Instance-dependent uniform tail bounds for empirical processes

Sohail Bahmani

October 1, 2024

Abstract

We formulate a uniform tail bound for empirical processes indexed by a class of functions, in terms of the individual deviations of the functions rather than the worst-case deviation in the considered class. The tail bound is established by introducing an initial "deflation" step to the standard generic chaining argument. The resulting tail bound is the sum of the complexity of the "deflated function class" in terms of a generalization of Talagrand's γ functional, and the deviation of the function instance, both of which are formulated based on the natural seminorm induced by the corresponding Cramér functions. Leveraging another less demanding natural seminorm, we also show similar bounds, though with implicit dependence on the sample size, in the more general case where finite exponential moments cannot be assumed. We also provide approximations of the tail bounds in terms of the more prevalent Orlicz norms or their "incomplete" versions under suitable moment conditions.

1 Introduction

Let $(X_i)_{i=1}^n$ be i.i.d. copies of a random variable X taking values in some space $\mathcal{X}$, and denote by $\mathbb{E}_n$ the expectation with respect to the empirical measure associated with the samples $(X_i)_{i=1}^n$. A central question of the theory of empirical processes is to find tail bounds for the empirical average $\mathbb{E}_n f(X) = n^{-1}\sum_{i=1}^n f(X_i)$ that hold uniformly for all functions f belonging to a given function class $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$.

Assuming that the functions in $\mathcal{F}$ are all zero-mean, the existing tail bounds in the literature typically assert that with probability at least $1 - e^{-r}$, for all $f \in \mathcal{F}$ we have

$$\mathbb{E}_n f \le (\mathsf{C}(\mathcal{F}) + \mathsf{S}_r(\mathcal{F}))o_n(1)\,,$$

where $\mathsf{C}(\mathcal{F})$ depends on some measure of "complexity" of the function class $\mathcal{F}$ (e.g., VC-dimension [VC71; Vap98], Rademacher complexity [vdVW12; KP00], or Talagrand's functional [Tal14]), and $\mathsf{S}_r(\mathcal{F})$ is some notion of the "worst-case deviation" of the functions $f \in \mathcal{F}$ at the confidence level e^{-r}. Our goal in this paper is to establish "instance-dependent" tail bounds in which the worst-case deviation above is replaced by the deviation of each particular function of the function class. It turns out that the instance-dependent tail bounds may provide some improvements in terms of the complexity term as well.

A closely related set of results are tail bounds for *ratio type* empirical processes. Giné, Koltchinskii, and Wellner [GKW03] and Giné and Koltchinskii [GK06] have developed such tail bounds for processes indexed by a class of $[0,1]$–bounded functions. In particular, various elaborate non-asymptotic tail bounds are derived in [GK06] by "slicing" the function class to sets of functions for which the variance proxy is nearly the same, and applying to each slice Talagrand's concentration inequality for uniformly bounded empirical processes.

Our inspiration is a recent result of Lugosi and Mendelson [LM23] on instance-dependent tail bounds for certain Gaussian processes. Lugosi and Mendelson [LM23] used this result as a benchmark to motivate their main goal which is robust mean estimation with optimal direction-dependent sub-Gaussian confidence intervals. Specifically, in the case of a Gaussian processes indexed by points in some centered Euclidean ball, [LM23, Proposition 1] derived refined tail bounds that depend on the standard deviation at any queried direction rather than the worst-case standard deviation (i.e., largest eigenvalue of the corresponding covariance matrix). Furthermore, the complexity of the entire class is replaced by a quantity which, depending on the confidence level and the spectrum of the covariance matrix, can be significantly smaller than the square root of the trace of the covariance matrix appearing in the standard bounds.

Section 2 provides a more precise statement of the problem of interest. The instance-dependent tail bounds under the assumption of finite exponential moments are presented in Section 3. As a complement to this section, our calculations in Appendix B to derive more explicit expressions in the more commonly used case of function classes in (exponential type) Orlicz spaces, can be of independent interest. In Section Section 4 we consider three illustrative examples. In Section 4.1 we discuss the problem studied by [LM23] in more details, and in Section 4.2 we use instance-dependent bounds to formulate confidence intervals for the m-th largest mean of a general Gaussian vector. Section 5 further generalizes the results of Section 3 to situations where the functions of interest are L_1 (with respect to the law of X), and particularly may not have finite exponential moments. As a corollary, these rather general bounds are made more explicit, especially in terms of the sample size, for functions with finite moments of every order which, again, do not necessarily have finite exponential moments.

2 Preliminaries and Problem Setup

Let $\mathcal{F}$ denote a finite but arbitrarily large subset[1] of a vector space $\mathbb{V}$ of *centered* functions from $\mathcal{X}$ to $\mathbb{R}$ whose *cumulant generating function* is finite in a neighborhood of the origin. Specifically, for every $f \in \mathbb{V}$ we have

$$\mathbb{E}\, f(X) = 0\,,$$

and $I_f = \{\lambda \in \mathbb{R} \colon \log \mathbb{E}\, e^{\lambda f(X)} < +\infty\}$, the domain of the corresponding cumulant generating function, contains 0 in its interior.

[1] In many situations infinite function classes can be considered as well, but a completely rigorous analysis for the problems of interest requires the measurability issues to be addressed, e.g., as in [Pol84, Appx. C].

For simplicity we assume that the zero function, denoted by 0, is also in $\mathcal{F}$. We also frequently use functions $T_r\colon \mathbb{V} \to \mathbb{R}_{\geq 0}$ that are defined for $r \geq 0$ as

$$T_r(g) \stackrel{\text{def}}{=} \inf_{\lambda \geq 0} \frac{r + \log \mathbb{E}\, e^{\lambda g(X)}}{\lambda}\,,$$

with the convention that for $r > 0$, if $\log \mathbb{E}\, e^{\lambda g(X)} = +\infty$, then the objective of the infimum is also infinite and the corresponding λ is implicitly excluded. These functions determine certain confidence intervals of interest and in fact are inverses of the *rate function*, a central object in the theory of large deviations [Var84; DZ10], associated with the random variable $g(X)$. We emphasize that the domain of $T_r(\cdot)$ is not restricted to $\mathcal{F}$, and as will be seen in the sequel we also apply $T_r(\cdot)$ to other functions in $\mathbb{V}$.

It is worth mentioning that $T_r(g)$ is a general substitute for many prevalent measures of "deviation" for a function $g \in \mathbb{V}$ at the confidence level e^{-r}. For example, if $g(X)$ has a sub-Gaussian distribution and the corresponding sub-Gaussian parameter is proportional to $\|g\|_{L_2}$, then we have $T_r(g) \lesssim \sqrt{r}\|g\|_{L_2}$.[2] Another common example is of bounded functions $g(\cdot)$, where using Bernstein-type bounds (see Lemma 4 in Appendix A) we can show that $T_r(g) \lesssim \sqrt{r}\|g\|_{L_2} + r\|g\|_{L_\infty}$. More generally, as detailed in Appendix B, for exponential-type Orlicz spaces, $T_r(g)$ can be bounded by the corresponding Orlicz norm.

The function $T_r(\cdot)$ has certain properties that are important in our derivations. We have collected these properties in the following lemma, which is proved in Appendix A to be self-contained. It is worth mentioning that more general alternatives to $T_r(\cdot)$ with similar properties can be defined easily using certain variational approximations of the corresponding quantile functions [Pin14, Theorem 2.4]. These variational approximations are important in concentration inequalities for sums of independent random variables (see, e.g., [Rio17] and [Mar21]). We use the mentioned less demanding alternatives of $T_r(\cdot)$ in Section 5 to state a more general, but less explicit, version of our results in Section 3.

Lemma 1 (Properties of $T_r(\cdot)$)**.** *The function $T_r(\cdot)$ has the following properties:*

(i) *$T_r(\cdot)$ is positive homogenous in the sense that $T_r(\alpha g) = \alpha T_r(g)$ for any $\alpha > 0$ and all functions $g \in \mathbb{V}$.*

(ii) *$T_0(g) = 0$ for all functions g.*

(iii) *The mapping $r \mapsto T_r(g)$, for $r > 0$ and any particular function $g \in \mathbb{V}$, is concave and subadditive.*

(iv) *The even envelope of $T_r(\cdot)$ defined as*

$$\overline{T}_r(g) = \max\{T_r(g), T_r(-g)\}\,, \tag{1}$$

is a seminorm.

For any fixed $f \in \mathbb{V}$ the following elementary lemma, which is essentially the well-known Chernoff bound, expresses a tail bound for $\mathbb{E}_n f$ in terms of $T_{r/n}(f)$. The proof is provided in the Appendix for completeness.

[2] Here and throughout, $P \lesssim Q$ is used as a shorthand for the inequality $P \leq cQ$ for some absolute constant $c > 0$.

Lemma 2. *With the definitions above, for any function $f \in \mathbb{V}$ (whose moment generating function has 0 in the interior of its domain), with probability at least $1 - e^{-r}$, we have*

$$\mathbb{E}_n f(X) \le T_{r/n}(f) \,.$$

It is natural to seek an extension of Lemma 2 that provides an upper tail bound for the random variable of the form

$$Z = \sup_{f \in \mathcal{F}} \left(\mathbb{E}_n f(X) - T_{r/n}(f) \right) ,$$

which translates to a uniform bound for $\mathbb{E}_n f$ that holds for every instance of $f \in \mathcal{F}$. It is often more convenient to work with tail bounds expressed in terms of some seminorm of f rather than the $T_r(f)$ which is not subadditive. A natural choice is $\overline{T}_r(\cdot)$ defined by (1), and consider the seminormed spaces $(\mathbb{V}, \overline{T}_r(\cdot))$ for $r \ge 0$, where the functions of $\mathcal{F}$ belong to. In this paper we focus on finding an upper tail bound for random variables of the form

$$Z = \sup_{f \in \mathcal{F}} \left(\mathbb{E}_n f(X) - \overline{T}_{r/n}(f) \right) .$$

We emphasize that we use the term "instance-dependent tail bounds" specifically to refer to the bounds that generalize the Chernoff bound for an individual function, to the entire class as described above. For example, the result of the standard generic chaining arguments can be expressed in a way that the tail bounds depend on the queried function f. However, the resulting bounds are in terms of the optimal choice of the so-called *admissible* subsets of the function class, and the term $\overline{T}_{r/n}(f)$, even with a crude multiplicative factor, is not guaranteed to appear in the bound.

3 Tail Bounds Assuming Finite Exponential Moments

We basically follow the *generic chaining* argument [Tal14] with an initial deflation of the function class that enables us to achieve the instance-dependence we aimed for. Furthermore, we use a "truncated chain" in our derivations similar to the approach of Dirksen [Dir15, Theorem 3.2], with the distinction that we derive the tail bounds directly without resorting to the polynomial moments as in [Dir15].

3.1 A Generalized γ functional

Let us define $\varrho_r(g, h) = \overline{T}_r(g - h)$ as a distance between a pair of functions $g, h \in \mathbb{V}$. With notation overloading, we also denote the distance of a function $g \in \mathbb{V}$ to a set of functions $\mathcal{H} \subseteq \mathbb{V}$ by

$$\varrho_r(g, \mathcal{H}) = \inf_{h \in \mathcal{H}} \overline{T}_r(g - h) \,. \tag{2}$$

Similar to the truncated variant of Talagrand's γ functionals introduced in [Dir15], for $\mathcal{A} \subseteq \mathbb{V}$, and $\underline{\ell} \in \mathbb{Z}_{\ge 0}$ we define

$$\gamma(\mathcal{A};\, r, \underline{\ell}, n) = \inf_{(\mathcal{A}_i)_{i \ge 0}} \sup_{a \in \mathcal{A}} \sum_{\ell \ge \underline{\ell}} \varrho_{(r + (r+1)2^{\ell - \underline{\ell}})/n}(a, \mathcal{A}_\ell) \,, \tag{3}$$

where the infimum is taken over an increasing *admissible* sequence $(\mathcal{A}_i)_{i\geq 0}$ of the subsets of $\mathcal{A}$ with $|\mathcal{A}_i| \leq 2^{2^i}$ for $i \geq 1$, and $\mathcal{A}_0 = \{0\}$. For $n = 1$, $\underline{\ell} \approx \log_2(r)$, and the approximation $\overline{T}_r(g) \leq r^{1/\alpha}\|g\|_{\psi_\alpha}$ with $\|\cdot\|_{\psi_\alpha}$ being a ψ_α Orlicz norm, defined below in Appendix B, the γ functional defined in (3) effectively reduces to the Talagrand's (truncated) γ_α functional. For a set $\mathcal{A}$, the Talagrand's γ_α functional with respect to the suitable pseudometric ρ is defined as

$$\gamma_\alpha(\mathcal{A}, \rho; \underline{\ell}) = \inf_{(\mathcal{A}_i)_{i\geq 0}} \sup_{a\in\mathcal{A}} \sum_{\ell\geq\underline{\ell}}^{\infty} 2^{\ell/\alpha}\rho(a, \mathcal{A}_\ell)\,,$$

where the infimum is again taken with respect to a sequence of admissible sets $(\mathcal{A}_i)_{i\geq 0}$. The importance of these types of functionals was first revealed by Talagrand's *majorizing measures theorem* [Tal87], whose appellation is due to the following essentially equivalent definition of $\gamma_\alpha(\mathcal{A}, \rho) = \gamma_\alpha(\mathcal{A}, \rho; 0)$:

$$\gamma_\alpha(\mathcal{A}, \rho) = \inf_{\mu} \sup_{a\in\mathcal{A}} \int_0^\infty \left(\log \frac{1}{\mu\left(\{b \in \mathcal{A}\colon \rho(b, a) \leq \varepsilon\}\right)}\right)^{1/\alpha} \mathrm{d}\varepsilon\,,$$

with the infimum taken over probability measures μ on $\mathcal{A}$ [Tal01]. The majorizing measures theorem confirms a conjecture due to Fernique [Fer75] that the expectation of the supremum of the centered Gaussian process indexed by $\mathcal{A}$, is equivalent to $\gamma_2(\mathcal{A}, \rho)$ up to constant factors, with ρ being the canonical pseudometric induced by the Gaussian process.

