Chakrabortty. Moving beyond sub-gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint arXiv:1804.02605*, 2018.

[Kin05] J. Kindler. A simple proof of Sion's minimax theorem. *The American Mathematical Monthly*, 112(4):356–358, 2005.

[KL17] V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 02 2017.

[LM00] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

[LM17] R. Latała and D. Matlak. Royen's proof of the Gaussian correlation inequality. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2014–2016*, pages 265–275. Springer, 2017.

[Pol] D. Pollard. Lectures notes on probability. `http://www.stat.yale.edu/~pollard/Books/Pttm`.

[Pol90] D. Pollard. *Empirical Processes: Theory and Applications*. NSF-CBMS regional conference series in probability and statistics. Institute of Mathematical Statistics and the American Statistical Association, 1990.

[RH23] P. Rigollet and J.-C. Hütter. High-dimensional statistics. *arXiv preprint arXiv:2310.19244*, 2023.

[Roy14] T. Royen. A simple proof of the Gaussian correlation conjecture extended to multivariate gamma distributions. *arXiv preprint arXiv:1408.1028*, 2014.

[Šid67] Z. Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American statistical association*, 62(318):626–633, 1967.

[Ver12] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge University Press, 2012.

[Wai19] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.