Minimax Linear Regression under the Quantile Risk

Ayoub El Hanchi   AELHAN@CS.TORONTO.EDU
Chris J. Maddison   CMADDIS@CS.TORONTO.EDU
Murat A. Erdogdu   ERDOGDU@CS.TORONTO.EDU
University of Toronto & Vector Institute

Abstract

We study the problem of designing minimax procedures in linear regression under the quantile risk. We start by considering the realizable setting with independent Gaussian noise, where for any given noise level and distribution of inputs, we obtain the *exact* minimax quantile risk for a rich family of error functions and establish the minimaxity of OLS. This improves on the lower bounds obtained by Lecué and Mendelson (2016) and Mendelson (2017) for the special case of square error, and provides us with a lower bound on the minimax quantile risk over larger sets of distributions.

Under the square error and a fourth moment assumption on the distribution of inputs, we show that this lower bound is tight over a larger class of problems. Specifically, we prove a matching upper bound on the worst-case quantile risk of a variant of the procedure proposed by Lecué and Lerasle (2020), thereby establishing its minimaxity, up to absolute constants. We illustrate the usefulness of our approach by extending this result to all p-th power error functions for $p \in (2, \infty)$.

Along the way, we develop a generic analogue to the classical Bayesian method for lower bounding the minimax risk when working with the quantile risk, as well as a tight characterization of the quantiles of the smallest eigenvalue of the sample covariance matrix.

Keywords: minimax procedures, linear regression, sample covariance matrix, quantile risk.

1. Introduction

We study the problem of designing minimax procedures in linear regression under the quantile risk over large classes of distributions. Specifically, for some $d \in \mathbb{N}$, there is an input random vector $X \in \mathbb{R}^d$ and an output random variable $Y \in \mathbb{R}$, and we are provided with $n \in \mathbb{N}$ i.i.d. samples $(X_i, Y_i)_{i=1}^n$ from their joint distribution P, with the goal of constructing a predictor of Y given X. We consider the set of linear predictors $\{x \mapsto \langle w, x \rangle \mid w \in \mathbb{R}^d\}$, and measure the error of a predictor $w \in \mathbb{R}^d$ on an input/output pair (X, Y) through $e(\langle w, X \rangle - Y)$ for an error function of our choice $e : \mathbb{R} \to \mathbb{R}$. We evaluate the overall error of a predictor $w \in \mathbb{R}^d$ through the expected error $E(w) := \mathrm{E}[e(\langle w, X \rangle - Y)]$, and define $\mathcal{E}(w) := E(w) - \inf_{v \in \mathbb{R}^d} E(v)$.

For a user-chosen failure probability $\delta \in (0, 1)$, we evaluate the performance of a procedure $\hat{w}_{n,\delta} : (\mathbb{R}^d \times \mathbb{R})^n \to \mathbb{R}^d$ on a particular distribution P through its quantile risk

$$R_{n,\delta}(P, \hat{w}_{n,\delta}) := Q_{\mathcal{E}(\hat{w}_{n,\delta})}(1 - \delta) = \inf\{t \geq 0 \mid \mathrm{P}(\mathcal{E}(\hat{w}_{n,\delta}) \leq t) \geq 1 - \delta\}, \tag{1}$$

where we shortened $\hat{w}_{n,\delta}((X_i, Y_i)_{i=1}^n)$ to $\hat{w}_{n,\delta}$. We consider the scenario where all that is known about P is that it belongs to a class of distributions $\mathcal{P}$ on $\mathbb{R}^d \times \mathbb{R}$. This justifies evaluating the overall performance of a procedure through its worst-case risk

$$R_{n,\delta}(\mathcal{P}, \hat{w}_{n,\delta}) := \sup_{P \in \mathcal{P}} R_{n,\delta}(P, \hat{w}_{n,\delta}).$$

Our goal is to characterize the minimax risk $R_{n,\delta}^*(\mathcal{P}) := \inf_{\hat{w}_{n,\delta}} R_{n,\delta}(\mathcal{P}, \hat{w}_{n,\delta})$ and design minimax procedures for rich classes of distributions and error functions.

Note on terminology. In this paper, we reserve the terms 'risk' and 'loss' to refer to the corresponding decision-theoretic concepts, see e.g. Lehmann and Casella (2006) for background on these notions. To avoid any confusion, we have used the terms 'error' and 'expected error' to refer to what is commonly called 'prediction loss' and 'prediction risk' in statistical learning theory.

Motivation. Our motivation for studying this problem has its roots in the work of Catoni (2012), who showed that the empirical mean is no longer minimax over the set of all distributions with finite variance under the square loss if one replaces the classical notion of risk, given by the expected loss, with the quantile risk, given by the $1 - \delta$ quantile of the loss, for any user-chosen failure probability $\delta \in (0, 1)$. Since then, minimax procedures were discovered for this problem (Devroye et al., 2016; Lee and Valiant, 2022), and there has been a lot of effort to construct minimax procedures under this new notion of risk for a variety of statistical problems (Lugosi and Mendelson, 2019a,b; Mendelson and Zhivotovskiy, 2020). We view our work as part of this larger effort.

Known results. To understand why previous results are insufficient to accomplish our stated goal, let us briefly review the most relevant ones. Most previous work has focused on the case of square error $e(t) = t^2/2$ (Audibert and Catoni, 2011; Hsu and Sabato, 2016; Lugosi and Mendelson, 2019a; Lecué and Lerasle, 2020). In this case, a natural class of distributions to consider is

$$\mathcal{P}_2(P_X, \sigma^2) := \{P \mid (X, Y) \sim P : X \sim P_X \text{ and } \operatorname{ess\,sup}(\mathrm{E}[\xi^2 \mid X]) \leq \sigma^2\}, \qquad (2)$$

for a given distribution P_X of inputs, noise level $\sigma^2 \in (0, \infty)$, and where $\xi := Y - \langle w^*, X \rangle$ is the noise and w^* is the unique minimizer of the expected error $E(w)$ under square error. The best lower bound on the minimax risk over this class has been obtained by considering the subclass

$$\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2) := \{P \mid (X, Y) \sim P : (X, \eta) \sim P_X \times \mathcal{N}(0, \sigma^2), Y = \langle w^*, X \rangle + \eta \text{ for } w^* \in \mathbb{R}^d\}.$$

The following results yield the best upper and lower bounds on the minimax risk over $\mathcal{P}_2(P_X, \sigma^2)$.

Proposition 1 (Lecué and Mendelson (2016); Mendelson (2017)) *Suppose that $e(t) = t^2/2$. There exist absolute constants $C, c > 0$ such that for all $\delta \in (0, 1/8)$, it holds that*

$$R_{n,\delta}^*(\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)) \geq \begin{cases} \infty & \text{if } n \leq d/C, \\ c \cdot \dfrac{\sigma^2(d + \log(1/\delta))}{n} & \text{otherwise.} \end{cases}$$

Proposition 2 (Oliveira and Resende (2023)) *Suppose that $e(t) = t^2/2$. There exists a procedure $\hat{w}_{n,\delta}$ and absolute constants $C, c > 0$ such that the following holds. If*

$$n \geq c \cdot \theta^2(P_X)(d + \log(1/\delta)), \quad \text{where} \quad \theta(P_X) := \sup_{w \in \mathbb{R}^d \setminus \{0\}} \frac{\mathrm{E}\big[\langle w, X \rangle^2\big]^{1/2}}{\mathrm{E}[|\langle w, X \rangle|]},$$

then

$$R_{n,\delta}(\mathcal{P}_2(P_X, \sigma^2), \hat{w}_{n,\delta}) \leq C \cdot \theta^2(P_X) \cdot \frac{\sigma^2 \cdot (d + \log(1/\delta))}{n}.$$

In the prescribed regime (n, δ) stated in Proposition 2, and on the set of distributions for which $\theta(P_X)$ is upper bounded by an absolute constant, the combination of Propositions 1 and 2 proves the minimaxity, up to an absolute constant, of the procedure in Proposition 2 over $\mathcal{P}_2(P_X, \sigma^2)$.

