

Variational Tail Bounds for Norms of Random Vectors and Matrices

Sohail Bahmani

SOHAIL.BAHMANI@GMAIL.COM

Editors: Steve Hanneke and Tor Lattimore

Abstract

We propose a variational tail bound for norms of random vectors and matrices under moment assumptions on their one-dimensional marginals. A simplified version of the bound that parametrizes the “aggregating distribution” using a certain pushforward of the Gaussian distribution is also provided. We apply the proposed method to reproduce some of the well-known bounds on norms of Gaussian random vectors, and also obtain dimension-free tail bounds for the Euclidean norm of random vectors with arbitrary moment profiles. Furthermore, we reproduce a dimension-free concentration inequality for sum of independent and identically distributed positive semidefinite matrices with sub-exponential marginals, and obtain a concentration inequality for the sample covariance matrix of sub-exponential random vectors. We also obtain a tail bound for the operator norm of a random matrix series whose random coefficients may have arbitrary moment profiles. Furthermore, we use coupling to formulate an abstraction of the proposed approach that applies more broadly. As a corollary, we derive a PAC-Bayesian-style bound in terms of a certain combination of the KL and Rényi divergences between the prior and posterior distributions.

Keywords: variational inequalities, tail bounds, (matrix) concentration inequalities, high-dimensional probability

1. Introduction

We revisit the fundamental problem of bounding norms of a random vector $X \in \mathbb{R}^d$. In particular, for a prescribed norm $\|\cdot\|$ we propose a new variational upper bound to $\|X\|$ assuming suitable conditions on the moments of the one-dimensional marginals of X .

By viewing $\|X\|$ as a supremum of linear functionals indexed by the unit ball of the dual norm $\|\cdot\|_*$, variety of techniques that are developed to bound supremum of stochastic processes can be used to bound $\|X\|$. For example, we can resort to *generic chaining* (Talagrand, 2014) or its variants to formulate an upper bound for $\|X\|$. If X has a Gaussian distribution, Talagrand’s celebrated *majorizing measures theorem* (Talagrand, 1987) guarantees that these upper bounds are optimal up to constant factors. However, the bounds obtained through generic chaining are usually abstract in nature and not easy to evaluate since one needs to determine an optimal (or nearly-optimal) *admissible sequence* of the subsets of the unit ball of $\|\cdot\|_*$. In cases where X is a sum of independent and identically distributed (i.i.d.) random vectors, methods developed in empirical processes theory can be applied. Namely, using the *symmetrization* (see, e.g., (Boucheron, Lugosi, and Massart, 2013, Lemma 11.4)) approach, the supremum of the corresponding empirical process can be bounded in terms of the *Rademacher complexity* of the set of linear functionals indexed by the unit ball of $\|\cdot\|_*$ (see (Mohri, Rostamizadeh, and Talwalkar, 2019, Ch. 3), (Shalev-Shwartz and Ben-David, 2014, Ch. 26), and (Boucheron et al., 2013, Ch. 3) for the definition of Rademacher complexity and the related historical remarks).

A distinct alternative to the mentioned methods is a variational approach, commonly known as the PAC-Bayesian argument (Shawe-Taylor and Williamson, 1997; McAllester, 1998), that is primarily developed and applied to analyze the generalization error in statistical learning theory (see,

e.g. (Langford and Shawe-Taylor, 2002; Catoni, 2007; Audibert and Bousquet, 2007; Lever et al., 2010)). More recently, PAC-Bayesian arguments are used to derive non-vacuous generalization bounds for neural networks that perfectly fit the training samples (Dziugaite and Roy, 2017; Zhou, Veitch, Austern, Adams, and Orbanz, 2019). In addition to their application in statistical learning theory, PAC-Bayesian arguments are also used in robust statistics (Audibert and Catoni, 2011; Catoni, 2012; Catoni and Giulini, 2017; Giulini, 2018) as well as non-asymptotic random matrix theory (Oliveira, 2016; Zhivotovskiy, 2024). In its most common formulation (see, e.g., (Alquier, 2024, Theorem 2.1)), to bound $\sup_{\theta \in \Theta} f_{\theta}(X)$ for a random variable X and given a parameter space Θ , the PAC-Bayesian argument uses the duality between the *cumulant generating function* and the *Kullback–Leibler (KL) divergence*. Specifically, the variational characterization of the KL divergence (Alquier, 2024, Lemma 2.2) guarantees that

$$\mathbb{E}_{\theta \sim \mathbb{P}}(f_{\theta}(x)) \leq \log \mathbb{E}_{\theta \sim \mathbb{P}_{\text{ref}}}(e^{f_{\theta}(x)}) + D_{\text{KL}}(\mathbb{P}, \mathbb{P}_{\text{ref}}), \quad (1)$$

where $D_{\text{KL}}(\mathbb{P}, \mathbb{P}_{\text{ref}})$ denotes the KL divergence between the arbitrary “posterior” \mathbb{P} and the reference probability measure or “prior” \mathbb{P}_{ref} assuming that \mathbb{P} is absolutely continuous with respect to \mathbb{P}_{ref} . The inequality (1) holds for any x , and in particular for $x = X$. Under suitable measurability conditions, $\mathbb{E}_{\theta \sim \mathbb{P}_{\text{ref}}}(e^{f_{\theta}(X)})$, as a random variable, can be bounded probabilistically using Markov’s inequality. It then suffices to choose a suitable reference measure and a suitable class \mathcal{P} of aggregating posteriors such that $\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\theta \sim \mathbb{P}}(f_{\theta}(x)) \approx \sup_{\theta \in \Theta} f_{\theta}(X)$ and $\sup_{\mathbb{P} \in \mathcal{P}} D_{\text{KL}}(\mathbb{P}, \mathbb{P}_{\text{ref}})$ is small in a certain precise sense. While the PAC-Bayesian arguments are usually based on the change of measure inequality (1), alternative change of measure inequalities that use other probability metrics and divergences are also considered in the literature (Ohnishi and Honorio, 2021; Picard-Weibel and Guedj, 2022).

Some components of the PAC-Bayesian arguments complicate and constrain the application of this method. For example, the fundamental inequality (1) is useful only if $f_{\theta}(X)$ has a finite exponential moment (in a neighborhood of the origin). More importantly, we need to have an approximation for the KL divergence term in (1) that is both accurate and “calculation-friendly”. These requirements add to the subtlety of choosing a suitable reference measure \mathbb{P}_{ref} and a suitable class of aggregating posteriors \mathcal{P} . For example, Zhivotovskiy (2024, Theorem 1) has established a matrix concentration inequality using the PAC-Bayesian argument in which the spectral norm is expressed as the supremum of a bilinear form, and the set \mathcal{P} consists of carefully chosen truncated non-isotropic Gaussian distributions for the left and right factors of the bilinear form.

Focusing on tail bounds for the norm of random vectors (or matrices), we propose an alternative variational bound that has simpler structure and applies under fewer constraints compared to the PAC-Bayesian bounds. For example, the standard PAC-Bayesian argument does not apply under the conditions of our results in Theorems 2, 6, and 7. The role of the main components of the proposed bound are also more transparent. Generalizations of the proposed bounds in terms more abstract form of couplings are presented in Appendix C. In particular, a PAC-Bayesian-style bound is obtained as a corollary in which a certain combination of the KL divergence and the *Rényi divergence* is used to measure the closeness of the posterior and prior distributions.

2. Problem Setup and the Main Lemma

Let $\|\cdot\|$ be a norm in \mathbb{R}^d whose dual is denoted by $\|\cdot\|_*$. Furthermore, denote the canonical centered unit balls of $\|\cdot\|$ and $\|\cdot\|_*$ respectively by \mathcal{B} and \mathcal{B}_* . Throughout, we denote the boundary of a set \mathcal{S}

by $\partial\mathcal{S}$. We also use the notation $\|u\|_M = \sqrt{u^\top M u}$ for any positive semidefinite matrix M . We also interpret $p!$ for $p \in \mathbb{R}$ as the standard extension of the factorial to real numbers using the Gamma function. We use the expression $\alpha \lesssim \beta$ to suppress some absolute constant $c > 0$ in the inequality $\alpha \leq c\beta$.

Let us also introduce some notation that we use to formulate our results. For a probability measure \mathbb{P}_0 over \mathbb{R}^d we denote the expectation with respect to $U \sim \mathbb{P}_0$ by \mathbb{E}_0 , and define

$$\nu(\mathbb{P}_0) \stackrel{\text{def}}{=} \sup_{x \in \partial\mathcal{B}} \frac{1}{\mathbb{P}_0(|\langle U, x \rangle| \geq 1)},$$

which is related to the convex geometric notions of convex cap coverings and convex floating bodies (Bárány and Larman, 1988; Schütt and Werner, 1990; Schütt, Thäle, Turchi, and Werner, 2025). The central part of our results is the following elementary lemma.

Lemma 1 *Given a random variable $X \in \mathbb{R}^d$, for all $u \in \mathbb{R}^d$ and $p \geq 1$ let*

$$M_X(u, p) \stackrel{\text{def}}{=} \mathbb{E}(|\langle u, X \rangle|^p).$$

Then, we have

$$\mathbb{E}(\|X\|^p) \leq \mathbb{E}_0(M_X(U, p)) \nu(\mathbb{P}_0). \quad (2)$$

Furthermore, for any $t \geq 0$, with probability at least $1 - e^{-t}$ we have

$$\|X\| \leq \inf_{p \geq 1, \mathbb{P}_0} e^{t/p} (\mathbb{E}_0(M_X(U, p)))^{1/p} \nu^{1/p}(\mathbb{P}_0). \quad (3)$$

Proof Observe that (3) follows from (2) by applying Markov's inequality, taking the p -th root, and optimizing with respect to $p \geq 1$ and \mathbb{P}_0 . So, it suffices to prove (2). For $x \in \mathbb{R}^d$ define

$$Q_0(x) \stackrel{\text{def}}{=} \mathbb{P}_0(|\langle U, x \rangle| \geq \|x\|),$$

which is invariant with respect to scaling of x , and observe that

$$\nu(\mathbb{P}_0) = \sup_{x \in \partial\mathcal{B}} \frac{1}{Q_0(x)}.$$

It follows from Markov's inequality that for any $p \geq 1$ and $x \in \mathbb{R}^d$ we have

$$\begin{aligned} \|x\|^p &\leq \mathbb{E}_0(|\langle U, x \rangle|^p) / Q_0(x) \\ &\leq \mathbb{E}_0(|\langle U, x \rangle|^p) \nu(\mathbb{P}_0). \end{aligned}$$

In particular, for $x = X$ we have

$$\|X\|^p \leq \mathbb{E}_0(|\langle U, X \rangle|^p) \nu(\mathbb{P}_0).$$

Then, (2) follows by taking the expectation with respect to X , and invoking Tonelli's theorem, whose application is justified because $(u, x) \mapsto |\langle u, x \rangle|^p$ is continuous and thus Borel measurable, i.e.,

$$\mathbb{E}(\|X\|^p) \leq \mathbb{E} \mathbb{E}_0(|\langle U, X \rangle|^p) \nu(\mathbb{P}_0) = \mathbb{E}_0(M_X(U, p)) \nu(\mathbb{P}_0).$$

■

Let us evaluate the general result of [Lemma 1](#) in some special cases where $M_X(u, p)$ is replaced by some explicit upper bound. Here, we focus on two simple examples for norms of Gaussian random vectors. We emphasize that the lemma can be applied under other suitable moment conditions as well. For example, we may assume that X whose covariance matrix is Σ , has *sub-exponential* marginals, i.e., for some $\eta > 0$ and all $p \geq 1$, $u \in \mathbb{R}^d$ we have

$$M_X(u, p) \leq p! (\eta \|u\|_\Sigma)^p. \quad (4)$$

In fact, in [Corollary 3](#) below we recover, up to constant factors, the results in ([Zhivotovskiy, 2024](#), Proposition 3) on the tail bounds for $\|X\|_2$ under the condition (4).

Example 1 (Polyhedral norm of a standard Gaussian random vector) *Suppose that $X \in \mathbb{R}^d$ is distributed as $\text{Normal}(0, I)$, for which we have*

$$M_X(u, 2p) \leq 2^p p! \|u\|_2^{2p}.$$

We examine the case where \mathcal{B} , the unit ball of the norm $\|\cdot\|$, is a symmetric polytope with nonempty interior. In this scenario, \mathcal{B}_ is also a polytope and we can write*

$$\|X\| = \max_{u \in \text{ext}(\mathcal{B}_*)} \langle u, X \rangle,$$

where

$$\text{ext}(\mathcal{B}_*) = \{\pm u_1, \dots, \pm u_N\},$$

is the set of extreme points of \mathcal{B}_ . Since any norm can be approximated by polyhedral norms to arbitrary precision, this example provides general insights about the nature of the bound (3).*

We choose \mathbb{P}_0 to be the law of

$$U = \operatorname{argmax}_{u \in \rho_0 \text{ext}(\mathcal{B}_*)} \langle u, G \rangle,$$

where $G \sim \text{Normal}(0, I)$ and $\rho_0 > 0$ is a parameter to be specified. Let $[N] \stackrel{\text{def}}{=} \{1, \dots, N\}$. For each $k \in [N]$ we define $c_k \geq 1$ as the smallest real number such that for every $x \in \partial\mathcal{B}$, for at least k different $i \in [N]$ we have $c_k \langle w, u_i \rangle \geq 1$. Formally,

$$c_k = \inf \left\{ c \geq 1 : \inf_{x \in \partial\mathcal{B}} \sum_{i \in [N]} \mathbb{1}(c|\langle u_i, x \rangle| \geq 1) \geq k \right\}. \quad (5)$$

In particular, $c_1 = 1$ due to the fact that \mathcal{B} is the polar set of \mathcal{B}_ . Furthermore, we define*

$$\pi_k = \min_{I \subseteq [N]: |I|=k} \sum_{i \in I} \mathbb{P}(U = \rho_0 u_i), \quad (6)$$

which is the smallest probability assigned to any k -subset of $\{\rho_0 u_1, \dots, \rho_0 u_N\}$ under \mathbb{P}_0 . Observe that π_k does not depend on the scaling parameter ρ_0 .