Evaluating or even finding a good approximation for a γ functional of a *general* set $\mathcal{A}$ can be challenging [Tal01; vHan18], and the only solution could be "guessing" an appropriate majorizing measure or an admissible sequence of subsets [Tal01]. By pulling the supremum into the summation in the definition of γ_α functional, the infimum over the admissible sets would be achieved with each $\mathcal{A}_i$ being a covering set of $\mathcal{A}$ of cardinality 2^{2^i}. This approximation describes the *Dudley's (entropy) integral inequality* (see, e.g., [Ver18, Theorem 8.1.3], [Dir15, equation 2.3]), i.e.,

$$\gamma_\alpha(\mathcal{A}, \rho) \lesssim_\alpha \int_0^\infty (\log N(\mathcal{A}, \rho, \varepsilon))^{1/\alpha}\,\mathrm{d}\varepsilon\,,$$

where $N(\mathcal{A}, \rho, \varepsilon)$ is the covering number of $\mathcal{A}$ with respect to ρ-balls of radius ε, and $\lesssim_\alpha$ is the usual inequality sign up to a (positive) constant factor depending only on α. If accurate estimates of the covering numbers of $\mathcal{A}$ are available, approximations of γ_α through Dudley's inequality are easy to compute. However, Dudley's inequality may not deliver sufficiently sharp approximations (see, e.g., [vHan18, Section 3.1]). The notable approach of van Handel [vHan18] improves on Dudley's inequality by replacing the entropy numbers of the entire set $\mathcal{A}$ by those of certain scale-dependent "thin" subsets of $\mathcal{A}$, imitating the multiscale form of γ_α. These thin subsets are "smoothed projections" of $\mathcal{A}$ expressed by minimizers of interpolation of the base metric and a given nonnegative functional at different scales [vHan18, Section 2.1]. The resulting approximation of γ_α is shown to be sharp in several nontrivial examples where Dudley's inequality yields rather loose approximations [vHan18, Section 3].

It is worth mentioning that the γ functional defined by (3) applies in more general settings than the standard γ_α functionals thanks to the less restricted form of the dependence of

the pseudometric $\varrho_r(\cdot,\cdot)$ on the "resolution scale" r. If $\mathcal{A}$, the function class of interest, is inhomogeneous in the sense that it contains functions with significantly different tail behavior, then the standard γ_α functionals might overestimate the size (or complexity) of $\mathcal{A}$. As an illustrative example, suppose that for some absolute constant $\eta > 0$ we have $\overline{T}_r(f) \approx \|f\|_{L_\infty} r + \eta \|f\|_{L_2} \sqrt{r}$ for all $f \in \mathbb{V}$, where approximation is in a multiplicative sense, and L_∞ and L_2 norms are defined with respect to the law of X. This form of dependence on the resolution scale cannot be reproduced by the γ_α functional or other similarly defined quantities where the resolution scale and the distance to an admissible set are decoupled. Measuring the distance with respect to the scale-insensitive norm $\|f\| = c_\infty \|f\|_{L_\infty} + c_2 \|f\|_{L_2}$ for arbitrary absolute constants $c_2, c_\infty \geq 0$, leads to a suboptimal upper bound $\overline{T}_r(f) \leq (c'_\infty r + c'_2 \sqrt{r})(c_\infty \|f\|_{L_\infty} + c_2 \|f\|_{L_2})$ with $c'_2, c'_\infty > 0$ being constants that may depend only on η.

3.2 Generic Chaining with a "Deflation" Step

The following theorem is our first main result.

Theorem 1. *Let $A\colon \mathcal{F} \to \mathcal{F}$ be a mapping such that*

$$\overline{T}_{(r+k)/n}(A[f]) \leq \overline{T}_{(r+k)/n}(f)\,, \quad \text{for all } f \in \mathcal{F}\,,$$

and

$$|A[\mathcal{F}]| \leq e^k\,,$$

for some nonnegative integer k, where $A[\mathcal{F}] = \{A[f]\colon f \in \mathcal{F}\}$ denotes the range of $A[\cdot]$. Furthermore, denote the "deflation" of $\mathcal{F}$ induced by $A[\cdot]$ by

$$\mathcal{A} = \{f - A[f]\colon f \in \mathcal{F}\}\,.$$

Setting $\underline{\ell} = \lfloor \log_2(r/3) \rfloor$ for $r \geq \log(2)$, with probability at least $1 - 2e^{-r}$, for all $f \in \mathcal{F}$ we have

$$\mathbb{E}_n f(X) - \overline{T}_{(r+k)/n}(f) - \min\left\{2\overline{T}_{(2r+1)/n}(f - A[f]),\, \mathrm{rad}_{(2r+1)/n}(\mathcal{A})\right\} \leq 2\gamma(\mathcal{A};\, r, \underline{\ell}, n)\,, \tag{4}$$

where $\gamma(\mathcal{A};\, r, \underline{\ell}, n)$ is defined as in (3), and $\mathrm{rad}_s(\mathcal{A}) = \max_{a \in \mathcal{A}} \overline{T}_s(a)$ denotes the radius of $\mathcal{A}$ measured by the seminorm $T_s(\cdot)$. The bound can further be optimized with respect to the mapping $A[\cdot]$, which both k and $\mathcal{A}$ depend on.

Let us pause here to make a few remarks about Theorem 1. First, the effectiveness of the deflation step becomes clear by observing that the result of the standard generic chaining argument can be reproduced by the possibly suboptimal choice of $A[f] = 0$ for all $f \in \mathcal{F}$ in (4). The admissible sequence in a standard generic chaining argument must cover $\mathcal{F}$, whereas in our formulation the admissible sequence must cover $\mathcal{A}$, the deflated version of $\mathcal{F}$. In particular, a desirable situation occurs when we can choose $A[\cdot]$ with $k \ll r$ such that

$\gamma(\mathcal{A}; r, \underline{\ell}, n)$ is smaller than $\gamma(\mathcal{F}; r, 0, n)$, and $\overline{T}_{(r+k)/n}(f)$ is close to $\overline{T}_{r/n}(f)$. Second, the assumption that $A[\mathcal{F}]$ is finite, is not essential; as can be seen below in the proof, it suffices to guarantee that $A[f] \le \overline{T}_{r/n+o_n(1)}(f)$ holds, with probability at least $\ge 1 - e^{-r}$, for all $f \in \mathcal{F}$. For example, in Proposition 1, this condition is shown to hold through the Gaussian concentration inequality. Third, the right-hand side of (4) is basically an upper bound for $\sup_{a\in\mathcal{A}} \mathbb{E}_n\, a(X)$ that holds with probability at least $1 - e^{-r}$. There are a few techniques to derive such upper bounds other than the generic chaining technique that we considered (see, e.g., [AB07] for a shortlist of the different techniques). The generic chaining has the advantage that it applies under rather general conditions, and in the case of Gaussian processes (as in the example of Section 4) yields sharp bounds. Finally, we can simply use the term $2\overline{T}_{(2r+1)/n}(f - A[f])$ instead of the "residual term" $\min\left\{2\overline{T}_{(2r+1)/n}(f - A[f]), \mathrm{rad}_{(2r+1)/n}(\mathcal{A})\right\}$ on the left-hand side of (4), since $\overline{T}_{(2r+1)/n}(f - A[f]) \le \mathrm{rad}_{(2r+1)/n}(\mathcal{A})$ and we lose no more than a factor 2 in the worst-case. The current formulation is meant to signify that if we merely need a uniform bound for the residual error, we may simply use $\mathrm{rad}_{(2r+1)/n}(\mathcal{A})$.

Proof of Theorem 1. As in (3), let $(\mathcal{A}_\ell)_{\ell\ge 0}$ be an increasing admissible sequence of subsets of $\mathcal{A}$ such that $\mathcal{A}_0 = \{0\}$. Let $c > 0$ denote a constant that we will specify later in the proof, and set $r_\ell = r + (r + c)2^{\ell - \underline{\ell}}$. Given $A[\cdot]$ and the sequence $(\mathcal{A}_\ell)_{\ell\ge 0}$, we can decompose every $f \in \mathcal{F}$ as

$$\begin{aligned} f &= A[f] + f - A[f] \\ &= A[f] + A_{\underline{\ell}}[f] + \sum_{\ell \ge \underline{\ell}} (A_{\ell+1}[f] - A_\ell[f])\,, \end{aligned}$$

where $A_\ell[f]$ denotes a function in $\mathcal{A}_\ell$ that is closest to $f - A[f]$ with respect to the seminorm $\overline{T}_{r_\ell}(\cdot)$, i.e.,

$$A_\ell[f] = \operatorname*{argmin}_{a\in\mathcal{A}_\ell} \overline{T}_{r_\ell/n}(f - A[f] - a)\,.$$

It follows from the above decomposition that

$$\begin{aligned} \mathbb{E}_n\, f(X) - \overline{T}_{(r+k)/n}(f) = \mathbb{E}_n\, A[f](X) - \overline{T}_{(r+k)/n}(f) + \\ \mathbb{E}_n\, A_{\underline{\ell}}[f](X) + \sum_{\ell\ge\underline{\ell}} \mathbb{E}_n (A_{\ell+1}[f] - A_\ell[f])(X)\,. \end{aligned} \tag{5}$$

Lemma 2 and a simple union bound guarantee that, with probability at least $1 - |A[\mathcal{F}]|e^{-r-k} \ge 1 - e^{-r}$, we have

$$\mathbb{E}_n\, A[f](X) \le \overline{T}_{(r+k)/n}(A[f])\,, \quad \text{for all } f \in \mathcal{F}\,. \tag{6}$$

Similarly, with probability at least $1 - 2^{2^{\underline{\ell}}} e^{-r_{\underline{\ell}}}$, we have

$$\mathbb{E}_n\, A_{\underline{\ell}}[f](X) \le \overline{T}_{r_{\underline{\ell}}/n}(A_{\underline{\ell}}[f])\,, \quad \text{for all } f \in \mathcal{F}\,. \tag{7}$$

Furthermore, for each index $\ell \ge \underline{\ell}$ there are at most $|\mathcal{A}_{\ell+1}|\,|\mathcal{A}_\ell| \le 2^{2^{\ell+1}} 2^{2^\ell} \le 2^{2^{\ell+2}}$ different functions $A_{\ell+1}[f] - A_\ell[f]$ as f varies in $\mathcal{F}$. Applying Lemma 2 and the union bound again it follows that, with probability at least $1 - 2^{2^{\ell+2}} e^{-r_\ell}$, we also have

$$\mathbb{E}_n\, (A_{\ell+1} - A_\ell)\,[f](X) \le \overline{T}_{r_\ell/n}((A_{\ell+1} - A_\ell)\,[f]) \quad \text{for all } f \in \mathcal{F}\,. \tag{8}$$

Putting (6), (7), and (8) back in the decomposition (5), with probability at least $1 - e^{-r} - 2^{2^{\underline{\ell}}} e^{-r_{\underline{\ell}}} - \sum_{\ell \geq \underline{\ell}} 2^{2^{\ell+2}} e^{-r_\ell}$, we have

$$
\begin{aligned}
&\mathbb{E}_n f(X) - \overline{T}_{(r+k)/n}(f) \\
&\leq \overline{T}_{(r+k)/n}(A[f]) - \overline{T}_{(r+k)/n}(f) + \overline{T}_{r_{\underline{\ell}/n}}(A_{\underline{\ell}}[f]) + \sum_{\ell \geq \underline{\ell}} \overline{T}_{r_\ell/n}\left(A_{\ell+1}[f] - A_\ell[f]\right) \\
&\leq 2\overline{T}_{r_{\underline{\ell}/n}}(f - A[f]) + \sum_{\ell \geq \underline{\ell}} \overline{T}_{r_\ell/n}\left(A_{\ell+1}[f] - A_\ell[f]\right) \\
&\leq 2\overline{T}_{r_{\underline{\ell}/n}}(f - A[f]) + \sum_{\ell \geq \underline{\ell}} \overline{T}_{r_\ell/n}\left(f - A[f] - A_\ell[f]\right) + \overline{T}_{r_\ell/n}\left(f - A[f] - A_{\ell+1}[f]\right) \\
&\leq 2\overline{T}_{r_{\underline{\ell}/n}}(f - A[f]) + 2\sum_{\ell \geq \underline{\ell}} \overline{T}_{r_\ell/n}\left(f - A[f] - A_\ell[f]\right) \,, \quad \text{for all } f \in \mathcal{F} \,,
\end{aligned}
$$

where the second inequality follows from the assumption $\overline{T}_{(r+k)/n}(A[f]) \leq \overline{T}_{(r+k)/n}(f)$ and the fact that

$$
\begin{aligned}
\overline{T}_{r_{\underline{\ell}/n}}(A_{\underline{\ell}}[f]) &\leq \overline{T}_{r_{\underline{\ell}/n}}(f - A[f] - A_{\underline{\ell}}[f]) + \overline{T}_{r_{\underline{\ell}/n}}(f - A[f]) \\
&\leq 2\overline{T}_{r_{\underline{\ell}/n}}(f - A[f]) \,,
\end{aligned} \tag{9}
$$

and the third and fourth inequalities respectively follow from part (iv) of Lemma 1 and the fact that $\overline{T}_r(f)$ inherits the monotonicity with respect to r from $T_r(f)$. Recalling the definition (2), on the same event we can write

$$
\begin{aligned}
\mathbb{E}_n f(X) - \overline{T}_{(r+k)/n}(f) - 2\overline{T}_{r_{\underline{\ell}/n}}(f - A[f]) &\leq 2\sum_{\ell \geq \underline{\ell}} \varrho_{r_\ell/n}\left(f - A[f], \mathcal{A}_\ell\right) \,, \quad \text{for all } f \in \mathcal{F} \\
&\leq 2\sup_{a \in \mathcal{A}} \sum_{\ell \geq \underline{\ell}} \varrho_{r_\ell/n}\left(a, \mathcal{A}_\ell\right) \,.
\end{aligned}
$$

Taking the infimum with respect to the admissible subsets $(\mathcal{A}_i)_{i \geq 0}$ on the right-hand side yields

$$
\mathbb{E}_n f(X) - \overline{T}_{(r+k)/n}(f) - 2\overline{T}_{r_{\underline{\ell}/n}}(f - A[f]) \leq 2\gamma(\mathcal{A}; r, \underline{\ell}, n) \,.
$$

Furthermore, if instead of the inequality (9) we use

$$
\sup_{f \in \mathcal{F}} \overline{T}_{r_{\underline{\ell}/n}}(A_{\underline{\ell}}[f]) \leq \mathrm{rad}_{r_{\underline{\ell}/n}}(\mathcal{A}) = \sup_{a \in \mathcal{A}} \overline{T}_{r_{\underline{\ell}/n}}(a) \,,
$$

the corresponding terms $2\overline{T}_{r_{\underline{\ell}/n}}(f - A[f])$ in the subsequent inequalities can all be replaced by $\mathrm{rad}_{r_{\underline{\ell}/n}}(\mathcal{A})$. Then (4) follows as the better of the two resulting bounds.