Unfortunately, this minimaxity result is unsatisfactory for two important reasons. First, the set of distributions for which $\theta(P_X)$ is bounded by an absolute constant is both difficult to characterize and too small to cover classes of problems of interest. Indeed, by using the relationship between $\theta(P_X)$ and the small-ball constants (Lecué and Lerasle, 2019), and using the lower bounds derived on the latter by Saumard (2018), it is possible to derive dimension-dependent lower bounds on $\theta(P_X)$ for standard linear regression problems with bounded inputs. Second, this minimaxity result is specific to the square error function. While procedures with guarantees have been studied for other error functions (Chinot et al., 2020), no lower bounds are known outside of Proposition 1.

Contributions.   Below we summarize our main results related to linear regression.

- We compute the *exact* minimax quantile risk over the class $\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)$ for a rich set of error functions, and show that OLS is minimax in this setting (Theorem 3). We deduce from this result the asymptotic minimax quantile risk over this class (Proposition 4).
- Focusing on the non-asymptotic setting with $e(t) = t^2/2$, we complement our exact computation with tight upper and lower bounds (Proposition 7). We then recover the lower bound of Proposition 1 and identify a setting in which it is tight (Corollary 9). We give an analogous result under the error function $e(t) = |t|^p/[p(p-1)]$ for $p \in (2, \infty)$ (Proposition 10).
- We then turn to finding minimax procedures on larger classes of distributions. For the square error, we establish the minimaxity, up to an absolute constant, of a variant of the min-max regression procedure (Audibert and Catoni, 2011; Lecué and Lerasle, 2020) over the class $\mathcal{P}_2(P_X, \sigma^2)$, and under a fourth moment assumption on P_X (Theorem 11).
- Finally, we study minimax linear regression under the error function $e(t) = |t|^p/[p(p-1)]$ for $p \in (2, \infty)$. Guided by our results, we identify a rich class of distributions analogous to $\mathcal{P}_2(P_X, \sigma^2)$, and show that the min-max regression procedure is minimax, up to a constant that depends only on p, and under a fourth moment assumption on P_X (Theorem 12).

Our contributions on linear regression are supported by the following more general results.

- We consider the quantile risk in full generality. We develop an analogue to the Bayesian method for lower bounding the minimax quantile risk (Theorem 13). We then prove that the minimaxity of procedures under the quantile risk is invariant to strictly increasing left-continuous transformations of the loss (Proposition 14).
- We illustrate the generality of our methods by applying them to two unrelated problems: multivariate mean estimation with Gaussian data, in which we recover a strengthening of the recent result of Depersin and Lecué (2022) (Proposition 15), and variance estimation with Gaussian data and known mean, where we show that, surprisingly, the sample variance is suboptimal, and design a new minimax estimator (Proposition 16).
- We conclude by studying the smallest eigenvalue of the sample covariance matrix. We prove a new tight asymptotic lower bound on its quantiles, and a nearly matching fully non-asymptotic upper bound (Proposition 17), both under a fourth moment assumption on P_X.

Organization   The rest of the paper is organized as follows. In Section 2, we present our results on the minimax quantile risk over the class $\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)$. In Section 3, we present new upper bounds on the worst-case quantile risk of the min-max regression procedure for the error functions $e(t) = |t|^p/[p(p-1)]$ for $p \in [2, \infty)$, showing its minimaxity over suitable classes of distributions up to constants. In Section 4 we study the quantile risk in full generality. Finally, in Section 5, we present our results on the smallest eigenvalue of the sample covariance matrix.

Notation. We call a function $f : \mathbb{R} \to \mathbb{R}$ increasing if $x \leq x'$ implies $f(x) \leq f(x')$. If $f : \mathbb{R} \to \mathbb{R}$ is an increasing function, we define its pseudo-inverse $f^- : [-\infty, \infty] \to [-\infty, \infty]$ by $f^-(y) := \inf\{x \in \mathbb{R} \mid f(x) \geq y\}$. For a random variable $X : \Omega \to \mathbb{R}$, we denote its CDF by F_X and its quantile function by $Q_X := F_X^-$. We allow random variables of the form $X : \Omega \to [0, \infty]$, but we restrict the definition of their CDFs to $[0, \infty)$. Without loss of generality, we assume throughout that the support of the distribution of inputs P_X is not contained in any hyperplane. We write $\Sigma = \mathrm{E}\big[XX^T\big]$ for the covariance matrix of the random vector X. We write $a \asymp b$ to mean that there exist absolute constants $C, c > 0$ such that $c \cdot b \leq a \leq C \cdot b$.

2. Minimax quantile risk over $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$

The following is the main result of this section.

Theorem 3 *Let P_X be a distribution on $\mathbb{R}^d$ and $\sigma^2 \in (0, \infty)$. Assume that e is strictly convex, differentiable, and symmetric i.e. $e(t) = e(-t)$ for all $t \in \mathbb{R}$. Define, for $(X, \eta) \sim P_X \times \mathcal{N}(0, \sigma^2)$,*

$$\widetilde{E}(\Delta) := \mathrm{E}[e(\langle \Delta, X \rangle + \eta)], \qquad \widetilde{\mathcal{E}}(\Delta) := \widetilde{E}(\Delta) - \widetilde{E}(0).$$

If P_X is such that $\widetilde{E}$ is finite everywhere and differentiable at 0 with $\nabla \widetilde{E}(0) = \mathrm{E}[\nabla e(\eta)]$, then

$$R_{n,\delta}^*(\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)) = Q_{\widetilde{\mathcal{E}}(Z)}(1 - \delta),$$

where the random variable Z is jointly distributed with $(X_i)_{i=1}^n \sim P_X^n$ such that $Z \mid (X_i)_{i=1}^n \sim \mathcal{N}(0, \frac{\sigma^2}{n}\widehat{\Sigma}_n^{-1})$ on the event that the sample covariance matrix $\widehat{\Sigma}_n := n^{-1}\sum_{i=1}^n X_i X_i^T$ is invertible, otherwise $\widetilde{\mathcal{E}}(Z) := \infty$. Moreover, all procedures satisfying the following condition are minimax

$$\hat{w}_{n,\delta}((X_i, Y_i)_{i=1}^n) \in \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n (\langle w, X_i \rangle - Y_i)^2.$$

We make a few remarks about this result before interpreting its content. First, Theorem 3 improves on the best known comparable result, Proposition 1, in two distinct ways: it provides the *exact* minimax risk over the class $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$ for the error function $e(t) = t^2/2$, and it generalizes this result to a rich collection of error functions. Second, and as can readily be seen from the proof, the strict convexity hypothesis on $e(t)$ in Theorem 3 can be weakened to the strict quasiconvexity of $E(w)$, and the strictness can be replaced by the existence of a unique minimizer of $E(w)$. Finally, the proof of Theorem 3 is based on the Bayesian method we develop in Theorem 13, an adaptation of an argument of Mourtada (2022), and Anderson's Lemma (e.g. Keener, 2010).

While exact, the result in Theorem 3 is both difficult to interpret and hard to manipulate. In particular, the dependence of the minimax risk on the problem parameters $(n, \delta, P_X, \sigma^2)$ as well as the error function e remains implicit in Theorem 3. This is not too surprising as the error function can interact with the parameters of the problems in quite complicated ways.

In the rest of this section, we develop tools to make these dependencies explicit. Specifically, in Section 2.1, we compute the asymptotic minimax risk for generic error functions as $n \to \infty$ and show that it takes on a simple form. In Section 2.2, we focus on the case of square error function, and identify a setting where the lower bound of Proposition 1 is tight. In Section 2.3, we extend this result to the case of the p-th power error function for $p \in (2, \infty)$.