Choosing $\rho_0 = c_k$ and recalling the definition of $\nu(\mathbb{P}_0)$, we have

$$\nu(\mathbb{P}_0) \leq \pi_k^{-1}.$$

Therefore, using the fact that $p! \leq p^p$, it follows from [Lemma 1](#) that

$$\|X\| \leq \inf_{p \geq 1, k \in [N]} e^{t/2p} \left(\sqrt{2p} \left(\sum_{i \in [N]} 2 \mathbb{P}_0(U = c_k u_i) \|u_i\|_2^{2p} \right)^{1/(2p)} c_k \pi_k^{-1/(2p)} \right), \quad (7)$$

with probability at least $1 - e^{-t}$. This bound can be further simplified to

$$\begin{aligned} \|X\| &\leq \inf_{p \geq 1, k \in [N]} e^{t/2p} \left(\sqrt{2p} c_k \pi_k^{-1/(2p)} \right) \max_{i \in [N]} \|u_i\|_2 \\ &\leq \sqrt{2e} \max_{i \in [N]} \|u_i\|_2 \min_{k \in [N]} c_k \max \left\{ 1, \sqrt{t - \log \pi_k} \right\}, \end{aligned} \quad (8)$$

For verification, we can recover, up to the constant factors, the well-known bound for the ℓ_∞ -norm of $X \sim \text{Normal}(0, I)$ which states that

$$\|X\|_\infty \leq \sqrt{2 \log(2d)} + \sqrt{t},$$

holds with probability at least $1 - e^{-t}$ (see, e.g., ([Boucheron et al., 2013](#), Example 5.7)). This special case is captured in our formulation, by choosing $N = d$ and u_1, \dots, u_d to be the canonical unit basis vectors in \mathbb{R}^d . Then, due to symmetry, we have $\mathbb{P}_0(U = c_k u_1) = \dots = \mathbb{P}_0(U = c_k u_d) = 1/(2d)$. Thus

$$\sum_{i \in [d]} 2 \mathbb{P}_0(U = c_k u_i) \|u_i\|_2^{2p} = 1,$$

and choosing $k = 1$ we have $c_1 = 1$ and $\pi_1 = 1/(2d)$. Therefore, the derived bound (7) simplifies to

$$\begin{aligned} \|X\|_\infty &\leq \inf_{p \geq 1} e^{t/2p} \left(\sqrt{2p} (2d)^{1/(2p)} \right) \\ &\leq \sqrt{2e(\log(2d) + t)}. \end{aligned}$$

2.1. Dimension-free tail bound for the Euclidean norm of random vectors with arbitrary moment profiles

Any norm can be approximated by polyhedral norms. However, using such an approximation is not always desirable because it ignores any special structure that the norm $\|\cdot\|$ might have, and may produce bounds that are not explicit in terms of the parameters that describe the law of the random vector. [Lemma 1](#) can still provide more explicit bounds in some special cases. We use [Lemma 1](#) to prove the following tail bound for the Euclidean norm of random vectors that have arbitrary ‘‘moment profiles’’. We emphasize that the standard PAC-Bayesian argument requires exponential moments to be finite, therefore it cannot, as a standalone method, address the general situation considered by the following theorem.

Theorem 2 Let $X \in \mathbb{R}^d$ be a centered random vector with covariance Σ and a non-decreasing moment profile $h: [1, \infty) \rightarrow [0, \infty]$ that for every $u \in \mathbb{R}^d$ and $p \geq 2$ satisfies

$$(\mathbb{E}(|\langle u, X \rangle|^p))^{1/p} \leq h(p) \|u\|_{\Sigma}. \quad (9)$$

Then for any $t > 0$, with probability at least $1 - e^{-t}$ we have

$$\|X\|_2 \leq \inf_{p \geq 2} 2p^{-1/2} h(p) \left(\text{tr}(\Sigma) + \frac{p}{2} \|\Sigma\|_{\text{op}} \right)^{1/2} e^{(t+\log(2))/p}. \quad (10)$$

Proof We use [Lemma 1](#) with \mathbb{P}_0 being the law of $U = \kappa_0 G$ for a certain $\kappa_0 > 0$ and a standard Gaussian vector $G \in \mathbb{R}^d$. The assumption on X provides the inequality

$$M_X(u, p) \leq h^p(p) \|u\|_{\Sigma}^p,$$

which together with [Lemma 11](#) yields

$$\begin{aligned} (\mathbb{E}_0(M_X(U, p)))^{1/p} &\leq \kappa_0 h(p) (\mathbb{E}(\|G\|_{\Sigma}^p))^{1/p} \\ &\leq \kappa_0 h(p) \left(\text{tr}(\Sigma) + \frac{p}{2} \|\Sigma\|_{\text{op}} \right)^{1/2}. \end{aligned}$$

Furthermore, recalling that \mathcal{B} is the unit ℓ_2 ball, for any $x \in \partial\mathcal{B}$ we can write

$$\mathbb{P}_0(|\langle U, x \rangle| \geq 1) = \mathbb{P}(\kappa_0 |g| \geq 1) = 2\Phi\left(-\frac{1}{\kappa_0}\right),$$

where $g \in \mathbb{R}$ is a standard Gaussian random variable, and $\Phi(\cdot)$ denotes the standard Gaussian cumulative distribution function (CDF). It follows from straightforward calculus and the inequality $2\pi < 3e$ that $4\Phi(-z) - e^{-2z^2/3}$ is decreasing in $z \geq 0$, which implies that for all $z \geq 0$ we have¹

$$\Phi(-z) \geq e^{-2z^2/3}/4. \quad (11)$$

Therefore, for all $x \in \mathcal{B}$ we have

$$\mathbb{P}_0(|\langle U, x \rangle| \geq 1) \geq e^{-2/(3\kappa_0^2)}/2,$$

and thus

$$\nu(\mathbb{P}_0) \leq 2e^{2/(3\kappa_0^2)}.$$

It then follows from [Lemma 1](#) (with the constraint $p \geq 2$ rather than $p \geq 1$) that with probability at least $1 - e^{-t}$ we have

$$\|X\|_2 \leq \inf_{p \geq 2, \kappa_0 > 0} \kappa_0 h(p) e^{2/(3p\kappa_0^2)} \left(\text{tr}(\Sigma) + \frac{p}{2} \|\Sigma\|_{\text{op}} \right)^{1/2} e^{(t+\log(2))/p}.$$

The bound (10) follows by choosing $\kappa_0 = 2/\sqrt{3p}$ and using the fact that $\sqrt{e/3} < 1$. ■

The following corollary provides more explicit bounds for random vectors that have sub-Gaussian or sub-exponential marginals.

1. Here we favored simplicity of the lower bound over its accuracy. More accurate lower bounds for the quantile function, or the *Mills ratio* of the standard Gaussian random variables [see e.g., ([Wainwright, 2019](#), Exercise 2.2) ([Boucheron et al., 2013](#), Exercise 7.8), ([Gasull and Utzet, 2014](#)), and references therein] can be used instead.

Corollary 3 *Under the conditions of [Theorem 2](#), if X has sub-Gaussian marginals, in the sense that $h(p) = \eta p^{1/2}$ in [\(9\)](#) for some $\eta \geq 1$, then for any $t > 0$, with probability at least $1 - e^{-t}$ we have*

$$\|X\|_2 \leq 6\eta \left(\text{tr}(\Sigma) + t\|\Sigma\|_{\text{op}} \right)^{1/2}. \quad (12)$$

Similarly, if X has sub-exponential marginals, i.e., $h(p) = \eta p$ for some $\eta \geq 1$, then for any $t \geq 1$, with probability at least $1 - e^{-t}$ we have

$$\|X\|_2 \leq 4\sqrt{e}\eta \left(\sqrt{t \text{tr}(\Sigma)} + t\|\Sigma\|_{\text{op}}^{1/2} \right). \quad (13)$$

Proof Let $r = \text{tr}(\Sigma)/\|\Sigma\|_{\text{op}}$. Applying [Theorem 2](#) in the sub-Gaussian case, with probability at least $1 - e^{-t}$ we have

$$\|X\|_2 \leq \inf_{p \geq 2} 2\eta \|\Sigma\|_{\text{op}}^{1/2} \left(r + \frac{p}{2} \right)^{1/2} e^{(t+\log(2))/p},$$

which by choosing $p = 2(t + \log(2) + \sqrt{r(t + \log(2))}) \geq 2(\log(2) + \sqrt{r \log(2)}) > 2$ yields

$$\begin{aligned} \|X\|_2 &\leq 2\eta \|\Sigma\|_{\text{op}}^{1/2} \left(r + t + \log(2) + \sqrt{r(t + \log(2))} \right)^{1/2} \sqrt{e} \\ &\leq \sqrt{6e}\eta \|\Sigma\|_{\text{op}}^{1/2} (r + t + \log(2))^{1/2} \\ &\leq 6\eta \|\Sigma\|_{\text{op}}^{1/2} (r + t)^{1/2} \\ &= 6\eta \left(\text{tr}(\Sigma) + t\|\Sigma\|_{\text{op}} \right)^{1/2}, \end{aligned}$$

where in the second line we used $2\sqrt{r(t + \log(2))} \leq r + t + \log(2)$, and in the third line we used the facts that $r \geq 1$ and $e(1 + \log(2)) < 6$.

In the sub-exponential case, [Theorem 2](#) guarantees, with probability at least $1 - e^{-t}$, that

$$\|X\|_2 \leq \inf_{p \geq 2} 2\eta \|\Sigma\|_{\text{op}}^{1/2} \left(pr + \frac{p^2}{2} \right)^{1/2} e^{(t+\log(2))/p}.$$

Since $t \geq 1$, choosing $p = 2t$ guarantees that $e^{(t+\log(2))/p} \leq \sqrt{2e}$ and thus

$$\begin{aligned} \|X\|_2 &\leq 4\sqrt{e}\eta \|\Sigma\|_{\text{op}}^{1/2} (tr + t^2)^{1/2} \\ &\leq 4\sqrt{e}\eta \left(\sqrt{t \text{tr}(\Sigma)} + t\|\Sigma\|_{\text{op}}^{1/2} \right), \end{aligned}$$

where the second line follows from the sub-additivity of the square root. ■

For random vectors with sub-Gaussian marginals, the tail bounds of the form [\(12\)](#) are well-known and can be derived using more standard methods (see, e.g., [\(Vershynin, 2018, Exercise 6.3.5\)](#)). If X is a centered Gaussian, the Gaussian concentration inequality ([Boucheron et al., 2013, Theorem 5.5](#)) provides a better approximation for the tail behavior of $\|X\|_2$ and guarantees that with probability at least $1 - e^{-t}$ we have

$$\|X\|_2 \leq \sqrt{\text{tr}(\Sigma)} + \sqrt{2\|\Sigma\|_{\text{op}}t}.$$

Therefore, with $h(p) = \sqrt{2p/\pi}$ being a valid moment profile for the Gaussian random vector X , (12) recovers the inequality above up to a factor of $6\sqrt{2/\pi} < 5$.

The tail bound (13), up to the constant factors, is equivalent to the tail bound due to Zhivotovskiy (2024, Proposition 3) for random vectors with sub-exponential marginals. As pointed out in (Zhivotovskiy, 2024), this bound applies to log-concave measures, including uniform measures on convex bodies, and as such it improves on a similar result due to Alesker (1995). Also, it can be viewed as a dimension-free specialization of a general result of Latała and Nayar (2020) to the case of the Euclidean norm of sub-exponential random vectors.

3. Simplified Bounds via Pushforward of Gaussians

The tail bound of Lemma 1 holds in rather general situations, but it is not entirely calculation-friendly, mainly due to its implicit form and the minimization with respect to \mathbb{P}_0 . Assuming that X has sub-Gamma marginals, by characterizing \mathbb{P}_0 as a certain pushforward of the standard Gaussian measure, the following theorem provides a bound that can be optimized using a few scalar parameters. In particular, one can recover the tail bound in Example 1, and (12) of Corollary 3 using this theorem. These analyses are provided in Appendix D.

Theorem 4 *Let $X \in \mathbb{R}^d$ be a zero-mean random variable with covariance matrix $\Sigma \succ 0$ and sub-Gamma marginals in the sense that for some constants $\eta_1, \eta_2 \geq 0$, for every $p \geq 1$ and $u \in \mathbb{R}^d$ we have*

$$\mathbb{E} \left(|\langle u, X \rangle|^{2p} \right) \leq p! (\eta_1 \|u\|_\Sigma)^{2p} + (2p)! (\eta_2 \|u\|_*)^{2p} .$$

Furthermore, for parameters $\rho_0 \geq 0$ and $\kappa_0 > 0$, let

$$f_0(x) = \inf_y \frac{\kappa_0}{2} \|x - y\|_{\Sigma^{-1/2}}^2 + \rho_0 \|y\| ,$$

With $(z)_+ \stackrel{\text{def}}{=} \max\{z, 0\}$ and $G \sim \text{Normal}(0, I)$ being a standard Gaussian random variable in \mathbb{R}^d , let

$$\underline{\lambda}_\Sigma(\kappa_0, \rho_0) = \inf_{x \in \partial \mathcal{B}} \mathbb{E} \left(\left(\left| \langle \nabla f_0(\Sigma^{1/2} G), x \rangle \right| - 1 \right)_+ \right) ,$$

and define

$$T_{\geq}(s) = 1 - T_{<}(s) = \mathbb{P} (\|G\|_* \geq s) ,$$

$$\begin{aligned} \Omega_\Sigma(p, \kappa_0, \rho_0) = \eta_1 \sqrt{p} \min \left\{ \rho_0 \|\Sigma\|_{\square}^{1/2}, \inf_{\tau \in (0,1]} \left[T_{\geq}^{1/(2p)} \left(\frac{\rho_0}{\tau \kappa_0} \right) \rho_0 \|\Sigma\|_{\square}^{1/2} + T_{<}^{1/(4p)} \left(\frac{\rho_0}{\tau \kappa_0} \right) \right. \right. \\ \left. \left. \cdot \kappa_0 \left((1 - \tau) \|\Sigma\|_{\text{op}}^{1/4} \sqrt{\text{tr}(\Sigma^{1/2})} + \tau \sqrt{\text{tr} \Sigma} + \sqrt{2p} \sqrt{\|\Sigma\|_{\text{op}}} \right) \right] \right\} + 2\eta_2 p \rho_0 , \end{aligned}$$

and

$$\bar{v}_\Sigma(\kappa_0, \rho_0) = \min \left\{ \frac{\rho_0}{\underline{\lambda}_\Sigma(\kappa_0, \rho_0)}, \frac{\kappa_0^2 \left(T_{\geq}(\rho_0/\kappa_0) \left(\|\Sigma^{-1}\|_{\text{op}} \|\Sigma\|_{\text{op}} - 1 \right) + 1 \right) \text{rad}^2(\mathcal{B})}{\underline{\lambda}_\Sigma^2(\kappa_0, \rho_0)} + 1 \right\} ,$$

where $\|M\|_{\square} = \sup_{u \in \mathcal{B}_*} u^T M u$ for any $d \times d$ symmetric matrix M , and $\text{rad}(\mathcal{B}) = \sup_{x \in \mathcal{B}} \|x\|_2$ denotes the radius of \mathcal{B} as measured by the ℓ_2 norm. Then, for any $t \geq 0$, with probability at least $1 - e^{-t}$, we have

$$\|X\| \leq \inf \left\{ e^{t/(2p)} \Omega_{\Sigma}(p, \kappa_0, \rho_0) \bar{\nu}_{\Sigma}^{1/(2p)}(\kappa_0, \rho_0) : p \geq 1, \rho_0 \geq 0, \kappa_0 > 0 \right\}. \quad (14)$$