To complete the proof, it suffices to show that for $c = \log(2) < 1$, and the prescribed $\underline{\ell} = \lfloor \log_2(r/3) \rfloor$, we have $2^{2^{\underline{\ell}}} e^{-r_{\underline{\ell}}} + \sum_{\ell \geq \underline{\ell}} 2^{2^{\ell+2}} e^{-r_\ell} \leq e^{-r}$. The specific choices of c and $\underline{\ell}$ ensures that for $\ell \geq \underline{\ell}$ we have

$$
2^{2^{\ell+2}} e^{-(r_\ell - r)} = \left(2^{2^{\underline{\ell}+2}} e^{-r-c}\right)^{2^{\ell-\underline{\ell}}}
$$

$$\leq 2^{-(2^{\ell-\underline{\ell}})}.$$

Furthermore, we have

$$2^{2^{\underline{\ell}}} e^{-(r_{\underline{\ell}}-r)} \leq \frac{1}{16}.$$

The desired inequality for the tail probability then follows as

$$\begin{aligned}
2^{2^{\underline{\ell}}} e^{-r_{\underline{\ell}}} + \sum_{\ell \geq \underline{\ell}} 2^{2^{\ell+2}} e^{-r_\ell} &= \left(2^{2^{\underline{\ell}}} e^{-(r_{\underline{\ell}}-r)} + \sum_{\ell \geq \underline{\ell}} 2^{2^{\ell+2}} e^{-(r_\ell - r)} \right) e^{-nr} \\
&\leq \left(\frac{1}{16} + \frac{1}{2} + \sum_{\ell > \underline{\ell}} 2^{-2^{\ell-\underline{\ell}}} \right) e^{-r} \\
&< \left(\frac{9}{16} + \frac{1/4}{1-1/4} \right) e^{-r} \\
&< e^{-r}.
\end{aligned}$$

□

4 Examples

In this section we consider two examples to further expose the structure and utility of instance-dependent bounds, and show that Theorem 1 provides optimal or nearly-optimal bounds. We show that, up to constant factors, Theorem 1 reproduces the bounds provided below in Propositions 1 and 2. Proofs of these propositions as stated are also provided in Appendix A.

4.1 Marginals of a Gaussian Vector

We first consider the case where the function class consists of *linear* functionals indexed by the centered unit Euclidean ball, i.e.,

$$\mathcal{F} = \{ x \mapsto \langle u, x \rangle \colon \ u \in \mathbb{R}^d, \|u\|_2 \leq 1 \},$$

and the law of the underlying random variable $X \in \mathbb{R}^d$ is Normal$(0, \Sigma)$. This scenario is studied in [LM23] who established the following.[3]

[3]The original statement in [LM23] uses slightly different formulation and notation. For example, the terms N, $\sigma(u)$, and $\log(1/\delta)$ in the original notation respectively correspond to n, $(u^\intercal \Sigma u)^{1/2}$, and r in our formulation. Furthermore, [LM23] considers a scaled version of the deviation term $\sqrt{2r/n}(u^\intercal \Sigma u)^{1/2}$ and effectively analyzes the upper and lower bounds for $\sup_{u\colon \|u\|_2 \leq 1} \langle u, X \rangle - C\sqrt{2r/n}(u^\intercal \Sigma u)^{1/2}$ for some absolute constant $C > 0$.

Proposition 1 (Lugosi and Mendelson [LM23, Proposition 1]). *Let* $X \sim \mathrm{Normal}(0, \Sigma/n)$ *be a random vector in* $\mathbb{R}^d$*, and denote the eigenvalues of the (scaled) covariance matrix* Σ *by* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$*. Furthermore, let*

$$S_k = \sup_{u \in \mathbb{R}^d \colon \|u\|_2 \leq 1} \mathbb{E}_n \langle u, X \rangle - \frac{\sqrt{2r} + \sqrt{k}}{\sqrt{n}} (u^\mathsf{T} \Sigma u)^{1/2} .$$

Then, for any nonnegative integer $k \leq d$*, with probability at least* $1 - 2e^{-r}$ *we have*[a]

$$S_k \leq \sqrt{\frac{\sum_{i=k+1}^d \lambda_i}{n}} + \sqrt{\frac{2r}{n} \lambda_{k+1}} . \tag{10}$$

Furthermore, with $k' = \lceil 6r + 3(\sqrt{2r} + \sqrt{k})^2 \rceil$*, with probability at least* $1 - 2e^{-r}$ *we have*

$$S_k \geq \sqrt{\frac{\sum_{i=k'+1}^d \lambda_i}{3n}} . \tag{11}$$

[a] We treat the summations whose lower index is larger than their upper index as empty summations that evaluate to zero.

To understand the significance of Proposition 1 as well as the role of the integer parameter k, it is worth comparing the derived instance-dependent bound to the conventional bounds. A standard approach to bound $\mathbb{E}_n \langle u, X \rangle$ uniformly for $\|u\|_2 \leq 1$ is to apply the *Gaussian concentration inequality* (see, e.g., [BLM13, Theorem 5.6]) to $\|\mathbb{E}_n X\|_2$, which, with probability at least $1 - e^{-r}$, guarantees that

$$\sup_{u \colon \|u\|_2 \leq 1} \mathbb{E}_n \langle u, X \rangle = \|\mathbb{E}_n X\|_2 \leq \sqrt{\frac{\mathrm{tr}(\Sigma)}{n}} + \sqrt{\frac{2r}{n} \|\Sigma\|_{\mathrm{op}}} ,$$

where $\mathrm{tr}(\cdot)$ and $\|\cdot\|_{\mathrm{op}}$, respectively, denote the trace and the operator norm of their matrix arguments. This bound pessimistically considers the worst-case deviation for all of the random variables $\mathbb{E}_n \langle u, X \rangle$. By setting $k = 0$ in (10), we can reproduce this pessimistic bound, except for an extra factor of 2 in front of the r-dependent term. A much better choice for k in the instance-dependent tail bound can be found as follows. For $\ell = 0, 1, \ldots, d$, let Σ_ℓ denote the best rank-ℓ approximation of Σ with respect to the operator norm, and denote the *effective rank* of $\Sigma - \Sigma_\ell$ by $d_\ell = \mathrm{tr}(\Sigma - \Sigma_\ell)/\|\Sigma - \Sigma_\ell\|_{\mathrm{op}}$, with the convention that $0/0 = 1$ at $\ell = d$. Furthermore, define

$$k_\star = \operatorname*{argmin}_{\ell = 0, 1, \ldots, d} \max \left\{ \frac{2r}{d_\ell}, \frac{\ell}{2} \right\} ,$$

and

$$C_\star = \max \left\{ \sqrt{\frac{2r}{d_{k_\star}}}, \sqrt{\frac{k_\star}{2r}} \right\} .$$

Then, setting $k = k_\star$ and straightforward manipulations of the tail bound in (10) yields the inequality

$$\mathbb{E}_n \langle u, X\rangle \le (C_\star + 1)\left(\sqrt{\frac{\operatorname{tr}(\Sigma - \Sigma_{k_\star})}{n}} + \sqrt{\frac{2r}{n}}(u^\intercal \Sigma u)^{1/2}\right).$$

Since r determines the confidence level of the tail bound, $k_\star$ and $C_\star$ only depend on this confidence level and the spectral characteristics of Σ. A favorable situation occurs when $C_\star$ is a small constant, which requires both $k_\star$ and $d_{k_\star}$ to be proportional to r.

We provide a slightly different proof of Proposition 1 that is more streamlined and makes the constant factors reasonably small and explicit. Our proof only invokes the Gaussian concentration inequality, whereas the original proof in [LM23] uses the Gaussian Poincaré inequality as well.

To put this special case in the general perspective, observe that, with $\mathbb{V}$ being the set of linear functionals over $\mathbb{R}^d$, the function class $\mathcal{F}$ consists of functions $f(x) = f_u(x) = \langle u, x\rangle$ with $\|u\|_2 \le 1$, and we have

$$T_r(f) = \overline{T}_r(f) = \sqrt{2r/n}\sqrt{u^\intercal \Sigma u} = \sqrt{2r/n}\|f\|_{L_2(X)}.$$

4.1.1 Reproducing (10) via Theorem 1

Let $\mathcal{B}_{k_0}$ denote the centered k_0-dimensional unit Euclidean ball in the span of the top k_0 eigenvectors of Σ, i.e., the column space of Σ_{k_0}. Furthermore, for a suitably small $\epsilon > 0$ let $\mathcal{N}_{\epsilon/2}$ denote an $\epsilon/2$-net of $\mathcal{B}_{k_0}$ with respect to the norm $\|u\|_\Sigma \overset{\text{def}}{=} (u^\intercal \Sigma u)^{1/2}$. Then, for $f_u(\cdot) = \langle u, \cdot\rangle \in \mathcal{F}$ we may choose

$$A[f_u] = f_{\widehat{u}_\epsilon},$$

where

$$\widehat{u}_\epsilon = \begin{cases} \left(1 - \frac{\epsilon}{2\|\widetilde{u}_{\epsilon/2}\|_\Sigma}\right)_+ \widetilde{u}_{\epsilon/2} & \text{if } \|\widetilde{u}_{\epsilon/2}\|_\Sigma > \|u\|_\Sigma, \\ \widetilde{u}_{\epsilon/2} & \text{otherwise}, \end{cases}$$

with

$$\widetilde{u}_{\epsilon/2} = \operatorname*{argmin}_{u' \in \mathcal{N}_{\epsilon/2}} \|u - u'\|_\Sigma.$$

This construction ensures that $\|u - \widehat{u}_\epsilon\|_\Sigma = \sqrt{\|u\|^2_{\Sigma - \Sigma_{k_0}} + \|u - \widehat{u}_\epsilon\|^2_{\Sigma_{k_0}}} \le \sqrt{\|u\|^2_{\Sigma - \Sigma_{k_0}} + \epsilon^2}$ and $\|\widehat{u}_\epsilon\|_\Sigma \le \|u\|_\Sigma$. With the choices made so far, we have

$$\mathcal{A} = \{f_u - f_{\widehat{u}_\epsilon} = \langle u - \widehat{u}_\epsilon, \cdot\rangle \colon \|u\|_2 \le 1\}.$$

With $\rho_2(x, y) \overset{\text{def}}{=} \|x - y\|_2$ denoting the normalized Euclidean metric, we have

$$\gamma(\mathcal{A}; r, \underline{\ell}, n) \lesssim n^{-1/2}\gamma_2(\mathcal{V}_{k_0} + \mathcal{V}^\perp_{k_0,\epsilon}, \rho_2; \underline{\ell})$$

$$\leq n^{-1/2}\gamma_2(\mathcal{V}_{k_0}, \rho_2; \underline{\ell}) + n^{-1/2}\gamma_2(\mathcal{V}_{k_0,\epsilon}^{\perp}, \rho_2; \underline{\ell}) \,,$$

where

$$\mathcal{V}_{k_0} = \left\{(\Sigma - \Sigma_{k_0})^{1/2} u \colon \|u\|_2 \leq 1\right\} \,,$$

and

$$\mathcal{V}_{k_0,\epsilon}^{\perp} = \left\{\Sigma_{k_0}^{1/2}(u - \widehat{u}_\epsilon) \colon \|u\|_2 \leq 1\right\} \,.$$

By the majorizing measures theorem [Tal14, Theorem 2.4.1], with $Z \sim \mathrm{Normal}(0, I)$ we have

$$\begin{aligned}
\gamma_2(\mathcal{V}_{k_0}, \rho_2; \underline{\ell}) &\lesssim \mathbb{E} \sup_{v \in \mathcal{V}_{k_0}} \langle v, Z \rangle \\
&\leq \mathbb{E} \left\| (\Sigma - \Sigma_{k_0})^{1/2} Z \right\|_2 \\
&\leq \sqrt{\sum_{i=k_0+1}^{d} \lambda_i} \,,
\end{aligned}$$

and

$$\begin{aligned}
\gamma_2(\mathcal{V}_{k_0,\epsilon}^{\perp}, \rho_2; \underline{\ell}) &\lesssim \mathbb{E} \sup_{v \in \mathcal{V}_{k_0,\epsilon}^{\perp}} \langle v, Z \rangle \\
&\leq \mathbb{E} \sup_{v \in \epsilon \mathcal{B}_{k_0}} \langle v, Z \rangle \\
&\leq \sqrt{k_0}\epsilon \,.
\end{aligned}$$

Furthermore, we have

$$\mathrm{rad}_{(2r+1)/n}(\mathcal{A}) \leq \sqrt{\frac{2r+1}{n}}(\sqrt{\lambda_{k_0+1}} + \epsilon) \,.$$

With these bounds at hand, invoking Theorem 1 with $k \geq \log(|\mathcal{N}_{\epsilon/2}|)$ guarantees that with probability at least $1 - 2e^{-r}$ for every u in the unit ℓ_2 ball we have

$$\langle u, X \rangle - \sqrt{\frac{2(r+k)}{n}} \|u\|_\Sigma \lesssim \frac{1}{\sqrt{n}} \left(\sqrt{\sum_{i=k_0+1}^{d} \lambda_i} + \sqrt{k_0}\epsilon \right) + \sqrt{\frac{r}{n}}(\sqrt{\lambda_{k_0+1}} + \epsilon) \qquad (12)$$

By a naïve approximation we have $|\mathcal{N}_{\epsilon/2}| \leq \left(1 + 4\sqrt{\lambda_1}/\epsilon\right)^{k_0}$. Therefore, we must have $\epsilon \geq 4\sqrt{\lambda_1}/(2^{-1/k_0} e^{k/k_0} - 1)$. In particular, if $k \geq k_0 \log(1 + 4\sqrt{\lambda_1/\lambda_{2k_0}})$, then we can choose $\epsilon = \min\{\sqrt{\lambda_{k_0+1}}, \sqrt{\sum_{i>k_0} \lambda_i / k_0}\}$ and (12) simplifies to

$$\langle u, X \rangle - \sqrt{\frac{2(r+k)}{n}} \|u\|_\Sigma \lesssim \sqrt{\frac{\sum_{i=k_0+1}^{d} \lambda_i}{n}} + \sqrt{\frac{r}{n}} \sqrt{\lambda_{k_0+1}} \,,$$

which, assuming that λ_1/λ_{2k_0} is a constant, is effectively (10) up to the constant factors.

4.2 Confidence Intervals for the "Middle-Ranked" Means of Correlated Gaussians

In this subsection we derive confidence intervals for the m-th largest mean of correlated Gaussian random variables, as another example where instance-dependent tail bounds can be applied. The proof of Proposition 2 provided in Appendix A, again relies on the Gaussian concentration inequality, as well as a bound on the expected supremum of canonical Gaussian processes over (symmetric) polytopes [Tal14, Proposition 2.4.16 and Theorem 2.4.18] (see also the discussion in [vHan18, Section 3.3]). These tools allow us to express the upper and lower bounds of the confidence interval in more explicit terms. We can basically recover Proposition 2 through Theorem 1 as explained at the end of this subsection.