2.1. General error functions

The following result shows that under a mild assumption on the error function, the asymptotic minimax risk is a pleasingly simple function of the parameters of the problem. In particular, this result shows that the lower bound of Proposition 1 is asymptotically tight.

Proposition 4 *Under the setup of Theorem 3, further assume that e is twice differentiable and $\widetilde{E}$ is twice differentiable at 0 with $\nabla^2 \widetilde{E}(0) = \mathrm{E}\big[\nabla^2 e(\eta)\big]$, and let $\alpha := \mathrm{E}[e''(\eta)]/2$. Then*

$$\lim_{n \to \infty} n \cdot R^*_{n,\delta}(\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)) = \sigma^2 \alpha \cdot Q_{\|Z\|_2^2}(1 - \delta) \asymp \sigma^2 \alpha \cdot [d + \log(1/\delta)],$$

where $Z \sim \mathcal{N}(0, I_{d \times d})$, and where the relation $\asymp$ holds when $\delta \in (0, 1/2)$.

Non-asymptotically, and with no more assumptions on the error function, it is difficult to say much more about the minimax risk than Proposition 4. However, determining when the minimax risk is infinite is tractable, as the next result shows.

Lemma 5 *Under the setup of Theorem 3, let $\varepsilon_n := \mathrm{P}\Big(\mathrm{rank}(\widehat{\Sigma}_n) < d\Big)$. Then*

$$R^*_{n,\delta}(\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)) = \infty \Leftrightarrow \delta \leq \varepsilon_n \quad \text{and} \quad \rho(P_X)^n \leq \varepsilon_n \leq \binom{n}{d-1} \rho(P_X)^{n-d-1},$$

where $\rho(P_X) := \sup_{w \in \mathbb{R}^d \setminus \{0\}} \mathrm{P}(\langle w, X \rangle = 0) < 1$.

The upper bound on ε_n as well as the statement $\rho(P_X) < 1$ in Lemma 5 are due to El Hanchi and Erdogdu (2023). At a high level, Lemma 5 says that the range of failure probabilities for which the risk is infinite gets exponentially small as a function of n. This is in sharp contrast with the result of Mourtada (2022) under the classical notion of risk and the square error, where it was shown that the minimax risk in that case is infinite for all sample sizes as soon as $\rho(P_X) > 0$.

2.2. Square error

We assume throughout this section that $e(t) = t^2/2$. We derive increasingly loose but more interpretable upper and lower bounds on the minimax risk in this setting. Our motivation is to better understand the influence of each of the parameters $(n, \delta, P_X, \sigma^2)$ of the problem on the minimax risk. Practically, the main result of this section is the identification of a general sufficient condition under which the lower bound in Proposition 1 is tight. With that achievement to look forward to, we start with a simple Corollary of Theorem 3.

Corollary 6 *Under the setup of Theorem 3,*

$$R^*_{n,\delta}(\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)) = \frac{\sigma^2}{2n} \cdot Q_{\|Z\|_2^2}(1 - \delta),$$

where the random variable Z is jointly distributed with $(X_i)_{i=1}^n \sim P_X^n$ such that $Z \mid (X_i)_{i=1}^n \sim \mathcal{N}(0, \widetilde{\Sigma}_n^{-1})$ on the event that the sample covariance matrix $\widehat{\Sigma}_n$ is invertible, and where $\widetilde{\Sigma}_n = \Sigma^{-1/2} \widehat{\Sigma}_n \Sigma^{-1/2}$ is the whitened sample covariance matrix; otherwise $\|Z\|_2^2 := \infty$.

Corollary 6 already makes explicit the dependence of the minimax risk on (n, σ^2), but the dependence on (P_X, δ) remains implicit. The next result is a step towards clarifying this relationship.

Proposition 7 *Under the setup of Theorem 3, and for all $\delta \in (0, (1 - \varepsilon_n)/4)$,*

$$
R^*_{n,\varepsilon_n+\delta}(\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2))
\begin{cases}
\leq 2 \cdot \dfrac{\sigma^2}{n} \Big[Q_{\text{Tr}(\widetilde{\Sigma}_n^{-1})}(1 - \varepsilon_n - \delta/2) + Q_W(1 - \varepsilon_n - \delta/2) \Big], \\[2ex]
\geq \dfrac{1}{6428} \cdot \dfrac{\sigma^2}{n} \Big[Q_{\text{Tr}(\widetilde{\Sigma}_n^{-1})}(1 - \varepsilon_n - 4\delta) + Q_W(1 - \varepsilon_n - 4\delta) \Big],
\end{cases}
$$

where we defined $\text{Tr}(\widetilde{\Sigma}_n^{-1}) := \infty$ when $\widetilde{\Sigma}_n$ is not invertible, and W is a random variable jointly distributed with $(X_i)_{i=1}^n \sim P_X^n$ and with conditional distribution $W \mid (X_i)_{i=1}^n \sim \text{Exp}(\lambda_{\min}(\widetilde{\Sigma}_n))$, with the convention that the exponential distribution $\text{Exp}(0)$ refers to the unit mass at ∞.

It is interesting to compare this result with the exact minimax risk under the classical notion of risk computed by Mourtada (2022), and given by $(\sigma^2/n) \cdot \mathbb{E}[\text{Tr}(\widetilde{\Sigma}_n^{-1})]$. Proposition 7 says that the minimax quantile risk is upper and lower bounded by a 'strong' term given by a quantile of $\text{Tr}(\widetilde{\Sigma}_n^{-1})$, and a 'weak' term governed by the distribution of $\lambda_{\min}(\widetilde{\Sigma}_n)$. Our next result shows that the lower bound from Proposition 7 improves on the one from Proposition 1.

Lemma 8 *Let $\delta \in (\varepsilon_n, 1)$. Then*

$$
\begin{aligned}
d \cdot (1 - \delta) \leq & Q_{\text{Tr}(\widetilde{\Sigma}_n^{-1})}(1 - \delta) \leq Q_{\lambda_{\max}(\widetilde{\Sigma}_n^{-1})}(1 - \delta) \cdot d, \\
\log(1/\delta) \leq & \quad Q_W(1 - \delta) \leq Q_{\lambda_{\max}(\widetilde{\Sigma}_n^{-1})}(1 - \delta/2) \cdot \log(2/\delta).
\end{aligned}
$$

This lemma further shows that a sufficient condition for the lower bound of Proposition 1 to be tight is the boundedness of $Q_{\lambda_{\max}(\widetilde{\Sigma}_n^{-1})}(1 - \delta/2)$ by an absolute constant. Under what conditions on (n, δ, P_X) does this hold ? Our results from Section 5 provide a satisfying answer.

Corollary 9 *Assume that P_X has fourth moments. If $\delta \in (0, 1/2)$ and*

$$
n \geq \max\left\{ 128[4\log(3d)\lambda_{\max}(S(P_X)) + R(P_X)\log(2/\delta)], \frac{\log(3d)}{18\lambda_{\max}(S(P_X))}, \frac{\log(2/\delta)}{R(P_X)} \right\},
$$

then

$$
R^*_{n,\delta}(\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)) \asymp \frac{\sigma^2(d + \log(1/\delta))}{n},
$$

where the parameters $S(P_X)$, $R(P_X)$ are as defined in (4).

Corollary 9 can be interpreted as a non-asymptotic version of Proposition 4 for the square error function. As we argue in Section 5, the fourth moment assumption is very natural in this setting, and the sample size restriction is, in a sense, optimal. The main restriction on the sample size comes from the first term, as both $\lambda_{\max}(S(P_X))$ and $R(P_X)$ are expected to be large.