It is worth mentioning that using duality we can show that the mapping $x \mapsto \nabla f_0(x)$ can be expressed equivalently as

$$\nabla f_0(x) = \underset{u \in \rho_0 \mathcal{B}_*}{\text{argmin}} \frac{1}{2} \left\| \kappa_0 \Sigma^{-1/2} x - u \right\|_{\Sigma^{1/2}}^2. \quad (15)$$

Specifically, $\nabla f_0(\Sigma^{1/2} G)$ is a projection of $\kappa_0 G$ onto $\rho_0 \mathcal{B}_*$ in the sense that

$$\nabla f_0(\Sigma^{1/2} G) = \underset{u \in \rho_0 \mathcal{B}_*}{\text{argmin}} \frac{1}{2} \left\| \kappa_0 G - u \right\|_{\Sigma^{1/2}}^2.$$

This characterization also shows that the law of $U = \nabla f_0(\Sigma^{1/2} G)$ is a mixture of a truncated Gaussian distribution on the interior of $\rho_0 \mathcal{B}_*$ and another distribution on the boundary of $\rho_0 \mathcal{B}_*$ that becomes more concentrated around the extreme points as κ_0 increases (relative to ρ_0).

Proof Taking $U = \nabla f_0(\Sigma^{1/2} G)$, it suffices to find the appropriate approximations for the terms that appear in (3). The dual representation of $f_0(x)$, i.e.,

$$f_0(x) = \sup_{u \in \rho_0 \mathcal{B}_*} \langle u, x \rangle - \frac{1}{2\kappa_0} \|u\|_{\Sigma^{1/2}}^2,$$

reveals that for all $x \in \mathbb{R}^d$ we have

$$\|\nabla f_0(x)\|_* \leq \rho_0. \quad (16)$$

Furthermore, [Lemma 13](#) below in [Appendix E](#) guarantees that

$$\left\| \nabla f_0(\Sigma^{1/2} G) \right\|_{\Sigma} \leq \|\Sigma\|_{\text{op}}^{1/4} (1 - \tau) \kappa_0 \|G\|_{\Sigma^{1/2}} + \tau \kappa_0 \|G\|_{\Sigma}.$$

Combining the bound above and (16) depending on whether $\|G\|_* < \rho_0/(\tau \kappa_0)$ or not, yields

$$\begin{aligned} \left\| \nabla f_0(\Sigma^{1/2} G) \right\|_{\Sigma} &\leq \mathbb{1}(\|G\|_* \geq \rho_0/(\tau \kappa_0)) \rho_0 \|\Sigma\|_{\square}^{1/2} \\ &\quad + \mathbb{1}(\|G\|_* < \rho_0/(\tau \kappa_0)) \left(\|\Sigma\|_{\text{op}}^{1/4} (1 - \tau) \kappa_0 \|G\|_{\Sigma^{1/2}} + \tau \kappa_0 \|G\|_{\Sigma} \right). \end{aligned}$$

Recalling the definitions of the shorthands $T_{\geq}(\cdot)$ and $T_{<}(\cdot)$, we can write

$$\begin{aligned} &\left(\mathbb{E} \left(\left\| \nabla f_0(\Sigma^{1/2} G) \right\|_{\Sigma}^{2p} \right) \right)^{1/(2p)} \\ &\leq T_{\geq}^{1/(2p)} \left(\frac{\rho_0}{\tau \kappa_0} \right) \rho_0 \|\Sigma\|_{\square}^{1/2} + \\ &T_{<}^{1/(4p)} \left(\frac{\rho_0}{\tau \kappa_0} \right) \left((1 - \tau) \kappa_0 \|\Sigma\|_{\text{op}}^{1/4} \left(\mathbb{E} \left(\|G\|_{\Sigma^{1/2}}^{4p} \right) \right)^{1/(4p)} + \tau \kappa_0 \left(\mathbb{E} \left(\|G\|_{\Sigma}^{4p} \right) \right)^{1/(4p)} \right), \end{aligned}$$

where we used the triangle inequality and the inequality $\|\xi_1 \xi_2\|_{L^{2p}} \leq \|\xi_1\|_{L^{4p}} \|\xi_2\|_{L^{4p}}$ with $\|\cdot\|_{L^q} = (\mathbb{E}(|\cdot|^q))^{1/q}$ denoting the L^q norm. Putting the derived bounds together, using [Lemma 11](#), optimizing with respect to $\tau \in (0, 1]$, and recalling the definition of $\Omega_\Sigma(p, \kappa_0, \rho_0)$, we obtain

$$\left(\mathbb{E}_0 \left(M_X(\nabla f_0(\Sigma^{1/2}G), 2p) \right) \right)^{1/(2p)} \leq \Omega_\Sigma(p, \kappa_0, \rho_0). \quad (17)$$

Furthermore, using the second moment method (see [Lemma 12](#) below in [Appendix E](#)) we have

$$\mathbb{P} \left(\left| \langle \nabla f_0(\Sigma^{1/2}G), x \rangle \right| \geq 1 \right) \geq \frac{\left(\mathbb{E} \left(\left| \langle \nabla f_0(\Sigma^{1/2}G), x \rangle \right| - 1 \right)_+^2 \right)}{\mathbb{E} \left(\left(\left| \langle \nabla f_0(\Sigma^{1/2}G), x \rangle \right| - 1 \right)_+^2 \right)},$$

and thus

$$\nu(\mathbb{P}_0) \leq \sup_{x \in \partial \mathcal{B}} \frac{\mathbb{E} \left(\left(\left| \langle \nabla f_0(\Sigma^{1/2}G), x \rangle \right| - 1 \right)_+^2 \right)}{\left(\mathbb{E} \left(\left| \langle \nabla f_0(\Sigma^{1/2}G), x \rangle \right| - 1 \right)_+ \right)^2}.$$

To further simplify the right-hand side of the inequality above we consider two cases. In the first case, for $x \in \partial \mathcal{B}$ we use the inequality

$$\mathbb{E} \left(\left(\left| \langle \nabla f_0(\Sigma^{1/2}G), x \rangle \right| - 1 \right)_+^2 \right) \leq \rho_0 \mathbb{E} \left(\left(\left| \langle \nabla f_0(\Sigma^{1/2}G), x \rangle \right| - 1 \right)_+ \right),$$

which implies

$$\nu(\mathbb{P}_0) \leq \sup_{x \in \partial \mathcal{B}} \frac{\rho_0}{\left(\mathbb{E} \left(\left| \langle \nabla f_0(\Sigma^{1/2}G), x \rangle \right| - 1 \right)_+ \right)} \leq \frac{\rho_0}{\underline{\Delta}_\Sigma(\kappa_0, \rho_0)}.$$

In the second case, the Gaussian Poincaré inequality² ([Boucheron et al., 2013](#), Theorem 3.20) yields

$$\text{var} \left(\left(\left| \langle \nabla f_0(\Sigma^{1/2}G), x \rangle \right| - 1 \right)_+ \right) \leq \mathbb{E} \left(\left\| \Sigma^{1/2} \nabla^2 f_0(\Sigma^{1/2}G) x \right\|_2^2 \right).$$

The definition of f_0 reveals that for a certain convex function g_0 we have $f_0 + g_0 = \kappa_0 \|\cdot\|_{\Sigma^{1/2}}^2/2$. It then follows from Alexandrov's differentiability theorem (see, e.g., ([Evans and Gariepy, 2025](#), Theorem 6.9), ([Villani, 2009](#), Theorems 14.25 and 14.1), and ([Rockafellar, 1999](#))) that $\nabla^2 f_0(\Sigma^{1/2}G) \preceq \kappa_0 \Sigma^{-1/2}$ almost surely. Particularly, $\nabla^2 f_0(\Sigma^{1/2}G) = \kappa_0 \Sigma^{-1/2}$ whenever $\kappa_0 \|G\|_* < \rho_0$. Therefore, we can write

$$\begin{aligned} & \text{var} \left(\left(\left| \langle \nabla f_0(\Sigma^{1/2}G), x \rangle \right| - 1 \right)_+ \right) \\ & \leq \mathbb{E} \left(\mathbb{1}(\|G\|_* \geq \rho_0/\kappa_0) \left\| \Sigma^{-1} \right\|_{\text{op}} \left\| \nabla^2 f_0(\Sigma^{1/2}G) \right\|_{\text{op}} \|x\|_2^2 + \mathbb{1}(\|G\|_* < \rho_0/\kappa_0) \kappa_0^2 \|x\|_2^2 \right) \\ & \leq \kappa_0^2 \left(T_{\geq}(\rho_0/\kappa_0) \left(\left\| \Sigma^{-1} \right\|_{\text{op}} \|\Sigma\|_{\text{op}} - 1 \right) + 1 \right) \|x\|_2^2, \end{aligned}$$

2. We apply the inequality to function $g \mapsto \left(\left| \langle \nabla f_0(\Sigma^{1/2}g), x \rangle \right| - 1 \right)_+$ which is weakly differentiable for $\kappa_0 < \infty$.

which in turn implies

$$\nu(\mathbb{P}_0) \leq \frac{\kappa_0^2 \left(T_{\geq}(\rho_0/\kappa_0) \left(\|\Sigma^{-1}\|_{\text{op}} \|\Sigma\|_{\text{op}} - 1 \right) + 1 \right) \text{rad}^2(\mathcal{B})}{\underline{\lambda}_{\Sigma}^2(\kappa_0, \rho_0)} + 1.$$

Therefore, recalling the definition of $\bar{\nu}_{\Sigma}(\kappa_0, \rho_0)$, we have shown that

$$\nu(\mathbb{P}_0) \leq \bar{\nu}_{\Sigma}(\kappa_0, \rho_0). \quad (18)$$

Then (14) follows by applying the inequalities (17) and (18) on the bound provided by Lemma 1. ■

4. A Concentration Inequality for Sum of Random Positive Semidefinite Matrices

We mentioned previously that we may easily adapt Lemma 1 to apply under different moment conditions. Furthermore, the special structure of the norm of interest $\|\cdot\|$ can be leveraged to adjust the choice of \mathbb{P}_0 in Lemma 1. As an example, we use Lemma 1 to recover, up to the constant factors, a result of Zhivotovskiy (2024, Theorem 1) that itself generalizes a concentration inequality due to Koltchinskii and Lounici (2017, Theorem 9) for sample covariance matrices of sub-Gaussian random vectors.³

Theorem 5 *Let Z_1, \dots, Z_n be i.i.d. copies of a random positive semidefinite matrix $Z \in \mathbb{R}^{\ell \times \ell}$ whose mean is $\Sigma = \mathbb{E} Z$ and for some parameter $\eta \geq 1$ and every $p \geq 1$ and $u \in \mathbb{R}^{\ell}$ obeys*

$$\left(\mathbb{E} (|u^{\top} Z u|^p) \right)^{1/p} \leq \eta p \|u\|_{\Sigma}^2.$$

Denoting the effective rank of Σ by

$$r_{\text{eff}}(\Sigma) \stackrel{\text{def}}{=} \frac{\text{tr} \Sigma}{\|\Sigma\|_{\text{op}}},$$

for any $t \geq 0$, with probability at least $1 - e^{-t}$, we have

$$\left\| \frac{1}{n} \sum_{i \in [n]} Z_i - \Sigma \right\|_{\text{op}} \lesssim \eta \|\Sigma\|_{\text{op}} \sqrt{\frac{2r_{\text{eff}}(\Sigma) + t}{n}} \max \left\{ 1, \sqrt{\frac{r_{\text{eff}}(\Sigma) + t/2}{n}} \right\}.$$

Proof Let $X = \sum_{i=1}^n (Z_i - \Sigma)$, and \mathbb{P}_0 be the law of $U = \kappa_0^2 G G^{\top}$ where G is a standard Gaussian random vector in \mathbb{R}^{ℓ} independent of $(Z_i)_{i \in [n]}$. With $\mathbb{M}_{\text{sym}}^{\ell}$ denoting the set of $\ell \times \ell$ real symmetric matrices, it follows from Lemma 1 that

$$\|X\|_{\text{op}} \leq \sup_{M \in \mathbb{M}_{\text{sym}}^{\ell} : \|M\|_{\text{op}}=1} e^{t/p} \kappa_0^2 \left(\mathbb{E} (|G^{\top} X G|^p) \right)^{1/p} \left(\mathbb{P} (\kappa_0^2 |G^{\top} M G| \geq 1) \right)^{-1/p}. \quad (19)$$

3. The emphasis in (Koltchinskii and Lounici, 2017) is on the concentration inequalities in possibly infinite dimensional spaces.