Our goal is to find an upper and lower bounds for the m-th largest entry of a parameter vector $\theta \in \mathbb{R}^d$ for $m = o(d)$. We are only given $\widehat{\theta} = \theta + X$, where X is a zero-mean Gaussian random variable with covariance $\Sigma = \mathbb{E}\, XX^\intercal$. We assume that Σ is known, and, without loss of generality, it is full-rank. For any vector v we denote by $v^\downarrow$ the vector of the entries of v sorted in decreasing order. Therefore, the m-th largest entry of a vector v can be expressed as $v^\downarrow_m$. Furthermore, for any subset S of $[d] \stackrel{\text{def}}{=} \{1, \dots, d\}$ let $v_S \in \mathbb{R}^{|S|}$ denote the restriction of v to the entries indexed by S. We also use the shorthand $\Sigma_S = \mathbb{E}\, X_S X_S^\intercal$, which is the same as Σ restricted to the rows and columns in S. By $\binom{[d]}{\ell}$, we denote the set of subsets of $[d]$ of size ℓ, and we write $\triangle^\ell$ to denote the unit simplex in $\mathbb{R}^\ell$

Perhaps the simplest approach for our problem is to use the inequality

$$\left|\theta^\downarrow_m - \widehat{\theta}^\downarrow_m\right| \le \left\|\theta - \widehat{\theta}\right\|_\infty = \|X\|_\infty ,$$

that suggests a confidence interval centered at the plug-in estimator $\widehat{\theta}^\downarrow_m$ whose width is no less than $2\|X\|_\infty$. The Gaussian concentration inequality then guarantees that

$$\|X\|_\infty \le \mathbb{E}\,\|X\|_\infty + \sqrt{2r}\max_{i\in[d]} \Sigma_{i,i}^{1/2} ,$$

with probability at least $1 - e^{-r}$. Furthermore, we can bound $\mathbb{E}\,\|X\|_\infty$, viewed as the expected supremum of a canonical Gaussian process over a (symmetric) polytope, using [Tal14, Proposition 2.4.16 and Theorem 2.4.18]. Denoting the i-th largest diagonal entry of Σ by $\Sigma^\downarrow_{i,i}$, for some constant $C > 0$ we have

$$\left|\theta^\downarrow_m - \widehat{\theta}^\downarrow_m\right| \le C \max_{i\in[d]} \sqrt{\Sigma^\downarrow_{i,i} \log(i+1)} + \sqrt{2r}\sqrt{\Sigma^\downarrow_{1,1}} . \tag{13}$$

Using the instance-dependent uniform tail bounds, a more refined confidence interval for $\theta^\downarrow_m$ can be established as follows. At the end of this subsection we explain how this proposition follows from Theorem 1 by modifying certain steps of the proof provided in the Appendix A.

Proposition 2. *Let $\widehat{\theta} = \theta + X$ be a noisy observation of a parameter $\theta \in \mathbb{R}^d$ with $X \sim \text{Normal}(0, \Sigma)$. Furthermore, let $m = o(d)$ be a positive integer[a], $r \in \mathbb{R}_{\ge 0}$, and*

$k \le \min\{r, m\}$ be a nonnegative integer. For any nonempty set $S \subseteq [d]$ denote by $\Sigma_{S,k}$ the best rank-k approximation of Σ_S with respect to the operator norm, and define the vector $\sigma = \sigma(S,k)$ such that $\sigma_i = \sqrt{(\Sigma_S - \Sigma_{S,k})_{i,i}}$ for $i \in [|S|]$. Then, defining

$$\sigma^*(S,k) \stackrel{\text{def}}{=} \max_{i \in [|S|]} \sigma_i^{\downarrow} \sqrt{\log(i+1)}\,,$$

$$Q_{S,\beta}(\vartheta) = \max_{u \in \triangle^{|S|}} \langle u, \vartheta \rangle - \beta \|u\|_{\Sigma_S}\,,$$

which implicitly depends on Σ_S, and

$$\beta_{r,m,k} = \sqrt{r + m + m\log(d/m) + k}\,,$$

with probability at least $1 - 4e^{-r}$, for some some universal constant $C > 0$ we have

$$\theta_m^{\downarrow} \ge \min_{S \in \binom{[d]}{d-m+1}} Q_{S,\beta_{r,m,k}}(\widehat{\theta}_S) - C\sigma^*(S,k) - \sqrt{2}\beta_{r,m,k}\|\sigma(S,k)\|_\infty\,, \tag{14}$$

and

$$\theta_m^{\downarrow} \le \max_{S \in \binom{[d]}{m}} -Q_{S,\beta_{r,m,k}}(-\widehat{\theta}_S) + C\sigma^*(S,k) + \sqrt{2}\beta_{r,m,k}\|\sigma(S,k)\|_\infty\,. \tag{15}$$

[a] The little o notation means that $m/d \to 0$ as $d \to \infty$

If in addition to the assumption $m = o(d)$, we have $m \lesssim r$ (i.e., $m \le cr$ for some fixed constant $c > 0$), the bounds above reproduce (13) up to an extra logarithmic factor for the term $\sqrt{\Sigma_{1,1}^{\downarrow}}$.

We also have the following minimax lower bounds for estimating $\theta_m^{\downarrow}$, whose proof is provided in Appendix A.

Proposition 3. *For $\kappa \ge 0$, let $\Theta = \{\theta \in \mathbb{R}^d \colon \left\|\Sigma^{-1/2}\theta\right\|_2 \le \kappa\}$ be a compact domain of parameters. With $\delta_m \ge 0$ defined as*

$$\delta_m = \sup_{\theta \in \Theta} \theta_m^{\downarrow} + \sup_{\eta \in \Theta} \eta_{d-m+1}^{\downarrow}\,.$$

For any estimator $g(\widehat{\theta})$ of $\theta_m^{\downarrow}$ we have

$$\sup_{\theta \in \Theta} \mathbb{E}(g(\widehat{\theta}) - \theta_m^{\downarrow})^2 \ge \frac{\delta_m^2}{8e\max\{1, 2\kappa^2\}}\,. \tag{16}$$

Furthermore, we have

$$\sup_{\theta \in \Theta} \mathbb{P}\left(|g(\widehat{\theta}) - \theta_m^{\downarrow}| > \frac{\delta_m}{3\max\{1, \sqrt{2}\kappa\}}\right) \ge \frac{1}{2e}\,.$$

Because of the complicated and implicit form of the expressions in (14) and (15), it is difficult to compare—in full generality—the width of the confidence interval provided by Proposition 2 and the minimax lower bound of Proposition 3. We only focus on the special case where Σ is diagonal. Furthermore, for the sake of simpler calculations we use the lower bound

$$Q_{S,\beta}(\widehat{\theta}_S) \geq \max_{i \in S} \left(\widehat{\theta}_i - \beta \sqrt{\Sigma_{i,i}} \right) .$$

The width of the confidence interval expressed by (14) and (15), which we denote by Δ_m, can be bounded as

$$\begin{aligned}
\Delta_m = \max_{S \in \binom{[d]}{m}, S' \in \binom{[d]}{d-m+1}} \Big(& - Q_{S,\beta_{r,m,k}}(-\widehat{\theta}_S) - Q_{S',\beta_{r,m,k}}(\widehat{\theta}_{S'}) \\
& + C(\sigma^*(S,k) + \sigma^*(S',k)) \\
& + \sqrt{2}\beta_{r,m,k}(\|\sigma(S,k)\|_\infty + \|\sigma(S',k)\|_\infty) \Big) \\
\leq \max_{S \in \binom{[d]}{m}, S' \in \binom{[d]}{d-m+1}} \Big(& \min_{i \in S, j \in S'} \widehat{\theta}_i - \widehat{\theta}_j + \beta_{r,m,k}(\sqrt{\Sigma_{i,i}} + \sqrt{\Sigma_{j,j}}) \\
& + C(\sigma^*(S,k) + \sigma^*(S',k)) \\
& + \sqrt{2}\beta_{r,m,k}(\|\sigma(S,k)\|_\infty + \|\sigma(S',k)\|_\infty) \Big) \\
\leq \max_{S \in \binom{[d]}{m}, S' \in \binom{[d]}{d-m+1}} \Big(& \min_{i \in S \cap S'} 2\beta_{r,m,k}\sqrt{\Sigma_{i,i}} + C(\sigma^*(S,k) + \sigma^*(S',k)) \\
& + \sqrt{2}\beta_{r,m,k}(\|\sigma(S,k)\|_\infty + \|\sigma(S',k)\|_\infty) \Big) ,
\end{aligned}$$

where the second inequality holds because $|S \cap S'| = |S| + |S'| - |S \cup S'| \geq 1$, and we can choose $i = j \in S \cap S'$. Furthermore, we have the inequalities

$$\max\left\{\|\sigma(S,k)\|_\infty, \|\sigma(S',k)\|_\infty\right\} \leq \sqrt{\Sigma^{\downarrow}_{1,1}} ,$$

and

$$\max\left\{\sigma^*(S,k), \sigma^*(S',k)\right\} \leq \max_{i \in [d]} \sqrt{\Sigma^{\downarrow}_{i+k,i+k} \log(i+1)} ,$$

using which we deduce

$$\Delta_m \leq 2C \max_{i \in [d-k]} \sqrt{\Sigma^{\downarrow}_{i+k,i+k} \log(i+1)} + (2 + 2\sqrt{2})\beta_{r,m,k}\sqrt{\Sigma^{\downarrow}_{1,1}} .$$

With Θ defined as in Proposition 3 we have

$$\sup_{\theta \in \Theta} \theta^{\downarrow}_m = \left(\sum_{i=1}^{m} \frac{1}{\Sigma^{\downarrow}_{i,i}} \right)^{-1/2} \kappa ,$$

and

$$\sup_{\eta\in\Theta}\eta^{\downarrow}_{d-m+1}=\left(\sum_{i=1}^{d-m+1}\frac{1}{\Sigma^{\downarrow}_{i,i}}\right)^{-1/2}\kappa\,.$$

Therefore, Proposition 3 implies that any confidence interval for $\theta^{\downarrow}_m$ with coverage probability no less than $1-1/(2e)$, should have a width equal to $C_\kappa\left(\sum_{i=1}^m\frac{1}{\Sigma^{\downarrow}_{i,i}}\right)^{-1/2}$ for some constant $C_\kappa\geq 0$ that may depend on κ. In particular, for any $\theta\in\Theta$ we have

$$\mathbb{P}\left(\Delta_m>C_\kappa\left(\sum_{i=1}^m\frac{1}{\Sigma^{\downarrow}_{i,i}}\right)^{-1/2}\right)\geq\frac{1}{2e}\,.$$

Choosing $r=1+\log(8)\approx 3$, we have also shown that

$$\mathbb{P}\left(\Delta_m>2C\max_{i\in[d-k]}\sqrt{\Sigma^{\downarrow}_{i+k,i+k}\log(i+1)}+(2+2\sqrt{2})\beta_{r,m,k}\sqrt{\Sigma^{\downarrow}_{1,1}}\right)\leq\frac{1}{2e}\,.$$

Then, if we define

$$p_m(\Sigma)\stackrel{\text{def}}{=}\sum_{i=1}^m\frac{\Sigma^{\downarrow}_{1,1}}{\Sigma^{\downarrow}_{i,i}}\,,$$

and

$$q_{m,k}(\Sigma)\stackrel{\text{def}}{=}\max_{i\in[d-k]}\frac{\Sigma^{\downarrow}_{i+k,i+k}}{\Sigma^{\downarrow}_{1,1}}\log(i+1)\,,$$

then Δ_m is optimal up to a factor polylog(d), if m, $p_m(\Sigma)$, and $q_{m,k}(\Sigma)$ are all bounded from above as polylog(d). Specifically, if m, $p_m(\Sigma)$, and $q_{m,k}(\Sigma)$ are all absolute constants, then Δ_m is optimal up to a constant factor.

4.2.1 Reproducing (14) and (15) via Theorem 1

Proof of (14) provided in Appendix A first expresses $\theta^{\downarrow}_m$ in a variational form as

$$\theta^{\downarrow}_m=\min_{S\in\binom{[d]}{d-m+1}}\max_{u\in\triangle^{d-m+1}}\langle u,\widehat{\theta}_S\rangle-\langle u,X_S\rangle\,.$$

Then it establishes (14) by leveraging a uniform instance-dependent tail bound for $\langle u,X_S\rangle$ and taking the union bound over $S\in\binom{[d]}{d-m+1}$. We only need to recover (22), the instance-dependent bound for $\langle u,X_S\rangle$, using Theorem 1. Therefore, the core of the argument is basically the same argument we used in Section 4.1.1 with some modifications.

Recalling that $\triangle^\ell$ denotes the unit simplex in $\mathbb{R}^\ell$, for any fixed $S\in\binom{[d]}{d-m+1}$ let

$$\mathcal{F}=\left\{f_u=\langle u,\cdot\rangle\colon u\in\triangle^{d-m+1}\right\}\,.$$

Furthermore, for a sufficiently small nonnegative integer k_0, let $\triangle_{k_0}^{d-m+1}$ denote the orthogonal projection of $\triangle^{d-m+1}$ onto the range of Σ_{S,k_0}. Taking $\mathcal{N}_{\epsilon/2}$ to be an $\epsilon/2$-net of $\triangle_{k_0}^{d-m+1}$ with respect to the metric induced by $\|\cdot\|_{\Sigma_S}$ let

$$\widetilde{u}_{\epsilon/2} = \operatorname*{argmin}_{u' \in \mathcal{N}_{\epsilon/2}} \|u - u'\|_{\Sigma_S} ,$$

and

$$\widehat{u}_\epsilon = \begin{cases} \left(1 - \frac{\epsilon}{2\|\widetilde{u}_{\epsilon/2}\|_{\Sigma_S}}\right)_+ \widetilde{u}_{\epsilon/2} & \text{if } \|\widetilde{u}_{\epsilon/2}\|_{\Sigma_S} > \|u\|_{\Sigma_S} , \\ \widetilde{u}_{\epsilon/2} & \text{otherwise} . \end{cases}$$

Then, we have

$$\mathcal{A} = \left\{ f_u - f_{\widehat{u}_\epsilon} = \langle u - \widehat{u}_\epsilon, \cdot \rangle \colon u \in \triangle^{d-m+1} \right\} ,$$

for which

$$\begin{aligned} \gamma(\mathcal{A}; r, \underline{\ell}, 1) &\lesssim \gamma_2(\mathcal{V}_{k_0} + \mathcal{V}_{k_0}^\perp, \rho_2; \underline{\ell}) \\ &\leq \gamma_2(\mathcal{V}_{k_0}, \rho_2; \underline{\ell}) + \gamma_2(\mathcal{V}_{k_0,\epsilon}^\perp, \rho_2; \underline{\ell}) , \end{aligned}$$

where again $\rho_2(x, y) = \|x - y\|_2$, and

$$\mathcal{V}_{k_0} = \left\{ (\Sigma_S - \Sigma_{S,k_0})^{1/2} u \colon u \in \triangle^{d-m+1} \right\} ,$$

and

$$\mathcal{V}_{k_0,\epsilon}^\perp = \left\{ \Sigma_{S,k_0}^{1/2} (u - \widehat{u}_\epsilon) \colon u \in \triangle^{d-m+1} \right\} .$$