2.3. p-th power error

The results of the last section are quite specific to the case of the square error, and it is a priori unclear how the minimax risk of other error functions can be studied non-asymptotically. Let us build on the observation that Corollary 9 is a non-asymptotic version of Proposition 4 for the square error. Can we do this for more general error functions ? The underlying proof idea of Proposition 4 is a simple second order Taylor expansion, which becomes exact as $n \to \infty$. If we have non-asymptotic control over the error in this expansion, we can carry out the argument behind Proposition 4 non-asymptotically. We implement this idea here, and conclude this section with the following non-asymptotic lower bound on the minimax risk under a p-th power error function.

Proposition 10 *Assume that $e(t) = |t|^p/[p(p-1)]$ for some $p \in (2, \infty)$. Under the setup of Theorem 3, and for $\delta \in (0, 1/2)$, we have*

$$R_{n,\delta}^*(\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)) \geq \frac{m(p-2)}{16(p-1)} \cdot \frac{\sigma^p[d + \log(1/\delta)]}{n}$$

where $m(p) := 2^{p/2-1}\Gamma(p/2-1)/\sqrt{\pi}$ is the p-th absolute moment of a standard normal variable.

3. Minimaxity of the min-max linear regression procedure

In this section, we establish the minimaxity of a variant of the popular min-max regression procedure, e.g. Audibert and Catoni (2011); Lecué and Lerasle (2020); Oliveira and Resende (2023) over suitably large classes of problems under the p-th power error functions $e(t) = |t|^p/[p(p-1)]$, for $p \in [2, \infty)$. Before stating our results, we briefly describe the construction of the procedure.

Let $\alpha, \beta \in \mathbb{R}$ such that $\alpha \leq \beta$ and define $\phi_{\alpha,\beta}(x) := \alpha \mathbb{1}_{(-\infty,\alpha)}(x) + x \mathbb{1}_{[\alpha,\beta]}(x) + \beta \mathbb{1}_{(\beta,\infty)}(x)$. For a real valued sequence $a := (a_i)_{i=1}^n$, define the sequence $a^* = (a_i^*)_{i=1}^n$ by $a_i^* := a_{\pi(i)}$ where π is a permutation that orders a increasingly. Fix $k \in \{1, \ldots, \lfloor n/2 \rfloor\}$, and define $\varphi_k[a] := \sum_{i=1}^n \phi_{a_{1+k}^*, a_{n-k}^*}(a_i)$. Given samples $(X_i, Y_i)_{i=1}^n$, and for $w, v \in \mathbb{R}^d$, define

$$\psi_k(w, v) := n^{-1}\varphi_k[(e(\langle w, X_i \rangle - Y_i) - e(\langle v, X_i \rangle - Y_i))_{i=1}^n],$$

and consider the procedure

$$\hat{w}_{n,k}((X_i, Y_i)_{i=1}^n) \in \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \max_{v \in \mathbb{R}^d} \psi_k(w, v). \tag{3}$$

3.1. Square error

Our first result shows that for the square error, and under appropriate conditions, the procedure (3) is minimax up to absolute constants over $\mathcal{P}_2(P_X, \sigma^2)$ when P_X has finite fourth moments.

Theorem 11 *Under the square error $e(t) = t^2/2$, let $\delta \in (0, 1/4)$ be such that $k := 8\log(4/\delta)$ is an integer satisfying $1 \leq k \leq \lfloor n/8 \rfloor$. Assume that P_X has finite fourth moments. If*

$$n \geq 800^2 \cdot (8\log(6d) \cdot [\lambda_{\max}(S(P_X)) + 1] + [R(P_X) + 1]\log(1/\delta)),$$

where $S(P_X)$ and $R(P_X)$ are as defined in (4), then

$$R_{n,\delta}(\mathcal{P}_2(P_X, \sigma^2), \hat{w}_{n,k}) \leq (100)^2 \cdot \frac{\sigma^2(d + \log(1/\delta))}{n}.$$

Compared to Proposition 2, the upper bound in Theorem 11 contains no distribution-dependence, showing the minimaxity of the procedure (3) up to an absolute constant by Proposition 1. On the other hand, we require the existence of fourth moments, which is more than what is required in Proposition 2. As we argue in Section 5 however, the fourth moment assumption is quite natural. We also note that the sample size restriction in Theorem 11 nearly matches that in Corollary 9, which as we discuss in Section 5 is optimal in a certain sense.

3.2. p-th power error

We now move to the case $p \in (2, \infty)$. The first difficulty we are faced with here is that it is a priori unclear what set of problems we should select that is both tractable and large enough to model realistic scenarios. Using our insights from Section 2, we propose the following analogue to the class $\mathcal{P}_2(P_X, \sigma^2)$, under the necessary conditions that P_X and the noise ξ have finite p-th moments

$$\mathcal{P}_p(P_X, \sigma^2, \mu) := \{P \mid (X, Y) \sim P : X \sim P_X \text{ and } (\star) \text{ holds.}\},$$

where $\mu \in (0, m(p) \cdot \sigma^{p-2}]$, $m(p)$ is as in Proposition 10, and $(\star)$ is the condition

$$\frac{\operatorname{ess\,sup}(\mathrm{E}[|\xi|^{2(p-1)} \mid X])}{\operatorname{ess\,inf}(\mathrm{E}[|\xi|^{p-2} \mid X])} \leq \frac{m(2p-2)}{m(p-2)} \cdot \sigma^p =: r(p) \quad \text{and} \quad \operatorname{ess\,inf}(\mathrm{E}[|\xi|^{p-2} \mid X]) \geq \mu, \quad (\star)$$

where $\xi := Y - \langle w^*, X \rangle$ and $w^* \in \mathbb{R}^d$ is the unique minimizer of the expected error $E(w)$. It is straightforward to check that $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2) \subset \mathcal{P}_p(P_X, \sigma^2, \mu)$, for all legal choices of μ.

While at first this seems like an overly special class of distributions, let us now argue that this far from the case. In fact, we claim that this class is a natural extension of $\mathcal{P}_2(P_X, \sigma^2)$. Firstly, by setting $p = 2$, we recover $\mathcal{P}_2(P_X, \sigma^2, \mu) = \mathcal{P}_2(P_X, \sigma^2)$ for all legal μ. Secondly, we note that $\mathcal{P}_p(P_X, \sigma^2, \mu) \subset \mathcal{P}_p(P_X, \sigma^2, \mu')$ whenever $\mu \geq \mu'$. Ideally, we would like to take as large a class as possible, which would correspond to the choice $\mu = 0$. Unfortunately, our bounds diverge in this setting. On the positive side however, this means that the upper constraint on μ is benign as the interesting regime is when μ is near zero. Finally, note that much like with the set $\mathcal{P}_2(P_X, \sigma^2)$, one can capture a large set of problems by varying σ^2. As an example, for any linear regression problem where the noise ξ is non-zero, symmetric, and independent of X, there exists (σ^2, μ) such that $\mathcal{P}_p(P_X, \sigma^2, \mu)$ contains this problem.

Remarkably, we show that the procedure (3) is minimax over this class under mild assumptions.