First, we bound $(\mathbb{E}(|G^\top X G|^p))^{1/p}$. With $\varepsilon_1, \dots, \varepsilon_n$ being a sequence of i.i.d. Rademacher random variables independent of everything else, the standard symmetrization argument guarantees

$$(\mathbb{E}(|u^\top X u|^p))^{1/p} \leq 2 \left(\mathbb{E} \left(\left| \sum_{i \in [n]} \varepsilon_i u^\top Z_i u \right|^p \right) \right)^{1/p}.$$

Since $\sum_{i \in [n]} \varepsilon_i u^\top Z_i u$ is a sum of independent symmetric random variables, for $p \geq 2$, it follows from a moment bound due to [Latała \(1997, Corollary 2\)](#) that

$$\begin{aligned} (\mathbb{E}(|u^\top X u|^p))^{1/p} &\lesssim \sup_{q \in [\max\{2, p/n\}, p]} \frac{p}{q} \left(\frac{n}{p}\right)^{1/q} (\mathbb{E}(|u^\top Z u|^q))^{1/q} \\ &\lesssim \sup_{q \in [\max\{2, p/n\}, p]} \eta p \left(\frac{n}{p}\right)^{1/q} \|u\|_\Sigma^2 \\ &\lesssim \eta \max\{\sqrt{np}, p\} \|u\|_\Sigma^2. \end{aligned} \quad (20)$$

In fact, by doubling the hidden constant, (20) holds for all $p \geq 1$, because for $p \in [1, 2]$ we have $(\mathbb{E}(|u^\top X u|^p))^{1/p} \leq (\mathbb{E}(|u^\top X u|^2))^{1/2}$ and $\max\{\sqrt{np}, p\} \leq \max\{\sqrt{2n}, 2\} \leq 2 \max\{\sqrt{np}, p\}$. Therefore, using (20), for any $p \geq 1$, together with [Lemma 11](#) we obtain

$$\begin{aligned} (\mathbb{E}(|G^\top X G|^p))^{1/p} &\lesssim \eta \max\{\sqrt{np}, p\} \left(\mathbb{E}(\|G\|_\Sigma^{2p}) \right)^{1/p} \\ &\lesssim \eta \max\{\sqrt{np}, p\} \left(\text{tr}(\Sigma) + p \|\Sigma\|_{\text{op}} \right). \end{aligned} \quad (21)$$

Next, we bound $\mathbb{P}(\kappa_0^2 |G^\top M G| \geq 1)$ for $M \in \mathbb{M}_{\text{sym}}^\ell$ with $\|M\|_{\text{op}} = 1$. We use a simple form of a general technique due to [Chatterjee \(2019, Lemma 1.2\)](#). Let $\tilde{G} = (I - vv^\top)G + \tilde{g}v$ where v is the singular vector of M corresponding to the singular value 1, and $\tilde{g} \sim \text{Normal}(0, 1)$ is independent of G . Since G and \tilde{G} are both standard Gaussian random vectors, we have

$$\begin{aligned} 2 \mathbb{P}(\kappa_0^2 |G^\top M G| \geq 1) &= \mathbb{P}(\kappa_0^2 |G^\top M G| \geq 1) + \mathbb{P}_0(\kappa_0^2 |\tilde{G}^\top M \tilde{G}| \geq 1) \\ &\geq \mathbb{P}(\kappa_0^2 |\tilde{G}^\top M \tilde{G} - G^\top M G| \geq 2) \\ &= \mathbb{P}(\kappa_0^2 |\langle \tilde{G}, v \rangle^2 - \langle G, v \rangle^2| \geq 2) \\ &= \mathbb{P}(\kappa_0^2 |\langle \tilde{G} + G, v \rangle| |\langle \tilde{G} - G, v \rangle| \geq 2). \end{aligned}$$

Observe that, by construction, $\langle \tilde{G} + G, v \rangle / \sqrt{2}$ and $\langle \tilde{G} - G, v \rangle / \sqrt{2}$ are independent standard Gaussian random variables. Therefore, the above inequality implies that

$$\begin{aligned} \mathbb{P}(\kappa_0^2 |G^\top M G| \geq 1) &\geq \frac{1}{2} \mathbb{P} \left(\frac{1}{\sqrt{2}} |\langle \tilde{G} + G, v \rangle| \geq \frac{1}{\kappa_0} \right) \mathbb{P} \left(\frac{1}{\sqrt{2}} |\langle \tilde{G} - G, v \rangle| \geq \frac{1}{\kappa_0} \right) \\ &= \frac{1}{2} (2\Phi(-\kappa_0^{-1}))^2 \\ &\geq \frac{1}{8} e^{-4/(3\kappa_0^2)}, \end{aligned} \quad (22)$$

where we used (11) on the third line. Applying the bounds (21) and (22) in (19) we obtain

$$\begin{aligned} \|X\|_{\text{op}} &\lesssim e^{t/p} \eta \max\{\sqrt{np}, p\} \kappa_0^2 \left(\text{tr}(\Sigma) + p \|\Sigma\|_{\text{op}} \right) 8^{1/p} e^{4/(3\kappa_0^2 p)} \\ &= C \eta \|\Sigma\|_{\text{op}} e^{(t+\log(8))/p} \sqrt{\max\{n, p\}} \left(p^{-1/2} r_{\text{eff}}(\Sigma) + p^{1/2} \right) \kappa_0^2 p e^{4/(3\kappa_0^2 p)}, \end{aligned}$$

where C represents a positive absolute constant. Minimizing the right-hand side of the inequality with respect to $\kappa_0 > 0$, choosing $p = r_{\text{eff}}(\Sigma) + t/\log(8)$, and using the facts that $r_{\text{eff}}(\Sigma) \geq 1$ and $\log(8) > 2$ we conclude that with probability at least $1 - e^{-t}$ we have

$$\left\| \frac{1}{n} \sum_{i \in [n]} Z_i - \Sigma \right\|_{\text{op}} = \frac{1}{n} \|X\|_{\text{op}} \lesssim \eta \|\Sigma\|_{\text{op}} \sqrt{\frac{2r_{\text{eff}}(\Sigma) + t}{n}} \max \left\{ 1, \sqrt{\frac{r_{\text{eff}}(\Sigma) + t/2}{n}} \right\}.$$

■

References

- Radosław Adamczak, Alexander E. Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of the American Mathematical Society*, 23(2):535–561, October 2009. ISSN 0894-0347. doi: 10.1090/S0894-0347-09-00650-X.
- Radosław Adamczak, Rafał Latała, Alexander E. Litvak, Krzysztof Oleszkiewicz, Alain Pajor, and Nicole Tomczak-Jaegermann. A short proof of Paouris’ inequality. *Canadian Mathematical Bulletin*, 57(1):3–8, 2014. doi: 10.4153/CMB-2012-014-5.
- Ishaq Aden-Ali. On the injective norm of sums of random tensors and the moments of Gaussian chaoses, 2025. arXiv preprint; [arXiv:2503.10580](https://arxiv.org/abs/2503.10580).
- Simeon Alesker. ψ_2 -estimate for the Euclidean norm on a convex body in isotropic position. In *Geometric Aspects of Functional Analysis*, pages 1–4. Birkhäuser Basel, 1995. ISBN 9783034890908. doi: 10.1007/978-3-0348-9090-8_1.
- Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends in Machine Learning*, 17(2):174–303, 2024. ISSN 1935-8237. doi: 10.1561/2200000100.
- Venkat Anantharam. A variational characterization of Rényi divergences. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 893–897, 2017. doi: 10.1109/ISIT.2017.8006657.
- Jean-Yves Audibert and Olivier Bousquet. Combining PAC-Bayesian and generic chaining bounds. *Journal of Machine Learning Research*, 8:863–889, 2007. ISSN 1532-4435.
- Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011. doi: 10.1214/11-AOS918.

- Afonso S. Bandeira, March T. Boedihardjo, and Ramon van Handel. Matrix concentration inequalities and free probability. *Inventiones mathematicae*, 234(1):419–487, June 2023. doi: 10.1007/s00222-023-01204-6.
- Afonso S. Bandeira, Giorgio Cipolloni, Dominik Schröder, and Ramon van Handel. Matrix concentration inequalities and free probability II. Two-sided bounds and applications, 2025. arXiv preprint; [arXiv:2406.11453](https://arxiv.org/abs/2406.11453).
- I. Bárány and D. G. Larman. Convex bodies, economic cap coverings, random polytopes. *Mathematika*, 35(2):274–291, 1988. doi: 10.1112/S0025579300015266.
- Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer International Publishing, 2017. doi: 10.1007/978-3-319-48311-5.
- Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, Oxford, February 2013. ISBN 978-0199535255.
- Tatiana Brailovskaya and Ramon van Handel. Universality and sharp matrix concentration inequalities. *Geometric and Functional Analysis*, 34(6):1734–1838, October 2024. ISSN 1420-8970. doi: 10.1007/s00039-024-00692-9.
- Olivier Catoni. PAC-Bayesian supervised classification: The thermodynamics of statistical learning. In *IMS Lecture Notes Monograph Series*, volume 56. Institute of Mathematical Statistics, 2007. doi: 10.1214/074921707000000391.
- Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012. doi: 10.1214/11-AIHP454.
- Olivier Catoni and Ilaria Giulini. Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression, 2017. arXiv preprint; [arXiv:1712.02747](https://arxiv.org/abs/1712.02747).
- Sourav Chatterjee. A general method for lower bounds on fluctuations of random variables. *The Annals of Probability*, 47(4):2140–2171, 2019. doi: 10.1214/18-aop1304.
- Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*. AUAI Press, 2017.
- Lawrence C. Evans and Ronald F. Gariepy. *Measure Theory and Fine Properties of Functions*. Textbooks in Mathematics Series. CRC Press LLC, 2nd ed. edition, 2025. ISBN 9781032946443. doi: 10.1201/9781003583004.
- Armengol Gasull and Frederic Utzet. Approximating Mills ratio. *Journal of Mathematical Analysis and Applications*, 420(2):1832–1853, 2014. ISSN 0022-247X. doi: 10.1016/j.jmaa.2014.05.034.
- Ilaria Giulini. Robust dimension-free Gram operator estimates. *Bernoulli*, 24(4B):3864–3923, November 2018. ISSN 1350-7265. doi: 10.3150/17-BEJ981.

- Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1), February 2017. ISSN 1350-7265. doi: 10.3150/15-BEJ730.
- John Langford and John Shawe-Taylor. PAC-Bayes & margins. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002.
- Rafał Latała. Estimation of moments of sums of independent real random variables. *The Annals of Probability*, 25(3):1502–1513, 1997. doi: 10.1214/aop/1024404522.
- Rafał Latała. Estimates of moments and tails of Gaussian chaoses. *The Annals of Probability*, 34(6):2315–2331, November 2006. ISSN 0091-1798. doi: 10.1214/009117906000000421.
- Rafał Latała. *On Some Problems Concerning Log-Concave Random Vectors*, pages 525–539. Springer New York, 2017. ISBN 9781493970056. doi: 10.1007/978-1-4939-7005-6_16.
- Rafał Latała and Piotr Nayar. Hadamard products and moments of random vectors. *Advances in Mathematics*, 375:107414, 2020. ISSN 0001-8708. doi: 10.1016/j.aim.2020.107414.
- Rafał Latała and Jakub. O. Wojtaszczyk. On the infimum convolution inequality. *Studia Mathematica*, 189(2):147–187, 2008.
- Joseph Lehec. *Moments of the Gaussian Chaos*, pages 327–340. Springer Berlin Heidelberg, September 2010. ISBN 9783642152177. doi: 10.1007/978-3-642-15217-7_13.
- Guy Lever, François Laviollette, and John Shawe-Taylor. Distribution-dependent PAC-Bayes priors. In Marcus Hutter, Frank Stephan, Vladimir Vovk, and Thomas Zeugmann, editors, *Algorithmic Learning Theory*, pages 119–133, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-16108-7.
- François Lust-Piquard. Inégalités de Khintchine dans c_p ($1 < p < \infty$). *C. R. Acad. Sci. Paris*, 303(7):289–292, August 1986.
- Françoise Lust-Piquard and Gilles Pisier. Non commutative Khintchine and Paley inequalities. *Arkiv för Matematik*, 29(1–2):241–260, December 1991. ISSN 0004-2080. doi: 10.1007/BF02384340.
- David A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT’98, pages 230–234, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 1581130570. doi: 10.1145/279943.279989.
- Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2019. ISBN 0262039400.
- Daniel Murawski. Comparing moments of real log-concave random variables, 2022. arXiv preprint; [arXiv:2211.05210](https://arxiv.org/abs/2211.05210).
- Yuki Ohnishi and Jean Honorio. Novel change of measure inequalities with applications to PAC-Bayesian bounds and Monte Carlo estimation. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1711–1719. PMLR, April 2021.

- Roberto Imbuzeiro Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166(3-4):1175–1194, September 2016. ISSN 1432-2064. doi: 10.1007/s00440-016-0738-9.
- Grigoris Paouris. Concentration of mass on convex bodies. *GFA Geometric And Functional Analysis*, 16(5):1021–1049, November 2006. ISSN 1420-8970. doi: 10.1007/s00039-006-0584-5.
- Antoine Picard-Weibel and Benjamin Guedj. On change of measure inequalities for f -divergences, 2022. arXiv preprint; [arXiv:2202.05568](https://arxiv.org/abs/2202.05568).
- Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2024.
- Maxim Raginsky and Igal Sason. Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends in Communications and Information Theory*, 10(1-2):1–246, 2013. ISSN 1567-2190. doi: 10.1561/01000000064.
- R Tyrrell Rockafellar. Second-order convex analysis. *Journal of Nonlinear and Convex Analysis*, 1:1–16, 1999.
- Carsten Schütt and Elisabeth Werner. The convex floating body. *Mathematica Scandinavica*, 66(2): 275–290, 1990. ISSN 00255521, 19031807.
- Carsten Schütt, Christoph Thäle, Nicola Turchi, and Elisabeth Werner. Weighted floating functions and weighted functional affine surface areas. *Transactions of the American Mathematical Society*, September 2025. ISSN 0002-9947. doi: 10.1090/tran/9439.
- Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, May 2014. ISBN 9781107298019. doi: 10.1017/cbo9781107298019.
- John Shawe-Taylor and Robert C. Williamson. A PAC analysis of a Bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory, COLT '97*, pages 2–9, New York, NY, USA, 1997. Association for Computing Machinery. ISBN 0897918916. doi: 10.1145/267460.267466.
- Michel Talagrand. Regularity of Gaussian processes. *Acta Mathematica*, 159(0):99–149, 1987. doi: 10.1007/bf02392556.
- Michel Talagrand. *Upper and Lower Bounds for Stochastic Processes*. Springer Berlin Heidelberg, 2014.
- Joel A. Tropp. Second-order matrix concentration inequalities. *Applied and Computational Harmonic Analysis*, 44(3):700–736, May 2018. ISSN 1063-5203. doi: 10.1016/j.acha.2016.07.005.
- Tim van Erven and Peter Harremoës. Rényi divergence and Kullback–Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014. doi: 10.1109/TIT.2014.2320500.
- Ramon van Handel. On the spectral norm of Gaussian random matrices. *Transactions of the American Mathematical Society*, 369(11):8161–8178, May 2017. ISSN 1088-6850. doi: 10.1090/tran/6922.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Cédric Villani. *Optimal Transport: Old and New*. Springer Berlin Heidelberg, 2009. ISBN 9783540710509. doi: 10.1007/978-3-540-71050-9.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Nikita Zhivotovskiy. Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *Electronic Journal of Probability*, 29:1–28, January 2024. ISSN 1083-6489. doi: 10.1214/23-ejp1021.