We again can invoke the majorizing measures theorem [Tal14, Theorem 2.4.1] as well as the bound on the entrywise maximum of a Gaussian random vector [Tal14, Proposition 2.4.16]; with $Z \sim \text{Normal}(0, I)$ we obtain

$$\begin{aligned} \gamma_2(\mathcal{V}_{k_0,\epsilon}^\perp, \rho_2; \underline{\ell}) &\lesssim \mathbb{E} \sup_{v \in \mathcal{V}_{k_0}} \langle v, Z \rangle \\ &\lesssim \sigma^*(S, k_0) , \end{aligned}$$

$$\begin{aligned} \gamma_2(\mathcal{V}_{k_0}, \rho_2; \underline{\ell}) &\lesssim \mathbb{E} \sup_{v \in V_{k_0,\epsilon}^\perp} \langle v, Z \rangle \\ &\leq \sqrt{k_0}\epsilon , \end{aligned}$$

and thereby

$$\gamma(\mathcal{A}; r, \underline{\ell}, 1) \lesssim \sigma^*(S, k_0) + \sqrt{k_0}\epsilon .$$

We also have

$$\begin{aligned}\mathrm{rad}_{2r+1}(\mathcal{A}) &\lesssim \sqrt{r} \sup_{u\in\triangle^{d-m+1}} \|u-\widehat{u}_\epsilon\|_{\Sigma_S} \\ &\le \sqrt{r} \sup_{u\in\triangle^{d-m+1}} (\|u\|_{\Sigma_S-\Sigma_{S,k_0}}+\epsilon) \\ &= \sqrt{r}\,(\|\sigma(S,k_0)\|_\infty+\epsilon)\,.\end{aligned}$$

Therefore, if $k \ge \log(|\mathcal{N}_{\epsilon/2}|) \ge |A[\mathcal{F}]|$, it follows from Theorem 1 that with probability at least $1-2e^{-r}$, for all $u\in\triangle^{d-m+1}$ we have

$$\langle u, X_S\rangle - \sqrt{2(r+k)}\|u\|_{\Sigma_S} \lesssim \sigma^*(S,k_0) + \sqrt{r}\|\sigma(S,k_0)\|_\infty + (\sqrt{r}+\sqrt{k_0})\epsilon\,.$$

By the approximation $|\mathcal{N}_{\epsilon/2}| \le (1+4\sqrt{\|\Sigma_S\|_{\mathrm{op}}}/\epsilon)^{k_0}$, it suffices to have $k \ge k_0\log(1+4\sqrt{\|\Sigma_S\|_{\mathrm{op}}}/\epsilon)$. In particular, using the fact that $\|\Sigma_S\|_{\mathrm{op}} \le \mathrm{tr}(\Sigma_S) \le (d-m+1)\|\sigma\|$ we can choose $\epsilon = \min\{\sigma^*/\sqrt{k_0}\,, \|\sigma\|_\infty\}$ and $k \ge k_0\log\Big(1+4\sqrt{\|\Sigma_S\|_{\mathrm{op}}}\max\{\sqrt{k_0}/\sigma^*, 1/\|\sigma\|_\infty\}\Big)$. Therefore, assuming that

$$\sqrt{\|\Sigma_S\|_{\mathrm{op}}}\max\{\sqrt{k_0}/\sigma^*, 1/\|\sigma\|_\infty\} \lesssim d\,,$$

we conclude that for $k \gtrsim k_0\log(d)$, with probability at least $1-2e^{-r}$, for all $u\in\triangle^{d-m+1}$ we have

$$\langle u, X_S\rangle - \sqrt{2(r+k)}\|u\|_{\Sigma_S} \lesssim \sigma^*(S,k_0) + \sqrt{r}\|\sigma(S,k_0)\|_\infty\,.$$

By union bound, with probability at least $1-2e^{-r}$, for all $S\in\binom{[d]}{d-m+1}$ and $u\in\triangle^{|S|}$ we have

$$\langle u, X_S\rangle - \sqrt{2(r+m+m\log(d/m)+k)}\|u\|_{\Sigma_S} \lesssim \sigma^*(S,k_0) + \sqrt{r+m+m\log(d/m)}\|\sigma(S,k_0)\|_\infty\,.$$

Using this inequality in variational expression for $\theta^\downarrow_m$ recovers (14) up to the constant factors. The derivations for the upper bound (15) can be carried out similarly by modifying the corresponding parts of the proof of Proposition 2.

4.2.2 An abstraction of the example

The instance-dependent bound was useful in this example thanks to the variational characterization of $\theta^\downarrow_m$. More generally, if we need to estimate $Y(\theta)$ given the noisy observation $\widehat{\theta} = \theta + X$, where the function $Y\colon \mathbb{R}^d \to \mathbb{R}$ is the minimum over $S\in\mathcal{S}$ of convex (lower-semicontinuous) functions $y_S(\cdot)$, i.e.,

$$Y(x) = \inf_{S\in\mathcal{S}} y_S(x)\,.$$

Expressing $y_S(\cdot)$ using its convex conjugate $y^*_S(\cdot)$, we have an equivalent definition

$$Y(x) = \inf_{S\in\mathcal{S}}\sup_u \langle u, x\rangle - y^*_S(u)\,.$$

Therefore,

$$Y(\theta) = \inf_{S \in \mathcal{S}} \sup_{u} \langle u, \widehat{\theta} \rangle - \langle u, X \rangle - y_S^*(u) \,,$$

which again is a variational formulation where the linear term $\langle u, X \rangle$ is exposed and can be approximated using instance-dependent tail bounds.

For example, if Θ is a $d_1 \times d_2$ real matrix with $d_2 \geq d_1$, the m-th largest singular value of Θ for $m \leq d_1$, denoted by $\sigma_m(\Theta)$, can be expressed as

$$\sigma_m(\Theta) = \inf_{S \subseteq \mathbb{R}^{d_1} \,:\, \dim(S) = d_1 - m + 1} \; \sup_{U \in \mathbb{R}^{d_1 \times d_2} \,:\, \|U\|_* \leq 1, \mathrm{range}(U) = S} \langle U, \Theta \rangle \,,$$

where the infimum is taken over $d_1 - m + 1$-dimensional subspaces of $\mathbb{R}^{d_1}$, and $\|U\|_*$ and $\mathrm{range}(U)$, respectively, denote the nuclear norm and the range (or column space) of the matrix U.

5 Tail Bounds Without the Exponential Moments

The results of Section 3 rely on the assumption that $\mathcal{F}$, the function class of interest, is a subset of (zero-mean) functions whose exponential moment is finite in a neighborhood of the origin. We may relax this assumption significantly by considering $\mathbb{V}$ to be the vector space of zero-mean functions in $L_1(X)$. Then, using a variational approximation of quantile functions [Pin14, Theorem 2.3] for $\mathbb{E}_n\, g(X)$, we can define the analog of $T_r(\cdot)$ as

$$T^\sharp_{r,n}(g) \stackrel{\mathrm{def}}{=} \inf_{t \in \mathbb{R}} t + e^r\, \mathbb{E}\left((\mathbb{E}_n\, g(X) - t)_+\right) \,, \tag{17}$$

where $(x)_+ = \max(x, 0)$ denotes the positive part of $x \in \mathbb{R}$. Similarly, we can define

$$\overline{T}^\sharp_{r,n}(g) = \max\left\{T^\sharp_{r,n}(g), T^\sharp_{r,n}(-g)\right\} \,,$$

which is a seminorm since it inherits convexity and subadditivity from the corresponding quantile approximation [Pin14, Theorem 2.3]. Equipped with the seminorm $\overline{T}^\sharp_{r,n}(\cdot)$, we can define the analogs of (2) and (3) respectively as

$$\varrho^\sharp r, n(g, \mathcal{H}) = \inf_{h \in \mathcal{H}} \overline{T}^\sharp_{r,n}(g - h) \,,$$

for any $\mathcal{H} \subseteq \mathbb{V}$, and

$$\gamma^\sharp(\mathcal{A};\, r, \underline{\ell}, n) = \inf_{(\mathcal{A}_i)_{i \geq 0}} \sup_{a \in \mathcal{A}} \sum_{\ell \geq \underline{\ell}} \varrho^\sharp_{(r + (r+1)2^{\ell - \underline{\ell}}), n}(a, \mathcal{A}_\ell) \,.$$

where, as in (3), $\mathcal{A} \subseteq \mathbb{V}$, $\underline{\ell} \geq 0$, and the infimum is taken over an increasing admissible sequence $(\mathcal{A}_i)_{i \geq 0}$ of the subsets of $\mathcal{A}$. The corresponding radius of $\mathcal{A}$ is also denoted by

$$\mathrm{rad}^\sharp_{r,n}(\mathcal{A}) = \max_{a \in \mathcal{A}} \overline{T}^\sharp_{r,n}(a) \,.$$

Therefore, we can refine Theorem 1 to the following theorem. We omit the proof as it is effectively the same as the proof of Theorem 1 with $\overline{T}_r(f)$ replaced by $\overline{T}^\sharp_{r,n}(f)$ for every $r \geq 0$ and $f \in \mathbb{V}$ that appear in the proof.

Theorem 2. *Let $A\colon \mathcal{F} \to \mathcal{F}$ be a mapping such that*

$$\overline{T}^{\sharp}_{r+k,n}(A[f]) \le \overline{T}^{\sharp}_{r+k,n}(f)\,, \quad \text{for all } f \in \mathcal{F}\,,$$

and

$$|A[\mathcal{F}]| \le e^k\,,$$

for some nonnegative integer k, where $A[\mathcal{F}] = \{A[f]\colon f \in \mathcal{F}\}$ denotes the range of $A[\cdot]$. Furthermore, let

$$\mathcal{A} = \{f - A[f]\colon f \in \mathcal{F}\}\,.$$

Setting $\underline{\ell} = \lfloor \log_2(r/3) \rfloor$ for $r \ge \log(2)$, with probability at least $1 - 2e^{-r}$, for all $f \in \mathcal{F}$ we have

$$\mathbb{E}_n f(X) - \overline{T}^{\sharp}_{r+k,n}(f) - \min\left\{2\overline{T}^{\sharp}_{2r+1,n}(f - A[f]),\, \mathrm{rad}^{\sharp}_{2r+1,n}(\mathcal{A})\right\} \le 2\gamma^{\sharp}(\mathcal{A};\, r, \underline{\ell}, n)\,.$$

The bound can further be optimized with respect to the mapping $A[\cdot]$, which both k and $\mathcal{A}$ depend on.

While Theorem 2 applies with minimal requirements thanks to the generality of the definition (17), it does not make the dependence on the sample size (i.e., n) transparent. To address this problem, the function class needs to be further restricted, allowing for an approximation of $T_{r,n}(g)$ that reveals the role of n. Results of this type already established in the literature, e.g., in [LT15] and [Men16], and in a specialized form in [MP12], by introducing a more refined "scale-sensitive" version of Talagrand's γ functional, merely assuming that the functions of interest have finite moments of any order. We can reproduce similar bounds from Theorem 2 using the following lemma.

Lemma 3. *Let $g \in \mathbb{V}$ be a zero-mean function with finite p-th moment for some $p \ge 2$. Then, we have*

$$\mathbb{E}\left((\mathbb{E}_n\, g(X) - t)_+\right) \le \left(2\sqrt{\frac{p}{n}}\|g\|_{\psi_{2,p}}\right)^p \frac{t^{-p+1}}{p-1}\,,$$

where[4]

$$\|g\|_{\psi_{2,p}} \stackrel{\text{def}}{=} \sup_{q\in[1,p]} \frac{\|g\|_{L_q}}{\sqrt{q}}\,. \tag{18}$$

Proof. For $t > 0$ we can apply Markov's inequality and Ginè-Zinn symmetrization (see, e.g., [Ver18, Lemma 6.4.2]) to obtain

$$\mathbb{E}\left((\mathbb{E}_n\, g(X) - t)_+\right) = \int_t^{\infty} \mathbb{P}\left(\mathbb{E}_n\, g(X) \ge y\right) \mathrm{d}y$$

[4]The defined norm is denoted by $\|\cdot\|_{(p)}$ in [Men16]. Viewing this norm as an "incomplete" sub-Gaussian norm, we use the more indicative notation $\|\cdot\|_{\psi_{2,p}}$ instead.

$$\leq \int_t^\infty \frac{\|\mathbb{E}_n\, g(X)\|_{L_p}^p}{y^p}\mathrm{d}y$$
$$\leq \left(\frac{2}{n}\right)^p \left\|\sum_{i=1}^n \varepsilon_i g(X_i)\right\|_{L_p}^p \frac{t^{-p+1}}{p-1}\,.$$

where $(\varepsilon_i)_{i\geq 1}$ is a sequence i.i.d. Rademacher random variables (independent of the X_is). Furthermore, the moments of $\sum_{i=1}^n \varepsilon_i g(X_i)$, as a sum of i.i.d. symmetric random variables, can be bounded using a result due to Latała [Lat97, Corollary 2] which yields

$$\left\|\sum_{i=1}^n \varepsilon_i g(X_i)\right\|_{L_p} \leq \sup\left\{\frac{p}{q}\left(\frac{n}{p}\right)^{1/q}\|g\|_{L_q}\colon\ \max(2,p/n)\leq q\leq p\right\}$$
$$\leq \sqrt{pn}\sup_{q\in[1,p]}\frac{\|g\|_{L_q}}{\sqrt{q}}\,.$$

Recalling the definition of $\|g\|_{\psi_{2,p}}$ in (18), the result follows by combining the above inequalities. □

Using Lemma 3 we can bound $T^\sharp_{r,n}(g)$ in terms of $\|g\|_{\psi_{2,p}}$. In particular, evaluating the argument of the infimum on the right-hand side of (17) at $t = 2\sqrt{p/n}e^{r/p}\|g\|_{\psi_{2,p}}$ reveals that

$$T^\sharp_{r,n}(g) \leq \frac{2p}{p-1}\sqrt{\frac{p}{n}}e^{r/p}\|g\|_{\psi_{2,p}}\,.$$

If $g(X)$ has a finite moment of order $p = r \geq 2$, then the above inequality reduces to

$$T^\sharp_{r,n}(g) \leq 4e\|g\|_{\psi_{2,r}}\sqrt{r/n}\,.$$

Therefore, if we further assume that the functions of interest have finite moments of arbitrary order, then

$$\gamma^\sharp(\mathcal{A};r,\underline{\ell},n) \lesssim \inf_{(\mathcal{A}_i)_{i\geq 0}}\sup_{a\in\mathcal{A}}\sum_{\ell\geq\underline{\ell}}\sqrt{\frac{r+(r+1)2^{\ell-\underline{\ell}}}{n}}\|a-\mathcal{A}_\ell\|_{\psi_{2,r+(r+1)2^{\ell-\underline{\ell}}}}\,,$$

where we use the shorthand $\|a-\mathcal{A}_\ell\|_{\psi_{2,r}}$ to denote the distance between $a\in\mathcal{A}$ and the set $\mathcal{A}_\ell$ with respect to $\|\cdot\|_{\psi_{2,r}}$. Choosing $\underline{\ell}$ as prescribed by Theorem 2, we have $r+(r+1)2^{\ell-\underline{\ell}} < 2^{\ell+4}$, thus

$$\gamma^\sharp(\mathcal{A};r,\underline{\ell},n) \lesssim \frac{\gamma^\flat(\mathcal{A};\underline{\ell})}{\sqrt{n}}\,,$$

where

$$\gamma^\flat(\mathcal{A};\underline{\ell}) \stackrel{\text{def}}{=} \inf_{(\mathcal{A}_i)_{i\geq 0}}\sup_{a\in\mathcal{A}}\sum_{\ell\geq\underline{\ell}}2^{\ell/2}\|a-\mathcal{A}_\ell\|_{\psi_{2,2^{\ell+4}}}\,,$$

and we have the following corollary.