Theorem 12 *Under the p-th power error $e(t) = |t|^p/[p(p-1)]$ for $p \in (2, \infty)$, let $\delta \in (0, 1)$ be such that $k := 8\log(4/\delta)$ is an integer satisfying $1 \leq k \leq \lfloor n/8 \rfloor$. Assume that P_X has finite fourth moments. If*

$$n \geq r^{\frac{p-2}{p-1}}(p)\mu^{-\frac{p}{p-1}} \cdot [8\log(6d)(\lambda_{\max}(S(P_X)) + 1) + (R(P_X) + 1)\log(1/\delta)]$$
$$+ (2400)^2 \cdot r(p)\mu^{-\frac{p}{p-2}}p^4 N^{\frac{2p}{p-2}}(P_X, p) \cdot [d + \log(4/\delta)]$$

where $r(p)$ and μ are as in $(\star)$, $S(P_X)$ and $R(P_X)$ are as in (4), and $N(P_X, p)$ is the norm equivalence constant between the L^p and L^2 norms induced by P_X on the set of linear functions on $\mathbb{R}^d$, given by $N(P_X, p) = \sup_{w \in \mathbb{R}^d \setminus \{0\}} \mathrm{E}[|\langle w, X \rangle|^p]^{1/p} / \mathrm{E}[\langle w, X \rangle^2]^{1/2}$, then

$$R_{n,\delta}(\mathcal{P}_p(P_X, \sigma^2, \mu), \hat{w}_{n,k}) \leq 120^2 \cdot K(p) \cdot \frac{\sigma^p[d + \log(1/\delta)]}{n},$$

where $K(p) := (p-1)^2 \cdot m(2p-2)/m(p-2)$.

Combining this result with Proposition 10 shows the minimaxity of the procedure (3) on this class of problems, up to a constant that depends only on p. The closest related result is due to El Hanchi and Erdogdu (2023) who studied the performance of ERM on linear regression under p-th power error.

Their result however is specific to ERM, and, as expected, only yields a polynomial dependence on $1/\delta$ instead of the needed $\log(1/\delta)$ to establish minimaxity. Our result combines the insights of that work with the proof techniques used to study the procedure (3) developed by Lugosi and Mendelson (2019a), as well as our new insights on how to leverage the fourth moment assumption to obtain absolute constants instead of distribution-dependent constants in the upper bound.

4. The quantile risk

In this section, we study the quantile risk in full generality. Our motivation for doing so is to provide the tools necessary for proving Theorem 3. The results we obtain are however more general and can be used to tackle other problems. We illustrate this with two examples at the end of the section.

Before we formulate our results, let us briefly introduce some basic decision-theoretic concepts. To avoid overloading the notation, the symbols we introduce here will be specific to this section. A decision problem has the following components: a set of possible observations $\mathcal{O}$, a subset $\mathcal{P}$ of probability measures on $\mathcal{O}$, a set of available actions $\mathcal{A}$, a loss function $\ell : \mathcal{P} \times \mathcal{A} \to \mathbb{R}$, and a decision rule $d : \mathcal{O} \to \mathcal{A}$. For a fixed distribution P, the performance of a decision rule is classically evaluated through its expected loss $\mathrm{E}[\ell(P, d(O))]$ where $O \sim P$. Here instead, for a user-chosen failure probability $\delta \in (0, 1)$, we evaluate the performance of a decision rule through its quantile risk $R_\delta(\ell, P, d) := Q_{\ell(P,d(O))}(1 - \delta)$. We define the associated worst-case and minimax risks by $R_\delta(\ell, d) := \sup_{P \in \mathcal{P}} R_\delta(\ell, P, d)$ and $R_\delta^*(\ell) := \inf_d R_\delta(\ell, d)$ respectively. Our aim is to develop methods to understand the minimax risk and establish the minimaxity of candidate decision rules.

4.1. A Bayesian criterion for minimaxity and an invariance principle

A classical way to establish the minimaxity of a decision rule is by computing its worst-case risk and showing that it matches the limit of a sequence of Bayes risks (Lehmann and Casella, 2006). The following result provides an analogue to this method when working under the quantile risk.

Theorem 13 *For a distribution π on $\mathcal{P}$, define $F^\pi_{\ell(P,d(O))}$ to be the cumulative distribution function of the random variable $\ell(P, d(O))$, where $P \sim \pi$ and $O \mid P \sim P$. Let $(\pi_k)_{k \in \mathbb{N}}$ be a sequence of distributions on $\mathcal{P}$. For any $t \in \mathbb{R}$, define*

$$p_{\ell,k}(t) := \sup_d F^{\pi_k}_{\ell(P,d(O))}(t).$$

Assume that the functions $(p_{\ell,k})_{k \in \mathbb{R}}$ are right-continuous and that the sequence is decreasing, i.e. $p_{\ell,k} \geq p_{\ell,k+1}$ and let $p_\ell := \inf_{k \in N} p_{\ell,k} = \lim_{k \to \infty} p_{\ell,k}$. If $\hat{d}$ is a decision satisfying

$$\sup_{P \in \mathcal{P}} R_\delta(\ell, P, \hat{d}) = p_\ell^-(1 - \delta),$$

where p_ℓ^- is the pseudo-inverse of p_ℓ, then $\hat{d} \in \operatorname{argmin}_d R_\delta(\ell, d)$, i.e. it is minimax.

We mention that instantiations of the arguments leading to Theorem 13 have been used by Depersin and Lecué (2022) and Gupta et al. (2023) to tackle specific problems. The general form we present here is new, and relies on new analytical results concerning collections of quantile functions.

In applications, it is desirable that the optimality of a decision rule depends as little as possible on the loss function, or conversely, that a single decision rule be minimax for as large a number of loss functions as possible. The following result shows that the minimaxity of a decision rule in the quantile risk is invariant to at least one form of transformation of the loss function.

Proposition 14 *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a strictly increasing left-continuous function, and define $\varphi(\infty) := \infty$ and $\varphi(-\infty) := -\infty$. Then $R_\delta(\varphi \circ \ell, P, d) = \varphi(R_\delta(\ell, P, d))$. Furthermore, if $R_\delta(\ell, d) < \infty$, then $R_\delta(\varphi \circ \ell, d) = \varphi(R_\delta(\ell, d))$. Finally, if $R_\delta^*(\ell) < \infty$, then*

$$d^* \in \operatorname*{argmin}_d R_\delta(\ell, d) \implies d^* \in \operatorname*{argmin}_d R_\delta(\varphi \circ \ell, d).$$

4.2. Mean estimation revisited

To exhibit the usefulness of the above results, we briefly revisit the problem of mean estimation under Gaussian data. This problem can be embedded in the above decision-theoretic setting as follows. The observations are random vectors $(X_i)_{i=1}^n \in (\mathbb{R}^d)^n$ for some $d, n \in \mathbb{N}$, the subset of distributions is the n-product of distributions in the class $\mathcal{P}_{\text{Gauss}}(\Sigma) := \{\mathcal{N}(\mu, \Sigma) \,|\, \mu \in \mathbb{R}^d\}$, for a fixed $\Sigma \in S_{++}^d$. The set of available actions is $\mathbb{R}^d$, and the loss function is given by $\ell(P^n, \mu) := e(\mu - \mu(P))$ for some error function e and where $\mu(P)$ is the mean of the distribution P. Finally, a decision rule is given by an estimator $\hat{\mu} : (\mathbb{R}^d)^n \to \mathbb{R}^d$. The following result gives the minimax quantile risk for this problem under a mild assumption on the error function e, generalizing the recent result of Depersin and Lecué (2022) which holds under stronger assumptions on e.

Proposition 15 *Assume that the error function e satisfies $e = \varphi \circ s$, where φ is a left-continuous strictly increasing function, and s is both quasiconvex, i.e. $s(tv + (1-t)u) \leq \max\{s(v), s(u)\}$ for all $t \in [0, 1]$ and $u, v \in \mathbb{R}^d$, and symmetric, i.e. $s(v) = s(-v)$. Then for all $\Sigma \in S_{++}^d$*

$$\inf_{\hat{\mu}} \sup_{P \in \mathcal{P}_{\text{Gauss}}(\Sigma)} R_\delta(\ell, P^n, \hat{\mu}) = Q_{e(Z)}(1 - \delta),$$

where $Z \sim \mathcal{N}(0, \Sigma/n)$. Furthermore, the sample mean $\hat{\mu}((X_i)_{i=1}^n) := n^{-1} \sum_{i=1}^n X_i$ is minimax.