Wenda Zhou, Victor Veitch, Morgane Austern, Ryan P. Adams, and Peter Orbanz. Non-vacuous generalization bounds at the ImageNet scale: a PAC-Bayesian compression approach. In *International Conference on Learning Representations*, 2019.

Appendix A. A Concentration Inequality for the Sample Covariance Matrix of Sub-exponential Random Vectors

Zhivotovskiy (2024, Theorem 3) also provides a concentration inequality for the sample covariance matrix of centered log-concave distributions. Following Adamczak et al. (2009), Zhivotovskiy (2024) approximates the quadratic form induced by the sample covariance matrix by a term obtained through a certain direction-dependent truncation, and a term formed by the “peaky” residuals. The PAC-Bayesian argument is then used to obtain the tail bound for the “truncated” term, whereas the tail bound for peaky term is obtained using the *decoupling-chaining* method (Talagrand, 2014, Proposition 9.4.2) as cited in (Zhivotovskiy, 2024), combined with the tail bound for the Euclidean norm of log-concave random vectors due to Paouris (2006, Theorem 1.1) (see also (Adamczak et al., 2014, Theorem 2)). In fact, Paouris’s inequality is crucial in the proof of Zhivotovskiy (2024, Theorem 3) in the sense that it is tailored to log-concave measures instead of merely using the sub-exponential property.⁴

Our next result, Theorem 6, provides a concentration inequality for the sample covariance matrix of random vectors with sub-exponential distribution. Similar to the model considered in Theorem 2, the marginals of the sample covariance matrix do not have finite exponential moments, which rules out the analysis based solely on the PAC-Bayesian argument. While not adapted for log-concave distributions, without resorting to truncation techniques and relying only on Lemma 1, we can partially recover the result of Zhivotovskiy (2024, Theorem 3) with a matching confidence level. Specifically, under the extra condition that the number of samples is at least the cube of the effective rank of the population covariance matrix, the tail bound we provide below in Theorem 6 is equivalent to the tail bound provided by Zhivotovskiy (2024, Theorem 3).

Theorem 6 *Let Z_1, \dots, Z_n be independent copies of a centered random vector $Z \in \mathbb{R}^d$ with covariance matrix $\Sigma = \mathbb{E}(ZZ^\top)$ and sub-exponential marginals, i.e., for a certain parameter $\eta \geq 1$*

4. Log-concave measures are “regular” (see, e.g., (Latała and Wojtaszczyk, 2008, Definition 3.4), (Latała, 2017, the statement following Definition 2), (Murawski, 2022)) which also implies that they are sub-exponential.

and every $u \in \mathbb{R}^d$ and $p \geq 1$ we have

$$(\mathbb{E}(|\langle u, Z \rangle|^p))^{1/p} \leq \eta p \|u\|_{\Sigma}. \quad (23)$$

Then, with $r = r_{\text{eff}}(\Sigma) = \text{tr}(\Sigma) / \|\Sigma\|_{\text{op}}$ denoting the effective rank of Σ , for any $t \geq 0$, with probability at least $1 - e^{-t}$ we have

$$\left\| \frac{1}{n} \sum_{i=1}^n Z_i Z_i^{\top} - \Sigma \right\|_{\text{op}} \lesssim \begin{cases} \eta^2 \|\Sigma\|_{\text{op}} \max \left\{ \frac{(r+t+\log(8))^2}{n}, \left(\frac{r+t+\log(8)}{n} \right)^{1/2} \right\}, & \text{if } r \leq n^{1/3} \\ \eta^2 \|\Sigma\|_{\text{op}} \max \left\{ \frac{(n^{1/3}+t+\log(8))^2}{n}, \frac{r(n^{1/3}+t+\log(8))}{n} \right\}, & \text{if } r > n^{1/3}. \end{cases}$$

Proof The proof is similar to the proof of [Theorem 5](#). Let $X = \sum_{i=1}^n (Z_i Z_i^{\top} - \Sigma)$ and $U = \kappa_0^2 G G^{\top}$ for a parameter $\kappa_0 > 0$ and a standard Gaussian random vector $G \in \mathbb{R}^d$ that is independent of everything else. We can apply [Lemma 1](#) as in the proof of [Theorem 5](#); modification is only required to approximate the moments of $|\langle U, X \rangle|$. Using the standard symmetrization argument, for any $u \in \mathbb{R}^d$ we have

$$\begin{aligned} (\mathbb{E}(|u^{\top} X u|^p))^{1/p} &= \left(\mathbb{E} \left(\left| \sum_{i \in [n]} (\langle u, Z_i \rangle^2 - \|u\|_{\Sigma}^2) \right|^p \right) \right)^{1/p} \\ &\leq 2 \left(\mathbb{E} \left(\left| \sum_{i \in [n]} \varepsilon_i \langle u, Z_i \rangle^2 \right|^p \right) \right)^{1/p}, \end{aligned}$$

where $\varepsilon_1, \dots, \varepsilon_n$ are Rademacher random variables independent of each other and everything else. It then follows from ([Latała, 1997](#), Corollary 2) that for $p \geq 2$ we have

$$\begin{aligned} \left(\mathbb{E} \left(\left| \sum_{i \in [n]} \varepsilon_i \langle u, Z_i \rangle^2 \right|^p \right) \right)^{1/p} &\lesssim \sup_{q \in [\max\{2, p/n\}, p]} \frac{p}{q} \left(\frac{n}{p} \right)^{1/q} \left(\mathbb{E}(|\langle u, Z \rangle|^{2q}) \right)^{1/q} \\ &\lesssim C_{n,p} \sup_{q \in [\max\{2, p/n\}, p]} q^{-2} \left(\mathbb{E}(|\langle u, Z \rangle|^{2q}) \right)^{1/q}, \end{aligned}$$

where

$$C_{n,p} \stackrel{\text{def}}{=} \sup_{q \in [\max\{2, p/n\}, p]} pq(n/p)^{1/q}.$$

Using the assumption [\(23\)](#), we obtain

$$(\mathbb{E}(|u^{\top} X u|^p))^{1/p} \lesssim \eta^2 C_{n,p} \|u\|_{\Sigma}^2.$$

The function $q \mapsto pq(n/p)^{1/q}$ has at most one stationary point over $q > 0$, which, if exists, is a minimum. Therefore, its supremum with respect to $q \in [\max\{2, p/n\}, p]$ is attained at one of the endpoints of the interval. For $p > n$ we have $(p/n)(n/p)^{n/p} < p(n/p)^{1/p} < p$. Consequently, to

approximate $C_{n,p}$ it suffices to consider $q = 2$ and $q = p$ as the maximizing endpoints, which yields $C_{n,p} \lesssim \max\{2\sqrt{np}, p^2\}$, and thus

$$(\mathbb{E}(|u^\top Xu|^p))^{1/p} \lesssim \eta^2 \max\{\sqrt{np}, p^2\} \|u\|_\Sigma^2.$$

It then follows from [Lemma 11](#) that

$$\begin{aligned} (\mathbb{E}(|G^\top XG|^p))^{1/p} &\lesssim \eta^2 \|\Sigma\|_{\text{op}} \max\{\sqrt{np}, p^2\} (r+p) \\ &\lesssim \eta^2 \|\Sigma\|_{\text{op}} \max\{rn^{1/2}p^{1/2}, n^{1/2}p^{3/2}, rp^2, p^3\}. \end{aligned}$$

We can again show [\(22\)](#) still holds following the same argument used in the proof of [Theorem 5](#). Therefore, we have

$$\|X\|_{\text{op}} \lesssim \eta^2 \|\Sigma\|_{\text{op}} \kappa_0^2 e^{4/(3\kappa_0^2 p)} e^{(t+\log(8))/p} \max\{rn^{1/2}p^{1/2}, n^{1/2}p^{3/2}, rp^2, p^3\}$$

which by optimizing with respect to $\kappa_0 > 0$ implies

$$\|X\|_{\text{op}} \lesssim \eta^2 \|\Sigma\|_{\text{op}} e^{(t+\log(8))/p} \max\{rn^{1/2}p^{-1/2}, n^{1/2}p^{1/2}, rp, p^2\}.$$

We consider two cases based on a comparison of r and n . In the first case we have $r \leq n^{1/3}$, which implies $r \leq r^{2/5}n^{1/5} \leq n^{1/3} \leq r^{-2}n$. It is then straightforward to verify that

$$\max\{rn^{1/2}p^{-1/2}, n^{1/2}p^{1/2}, rp, p^2\} = \begin{cases} p^2, & \text{for } p \geq n^{1/3} \\ n^{1/2}p^{1/2}, & \text{for } r \leq p \leq n^{1/3} \\ rn^{1/2}p^{-1/2}, & \text{for } 2 \leq p \leq r. \end{cases}$$

Therefore, choosing $p = r + t + \log(8)$ we have

$$\|X\|_{\text{op}} \lesssim \eta^2 \|\Sigma\|_{\text{op}} \max\left\{(r+t+\log(8))^2, n^{1/2}(r+t+\log(8))^{1/2}\right\}. \quad (24)$$

Similarly, in the second case we have $r > n^{1/3}$, which implies $r > r^{2/5}n^{1/5} > n^{1/3} > r^{-2}n$. Then we can verify that

$$\max\{rn^{1/2}p^{-1/2}, n^{1/2}p^{1/2}, rp, p^2\} = \begin{cases} p^2, & \text{for } p \geq r \\ rp, & \text{for } n^{1/3} \leq p \leq r \\ rn^{1/2}p^{-1/2}, & \text{for } p \leq n^{1/3}. \end{cases}$$

Choosing $p = n^{1/3} + t + \log(8)$ again we have

$$\|X\|_{\text{op}} \lesssim \eta^2 \|\Sigma\|_{\text{op}} \max\left\{\left(n^{1/3} + t + \log(8)\right)^2, r\left(n^{1/3} + t + \log(8)\right)\right\}. \quad (25)$$

Recalling that $n^{-1} \sum_{i=1}^n Z_i Z_i^\top - \Sigma = n^{-1} X$, the result follows from [\(24\)](#) and [\(25\)](#). ■

Appendix B. A Tail Bound for the Operator Norm of Random Matrix Series

A classical model studied in random matrix theory is the random matrix series model

$$X = \sum_{i=1}^n \xi_i A_i.$$

where the matrices $A_1, \dots, A_n \in \mathbb{R}^{d_1 \times d_2}$ are deterministic and the coefficients ξ_1, \dots, ξ_n are random variables. In particular, if ξ_i s are i.i.d. standard Gaussian (or Rademacher) random variables the *non-commutative Khintchine inequality* (Lust-Piquard, 1986; Lust-Piquard and Pisier, 1991) provides a bound on the *Schatten- q* norms, including the operator norm, of X ; in its prevalent formulation for $d \times d$ symmetric matrices A_i (see, e.g., (Tropp, 2018, Corollary 2.4)) it guarantees that

$$\mathbb{E} \|X\|_{\text{op}} \lesssim \sqrt{\log d} \left\| \sum_{i=1}^n A_i^2 \right\|_{\text{op}}^{1/2}.$$

The suboptimal $O(\sqrt{\log(d)})$ multiplicative factor is eliminated recently in (Bandeira, Boedihardjo, and van Handel, 2023; Bandeira, Cipolloni, Schröder, and van Handel, 2025) which, borrowing ideas from *free probability*, derived a more refined inequality by showing that the spectrum of X is closely approximated by the spectrum of a certain corresponding infinite-dimensional deterministic operator X_{free} . In particular, Bandeira et al. (2025, Theorem 1.1) showed that

$$\left| \mathbb{E} \left(\|X\|_{\text{op}} \right) - \|X_{\text{free}}\| \right| \lesssim \tilde{v}(X) (\log(d))^{3/4}, \quad (26)$$

where $\|X_{\text{free}}\| \leq 2 \left\| \sum_{i \in [n]} A_i^2 \right\|_{\text{op}}^{1/2}$ and $\tilde{v}(X) = \left\| \sum_{i \in [n]} A_i^2 \right\|_{\text{op}}^{1/4} \sup_{w \in \mathbb{R}^n: \|w\|_2 \leq 1} \left\| \sum_{i=1}^n w_i A_i \right\|_{\text{F}}^{1/2}$.

If $\left\| \sum_{i \in [n]} A_i^2 \right\|_{\text{op}}^{1/2} \gg (\log(d))^\beta \sup_{w \in \mathbb{R}^n: \|w\|_2 \leq 1} \left\| \sum_{i=1}^n w_i A_i \right\|_{\text{F}}$, for $\beta \geq 1$ the bound (26) dominates the non-commutative Khintchine inequality, and further becomes dimension-free for $\beta \geq 3/2$. Brailovskaya and van Handel (2024) further showed the *universality* of these inequalities by considering X to be a sum of independent random matrices and relating its spectrum to the spectrum of a Gaussian matrix whose mean and covariance operator match those of X .

The following theorem considers a more general model that accommodates the cases where the coefficients $(\xi_i)_{i \in [n]}$ have heavier tail and perhaps are not even independent.