Corollary 1. *Let $A\colon \mathcal{F} \to \mathcal{F}$ be a mapping such that*

$$\|A[f]\|_{\psi_{2,r+k}} \le \|f\|_{\psi_{2,r+k}}\,, \quad \text{for all } f \in \mathcal{F}\,,$$

and

$$|A[\mathcal{F}]| \le e^k\,,$$

for some nonnegative integer k, where $A[\mathcal{F}] = \{A[f]\colon f \in \mathcal{F}\}$ denotes the range of $A[\cdot]$. Furthermore, let

$$\mathcal{A} = \{f - A[f]\colon f \in \mathcal{F}\}\,.$$

Setting $\underline{\ell} = \lfloor \log_2(r/3) \rfloor$ for $r \ge 2$, with probability at least $1 - 2e^{-r}$, for all $f \in \mathcal{F}$ we have

$$\mathbb{E}_n f(X) - 4e\sqrt{\frac{r+k}{n}}\|f\|_{\psi_{2,r+k}} - 4e\sqrt{\frac{2r+1}{n}}\|f - A[f]\|_{\psi_{2,2r+1}} \lesssim \frac{\gamma^\flat(\mathcal{A}; \underline{\ell}, n)}{\sqrt{n}}\,.$$

A Remaining Lemmas and Proofs

Proofs of Sections 1 and 2

Proof of Lemma 1. Part (i) of the lemma follows from a straightforward change of variable.

For part (ii), we have

$$\begin{aligned} T_0(g) &= \inf_{\lambda \ge 0} \frac{\log \mathbb{E}\, e^{\lambda g(X)}}{\lambda} \\ &\le \lim_{\lambda \downarrow 0} \frac{\log \mathbb{E}\, e^{\lambda g(X)}}{\lambda} \\ &= \lim_{\lambda \downarrow 0} \frac{\mathbb{E}\left(g(X) e^{\lambda g(X)}\right)}{\mathbb{E}\, e^{\lambda g(X)}} \\ &= 0\,, \end{aligned}$$

where the third and fourth line respectively follow from the l'Hôpital's rule and the assumption that $g(X)$ is zero-mean. However, by Jensen's inequality we have $\mathbb{E}\, e^{\lambda g(X)} \ge e^{\lambda\, \mathbb{E}\, g(X)} = 1$, which means that $T_0(g) \ge 0$. Therefore, we must have

$$T_0(g) = 0\,.$$

For part (iii) observe that $T_r(g)$ can be equivalently expressed as

$$\begin{aligned} T_r(g) &= \inf_{\theta \ge 0} \left(\theta r + \theta \log \mathbb{E}\, e^{g(X)/\theta}\right) \\ &= -\sup_{\theta \ge 0} \left(-\theta r - \theta \log \mathbb{E}\, e^{g(X)/\theta}\right)\,. \end{aligned}$$

Since the supremum is the convex conjugate of $\theta \mapsto \theta \log \mathbb{E}\, e^{g(X)/\theta}$ evaluated at $-r$, we conclude that $r \mapsto T_r(g)$ is concave. The proved concavity together with part (ii) of the lemma, guarantee that for all $r, s \geq 0$ we have

$$\begin{aligned}\frac{r}{r+s}T_{r+s}(g) &= \frac{r}{r+s}T_{r+s}(g) + \frac{s}{r+s}T_0(g) \\ &\leq T_r(g)\,,\end{aligned}$$

and

$$\begin{aligned}\frac{s}{r+s}T_{r+s}(g) &= \frac{s}{r+s}T_{r+s}(g) + \frac{r}{r+s}T_0(g) \\ &\leq T_s(g)\,,\end{aligned}$$

which add up to

$$T_{r+s}(g) \leq T_r(g) + T_s(g)\,,$$

proving the subadditivity of $r \mapsto T_r(g)$.

To prove part (iv) we readily have $\overline{T}_r(0) = 0$, and

$$\begin{aligned}\overline{T}_r(\alpha g) &= \max\{T_r(\alpha g), T_r(-\alpha g)\} \\ &= |\alpha| \overline{T}_r(g)\,,\end{aligned}$$

for every $f \in \mathbb{V}$ and nonzero real number α. Therefore, it suffices to show that $\overline{T}_r(g)$ is convex in $f \in \mathbb{V}$. We show that $T_r(\cdot)$ is convex, which implies the convexity of $\overline{T}_r(\cdot)$. Let us define $\kappa(g) = \log \mathbb{E}\, e^{g(X)}$ for $f \in \mathbb{V}$, and denote by $\mathbb{V}^*$ the dual space of $\mathbb{V}$, i.e., the space of linear functionals on $\mathbb{V}$ that are bounded in the sup norm. It follows from the Hölder's inequality that $\kappa(\cdot)$ is convex. We also define the convex conjugate of $\kappa(\cdot)$ as

$$\kappa^*(w) = \sup_{f \in \mathbb{V}} \langle w, f \rangle - \kappa(g)\,,$$

for every $w \in \mathbb{V}^*$. It can be shown that $\kappa(\cdot)$ is also lower semi-continuous which guarantees $\kappa(g) = \sup_{w \in \mathbb{V}^*} \langle w, g \rangle - \kappa^*(w)$ for all $g \in \mathbb{V}$. Our goal is to show that

$$T_r(g) = \sup_{w \in \mathbb{V}^*\colon \kappa^*(w) \leq r} \langle w, g \rangle\,, \tag{19}$$

which clearly proves the convexity of $T_r(g)$. For $r = 0$ the identity (19) holds trivially as $T_0(g) = 0$ for all $g \in \mathbb{V}$. Then, without loss of generality we may assume that $r > 0$ and write the right-hand side of (19) as

$$\sup_{w \in \mathbb{V}^*\colon \kappa^*(w) \leq r} \langle w, g \rangle = \sup_{w \in \mathbb{V}^*} \inf_{\gamma \geq 0} \langle w, g \rangle - \gamma(\kappa^*(w) - r)\,.$$

Straightforward calculations show that $\kappa^*(0) = 0 < r$. Therefore, the *Slater's condition* is satisfied, and by invoking strong duality we can write

$$\sup_{w \in \mathbb{V}^*\colon \kappa^*(w) \leq r} \langle w, g \rangle = \sup_{w \in \mathbb{V}^*} \inf_{\gamma \geq 0} \langle w, g \rangle - \gamma(\kappa^*(w) - r)$$

$$
\begin{aligned}
&= \inf_{\gamma \geq 0} \sup_{w \in \mathbb{V}^*} \langle w, g \rangle - \gamma(\kappa^*(w) - r) \\
&= \inf_{\gamma \geq 0} \gamma \kappa(\gamma^{-1} g) + \gamma r \\
&= T_r(g)\,,
\end{aligned}
$$

where the last equation follows by the change of variable $\lambda = 1/\gamma$. □

Proof of Lemma 2. By the standard Chernoff bound, for any $T > T_{r/n}(f)$ we have

$$
\mathbb{P}\left(\mathbb{E}_n f(X) > T\right) \leq \inf_{\lambda \geq 0} e^{n \log \mathbb{E}\, e^{\lambda f(X)} - n\lambda T}\,.
$$

It follows from the definition of $T_{r/n}(\cdot)$ that there exists $\lambda' \geq 0$ such that

$$
T_{r/n}(f) \leq \frac{r/n + \log \mathbb{E}\, e^{\lambda' f(X)}}{\lambda'} < T\,.
$$

Therefore, we deduce

$$
\begin{aligned}
\mathbb{P}\left(\mathbb{E}_n f(X) > T\right) &\leq e^{\log \mathbb{E}\, e^{\lambda' f(X)} - \lambda' T} \\
&\leq e^{-r}\,.
\end{aligned}
$$

and consequently

$$
\begin{aligned}
\mathbb{P}\left(\mathbb{E}_n f(X) \leq T_r(f)\right) &= \lim_{T \downarrow T_{r/n}(f)} \mathbb{P}\left(\mathbb{E}_n f(X) \leq T\right) \\
&\geq 1 - e^{-r}\,.
\end{aligned}
$$

□

Proofs of Section 4

Proof of Proposition 1. Let $G \sim \mathrm{Normal}(0, \mathrm{I})$ be a standard normal random vector. Clearly, $\mathbb{E}_n \langle u, X \rangle = \langle \frac{1}{\sqrt{n}} \Sigma^{1/2} u, G \rangle$ in distribution, with $\Sigma^{1/2}$ denoting the symmetric square root of the covariance matrix Σ. Let Σ_k denote the best rank-k approximation of Σ with respect to the operator norm, and let $\pi_k G$ denote the orthogonal projection of G onto the range of Σ_k. We have

$$
\begin{aligned}
&\langle \frac{1}{\sqrt{n}} \Sigma^{1/2} u, G \rangle - \frac{\sqrt{2r} + \sqrt{k}}{\sqrt{n}} (u^\intercal \Sigma u)^{1/2} \\
&\leq \langle \frac{1}{\sqrt{n}} \Sigma^{1/2} u, G - \pi_k G \rangle + \langle \frac{1}{\sqrt{n}} \Sigma^{1/2} u, \pi_k G \rangle - \frac{\sqrt{2r} + \sqrt{k}}{\sqrt{n}} (u^\intercal \Sigma_k u)^{1/2} \\
&\leq \frac{1}{\sqrt{n}} \left\| \Sigma^{1/2} (G - \pi_k G) \right\|_2 + \sqrt{\frac{u^\intercal \Sigma_k u}{n}} \left(\|\pi_k G\|_2 - \sqrt{2r} - \sqrt{k} \right)\,,
\end{aligned}
$$

where the second line follows from the fact that $u^\intercal \Sigma_k u \leq u^\intercal \Sigma u$, and the third line follows from the Cauchy–Schwarz inequality applied to each of the inner products. Using the Gaussian concentration inequality, with probability at least $1 - e^{-r}$ we have

$$
\left\| \Sigma^{1/2} (G - \pi_k G) \right\|_2 \leq \sqrt{\mathrm{tr}(\Sigma - \Sigma_k)} + \sqrt{\|\Sigma - \Sigma_k\|_{\mathrm{op}}}\, \sqrt{2r}\,,
$$

and similarly, with probability at least $1 - e^{-r}$,

$$\|\pi_k G\|_2 \le \sqrt{k} + \sqrt{2r}\,.$$

The upper bound for S_k in (10) follows by combining the three derived inequalities and using the identities $\mathrm{tr}(\Sigma - \Sigma_k) = \sum_{i=k+1}^{d} \lambda_i$ and $\|\Sigma - \Sigma_k\|_{\mathrm{op}} = \lambda_{k+1}$.

To prove the lower bound for S_k, first observe that if $6r + 3(\sqrt{2r} + \sqrt{k})^2 > d$ then (11) holds trivially as its right-hand side vanishes to zero. Therefore, without loss of generality we may assume that $6r + 3(\sqrt{2r} + \sqrt{k})^2 \le d$. We can express S_k by its dual representation as

$$\begin{aligned}
S_k &= \sup_u \inf_x \langle u, X - x\rangle - \frac{\sqrt{2r} + \sqrt{k}}{\sqrt{n}} \left(u^\intercal \Sigma u\right)^{1/2} + \|x\|_2 \\
&= \inf_x \sup_u \langle u, X - x\rangle - \frac{\sqrt{2r} + \sqrt{k}}{\sqrt{n}} \left(u^\intercal \Sigma u\right)^{1/2} + \|x\|_2 \\
&= \inf_{x^\intercal \Sigma^{-1} x \le (\sqrt{2r}+\sqrt{k})^2/n} \|X - x\|_2\,,
\end{aligned}$$

where we used the strong duality on the second line, which holds by the Slater's condition. Using the strong duality again to simplify S_k^2, we have

$$\begin{aligned}
S_k^2 &= \inf_x \sup_{\beta \ge 0} \|X - x\|_2^2 + \beta \left(x^\intercal \Sigma^{-1} x - \frac{(\sqrt{2r} + \sqrt{k})^2}{n} \right) \\
&= \sup_{\beta \ge 0} \inf_x \|X - x\|_2^2 + \beta \left(x^\intercal \Sigma^{-1} x - \frac{(\sqrt{2r} + \sqrt{k})^2}{n} \right) \\
&= \sup_{\beta \ge 0} \left\| \left(I - \left(I + \beta \Sigma^{-1} \right)^{-1} \right) X \right\|_2^2 + \beta \left(\left\| \left(I + \beta \Sigma^{-1} \right)^{-1} X \right\|_{\Sigma^{-1}}^2 - \frac{(\sqrt{2r} + \sqrt{k})^2}{n} \right) \\
&= \sup_{\beta \ge 0} X^\intercal \left(I - \left(I + \beta \Sigma^{-1} \right)^{-1} \right) X - \frac{\beta (\sqrt{2r} + \sqrt{k})^2}{n} \\
&\overset{\text{dist.}}{=} \sup_{\beta \ge 0} \sum_{i=1}^{d} \frac{\beta \lambda_i}{n(\lambda_i + \beta)} g_i^2 - \frac{\beta (\sqrt{2r} + \sqrt{k})^2}{n}\,,
\end{aligned}$$

where g_i's are i.i.d. standard Gaussian random variables. With $a_i = \beta \lambda_i / (n(\beta + \lambda_i))$ for $i \in [d]$, for any fixed $\beta \ge 0$, using the Chernoff bound and the formula for the moment-generating function of g_i^2, with probability at least $1 - e^{-r}$, we have

$$\sum_{i=1}^{d} a_i g_i^2 \ge \sup_{c \ge 0} \frac{-r + \sum_{i=1}^{d} \log(1 + 2 c a_i)/2}{c}\,.$$

Therefore, we can guarantee with the same probability that

$$S_k^2 \ge \sup_{\beta \ge 0} \sup_{c \ge 0} \frac{-r + \sum_{i=1}^{d} \log\left(1 + \frac{2c\beta\lambda_i}{n(\beta + \lambda_i)}\right)/2}{c} - \frac{\beta (\sqrt{2r} + \sqrt{k})^2}{n}\,.$$