4.3. Minimax estimation of the variance of a Gaussian with known mean

As a second application of our results, we consider the problem of variance estimation with one-dimensional Gaussian data. For this problem, the observations are random variables $(X_i)_{i=1}^n \in \mathbb{R}^n$ for some $n \in \mathbb{N}$, the subset of distributions is the n-product of distributions in the class $\mathcal{P}_{\text{Gauss}}(\mu) := \{\mathcal{N}(\mu, \sigma^2) \,|\, \sigma \in (0, \infty)\}$, for a fixed μ. The set of available actions is $(0, \infty)$, and we consider the following loss function which captures the problem of estimating σ^2 in relative error: $\ell(P^n, \sigma^2) := \left|\log(\sigma^2(P)/\sigma^2)\right|$ where $\sigma^2(P)$ is the variance of the distribution P. Finally, a decision rule is given by an estimator $\hat{\sigma}^2 : \mathbb{R}^n \to (0, \infty)$. Using Theorem 13, we obtain the following result.

Proposition 16 *For $\alpha \in (0, \infty)$ and $Z \sim \text{Inv-Gamma}(\alpha, \alpha)$, define $p_\alpha : (0, \infty) \to [0, 1]$ by*

$$p_\alpha(t) := P\left(\frac{1 - \exp(-2t)}{2t} \leq Z \leq \frac{\exp(2t) - 1}{2t}\right).$$

Then for all $\mu \in \mathbb{R}$

$$\inf_{\hat{\sigma}^2} \sup_{P \in \mathcal{P}_{\text{Gauss}}(\mu)} R_\delta(\ell, P^n, \hat{\sigma}^2) = p_{n/2}^-(1 - \delta).$$

Furthermore, the sample variance is not minimax and the estimator

$$\hat{\sigma}^2((X_i)_{i=1}^n) := \frac{\sum_{i=1}^n (X_i - \mu)^2}{n} \phi\left(p_{n/2}^-(1 - \delta)\right)$$

is minimax, where $\phi(x) := \sinh(x)/x$, and $p_{n/2}^-$ is the pseudo-inverse of $p_{n/2}$.

Surprisingly, Proposition 16 shows that the sample variance is suboptimal under the quantile risk, but that a careful reweighting of it is. We note that as $n \to \infty$, the weight converges to 1, so that the sample variance is asymptotically minimax. We are not aware of a similar result in the literature.

5. Smallest eigenvalue of the sample covariance matrix

The results of Sections 2 and 3, and in particular the sample size conditions in Corollary 9 and Theorems 11 and 12, rely on new high probability lower bounds on the smallest eigenvalue of the sample covariance matrix we describe in this section. We briefly formalize our problem, we then state our main results, and conclude this section by discussing and relating them to the literature.

Let P_X be a distribution on $\mathbb{R}^d$ with finite second moments, $X \sim P_X$, and denote by $\Sigma := \mathrm{E}\big[XX^T\big]$ its covariance matrix. For i.i.d. samples $(X_i)_{i=1}^n \sim P_X^n$, define the sample covariance matrix $\widehat{\Sigma}_n := n^{-1} \sum_{i=1}^n X_i X_i^T$. In this section, we want to identify how close $\widehat{\Sigma}_n$ is to Σ in a relative error sense and in a one-sided fashion. Specifically, we want to characterize the quantiles of the random variable $\lambda_{\max}(I - \Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2}) = 1 - \lambda_{\min}(\Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2})$. To ease notation, we introduce the whitened random vector $\widetilde{X} := \Sigma^{-1/2}X$. Notice that $\mathrm{E}[\widetilde{X}\widetilde{X}^T] = I_{d \times d}$, and that $\widetilde{\Sigma}_n := n^{-1} \sum_{i=1}^n \widetilde{X}_i \widetilde{X}_i^T = \Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2}$. We want to understand the quantiles of $1 - \lambda_{\min}(\widetilde{\Sigma}_n)$.

As already mentioned, our motivation for studying this problem stems from the fact that upper bounds on these quantiles form a crucial step in the analysis of linear regression in general, e.g. Oliveira (2016); Mourtada (2022), and in particular our results from Section 2 and 3.

We are now ready to state our results. Define the following variance-like parameters

$$S(P_X) := \mathrm{E}\left[\left(\widetilde{X}\widetilde{X}^T - I\right)^2\right], \qquad R(P_X) := \sup_{v \in S^{d-1}} \mathrm{E}\left[\left(\langle v, \widetilde{X}\rangle^2 - 1\right)^2\right]. \qquad (4)$$

Our first result is the following proposition, which provides an asymptotic lower bound on the quantiles of $1 - \lambda_{\min}(\widetilde{\Sigma}_n)$, and a nearly matching non-asymptotic upper bound.

Proposition 17 *Assume that P_X has finite fourth moments. Then for all $\delta \in (0, 0.1)$,*

$$\lim_{n \to \infty} \sqrt{n} \cdot Q_{1 - \lambda_{\min}(\widetilde{\Sigma}_n)}(1 - \delta) \geq \frac{1}{40} \cdot \left(\sqrt{\lambda_{\max}(S(P_X))} + \sqrt{R(P_X)\log(1/\delta)}\right).$$

Furthermore, for all $n \in \mathbb{N}$ and $\delta \in (0, 1)$,

$$Q_{1 - \lambda_{\min}(\widetilde{\Sigma}_n)}(1 - \delta) \leq \sqrt{\frac{8\lambda_{\max}(S(P_X))\log(3d)}{n}} + \sqrt{\frac{2R(P_X)\log(1/\delta)}{n}} + \frac{(2\log(3d) + 4\log(1/\delta))}{3n}.$$

Our second result extends the upper bound in Proposition 17 to the case where $\lambda_{\min}(\widetilde{\Sigma}_n) = \inf_{v \in S^{d-1}} n^{-1} \sum_{i=1}^n \langle v, \widetilde{X}_i\rangle^2$ is subject to a direction dependent adversarial truncation. This result is needed in our analyses of Section 3, from which we recall the definition of a^* for a sequence a.

Proposition 18 *Let $\delta \in (0, 1/2)$ such that $k = 8\log(2/\delta)$ is an integer satisfying $1 \leq k \leq \lfloor n/2 \rfloor$. For $(v, i) \in S^{d-1} \times [n]$, define $Y_{i,v} := \langle v, \widetilde{X}_i\rangle^2$ and $\overline{\lambda}_{\min}(\widetilde{\Sigma}_n) := \inf_{v \in S^{d-1}} n^{-1} \sum_{i=k+1}^{n-k} Y_{i,v}^*$. Then, if $n \geq 8\log(6d)$,*

$$Q_{(1 - 2k/n) - \overline{\lambda}_{\min}(\widetilde{\Sigma}_n)}(1 - \delta) \leq 100\left(\sqrt{\frac{8\log(6d)(\lambda_{\max}(S(P_X)) + 1)}{n}} + \sqrt{\frac{(R(P_X) + 1)\log(1/\delta)}{n}}\right).$$

11

Comparison with existing results. To the best of our knowledge, the only known lower bound, asymptotic or not, on the quantiles of $1 - \lambda_{\min}(\widetilde{\Sigma}_n)$ is due to Mourtada (2022, Proposition 4). This bound however is distribution-free and decays fast, as $\log(1/\delta)/n$. In terms of upper bounds comparable to that of Proposition 17, the closest result we are aware of is due to Oliveira (2016) (see also Zhivotovskiy (2024)), who proved $\sqrt{n}Q_{1-\lambda_{\min}(\widetilde{\Sigma}_n)}(1-\delta) \lesssim \sqrt{(R(P_X)+1)[d + \log(1/\delta)]}$. In general, our upper bound and theirs are not comparable, and when combined they yield the best of both. Nevertheless, we suspect that our bound is better on heavy-tailed problems. Indeed, by Jensen's inequality, it is not hard to see that $\lambda_{\max}(S(P_X)) \leq R(P_X) \cdot d$, so our upper bound from Proposition 17 can be at most worse by $\sqrt{\log(d)}$. This occurs when X is a centred Gaussian, for which it is known that Oliveira's bound is tight (Koltchinskii and Lounici, 2017). On the other hand, consider X with centred independent coordinates, and where the first coordinate has kurtosis $\kappa \gg 1$, while the other coordinates have constant kurtosis. Then Oliveira's bound scales as $\sqrt{\kappa \cdot d}$, while ours scales as $\sqrt{\kappa \cdot \log(d)}$. Finally, versions of Proposition 18 that mimic Oliveira's bound can be deduced from the recent literature (Abdalla and Zhivotovskiy, 2023; Oliveira and Rico, 2022). The same considerations apply when comparing these results.