Theorem 7 *Let $\xi \in \mathbb{R}^n$ be a centered isotropic random vector such that for all $v \in \mathbb{R}^n$ we have*

$$(\mathbb{E} (|\langle \xi, v \rangle|^p))^{1/p} \leq h(p) (\mathbb{E} (|g|^p))^{1/p} \|v\|_2, \quad (27)$$

where $g \sim \text{Normal}(0, 1)$ and the function $h : [1, \infty) \mapsto (0, \infty]$ is a moment profile of ξ relative to the standard Gaussian. Furthermore, given matrices $A_1, \dots, A_n \in \mathbb{R}^{d_1 \times d_2}$ let

$$X = \sum_{i=1}^n \xi_i A_i.$$

For any $t \geq 0$, with probability at least $1 - e^{-t}$ we have

$$\|X\|_{\text{op}} \lesssim \inf_{p \geq 2} e^{t/p} h(p) \left(\sigma_* p^{1/2} + \sigma + v + \sigma_\diamond p^{-1/2} \right), \quad (28)$$

where

$$\begin{aligned} \sigma_* &= \sup_{w \in \mathbb{R}^n: \|w\|_2 \leq 1} \left\| \sum_{i=1}^n w_i A_i \right\|_{\text{op}}, & v &= \sup_{w \in \mathbb{R}^n: \|w\|_2 \leq 1} \left\| \sum_{i=1}^n w_i A_i \right\|_{\text{F}}, \\ \sigma &= \left\| \sum_{i=1}^n A_i^\top A_i \right\|_{\text{op}}^{1/2} + \left\| \sum_{i=1}^n A_i A_i^\top \right\|_{\text{op}}^{1/2}, & \sigma_\diamond &= \left(\sum_{i=1}^n \|A_i\|_{\text{F}}^2 \right)^{1/2}. \end{aligned}$$

In the special case of symmetric Gaussian series matrices (i.e., $\xi_i \stackrel{\text{i.i.d.}}{\sim} \text{Normal}(0, 1)$) we have $h(p) = 1$, and choosing $p = 2t + 2\sigma_\diamond/\sigma_* \geq 2$ in [Theorem 7](#) to obtain $\|X\|_{\text{op}} \lesssim \sigma + v + \sigma_\diamond^{1/2} \sigma_*^{1/2} + \sqrt{2t} \sigma_*$ with probability at least $1 - e^{-t}$. The bound

$$\mathbb{E} \|X\|_{\text{op}} \lesssim \sigma + v + \sigma_\diamond^{1/2} \sigma_*^{1/2}$$

can be extracted by integrating the tail bound (or directly applying the argument used to prove [Lemma 1](#)). Using the PAC-Bayesian argument, [Aden-Ali \(2025, Theorem 1.2\)](#) obtained a bound for the expected *injective norm*⁵ of sub-Gaussian random tensor series (where A_i s are tensors and ξ_i s are sub-Gaussian). For matrices, the bound (28) and that of [\(Aden-Ali, 2025\)](#) are the same up to constant factors. These bounds do not have the sub-optimal $\sqrt{\log d}$ multiplicative factor of the non-commutative Khintchine inequality, but they are generally weaker than the more refined inequality [\(26\)](#). For example, for series with symmetric matrices we may have

$$\sigma_\diamond \sigma_* = \left(\sum_{i=1}^n \|A_i\|_{\text{F}}^2 \right)^{1/2} \sup_{w: \|w\|_2 \leq 1} \left\| \sum_{i=1}^n w_i A_i \right\|_{\text{op}} \gg \left\| \sum_{i=1}^n A_i^2 \right\|_{\text{op}}^{1/2} \sup_{w: \|w\|_2 \leq 1} \left\| \sum_{i=1}^n w_i A_i \right\|_{\text{F}} = \frac{\sigma v}{2},$$

where the left-hand side can be as large as \sqrt{d} times the right-hand side (e.g., for $n = d$ and $A_i = e_i \otimes e_i$ for all $i \in [n]$).

The advantage of [Theorem 7](#) is that it addresses the situations where ξ can have a general (relative) moment profile and does not impose the independence condition on the coordinates of ξ . For example, if ξ is sub-exponential, even with independent coordinates, the standard PAC-Bayesian argument, as used in [\(Aden-Ali, 2025\)](#), cannot be employed since the analysis involves random variables that do not have finite exponential moments.

Proof of Theorem 7 For $i \in [2]$, let $G_i \in \mathbb{R}^{d_i}$ be independent standard Gaussian random vectors. Furthermore, for parameter $\kappa_0 > 0$, let \mathbb{P}_0 to be the law of

$$U = \kappa_0 G_1 \otimes G_2.$$

5. For a tensor $T \in \mathbb{R}^{d_1 \times \dots \times d_k}$ the injective norm of T induced by the q -norm is $\|T\|_{\text{V}_q} = \sup\{\langle u_1 \otimes \dots \otimes u_k, T \rangle: \forall i \in [k] u_i \in \mathbb{R}^{d_i}, \|u_i\|_q \leq 1\}$

To apply [Lemma 1](#) we first find an upper bound for $(\mathbb{E}_0 (M_X(U, p)))^{1/p}$. Let $X^\natural = \sum_{i \in [n]} \omega_i A_i$ where ω_i s are i.i.d. standard Gaussian random variables. Observe that under the assumption [\(27\)](#) we have $M_X(u, p) \leq h^p(p) M_{X^\natural}(u, p)$ for every $u \in \mathbb{R}^{d_1 \times d_2}$. Therefore, we have

$$\begin{aligned} (\mathbb{E}_0 (M_X(U, p)))^{1/p} &\leq h(p) (\mathbb{E}_0 (M_{X^\natural}(U, p)))^{1/p} \\ &= h(p) \left(\mathbb{E} \left(\left| \langle U, X^\natural \rangle \right|^p \right) \right)^{1/p}. \end{aligned}$$

Let $A_{[n]} \in \mathbb{R}^{d_1 \times d_2 \times n}$ be the tensor formed by stacking A_i s along a third axis. With $G_3 = [\omega_1, \dots, \omega_n]^\top$ and $p \geq 2$, applying the bound on the moments of Gaussian chaoses due to [Latała \(2006, Theorem 1\)](#) (see also [\(Lehec, 2010\)](#)) to the Gaussian chaos $\langle U, X^\natural \rangle = \kappa_0 \langle G_1 \otimes G_2 \otimes G_3, A_{[n]} \rangle$ yields

$$\left(\mathbb{E} \left(\left| \langle U, X^\natural \rangle \right|^p \right) \right)^{1/p} \lesssim \kappa_0 \left(\sigma_* p^{3/2} + (\sigma + v)p + \sigma_\diamond p^{1/2} \right).$$

Therefore, we have

$$(\mathbb{E}_0 (M_X(U, p)))^{1/p} \lesssim \kappa_0 h(p) \left(\sigma_* p^{3/2} + (\sigma + v)p + \sigma_\diamond p^{1/2} \right). \quad (29)$$

Furthermore, for $x \in \mathbb{R}^{d_1 \times d_2}$ with $\|x\|_{\text{op}} = 1$, let $(v_i \in \mathbb{R}^{d_i})_{i \in [2]}$ be the top singular vectors of x for which $\langle v_1 \otimes v_2, x \rangle = 1$. Also, for $i \in [2]$ let $\tilde{G}_i = (I_{d_i} - v_i v_i^\top) G_i + \tilde{g}_i v_i$ where I_{d_i} is the $d_i \times d_i$ identity matrix and \tilde{g}_i is a standard Gaussian random variable independent of everything else. Clearly, $\tilde{G}_i \sim \text{Normal}(0, I_{d_i})$ and $G_i - \tilde{G}_i \stackrel{d}{=} \sqrt{2} g_i v_i$ where $\stackrel{d}{=}$ denotes equivalence in distribution and $g_i \sim \text{Normal}(0, 1)$. Let $Z_1 = U$, $\tilde{Z}_1 = \kappa_0 \tilde{G}_1 \otimes G_2$,

$$\begin{aligned} Z_2 &= \kappa_0 (G_1 - \tilde{G}_1) \otimes G_2, \\ \tilde{Z}_2 &= \kappa_0 (G_1 - \tilde{G}_1) \otimes \tilde{G}_2, \end{aligned}$$

and

$$\begin{aligned} Z_3 &= \kappa_0 (G_1 - \tilde{G}_1) \otimes (G_2 - \tilde{G}_2) \\ &\stackrel{d}{=} 2\kappa_0 g_1 g_2 v_1 \otimes v_2, \end{aligned}$$

and observe that $Z_{i+1} = Z_i - \tilde{Z}_i$. For each $i \in [2]$ we have $Z_i \stackrel{d}{=} \tilde{Z}_i$ and thereby

$$\begin{aligned} 2 \mathbb{P} (|\langle Z_i, x \rangle| \geq 1) &= \mathbb{P} (|\langle Z_i, x \rangle| \geq 1) + \mathbb{P} \left(\left| \langle \tilde{Z}_i, x \rangle \right| \geq 1 \right) \\ &\geq \mathbb{P} \left(\left| \langle Z_i, x \rangle \right| + \left| \langle \tilde{Z}_i, x \rangle \right| \geq 2 \right) \\ &\geq \mathbb{P} (|\langle Z_{i+1}, x \rangle| \geq 2). \end{aligned}$$

Therefore, we obtain

$$\mathbb{P} (|\langle Z_1, x \rangle| \geq 1) \geq 4^{-1} \mathbb{P} (|\langle Z_3, x \rangle| \geq 4).$$

Recalling the definitions of Z_1 and Z_3 , and the fact that $\langle v_1 \otimes v_2, x \rangle = \|x\|_{\text{op}} = 1$ by choice, it follows that

$$\begin{aligned} \mathbb{P}_0(|\langle U, x \rangle| \geq 1) &\geq 4^{-1} \mathbb{P}(\kappa_0 |g_1 g_2| \geq 2) \\ &\geq 4^{-1} \mathbb{P}\left(|g_i| \geq \sqrt{2}\kappa_0^{-1/2} \text{ for all } i \in [2]\right) \\ &= \prod_{i \in [2]} \Phi(-\sqrt{2}\kappa_0^{-1/2}) \\ &\geq 4^{-2} e^{-8/(3\kappa_0)}, \end{aligned}$$

where again we used (11) in the fourth line. Consequently $\nu(\mathbb{P}_0) \leq 16e^{8/(3\kappa_0)}$. In view of (29), it follows from Lemma 1 that with probability at least $1 - e^{-t}$ we have

$$\|X\|_{\text{op}} \lesssim \inf_{p \geq 2, \kappa_0 > 0} e^{t/p} 16^{1/p} e^{8/(3p\kappa_0)} \kappa_0 h(p) \left(\sigma_* p^{3/2} + (\sigma + v)p + \sigma_\diamond p^{1/2} \right).$$

Simple calculus shows that the choice of $\kappa_0 = 8/(3p)$ is optimal. Evaluating the objective at this value yields (28). \blacksquare

Appendix C. Further Abstraction via Coupling

The argument behind Lemma 1, due to its simplicity, can be applied in a much broader scope. Instead of bounding norms of random vectors or matrices, in principle, we can derive tail bounds or moment bounds for rather complicated functions of a random variable, e.g., supremum over certain class of functionals of a random variable. The facts that $\|X\|$ is compared with a linear form $\langle U, X \rangle$, and U is taken to be independent of X are other implicit limitations of Lemma 1. The following lemma avoids such limitations and provides a more abstract generalization of Lemma 1 by expressing the bounds in terms of a coupling with X . In particular, it provides an *exact* variational formulation of the moments of interest.

Lemma 8 *Let $F: \mathbb{R}^d \rightarrow \mathbb{R}$ be a function that has a finite moment of order p with respect to the random variable $X \in \mathbb{R}^d$. Consider a coupling (X, Y) where the random variable $Y \in \mathbb{R}$, which may depend on X , is such that $(Y)_+$ has a finite moment. With $\text{ess supp}(X)$ denoting the essential support of X , define*

$$\nu(\mathbb{P}(Y | X)) = \nu_F(\mathbb{P}(Y | X)) \stackrel{\text{def}}{=} \sup_{x \in \text{ess supp}(X)} \frac{1}{\mathbb{P}(Y \geq F(X) | X = x)} \quad (30)$$

Then, for any $p \geq 1$ we have

$$\left(\mathbb{E}\left(\left(F(X)_+\right)^p\right)\right)^{1/p} = \inf_{\mathbb{P}(Y|X)} \left(\mathbb{E}\left(\left(Y_+\right)^p\right)\right)^{1/p} \nu^{1/p}(\mathbb{P}(Y | X)). \quad (31)$$

Furthermore, for any $t \geq 0$ with probability at least $1 - e^{-t}$ we have

$$F(X) \leq \inf_{\substack{\mathbb{P}(Y|X) \\ b \in \mathbb{R} \\ p \geq 1}} b + e^{t/p} \left(\mathbb{E}\left(\left(Y - b\right)_+^p\right)\right)^{1/p} \nu^{1/p}(\mathbb{P}(Y | X)). \quad (32)$$

More precisely, for any $s \in \mathbb{R}$ we have the following inequality

$$\mathbb{P}(F(X) \geq s) \leq \nu(\mathbb{P}(Y | X)) \mathbb{P}(Y \geq s), \quad (33)$$

using which the quantiles of $F(X)$ can be bounded in terms of the quantiles of Y .

Proof By Markov's inequality we have

$$\begin{aligned} \mathbb{P}(Y \geq F(X) | X) &\leq \mathbb{P}((Y - b)_+ \geq (F(X) - b)_+ | X) \\ &\leq \frac{\mathbb{E}((Y - b)_+^p | X)}{(F(X) - b)_+^p}, \end{aligned}$$

and thus

$$\begin{aligned} (F(X) - b)_+^p &\leq \frac{\mathbb{E}((Y - b)_+^p | X)}{\mathbb{P}(Y \geq F(X) | X)} \\ &\leq \frac{\mathbb{E}((Y - b)_+^p | X)}{\inf_{x \in \text{ess sup}(X)} \mathbb{P}(Y \geq F(X) | X = x)} \\ &= \mathbb{E}((Y - b)_+^p | X) \nu(\mathbb{P}(Y | X)). \end{aligned}$$

Taking the expectation with respect to X we obtain

$$\mathbb{E}((F(X) - b)_+^p) \leq \mathbb{E}((Y - b)_+^p) \nu(\mathbb{P}(Y | X)). \quad (34)$$

Then, (31) follows by setting $b = 0$ and taking the infimum with respect to $\mathbb{P}(Y | X)$ and observing that for $Y = F(X)$ we have $\nu(\mathbb{P}(Y | X)) = 1$. The tail bound (32) follows from (34), Markov's inequality, and the inequality $F(X) \leq (F(X) - b)_+ + b$.