Recall that $k' = \lceil 6r + 3(\sqrt{2r} + \sqrt{k})^2 \rceil$. Choosing $\beta = \lambda_{k'}$, and $c = n/(2\beta)$ we have

$$\begin{aligned} S_k^2 &\geq \frac{-r + \sum_{i=1}^d \log(1 + \frac{\lambda_i}{\lambda_{k'} + \lambda_i})/2}{n/(2\lambda_{k'})} - \frac{(\sqrt{2r} + \sqrt{k})^2}{n} \lambda_{k'} \\ &\geq \frac{\lambda_{k'} \sum_{i=1}^d \log(1 + \frac{\lambda_i}{\lambda_{k'} + \lambda_i})}{n} - \frac{2r + (\sqrt{2r} + \sqrt{k})^2}{n} \lambda_{k'} \\ &\geq \frac{\sum_{i=1}^d \lambda_{k'} \lambda_i / (\lambda_{k'} + 2\lambda_i)}{n} - \frac{2r + (\sqrt{2r} + \sqrt{k})^2}{n} \lambda_{k'} \,, \end{aligned}$$

where we used the inequality $\log(1 + x) \geq x/(x + 1)$ for $x \geq 0$ on the last line. Splitting the sum into a sum over $i \leq k'$, and a sum over $i > k'$, we have

$$\sum_{i \leq k'} \lambda_{k'} \lambda_i / (\lambda_{k'} + 2\lambda_i) \geq \frac{k'}{3} \lambda_{k'} \,,$$

and

$$\sum_{i > k'} \lambda_{k'} \lambda_i / (\lambda_{k'} + 2\lambda_i) \geq \frac{1}{3} \sum_{i > k'} \lambda_i \,.$$

Therefore, we have

$$S_k^2 \geq \frac{1}{3n} \sum_{i=k'+1}^d \lambda_i \,.$$

□

Proof of Proposition 2. We express $\theta_m^{\downarrow}$ in an equivalent min-max variational form as

$$\begin{aligned} \theta_m^{\downarrow} &= \min_{S \in \binom{[d]}{d-m+1}} \max_{u \in \triangle^{d-m+1}} \langle u, \theta_S \rangle \\ &= \min_{S \in \binom{[d]}{d-m+1}} \max_{u \in \triangle^{d-m+1}} \langle u, \widehat{\theta}_S \rangle - \langle u, X_S \rangle \,. \end{aligned} \tag{20}$$

For the prescribed nonnegative integer $k \leq r$, let π_S and $\pi_S^{\perp}$, respectively, denote the orthogonal projections onto the range and the nullspace of $\Sigma_{S,k}$. Then, with $G \sim \mathrm{Normal}(0, I)$ we can write

$$\begin{aligned} \langle u, \pi_S X_S \rangle &\overset{\text{dist.}}{=} \langle \Sigma_S^{1/2} \pi_S u, \pi_S G \rangle \\ &\leq \left\| \Sigma_S^{1/2} \pi_S u \right\|_2 \|\pi_S G\|_2 \\ &= \|u\|_{\Sigma_{S,k}} \|\pi_S G\|_2 \\ &\leq \|u\|_{\Sigma_S} \|\pi_S G\|_2 \,. \end{aligned}$$

By the Gaussian concentration inequality, with probability at least $1 - e^{-r}$, we have

$$\|\pi_S G\|_2 \leq \sqrt{2(r + k)} \,,$$

thereby, on the same event, for all $u \in \mathbb{R}^{d-m+1}$ we have

$$\langle u, \pi_S X_S \rangle \le \|u\|_{\Sigma_S} \sqrt{2(r+k)}\,. \tag{21}$$

Furthermore, recalling the definition of $\sigma = \sigma(S,k)$, with probability at least $1-e^{-r}$ we have

$$\begin{aligned}\langle u, \pi_S^\perp X_S \rangle &\le \|u\|_1 \big\|\pi_S^\perp X_S\big\|_\infty \\ &\le \|u\|_1 (\mathbb{E}(\big\|\pi_S^\perp X_S\big\|_\infty) + \sqrt{2r}\|\sigma(S,k)\|_\infty) \\ &\le \|u\|_1 (C\sigma^*(S,k) + \sqrt{2r}\|\sigma(S,k)\|_\infty)\,,\end{aligned}$$

where the second line follows from the Gaussian concentration inequality, and the third line follows from [Tal14, Proposition 2.4.16 and the remarks after Theorem 2.4.18] for some absolute constant $C > 0$. Adding the derived inequalities, with probability at least $1-2e^{-r}$, for all $u \in \mathbb{R}^{d-m+1}$, we can guarantee

$$\begin{aligned}\langle u, X_S \rangle &= \langle u, \pi_S^\perp X_S \rangle + \langle u, \pi_S X_S \rangle \\ &\le \|u\|_1 (C\sigma^*(S,k) + \sqrt{2r}\|\sigma(S,k)\|_\infty) + \|u\|_{\Sigma_S} \sqrt{2(r+k)}\,.\end{aligned} \tag{22}$$

Applying this bound in (20), for any fixed $S \in \binom{[d]}{d-m+1}$, with probability at least $1-2e^{-r}$, we have

$$\max_{u \in \triangle^{d-m+1}} \langle u, \theta_S \rangle \ge \max_{u \in \triangle^{d-m+1}} \langle u, \widehat{\theta}_S \rangle - (C\sigma^*(S,k) + \sqrt{2r}\|\sigma(S,k)\|_\infty) - \|u\|_{\Sigma_S} \sqrt{2(r+k)}\,.$$

To obtain a lower bound for $\theta_m^\downarrow$, we can choose S to be the indices of the $d-m+1$ smallest entries of θ. But to be truly agnostic to the choice of θ, we need to invoke the union bound and minimize the lower bound with respect to $S \in \binom{[d]}{d-m+1}$, at the cost of increasing r by $m + m\log(d/m) > \log \binom{d}{m-1}$. The resulting inequality is then

$$\begin{aligned}\theta_m^\downarrow \ge \min_{S \in \binom{[d]}{d-m+1}} \max_{u \in \triangle^{d-m+1}} \Big(&\langle u, \widehat{\theta}_S \rangle - (C\sigma^*(S,k) + \sqrt{2(r+m+m\log(d/m))}\|\sigma(S,k)\|_\infty) \\ &- \|u\|_{\Sigma_S} \sqrt{2(r+m+m\log(d/m)+k)} \Big)\,,\end{aligned}$$

which, by identifying the expressions of $\beta_{r,m,k}$ and $Q_{S,\beta_{r,m,k}}(\cdot)$, is equivalent to (14). To establish the upper bound (15), observe that $\theta_m^\downarrow = -(-\theta)_{d-m+1}^\downarrow$, which allows us to reuse the inequalities above to derive an upper bounds for $\theta_m^\downarrow$ through the lower bound for $(-\theta)_{d-m+1}^\downarrow$. □

It is worth mentioning that for $k = 0$, the left-hand side of (21) vanishes, thereby we can improve the inequality (22) to

$$\langle u, X_S \rangle \le \|u\|_1 (C\sigma^*(S,0) + \sqrt{2r}\|\sigma(S,0)\|_\infty)\,.$$

Consequently, for $k = 0$ the corresponding bounds are in fact

$$\theta_m^\downarrow \ge \min_{S \in \binom{[d]}{d-m+1}} \Big(\max_{i \in S} \widehat{\theta}_i - C\sigma^*(S,0) - \sqrt{2(r+m+m\log(d/m))}\|\sigma(S,0)\|_\infty \Big)\,,$$

and

$$\theta_m^{\downarrow} \le \max_{S \in \binom{[d]}{m}} \left(\min_{i \in S} \widehat{\theta}_i + C\sigma^*(S,0) + \sqrt{2(r+m+m\log(d/m))}\|\sigma(S,0)\|_\infty \right) .$$

Proof of Proposition 3. Le Cam's two point method [PW24, Theorem 31.1] (see also [Yu97, Lemma 1]) guarantees that

$$\sup_{\theta\in\Theta} \mathbb{E}(g(\widehat{\theta}) - \theta_m^{\downarrow})^2 \ge \sup_{\theta,\eta\in\Theta} \frac{1}{4}(\theta_m^{\downarrow} - \eta_m^{\downarrow})^2 \left(1 - D_{\mathrm{TV}}(\mathbb{P}_\theta, \mathbb{P}_\eta)\right) ,$$

where $D_{\mathrm{TV}}(\cdot,\cdot)$ denotes the total variation distance, and $\mathbb{P}_\theta = \mathrm{Normal}(\theta, \Sigma)$ and $\mathbb{P}_\eta = \mathrm{Normal}(\eta, \Sigma)$. The "simplified" Bretagnolle–Huber inequality [Tsy08, Equation 2.25] (see also [Can23] for a broader context) guarantees that

$$D_{\mathrm{TV}}(\mathbb{P}_\theta, \mathbb{P}_\eta) \le 1 - \frac{1}{2} e^{-(\theta-\eta)^\mathsf{T}\Sigma^{-1}(\theta-\eta)/2} , \tag{23}$$

using which we obtain

$$\begin{aligned}
\sup_{\theta\in\Theta} \mathbb{E}(g(\widehat{\theta}) - \theta_m^{\downarrow})^2 &\ge \sup_{\theta,\eta\in\Theta} \frac{1}{8}(\theta_m^{\downarrow} - \eta_m^{\downarrow})^2 e^{-(\theta-\eta)^\mathsf{T}\Sigma^{-1}(\theta-\eta)/2} \\
&= \sup_{\theta,\eta\in\Theta} \sup_{b\in[0,1]} \frac{1}{8}(\theta_m^{\downarrow} - \eta_m^{\downarrow})^2 b e^{-b(\theta-\eta)^\mathsf{T}\Sigma^{-1}(\theta-\eta)/2} \\
&\ge \frac{1}{8e\max\{1, 2\kappa^2\}} \sup_{\theta,\eta\in\Theta} (\theta_m^{\downarrow} - \eta_m^{\downarrow})^2 ,
\end{aligned}$$

where the second line follows from the fact that Θ is star-shaped, and the third line follows from the inequality $\max_{b\in[0,1]} b e^{-bz} \ge e^{-1}/\max\{1, z\}$ for $z \ge 0$. We can derive (16) from this lower bound using the fact that

$$\begin{aligned}
\sup_{\theta,\eta\in\Theta} (\theta_m^{\downarrow} - \eta_m^{\downarrow})^2 &= \left(\sup_{\theta\in\Theta} \theta_m^{\downarrow} - \inf_{\eta\in\Theta} \eta_m^{\downarrow} \right)^2 \\
&= \left(\sup_{\theta\in\Theta} \theta_m^{\downarrow} + \sup_{\eta\in\Theta} \eta_{d-m+1}^{\downarrow} \right)^2 ,
\end{aligned}$$

where the latter equation follows from the symmetry of the set Θ, and the fact that $-\eta_m^{\downarrow} = (-\eta)_{d-m+1}^{\downarrow}$.

Furthermore, it follows from the definition of the total variation distance and (23) that for any $c \ge 0$ we have

$$\left| \mathbb{P}(|g(\widehat{\theta}) - \theta_m^{\downarrow}| > c) - \mathbb{P}(|g(\widehat{\eta}) - \theta_m^{\downarrow}| > c) \right| \le 1 - \frac{1}{2} e^{-(\theta-\eta)^\mathsf{T}\Sigma^{-1}(\theta-\eta)/2} .$$

In particular, for any $c \le |\theta_m^{\downarrow} - \eta_m^{\downarrow}|/2$, together with the inequality

$$\mathbb{P}(|g(\widehat{\eta}) - \theta_m^{\downarrow}| > \frac{1}{2}|\theta_m^{\downarrow} - \eta_m^{\downarrow}|) \ge \mathbb{P}(|g(\widehat{\eta}) - \eta_m^{\downarrow}| < \frac{1}{2}|\theta_m^{\downarrow} - \eta_m^{\downarrow}|) ,$$

which follows from the triangle inequality, we obtain

$$\mathbb{P}(|g(\widehat{\theta}) - \theta_m^{\downarrow}| \geq c) + \mathbb{P}(|g(\widehat{\eta}) - \eta_m^{\downarrow}| \geq c) \geq \frac{1}{2} e^{-(\theta-\eta)^{\mathsf{T}}\Sigma^{-1}(\theta-\eta)/2}\,.$$

Therefore, if there exists a pair $\theta, \eta \in \Theta$ such that $(\theta-\eta)^{\mathsf{T}}\Sigma^{-1}(\theta-\eta) \leq 2$ and $|\theta_m^{\downarrow} - \eta_m^{\downarrow}|/2 \geq c$, then

$$\sup_{\theta\in\Theta} \mathbb{P}(|g(\widehat{\theta}) - \theta_m^{\downarrow}| \geq c) \geq \frac{1}{2e}\,.$$

The desired result follows by setting $c = \sup_{\theta,\eta\in\Theta} |\theta_m^{\downarrow} - \eta_m^{\downarrow}|/(3\max\{1, \sqrt{2}\kappa\})$ which meets the required conditions. □

Lemma 4. *Let $Y \in [-1, 1]$ be a zero-mean random variable. Then, we have*

$$\inf_{\lambda\geq 0} \frac{r + \log \mathbb{E}\, e^{\lambda Y}}{\lambda} \leq \frac{1}{3} r + \sqrt{2\,\mathbb{E}(Y^2) r}\,.$$

Proof. For all $\lambda \in [-3, 3]$, we have

$$\begin{aligned}
\mathbb{E}\, e^{\lambda Y} &= 1 + \sum_{m=2}^{\infty} \frac{\mathbb{E}\, Y^m}{m!} \lambda^m \\
&\leq 1 + \sum_{m=2}^{\infty} \frac{\mathbb{E}\, Y^2}{m!} |\lambda|^m \\
&\leq 1 + \frac{\mathbb{E}\, Y^2}{2} \sum_{m=2}^{\infty} \lambda^2 \left(\frac{|\lambda|}{3}\right)^{m-2} \\
&\leq 1 + \frac{\mathbb{E}\, Y^2}{2} \cdot \frac{\lambda^2}{1 - |\lambda|/3}\,.
\end{aligned}$$

Since $\log(1 + u) \leq u$ for all $u > -1$, it follows that

$$\begin{aligned}
\inf_{\lambda\geq 0} \frac{r + \log \mathbb{E}\, e^{\lambda Y}}{\lambda} &\leq \inf_{\lambda\in[0,3]} \frac{r + \mathbb{E}(Y^2)\lambda^2/(2 - 2\lambda/3)}{\lambda} \\
&\leq \frac{1}{3} r + \sqrt{2\,\mathbb{E}(Y^2) r}\,,
\end{aligned}$$

where the second line follows by evaluating the argument of the infimum at $\lambda = 3\sqrt{r}/(\sqrt{r} + 3\sqrt{\mathbb{E}(Y^2)/2})$. □

B Bounding $T_r(f)$ in Orlicz Spaces

The purpose of this subsection is to approximate $T_r(f)$ for $f \in \mathbb{V}$, in situations where $\mathbb{V}$ is an *Orlicz space* of *exponential type*. Orlicz spaces are one of the important function spaces studied in functional analysis and probability theory. These function spaces can be described

by their corresponding *Orlicz norms*. For a convex increasing function $\psi\colon [0,\infty) \to [0,\infty)$ with $\psi(0) = 0$ the ψ-Orlicz norm of a random variable Y is defined as

$$\|Y\|_\psi \stackrel{\text{def}}{=} \inf\left\{u > 0\colon \mathbb{E}\,\psi\left(\frac{|Y|}{u}\right) \le 1\right\}.$$

Special cases are the usual p-norms for $p \ge 1$, the sub-Gaussian norm, and the sub-exponential norm, respectively, corresponding to $\psi(t) = t^p$, $\psi(t) = e^{t^2} - 1$, and $\psi(t) = e^t - 1$. Other interesting cases are the Bernstein–Orlicz norm corresponding to

$$\psi(t) = e^{\left(\sqrt{1+2Lt}-1\right)^2/L^2} - 1\,,$$

for some parameter $L > 0$, introduced by van de Geer and Lederer [vdGL12], as well as the Bennett–Orlicz norm corresponding to

$$\psi(t) = e^{2((1+Lt)\log(1+Lt)-Lt)/L^2} - 1\,,$$

for some parameter $L > 0$, introduced by Wellner [Wel17].