On the fourth moment assumption. We carried out our analysis under a fourth moment assumption on P_X. We argue here that this is in some sense the most natural assumption to study this problem. Indeed, recall the fact that $\widetilde{\Sigma}_n$ is an empirical average of the random matrix $\widetilde{X}\widetilde{X}^T$. Therefore, by the law of large numbers, $\widetilde{\Sigma}_n \xrightarrow{d} I_{d \times d}$, and by the continuous mapping theorem, $\lambda_{\min}(\widetilde{\Sigma}_n) \xrightarrow{d} 1$ as $n \to \infty$. To say something about the rate of this convergence, the most natural assumption to make is that the entries of the random matrix XX^T have finite variance so that the CLT holds. This is equivalent to assuming that P_X has fourth moments.

On the critical sample size. Our main application of Propositions 17 and 18 is in providing upper bounds on the critical sample size $n^*(P_X, \delta) := \min\left\{n \in \mathbb{N} \,\middle|\, Q_{1-\lambda_{\min}(\widetilde{\Sigma}_n)}(1-\delta/2) \leq 1/4\right\}$. In particular, these upper bounds correspond to the sample size restrictions in Corollary 9 and Theorems 11 and 12. We claimed after the statement of these results that these restrictions were in some sense optimal; we expand on this here. Let $L := \lim_{n \to \infty} \sqrt{n} \cdot Q_{1-\lambda_{\min}(\widetilde{\Sigma}_n)}(1-\delta)$, and define $n_0(P_X, \delta) := \min\left\{n \in \mathbb{N} \,\middle|\, m \geq n \Rightarrow Q_{1-\lambda_{\min}(\widetilde{\Sigma}_m)}(1-\delta/2) \geq \frac{L}{4\sqrt{m}}\right\}$. If $n_0(P_X, \delta) \leq n^*(P_X, \delta)$, then we can reverse the above-mentioned upper bounds using the first item of Proposition 17, up to a $\sqrt{\log(d)}$ factor. In words, if the critical sample size is determined by the asymptotic behaviour of the $1 - \delta/2$ quantile of $1 - \lambda_{\min}(\widetilde{\Sigma}_n)$, then our bounds on the critical sample size are tight. When the hypothesis in this last statement is true remains unclear however. The choice of the constant $1/4$ in the above argument is arbitrary, and can be replaced with any absolute constant.

6. Conclusion

In this paper, we studied minimax linear regression under the quantile risk. We gave an in-depth characterization of the minimax risk over the Gaussian class $\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)$, and leveraged these results to establish the minimaxity, up to absolute constants, of the min-max regression procedure for p-norm regression problems. While the problem of estimation with high confidence has been studied intensely recently, we are not aware of its formalization through quantiles as was done in this paper. We hope this new perspective proves fruitful in advancing both our understanding of learning problems and our ability to design efficient solutions for them.

Acknowledgments

References

Pedro Abdalla and Nikita Zhivotovskiy. Covariance Estimation: Optimal Dimension-free Guarantees for Adversarial Corruption and Heavy Tails, July 2023.

Deeksha Adil, Rasmus Kyng, Richard Peng, and Sushant Sachdeva. Fast Algorithms for ℓ_p-Regression, October 2023.

Jean-Yves Audibert and Olivier Catoni. Robust Linear Least Squares Regression. *The Annals of Statistics*, 2011.

Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, January 2002. doi: 10.1016/S1631-073X(02)02292-6.

Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, November 2012. doi: 10.1214/11-AIHP454.

Olivier Catoni. Pac-bayesian bounds for the gram matrix and least squares regression with a random design. *arXiv preprint arXiv:1603.05229*, 2016.

Geoffrey Chinot, Guillaume Lecué, and Matthieu Lerasle. Robust statistical learning with Lipschitz and convex loss functions. *Probability Theory and Related Fields*, April 2020. doi: 10.1007/s00440-019-00931-3.

Jules Depersin and Guillaume Lecué. Optimal robust mean and location estimation via convex programs with respect to any pseudo-norms. *Probability Theory and Related Fields*, August 2022. doi: 10.1007/s00440-022-01127-y.

Luc Devroye, Matthieu Lerasle, Gabor Lugosi, and Roberto I. Oliveira. Sub-Gaussian mean estimators. *The Annals of Statistics*, December 2016. doi: 10.1214/16-AOS1440.

Ayoub El Hanchi and Murat A. Erdogdu. Optimal Excess Risk Bounds for Empirical Risk Minimization on p-Norm Linear Regression. In *Thirty-Seventh Conference on Neural Information Processing Systems*, November 2023.

Shivam Gupta, Jasper C. H. Lee, Eric Price, and Paul Valiant. Minimax-Optimal Location Estimation. In *Thirty-Seventh Conference on Neural Information Processing Systems*, November 2023.

Daniel Hsu and Sivan Sabato. Loss Minimization and Parameter Estimation with Heavy Tails. *Journal of Machine Learning Research*, 2016.

Robert W Keener. *Theoretical statistics: Topics for a core course*. Springer Science & Business Media, 2010.

T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, May 2005. doi: 10.1214/009117905000000044.

Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, February 2017. doi: 10.3150/15-BEJ730.

Guillaume Lecué and Matthieu Lerasle. Learning from MOM's principles: Le Cam's approach. *Stochastic Processes and their Applications*, November 2019. doi: 10.1016/j.spa.2018.11.024.

Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: Theory and practice. *The Annals of Statistics*, April 2020. doi: 10.1214/19-AOS1828.

Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes : Upper and minimax bounds, September 2016.

Jasper C.H. Lee and Paul Valiant. Optimal Sub-Gaussian Mean Estimation $\mathbb{R}$. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, Denver, CO, USA, February 2022. doi: 10.1109/FOCS52979.2021.00071.

Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

Gábor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, December 2019a. doi: 10.4171/JEMS/937.

Gábor Lugosi and Shahar Mendelson. Sub-Gaussian estimators of the mean of a random vector. *The Annals of Statistics*, April 2019b. doi: 10.1214/17-AOS1639.

Gábor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: The optimality of trimmed mean. *The Annals of Statistics*, February 2021. doi: 10.1214/20-AOS1961.

Shahar Mendelson. "Local" vs. "global" parameters—breaking the Gaussian complexity barrier. *The Annals of Statistics*, October 2017. doi: 10.1214/16-AOS1510.

Shahar Mendelson and Nikita Zhivotovskiy. Robust covariance estimation under $L_4 - L_2$ norm equivalence. *The Annals of Statistics*, June 2020. doi: 10.1214/19-AOS1862.

Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics*, August 2022. doi: 10.1214/22-AOS2181.

Roberto I. Oliveira and Lucas Resende. Trimmed sample means for robust uniform mean estimation and regression, February 2023.

Roberto I. Oliveira and Zoraida F. Rico. Improved covariance estimation: Optimal robustness and sub-Gaussian guarantees under heavy tails, September 2022.