To prove (33), observe that

$$\begin{aligned} \mathbb{P}(Y \geq s) &= \mathbb{E}(\mathbb{1}(Y \geq s)) \\ &\geq \mathbb{E} \mathbb{E}(\mathbb{1}(Y \geq F(X)) \mathbb{1}(F(X) \geq s) | X) \\ &= \mathbb{E}(\mathbb{P}(Y \geq F(X) | X) \mathbb{1}(F(X) \geq s)) \\ &\geq \mathbb{P}(F(X) \geq s) / \nu(\mathbb{P}(Y | X)), \end{aligned}$$

multiplying both sides by $\nu(\mathbb{P}(Y | X))$ yields the claimed inequality. ■

While the choice of $Y = F(X)$ attains the infimum in (31), interesting situations occur when a nearly optimal bound can be achieved by restricted choices of the coupling (X, Y) . An obvious practical requirement for suitable couplings is that $\mathbb{E}(Y^p)$ and $\nu(\mathbb{P}(Y | X))$ must be easy-to-approximate. For example, Lemma 1, as a special case of Lemma 8 with $F(X) = \|X\|$, only considers $Y = |\langle U, X \rangle|$ where $U \sim \mathbb{P}_0$ is independent of X . A more refined choice is a *Gibbs measure* with a suitable potential function as the conditional law of U given X . It is also worth mentioning that Lemma 8 uses a slightly refined definition of $\nu(\mathbb{P}_0)$ of Lemma 1 in cases where $X/\|X\|$ is (almost surely) supported on a *proper* subset of $\partial\mathcal{B}$.

We can also use a similar argument to bound the expectation of supremum over functions of X that belong to a prescribed class of functions. The following proposition provides such a bound.

Proposition 9 *Given an index set Θ let $(F_\theta: \mathbb{R}^d \rightarrow \mathbb{R})_{\theta \in \Theta}$ be a set of functions that each have a finite moment with respect to the random variable $X \in \mathbb{R}^d$. With $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ being a coupling induced by the conditional distribution $\mathbb{P}(Y | X)$ such that $(Y)_+$ has a finite moment, and recalling the definition of $\nu_F(\cdot)$ in (30), we have*

$$\mathbb{E} \left(\sup_{\theta \in \Theta} F_\theta(X) \right) \leq \inf_{p \geq 1, \mathbb{P}(Y|X)} \left(\mathbb{E} \left((Y)_+^p \right) \right)^{1/p} \sup_{\theta \in \Theta} \nu_{F_\theta}^{1/p}(\mathbb{P}(Y | X)). \quad (35)$$

Proof Following the same argument as in the proof of Lemma 8, for all $\theta \in \Theta$ we have

$$(F_\theta(X))_+^p \leq \mathbb{E} \left((Y)_+^p | X \right) \nu_{F_\theta}(\mathbb{P}(Y | X)).$$

Taking the p -th root on both sides and then maximizing with respect to $\theta \in \Theta$ we get

$$\sup_{\theta \in \Theta} (F_\theta(X))_+ \leq \left(\mathbb{E} \left((Y)_+^p | X \right) \right)^{1/p} \sup_{\theta \in \Theta} \nu_{F_\theta}^{1/p}(\mathbb{P}(Y | X)).$$

Then, (35) follows by taking the expectation with respect to X , applying Jensen's inequality for the concave function $z \mapsto z^{1/p}$, and using the fact that $\sup_{\theta \in \Theta} (F_\theta(x))_+ \geq \sup_{\theta \in \Theta} F_\theta(x)$ for all $x \in \mathbb{R}^d$. \blacksquare

The following corollary, considers a situation where the bound above leads to a more recognizable PAC-Bayesian-style bound in which the corresponding $\nu_{F_\theta}(\mathbb{P}(Y | X))$ is approximated by a certain combination of the KL divergence and the Rényi divergence (see, e.g., (Raginsky and Sason, 2013, Section 3.2.3), (Polyanskiy and Wu, 2024, Section 7.12), (Anantharam, 2017), and (van Erven and Harremos, 2014) for its properties and relation with the KL divergence) between the prior and posterior measures. Recall that for two probability measures μ and μ_{ref} with $\mu \ll \mu_{\text{ref}}$ the Rényi divergence of order $\alpha \in \mathbb{R}_{>0} \setminus \{1\}$ of μ relative to μ_{ref} is

$$D_\alpha(\mu, \mu_{\text{ref}}) = \frac{1}{\alpha - 1} \log \mathbb{E}_{\mu_{\text{ref}}} \left(\left(\frac{d\mu}{d\mu_{\text{ref}}} \right)^\alpha \right).$$

Corollary 10 *Let $X \sim \mu_X$ be a random variable in \mathbb{R}^d , \mathcal{M} be a set of probability measures on some sample space \mathcal{Z} , and $f: \mathbb{R}^d \times \mathcal{Z} \rightarrow \mathbb{R}$ be a function that for each $\mu \in \mathcal{M}$ has a finite $L^p(\mu_X \times \mu)$ norm for some $p \geq 1$. Denote the median⁶ of $f(x, Z)$ with respect to $Z \sim \mu$ by*

$$F_\mu(x) = \underset{Z \sim \mu}{\text{med}}(f(x, Z)).$$

Then, for any reference measure $\mu_{\text{ref}} \in \mathcal{M}$ and $\alpha > 1$ we have

$$\mathbb{E}_{\mu_X} \left(\sup_{\mu \in \mathcal{M}: \mu \ll \mu_{\text{ref}}} F_\mu(X) \right) \leq \inf_{p \geq 1} \left(\mathbb{E}_{\mu_X \times \mu_{\text{ref}}} \left((f(X, Z))_+^p \right) \right)^{1/p} \sup_{\mu \in \mathcal{M}: \mu \ll \mu_{\text{ref}}} e^{\widehat{D}_{\alpha, \text{KL}}(\mu, \mu_{\text{ref}})/p},$$

where

$$\widehat{D}_{\alpha, \text{KL}}(\mu, \mu_{\text{ref}}) = \min \left\{ D_\alpha(\mu, \mu_{\text{ref}}) + \frac{\alpha}{\alpha - 1} \log(2), 2D_{\text{KL}}(\mu, \mu_{\text{ref}}) + 2 \log(2) \right\}.$$

6. We chose the median instead of any other fixed quantile for simplicity and to avoid introducing extra parameters.

Proof For any μ_{ref} -measurable event \mathcal{E} , it follows from the definition of the Rényi divergence for $\alpha > 1$ and Hölder's inequality that

$$\begin{aligned} \mu(\mathcal{E}) &= \mathbb{E}_{\mu_{\text{ref}}} \left(\frac{d\mu}{d\mu_{\text{ref}}} \cdot \mathbb{1}_{\mathcal{E}} \right) \\ &\leq \left(\mathbb{E}_{\mu_{\text{ref}}} \left(\left(\frac{d\mu}{d\mu_{\text{ref}}} \right)^\alpha \right) \right)^{1/\alpha} (\mathbb{E}_{\mu_{\text{ref}}} (\mathbb{1}_{\mathcal{E}}))^{(\alpha-1)/\alpha} \\ &= e^{(\alpha-1)D_\alpha(\mu, \mu_{\text{ref}})/\alpha} (\mu_{\text{ref}}(\mathcal{E}))^{(\alpha-1)/\alpha} . \end{aligned}$$

In particular, for $\mathcal{E} = \mathcal{E}_x = \{f(x, Z) \geq F_\mu(x)\}$, we get

$$\begin{aligned} \mathbb{P}_{Z \sim \mu_{\text{ref}}} (f(x, Z) \geq F_\mu(x)) &\geq (\mathbb{P}_{Z \sim \mu} (f(x, Z) \geq F_\mu(x)))^{\alpha/(\alpha-1)} e^{-D_\alpha(\mu, \mu_{\text{ref}})} \\ &\geq 2^{-\alpha/(\alpha-1)} e^{-D_\alpha(\mu, \mu_{\text{ref}})} , \end{aligned} \quad (36)$$

where the second line follows from the definition of $F_\mu(\cdot)$.

We can obtain an alternative lower bound for $\mu_{\text{ref}}(\mathcal{E}) = \mathbb{P}_{Z \sim \mu_{\text{ref}}} (f(x, Z) \geq F_\mu(x))$ in terms of the KL divergence between μ and μ_{ref} . Recalling the choice of \mathcal{E} , it follows from the *data processing inequality* (see, e.g., (Polyanskiy and Wu, 2024, Theorem 7.4)) that

$$\begin{aligned} D_{\text{KL}}(\mu, \mu_{\text{ref}}) &\geq D_{\text{KL}}(\text{Bernoulli}(\mu(\mathcal{E})), \text{Bernoulli}(\mu_{\text{ref}}(\mathcal{E}))) \\ &= -\frac{1}{2} \log (4\mu_{\text{ref}}(\mathcal{E}) (1 - \mu_{\text{ref}}(\mathcal{E}))) , \end{aligned}$$

or equivalently

$$|1 - 2\mu_{\text{ref}}(\mathcal{E})| \leq \sqrt{1 - e^{-2D_{\text{KL}}(\mu, \mu_{\text{ref}})}} ,$$

from which we can extract the one-sided inequality

$$\begin{aligned} \mathbb{P}_{Z \sim \mu_{\text{ref}}} (f(x, Z) \geq F_\mu(x)) &\geq \frac{1}{2} \left(1 - \sqrt{1 - e^{-2D_{\text{KL}}(\mu, \mu_{\text{ref}})}} \right) \\ &\geq \frac{1}{4} e^{-2D_{\text{KL}}(\mu, \mu_{\text{ref}})} . \end{aligned} \quad (37)$$

Combining (36) and (37) we deduce

$$\nu_{F_\mu}(\mathbb{P}(Y | X)) \leq \min \left\{ 2^{\alpha/(\alpha-1)} e^{D_\alpha(\mu, \mu_{\text{ref}})}, 4e^{2D_{\text{KL}}(\mu, \mu_{\text{ref}})} \right\} = e^{\widehat{D}_{\alpha, \text{KL}}(\mu, \mu_{\text{ref}})} .$$

The claim follows by applying [Proposition 9](#) with $Y = f(X, Z)$ for $Z \sim \mu_{\text{ref}}$ and treating μ as the index for $F_\mu(\cdot)$. ■

Below we provide a simple application of [Proposition 9](#).

Example 2 Given $\sigma_1 \geq \dots \geq \sigma_d > 0$, for $i = 1, \dots, d$ let $X_i \sim \text{Normal}(0, \sigma_i^2)$ be independent centered Gaussian random variables. Our goal is to bound $\mathbb{E}(\|X\|_\infty)$ where $X = [X_i]_{i=1}^d \in \mathbb{R}^d$. Let $\Theta = [d]$ and for $\theta \in \Theta$ define $F_\theta: \mathbb{R}^d \rightarrow \mathbb{R}$ as $F_\theta(x) = |x_\theta| - b$ for some parameter $b \geq 0$.

Furthermore, take J to be uniformly distributed over $[d]$ independent of everything else and let $Y = |X_J| - b$. We have $\nu_{F_\theta}(\mathbb{P}(Y \mid X)) \leq d$ and it follows from [Proposition 9](#) that

$$\begin{aligned} \mathbb{E} \left(\max_{i \in [d]} |X_i| - b \right) &\leq d^{1/p} \left(\mathbb{E} \left((Y)_+^p \right) \right)^{1/p} \\ &= \left(\sum_{j=1}^d \mathbb{E} \left((|X_j| - b)_+^p \right) \right)^{1/p}. \end{aligned}$$

Each of the summands on the right-hand side can be bounded as

$$\begin{aligned} \mathbb{E} \left((|X_j| - b)_+^p \right) &= \int_0^\infty \mathbb{P}(|X_j| \geq s + b) p s^{p-1} ds \\ &\leq \int_0^\infty \inf_{\lambda \geq 0} 2e^{\lambda^2 \sigma_i^2 / 2 - \lambda(s+b)} p s^{p-1} ds \\ &= \int_0^\infty 2e^{-(s+b)^2 / (2\sigma_i^2)} p s^{p-1} ds \\ &\leq 2\sigma_i^p e^{-b^2 / (2\sigma_i^2)} \left(\frac{p}{2} \right)!. \end{aligned}$$

Combining the derived inequalities and optimizing with respect to $p \geq 1$ we obtain

$$\mathbb{E}(\|X\|_\infty) \leq b + \inf_{p \geq 1} \sqrt{2p} \left(\sum_{i=1}^d \sigma_i^p e^{-b^2 / (2\sigma_i^2)} \right)^{1/p}.$$

Since the optimal choice of $b \geq 0$ and $p \geq 1$ cannot be expressed explicitly as a function of the σ_i s, we resort to approximation. Let

$$\sigma_* = \max_{i \in [d]} \sigma_i \sqrt{\log(i+1)},$$

and choose $b = c\sigma_*$ for some coefficient $c > 0$ to obtain

$$\mathbb{E}(\|X\|_\infty) \leq c\sigma_* + \inf_{p \geq 1} \sqrt{2p} \sigma_1 \left(\sum_{i=1}^d (i+1)^{-c^2/2} \right)^{1/p}.$$

In particular, for $c = 2$ we have

$$\mathbb{E}(\|X\|_\infty) < 2\sigma_* + \sqrt{2}\sigma_1,$$

which, in view of the fact that $\sigma_1 \leq \sigma_* / \sqrt{\log 2}$, coincides with the tight estimate of $\mathbb{E}(\|X\|_\infty)$ (see, e.g., ([van Handel, 2017, Section 2.1](#))). Applying the Gaussian concentration inequality we also obtain the moment and tail bounds for $\|X\|_\infty$.