To express the general bounds presented in Theorem 1 when the underlying metric of interest imposed on $\mathcal{F}$ is induced by an Orlicz ψ-norm, it suffices to bound $T_r(f)$ in terms of $\|f\|_\psi \stackrel{\text{def}}{=} \|f(X)\|_\psi$. The following simple lemma can provide such bounds.

Lemma 5. *For every $f \in \mathbb{V}$ we have*

$$T_r(f) \le \inf_{\lambda \ge 0} \frac{r + \log\left(1 + \int_0^\infty 2\lambda\left(e^{\lambda t} - 1\right)/\left(\psi(t) + 1\right)\mathrm{d}t\right)}{\lambda}\,\|f\|_\psi\,. \tag{24}$$

Proof. Without loss of generality we may assume $f \neq 0$. The inequality follows by bounding the moment generating function of the zero-mean random variable $Y = f(X)/\|f(X)\|_\psi$, which has a unit ψ-Orlicz norm, as

$$\begin{aligned}
\mathbb{E}\,e^{\lambda Y} &= \mathbb{E}\left(e^{\lambda Y} - \lambda Y\right)\\
&\le \mathbb{E}\left(e^{\lambda|Y|} - \lambda|Y|\right)\\
&= 1 + \int_0^\infty \lambda\,\mathbb{P}\left(|Y| > t\right)\left(e^{\lambda t} - 1\right)\mathrm{d}t\\
&\le 1 + \int_0^\infty \lambda\left(\mathbb{E}\,\psi(|Y|) + 1\right)\left(e^{\lambda t} - 1\right)/\left(\psi(t) + 1\right)\mathrm{d}t\\
&= 1 + \int_0^\infty 2\lambda\left(e^{\lambda t} - 1\right)/\left(\psi(t) + 1\right)\mathrm{d}t\,.
\end{aligned}$$

□

For exponential type Orlicz norms, defined below, we have the following proposition that provides a more explicit approximation for $T_r(f)$ in terms of $\|f\|_\psi$.

Proposition 4. *Let $\|\cdot\|_\psi$ be an Orlicz norm of exponential type, meaning that*

$$\psi(t) = e^{\phi(t)} - 1$$

for a convex and increasing function $\phi\colon [0,\infty) \to [0,\infty)$ *with* $\phi(0) = 0$. *Furthermore, let* $\phi^*(\cdot)$ *denote the convex conjugate of* $\phi(\cdot)$, *i.e.,*

$$\phi^*(\lambda) = \sup_{t\geq 0} \left(\lambda t - \phi(t)\right) .$$

If for some $M > 0$ *we have*

$$\inf_{\lambda\geq 0} \frac{e^{\phi^*(\lambda)} - 1}{\lambda^2} \geq M \int_0^\infty t e^{-\phi(t)/2} \mathrm{d}t \,, \tag{25}$$

then for every $f \in \mathbb{V}$ *we have*

$$T_r(f) \leq \max\{3, 3/\sqrt{2M}\} \phi^{-1}(2r/3) \left\|f\right\|_\psi . \tag{26}$$

Proof. Since $\phi^*(\cdot)$ is the convex conjugate of $\phi(\cdot)$, for every $\lambda, t \geq 0$ we can write

$$\lambda t \leq \frac{1}{2}\phi(t) + \frac{1}{2}\phi^*(2\lambda) .$$

Applying this bound to (24) of Lemma 5 we have

$$\begin{aligned} T_r(f) &\leq \inf_{\lambda\geq 0} \frac{r + \log\left(1 + \int_0^\infty 2\lambda \left(1 - e^{-\lambda t}\right) e^{\lambda t - \phi(t)} \mathrm{d}t\right)}{\lambda} \left\|f\right\|_\psi \\ &\leq \inf_{\lambda\geq 0} \frac{r + \log\left(1 + 2\lambda^2 e^{\phi^*(2\lambda)/2} \int_0^\infty t e^{-\phi(t)/2} \mathrm{d}t\right)}{\lambda} \left\|f\right\|_\psi , \end{aligned}$$

where we also used the inequality $1 - e^{-\lambda t} \leq \lambda t$. It follows from (25) that

$$2\lambda^2 \int_0^\infty t e^{-\phi(t)/2} \mathrm{d}t \leq e^{\phi^*(\sqrt{2/M}\lambda)} - 1 .$$

Then, using the fact that $\phi^*(\cdot)$ is nonnegative, we have

$$T_r(f) \leq \inf_{\lambda\geq 0} \frac{r + \phi^*(2\lambda)/2 + \phi^*(\sqrt{2/M}\lambda)}{\lambda} \left\|f\right\|_\psi .$$

Furthermore, because $\phi^*(\cdot)$ is increasing, we can write

$$\phi^*(2\lambda)/2 + \phi^*(\sqrt{2/M}\lambda) \leq \frac{3}{2}\phi^*\left(\max\{2, \sqrt{2/M}\}\lambda\right) .$$

Therefore, we conclude that

$$\begin{aligned} T_r(f) &\leq \inf_{\lambda\geq 0} \frac{r + \frac{3}{2}\phi^*\left(\max\{2, \sqrt{2/M}\}\lambda\right)}{\lambda} \left\|f\right\|_\psi \\ &= \max\{3, 3/\sqrt{2M}\} \phi^{-1}(2r/3) \left\|f\right\|_\psi . \end{aligned}$$

□

It is worth mentioning that the constants appearing in the proposition are not necessarily optimal. In fact, the result may be improved for example by using the bound $1 - e^{-\lambda t} \leq \min\{\lambda t, 1\}$ instead of the inequality $1 - e^{-\lambda t} \leq \lambda t$ that is used in the current proof. We did not pursue these refinements intending to obtain relatively simpler expressions.

Let us quantify the result of Proposition 4 when $\|\cdot\|_\psi$ is the sub-Gaussian Orlicz norm, and when it is the Bernstein–Orlicz norm. In the sub-Gaussian case, we have $\phi(t) = t^2$ and $\phi^*(\lambda) = \mathbb{1}(\lambda \geq 0)\lambda^2/4$. It is easy to verify that (25) holds with $M = 1/4$. Therefore, for the sub-Gaussian Orlicz norm, (26) reduces to

$$T_r(f) \leq \sqrt{12r}\, \|f\|_\psi \,.$$

In the case of Bernstein–Orlicz norm, $\phi(t) = (\sqrt{1+2Lt} - 1)^2/L^2$. By the change of variable $t = ((Lu+1)^2 - 1)/(2L)$ and using standard Gaussian integral formulas we can calculate the integral on the right-hand side of (25) as

$$\int_0^\infty t e^{-\left(\sqrt{1+2Lt}-1\right)^2/(2L^2)} \mathrm{d}t = \sqrt{\frac{\pi}{8}} L + 1 \,.$$

Furthermore, with some straightforward calculations we can show that the convex conjugate of $\phi(\cdot)$ is

$$\phi^*(\lambda) = \begin{cases} 0 \,, & \lambda < 0 \,, \\ \frac{\lambda^2}{4(1-L\lambda/2)} \,, & \lambda \in [0, 2/L) \,, \\ \infty \,, & \lambda > 2/L \,. \end{cases}$$

Therefore, for $\lambda \geq 0$, we have

$$e^{\phi^*(\lambda)} - 1 \geq \phi^*(\lambda) \geq \frac{\lambda^2}{4} \,.$$

Consequently, (25) holds if

$$M = \frac{1}{\sqrt{2\pi} L + 4} \,,$$

for which (26) reduces to

$$\begin{aligned} T_r(f) &\leq 3(\sqrt{\pi/2} L + 2)^{1/2} \phi^{-1}(2r/3)\, \|f\|_\psi \\ &= (\sqrt{\pi/2} L + 2)^{1/2} \left(Lr + \sqrt{6r} \right) \|f\|_\psi \,. \end{aligned}$$

References

[AB07] J.-Y. Audibert and O. Bousquet. "Combining PAC-Bayesian and generic chaining bounds". *Journal of Machine Learning Research* 8 (2007), pp. 863–889 (cit. on p. 7).

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford: Oxford University Press, Feb. 2013 (cit. on p. 10).

[Can23] C. L. Canonne. *A short note on an inequality between KL and TV*. 2023. arXiv: 2202.07198 [math.PR] (cit. on p. 28).

[DZ10] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*. Springer Berlin Heidelberg, 2010 (cit. on p. 3).

[Dir15] S. Dirksen. "Tail bounds via generic chaining". *Electronic Journal of Probability* 20 (2015) (cit. on pp. 4, 5).

[Fer75] X. Fernique. "Regularite des trajectoires des fonctions aleatoires Gaussiennes" (1975), pp. 1–96 (cit. on p. 5).

[GK06] E. Giné and V. Koltchinskii. "Concentration inequalities and asymptotic results for ratio type empirical processes". *The Annals of Probability* 34: 3 (May 2006), pp. 1143–1216 (cit. on p. 2).

[GKW03] E. Giné, V. Koltchinskii, and J. A. Wellner. "Ratio limit theorems for empirical processes". *Stochastic Inequalities and Applications*. 2003, pp. 249–278 (cit. on p. 2).

[KP00] V. Koltchinskii and D. Panchenko. "Rademacher processes and bounding the risk of function learning". *High Dimensional Probability II*. 2000, pp. 443–457 (cit. on p. 1).

[Lat97] R. Latała. "Estimation of moments of sums of independent real random variables". *The Annals of Probability* 25: 3 (1997), pp. 1502–1513 (cit. on p. 21).

[LT15] R. Latała and T. Tkocz. "A note on suprema of canonical processes based on random variables with regular moments". *Electronic Journal of Probability* 20 (2015), pp. 1–17 (cit. on p. 20).

[LM23] G. Lugosi and S. Mendelson. "Multivariate mean estimation with direction-dependent accuracy". *Journal of the European Mathematical Society* (Jan. 2023) (cit. on pp. 2, 9–11).

[Mar21] A. Marchina. "Concentration inequalities for suprema of unbounded empirical processes". *Annales Henri Lebesgue* 4 (Aug. 2021), pp. 831–861 (cit. on p. 3).

[Men16] S. Mendelson. "Upper bounds on product and multiplier empirical processes". *Stochastic Processes and their Applications* 126: 12 (2016). In Memoriam: Evarist Giné, pp. 3652–3680 (cit. on p. 20).

[MP12] S. Mendelson and G. Paouris. "On generic chaining and the smallest singular value of random matrices with heavy tails". *Journal of Functional Analysis* 262: 9 (2012), pp. 3775–3811 (cit. on p. 20).

[Pin14] I. Pinelis. "An optimal three-way stable and monotonic spectrum of bounds on quantiles: A spectrum of coherent measures of financial risk and economic inequality". *Risks* 2: 3 (2014), pp. 349–392 (cit. on pp. 3, 19).

[Pol84] D. Pollard. *Convergence of stochastic processes*. Springer New York, 1984 (cit. on p. 2).

[PW24] Y. Polyanskiy and Y. Wu. *Information theory: from coding to learning*. Cambridge University Press, 2024 (cit. on p. 28).

[Rio17] E. Rio. "About the constants in the Fuk-Nagaev inequalities". *Electronic Communications in Probability* 22 (2017), pp. 1–12 (cit. on p. 3).

[Tal87] M. Talagrand. "Regularity of Gaussian processes". *Acta Mathematica* 159: 0 (1987), pp. 99–149 (cit. on p. 5).

[Tal01] M. Talagrand. "Majorizing measures without measures". *The Annals of Probability* 29: 1 (Feb. 2001) (cit. on p. 5).

[Tal14] M. Talagrand. *Upper and lower bounds for stochastic processes*. Springer Berlin Heidelberg, 2014 (cit. on pp. 1, 4, 12, 13, 17, 27).

[Tsy08] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer-Verlag GmbH, Oct. 22, 2008 (cit. on p. 28).

[vdGL12] S. van de Geer and J. Lederer. "The Bernstein–Orlicz norm and deviation inequalities". *Probability Theory and Related Fields* 157: 1-2 (Oct. 2012), pp. 225–250 (cit. on p. 30).

[vdVW12] A. van der Vaart and J. Wellner. *Weak convergence and empirical processes*. Springer New York, 2012 (cit. on p. 1).

[vHan18] R. van Handel. "Chaining, interpolation, and convexity". *Journal of the European Mathematical Society* 20: 10 (July 2018), pp. 2413–2435 (cit. on pp. 5, 13).

[VC71] V. N. Vapnik and A. Y. Chervonenkis. "On the uniform convergence of relative frequencies of events to their probabilities". *Theory of Probability & Its Applications* 16: 2 (Jan. 1971), pp. 264–280 (cit. on p. 1).

[Vap98] V. Vapnik. *Statistical learning theory*. New York: Wiley, 1998 (cit. on p. 1).

[Var84] S. R. S. Varadhan. *Large deviations and applications*. Society for Industrial and Applied Mathematics, Jan. 1984 (cit. on p. 3).

[Ver18] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018 (cit. on pp. 5, 20).

[Wel17] J. A. Wellner. "The Bennett–Orlicz norm". *Sankhya A* 79: 2 (May 2017), pp. 355–383 (cit. on p. 30).

[Yu97] B. Yu. "Assouad, Fano, and Le Cam". *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*. New York, NY, 1997, pp. 423–435 (cit. on p. 28).