Appendix D. Recovering [Example 1](#) and (12) via [Theorem 4](#)

We can recover the bound (8) in [Example 1](#) from [Theorem 4](#) up to a factor of order $\sqrt{\log c_k}$ that can be ignored if c_k is not large relative to π_k^{-1} . The parameters of the problem in this example are

$\eta_1 = \sqrt{2}$, $\eta_2 = 0$, and $\Sigma = I$. By taking $\kappa_0 \rightarrow \infty$, we have $f_0(x) = \rho_0 \|x\|$ which corresponds to the choice

$$U = \nabla f_0(G) = \operatorname{argmax}_{u \in \rho_0 \{\pm u_1, \dots, \pm u_N\}} \langle u, G \rangle,$$

as considered in [Example 1](#). Recalling the definitions (5) and (6), we focus on the case $\rho_0 = 2c_k$ for some $k \in [N]$. Because

$$\|\Sigma\|_{\square} = \sup_{u \in \mathcal{B}_*} u^\top \Sigma u = \sup_{u \in \mathcal{B}_*} \|u\|_2^2 = \max_{i \in [N]} \|u_i\|_2^2,$$

we have

$$\Omega_{\Sigma}(p, \infty, 2c_k) \leq \eta_1 \sqrt{p} \rho_0 \|\Sigma\|_{\square}^{1/2} = 2\sqrt{2p}c_k \max_{i \in [N]} \|u_i\|_2^2.$$

Next, we bound $\bar{\nu}_{\Sigma}(\infty, 2c_k) \leq \rho_0 / \lambda_{\Sigma}(\infty, 2c_k)$. Let $I_k(x) = \{i \in [N] : c_k |\langle u_i, x \rangle| \geq 1\}$. We have

$$\begin{aligned} \lambda_{\Sigma}(\infty, 2c_k) &= \inf_{x \in \partial B} \sum_{i \in [N]} 2 \mathbb{P}_0(U = 2c_k u_i) (2c_k |\langle u_i, x \rangle| - 1)_+ \\ &\geq \inf_{x \in \partial B} \sum_{i \in I_k(x)} 2 \mathbb{P}_0(U = 2c_k u_i) c_k |\langle u_i, x \rangle| \\ &\geq 2\pi_k, \end{aligned}$$

using which we obtain the bound

$$\bar{\nu}_{\Sigma}(\infty, 2c_k) \leq \frac{c_k}{\pi_k}.$$

Therefore, the bound (14) provided by [Theorem 4](#) implies

$$\begin{aligned} \|X\| &\leq 2 \max_{i \in [N]} \|u_i\|_2^2 \inf_{p \geq 1, k \in [N]} e^{t/(2p)} \sqrt{2p} c_k \left(\frac{c_k}{\pi_k} \right)^{1/(2p)} \\ &\leq \sqrt{8e} \max_{i \in [N]} \|u_i\|_2^2 \min_{k \in [N]} c_k \max \left\{ 1, \sqrt{t - \log(\pi_k) + \log(c_k)} \right\}. \end{aligned}$$

Let $k_{\star} \in [N]$ be the minimizer of $\min_{k \in [N]} c_k \max \left\{ 1, \sqrt{t - \log(\pi_k)} \right\}$ in (8). If we have $\log c_{k_{\star}} \lesssim t - \log \pi_{k_{\star}}$, our derivations confirm that, up to constant factors, [Theorem 4](#) recovers the bound (8) of [Example 1](#).

Next we show that the tail bound (12), which applies when X has sub-Gaussian marginals, can be recovered from [Theorem 4](#). We have $\eta_1 = \eta$ and again $\eta_2 = 0$. Also, recall that in this example \mathcal{B} and \mathcal{B}_* coincide with the unit ℓ_2 -ball. By taking $\rho_0 \rightarrow +\infty$ we have $f_0(x) = \kappa_0 \|x\|_{\Sigma^{-1/2}}^2 / 2$, and thus

$$U = \nabla f_0(\Sigma^{1/2} G) = \kappa_0 G.$$

Furthermore, because $\|G\|_*$ has a sub-Gaussian tail we have $\lim_{\rho_0 \rightarrow +\infty} T_{\geq}(\rho_0 / \tau \kappa_0) \rho_0 = 0$. Therefore, choosing $\tau = 1$ in the definition of $\Omega_{\Sigma}(p, \kappa_0, \rho_0)$ we obtain

$$\Omega_{\Sigma}(p, \kappa_0, \infty) = \eta \sqrt{p} \kappa_0 \left(\sqrt{\operatorname{tr}(\Sigma)} + \sqrt{2p \|\Sigma\|_{\operatorname{op}}} \right).$$

Furthermore, since $\langle G, x \rangle$ has the standard Gaussian distribution we can write

$$\begin{aligned} \lambda_{\Sigma}(\kappa_0, \infty) &= \mathbb{E} \left((\kappa_0 |\langle G, x \rangle| - 1)_+ \right) = \frac{2}{\sqrt{2\pi}} \int_{1/\kappa_0}^{\infty} (\kappa_0 z - 1) e^{-z^2/2} dz \\ &\geq \frac{2}{\sqrt{2\pi}} \int_{1/\kappa_0}^{\infty} (\kappa_0 z - 1) e^{-((z-\kappa_0^{-1})^2 + \kappa_0^{-2})} dz \\ &= \frac{2\kappa_0}{\sqrt{2\pi}} e^{-\kappa_0^{-2}} \int_0^{\infty} s e^{-s^2} ds \\ &= \frac{\kappa_0}{\sqrt{2\pi}} e^{-\kappa_0^{-2}} \end{aligned}$$

where we used the inequality $z^2/2 \leq (z-\kappa_0^{-1})^2 + \kappa_0^{-2}$ on the second line, and the change of variable $s = z - \kappa_0^{-1}$ on the fourth line. Applying the inequality above in the definition of $\bar{\nu}_{\Sigma}(\kappa_0, \infty)$ yields

$$\bar{\nu}_{\Sigma}(\kappa_0, \infty) \leq 2\pi e^{1/\kappa_0^2} + 1 < 8e^{1/\kappa_0^2}.$$

Therefore, the bound (14) of [Theorem 4](#) implies

$$\begin{aligned} \|X\| = \|X\|_2 &\leq \inf_{p \geq 1, \kappa_0 > 0} e^{t/(2p)} \eta \sqrt{p} \kappa_0 \left(\sqrt{\text{tr}(\Sigma)} + \sqrt{2p \|\Sigma\|_{\text{op}}} \right) 8^{1/(2p)} e^{1/(2p\kappa_0^2)} \\ &= \sqrt{e} \eta \inf_{p \geq 1} e^{(t+\log 8)/(2p)} \left(\sqrt{\text{tr}(\Sigma)} + \sqrt{2p \|\Sigma\|_{\text{op}}} \right) \\ &\leq 4\sqrt{e} \eta \left(3 \text{tr}(\Sigma) + \frac{2}{3 \log 2} \|\Sigma\|_{\text{op}} t \right)^{1/2}, \end{aligned}$$

where the third line follows by evaluating the argument of the infimum at $p = 1 + t/\log 8$, the Cauchy–Schwarz inequality, and the fact that $\|\Sigma\|_{\text{op}} \leq \text{tr}(\Sigma)$. Therefore, up to the constant factors, we have recovered the bound (12) through [Theorem 4](#).

Appendix E. Supporting Lemmas

Lemma 11 *Let $B \in \mathbb{R}^{\ell \times \ell}$ be a symmetric matrix and $G \in \mathbb{R}^{\ell}$ be a standard Gaussian random variable. Then, for $p \geq 1$ we have*

$$\mathbb{E} (|G^{\top} B G|^p) \leq \|B\|_*^p \prod_{i=1}^{\lfloor p \rfloor} \left(1 + \frac{2(p-i) \|B\|_{\text{op}}}{\|B\|_*} \right), \quad (38)$$

where $\|B\|_*$ denotes the nuclear norm of B . The inequality above can be simplified to

$$\mathbb{E} (|G^{\top} B G|^p) \leq \left(\|B\|_* + p \|B\|_{\text{op}} \right)^p.$$

Proof With B_+ and B_- respectively denoting the positive semidefinite parts of B and $-B$, let $B_{\dagger} = B_+ + B_-$ which is a positive semidefinite matrix itself. It follows from the triangle inequality that

$$|G^{\top} B G| \leq G^{\top} B_{\dagger} G.$$

Using the Gaussian integration by parts (or Stein's lemma) we have

$$\begin{aligned}
 \mathbb{E} (|G^\top B G|^p) &\leq \mathbb{E} \left((G^\top B_{\natural} G)^p \right) \\
 &= \text{tr}(B_{\natural}) \mathbb{E} \left((G^\top B_{\natural} G)^{p-1} \right) + 2(p-1) \mathbb{E} \left((G^\top B_{\natural}^2 G) (G^\top B_{\natural} G)^{p-2} \right) \\
 &\leq \left(\text{tr}(B_{\natural}) + 2(p-1) \|B_{\natural}\|_{\text{op}} \right) \mathbb{E} \left((G^\top B_{\natural} G)^{p-1} \right) \\
 &\leq \prod_{i=1}^{\lfloor p \rfloor} \left(\text{tr}(B_{\natural}) + 2(p-i) \|B_{\natural}\|_{\text{op}} \right) \mathbb{E} \left((G^\top B_{\natural} G)^{p-\lfloor p \rfloor} \right),
 \end{aligned}$$

where the third line follows from the fact that $B_{\natural}^2 \preceq \|B_{\natural}\|_{\text{op}} B_{\natural}$, and the fourth line holds by recursion. Since $p - \lfloor p \rfloor \in [0, 1)$ and $z \mapsto z^{p-\lfloor p \rfloor}$ is concave over $z \geq 0$, Jensen's inequality guarantees that

$$\mathbb{E} \left((G^\top B_{\natural} G)^{p-\lfloor p \rfloor} \right) \leq (\text{tr}(B_{\natural}))^{p-\lfloor p \rfloor}.$$

Combining the derived inequalities, we obtain

$$\mathbb{E} (|G^\top B G|^p) \leq (\text{tr}(B_{\natural}))^p \prod_{i=1}^{\lfloor p \rfloor} \left(1 + \frac{2(p-i) \|B_{\natural}\|_{\text{op}}}{\text{tr}(B_{\natural})} \right)$$

The result follows by observing that $\|B_{\natural}\|_{\text{op}} = \|B\|_{\text{op}}$ and $\text{tr}(B_{\natural}) = \|B\|_*$.

To obtain the simplified version of (38), we can apply the AM–GM inequality to the right-hand side of (38) which yields

$$\begin{aligned}
 \mathbb{E} (|G^\top B G|^p) &= \|B\|_*^p \left(1 + (2p-1-\lfloor p \rfloor) \frac{\|B\|_{\text{op}}}{\|B\|_*} \right)^{\lfloor p \rfloor} \\
 &\leq \left(\|B\|_* + p \|B\|_{\text{op}} \right)^p,
 \end{aligned}$$

as desired. ■

Lemma 12 (Second Moment Method) *Let $\xi \in \mathbb{R}$ be a random variable. For any $s \in \mathbb{R}$ we have*

$$\mathbb{P}(\xi \geq s) \geq \frac{(\mathbb{E}(\xi - s)_+)^2}{\mathbb{E}((\xi - s)_+^2)},$$

where $(x)_+ \stackrel{\text{def}}{=} \max\{x, 0\}$. In particular, if $\mathbb{E} \xi \geq 0$, then for any $\rho \in (0, 1)$ we have

$$\mathbb{P}(\xi \geq \rho \mathbb{E} \xi) \geq \frac{(1-\rho)^2 (\mathbb{E} \xi)^2}{\text{var}(\xi) + (1-\rho)^2 (\mathbb{E} \xi)^2}.$$

Proof By the Cauchy–Schwarz inequality we have

$$\begin{aligned}
 \mathbb{P}(\xi \geq s) \mathbb{E}((\xi - s)_+^2) &= \mathbb{E}(\mathbb{1}(\xi \geq s)) \mathbb{E}((\xi - s)_+^2) \\
 &\geq (\mathbb{E}((\xi - s)_+))^2,
 \end{aligned}$$

which yields the main inequality. If we further know that $\mathbb{E} \xi \geq 0$, then choosing $s = \rho \mathbb{E} \xi$ yields

$$\begin{aligned} \mathbb{P}(\xi \geq \rho \mathbb{E} \xi) &\geq \frac{(\mathbb{E}(\xi - \rho \mathbb{E} \xi)_+)^2}{\mathbb{E}((\xi - \rho \mathbb{E} \xi)^2)} \\ &\geq \frac{(1 - \rho)^2 (\mathbb{E} \xi)^2}{\text{var}(\xi) + (1 - \rho)^2 (\mathbb{E} \xi)^2}, \end{aligned}$$

where the second line follows by applying the Jensen's inequality in the numerator and using the assumption $\mathbb{E} \xi \geq 0$. \blacksquare

Lemma 13 *Recall the gradient of f_0 given by (15). For every $x \in \mathbb{R}^d$ and $\tau \in (0, 1]$, we have*

$$\|\nabla f_0(x)\|_{\Sigma} \leq \|\Sigma\|_{\text{op}}^{1/4} (1 - \tau) \kappa_0 \left\| \Sigma^{-1/2} x \right\|_{\Sigma^{1/2}} + \tau \kappa_0 \|x\|_2.$$

Proof We can view (15) as a projection (in a Hilbert space with norm $\|\cdot\|_{\Sigma^{1/2}}$). Therefore, using the fact that projections onto closed convex sets are *firmly nonexpansive* (Bauschke and Combettes, 2017, Proposition 4.8) and thus define contractions, for all $x, y \in \mathbb{R}^d$ we have

$$\|\nabla f_0(x) - \nabla f_0(y)\|_{\Sigma^{1/2}} \leq \left\| \kappa_0 \Sigma^{-1/2} (x - y) \right\|_{\Sigma^{1/2}}. \quad (39)$$

Therefore, if for $\tau \in (0, 1]$ we have $\tau \kappa_0 \|\Sigma^{-1/2} x\|_* < \rho_0$ which guarantees $\nabla f_0(\tau x) = \tau \kappa_0 \Sigma^{-1/2} x$, by choosing $y = \tau x$ we can write

$$\begin{aligned} \|\nabla f_0(x)\|_{\Sigma} &\leq \|\nabla f_0(x) - \nabla f_0(\tau x)\|_{\Sigma} + \|\nabla f_0(\tau x)\|_{\Sigma} \\ &= \|\nabla f_0(x) - \nabla f_0(\tau x)\|_{\Sigma} + \tau \kappa_0 \|x\|_2 \\ &\leq \|\Sigma\|_{\text{op}}^{1/4} \|\nabla f_0(x) - \nabla f_0(\tau x)\|_{\Sigma^{1/2}} + \tau \kappa_0 \|x\|_2 \\ &\leq \|\Sigma\|_{\text{op}}^{1/4} (1 - \tau) \kappa_0 \left\| \Sigma^{-1/2} x \right\|_{\Sigma^{1/2}} + \tau \kappa_0 \|x\|_2, \end{aligned}$$

where used the triangle inequality in the first line, the fact that $\Sigma \preceq \|\Sigma\|_{\text{op}}^{1/2} \Sigma^{1/2}$ in the third line, and (39) in the fourth line. \blacksquare