Roberto Imbuzeiro Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 2016. doi: 10.1007/s00440-016-0738-9.

Adrien Saumard. On optimality of empirical risk minimization in linear aggregation. *Bernoulli*, August 2018. doi: 10.3150/17-BEJ925.

Joel A. Tropp. The Expected Norm of a Sum of Independent Random Matrices: An Elementary Approach, October 2015a.

Joel A. Tropp. An Introduction to Matrix Concentration Inequalities, January 2015b.

Ramon Van Handel. Structured random matrices. *Convexity and concentration*, pages 107–156, 2017.

Nikita Zhivotovskiy. Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *Electronic Journal of Probability*, January 2024. doi: 10.1214/23-EJP1021.

Contents

Appendix A. Preliminaries

A.1. Pseudo-inverses and quantile function

We say a function $f : \mathbb{R} \to \mathbb{R}$ is increasing if $x < y$ implies $f(x) \leq f(y)$, and strictly increasing if $x < y$ implies $f(x) < f(y)$. For a function $f : \mathbb{R} \to \mathbb{R}$, we define $\mathrm{Im}(f) = \{f(x) \,|\, x \in \mathbb{R}\}$.

Definition 19 *Let $f : \mathbb{R} \to \mathbb{R}$ be an increasing function. We define $f^- : [-\infty, \infty] \to [-\infty, \infty]$, the pseudo-inverse of f, by*

$$f^-(y) := \inf\{x \in \mathbb{R} \,|\, f(x) \geq y\},$$

with the conventions $\inf \varnothing := \infty$ and $\inf \mathbb{R} := -\infty$.

Lemma 20 *The following holds.*

- *Let $f : \mathbb{R} \to \mathbb{R}$ be an increasing function. Then for all $x \in \mathbb{R}$, $f^-(f(x)) \leq x$.*

- *Let f, g be increasing functions from $\mathbb{R}$ to $\mathbb{R}$. If $f \geq g$ then $f^- \leq g^-$.*

- *Let I be an index set and let $\{f_i\}_{i \in I}$ be a collection of increasing functions from $\mathbb{R}$ to $\mathbb{R}$. Then*

$$\left(\sup_{i \in I} f_i\right)^- \leq \inf_{i \in I} f_i^- \leq \sup_{i \in I} f_i^- \leq \left(\inf_{i \in I} f_i\right)^-.$$

- *Let $f : \mathbb{R} \to \mathbb{R}$ be a strictly increasing function, so that it is a bijection from $\mathbb{R}$ to $\mathrm{Im}(f)$. Denote by $f^{-1} : \mathrm{Im}(f) \to \mathbb{R}$ the inverse of f. Then for all $y \in \mathrm{Im}(f)$, $f^-(y) = f^{-1}(y)$.*

- *Let $f : \mathbb{R} \to \mathbb{R}$ be an increasing right-continuous function. Then for all $y \in [-\infty, \infty]$,*

$$f^-(y) = \min\{x \in \mathbb{R} \,|\, f(x) \geq y\},$$

  with the conventions $\min \varnothing := \infty$ and $\min \mathbb{R} := -\infty$.

- *Let I be an index set and let $\{f_i\}_{i \in I}$ be a collection of increasing right-continuous functions from $\mathbb{R}$ to $\mathbb{R}$. Then*

$$\sup_{i \in I} f_i^- = \left(\inf_{i \in I} f_i\right)^-.$$

- *Let $f : \mathbb{R} \to \mathbb{R}$ be increasing and right-continuous, and let $g : \mathbb{R} \to \mathbb{R}$ be increasing. Then*

$$(f \circ g)^- = g^- \circ f^-.$$

- *Let $(f_k)_{k \in \mathbb{N}}$ be a decreasing sequence of increasing right-continuous functions, and assume that $f_n \to f$ pointwise as $n \to \infty$. Then,*

$$\sup_{n \in \mathbb{N}} f_n^{-1} = f^{-1}.$$

Proof

- We have, since $f(x) \geq f(x)$,

$$f^-(f(x)) = \inf\{z \in \mathbb{R} \mid f(z) \geq f(x)\} \leq x.$$

- Fix $y \in [-\infty, \infty]$, and define $S_f := \{x \in \mathbb{R} \mid f(x) \geq y\}$, and $S_g := \{x \in \mathbb{R} \mid g(x) \geq y\}$. We claim that $S_g \subset S_f$ from which the statement follows. If $S_g = \varnothing$, the statement follows trivially. Otherwise let $x \in S_g$. Then $f(x) \geq g(x) \geq y$, so $x \in S_f$.

- We prove the last inequality. The first follows from a similar argument. By definition, for all $j \in I$, $\sup_{i \in I} f_i \geq f_j$. Applying the second item yields that for all $j \in I$, $(\sup_{i \in I} f_i)^- \leq f_j^-$. Taking the infimum over $j \in I$ yields the result.

- Let $y \in \mathrm{Im}(f)$ and let $S := \{x \in \mathbb{R} \mid f(x) \geq y\}$. We claim that $f^{-1}(y) = \min S$ from which the claim follows since then $f^-(y) = \min S$. Indeed, since $f(f^{-1}(y)) = y$, we have $f^{-1}(y) \in S$. Now suppose that there exists an $x \in S$ such that $f^{-1}(y) > x$. Since f is strictly increasing, we would have $y = f(f^-(y)) > f(x)$, which contradicts the fact that $x \in S$. Therefore $f^{-1}(y) \leq x$ for all $x \in S$. This proves that $f^{-1}(y) = \min S$.

- The statement holds trivially if $f^-(y) \in \{-\infty, \infty\}$. Otherwise, $f^-(y) \in \mathbb{R}$, and by definition of the infimum, for all $k \in \mathbb{N}$, we have $x_k := f^-(y) + 1/k \in S$ and therefore $f(x_k) \geq y$. Furthermore, $\lim_{k \to \infty} x_k = f^{-1}(y)$ and $x_k > f^{-1}(y)$, so by the right-continuity of f we obtain

$$f(f^{-1}(y)) = \lim_{k \to \infty} f(x_k) \geq y.$$

Therefore $f^{-1}(y) \in S$ which implies $f^{-1}(y) = \min S$.

- The inequality ($\leq$) is covered by the third item, therefore it is enough to prove the inequality ($\geq$). Let $y \in [-\infty, \infty]$. We claim that

$$\sup_{i \in I} f_i^-(y) \geq \left(\inf_{i \in I} f_i\right)^- (y).$$

The statement follows trivially if $\sup_{i \in I} f_i^-(y) = \infty$. Otherwise, we have $f_i^-(y) < \infty$ for all $i \in I$. If $\sup_{i \in I} f_i^-(y) = -\infty$, then $f_i^-(y) = -\infty$ for all $i \in I$, which implies that $f_i(x) \geq y$ for all $x \in \mathbb{R}$. This in turn implies that for all $x \in \mathbb{R}$, $\inf_{i \in I} f_i(x) \geq y$ and therefore $(\inf_{i \in I} f_i)^-(y) = -\infty$. It remains to consider the case where $\sup_{i \in I} f_i^-(y) \in \mathbb{R}$. We claim that

$$\inf_{i \in I} f_i\left(\sup_{j \in I} f_j^-(y)\right) \geq y$$

from which the main claim follows by definition of the pseudo-inverse. Indeed, let $i \in I$. If $f_i^-(y) \in \mathbb{R}$, then we have

$$f_i\left(\sup_{j \in I} f_j^-(y)\right) \geq f_i(f_i^-(y)) \geq y$$

where the first inequality holds since f_i is increasing, and the second by the fifth item and the fact that $f_i^-(y) \in \mathbb{R}$. Otherwise $f_i^-(y) = -\infty$, and therefore $f_i(x) \geq y$ for all $x \in \mathbb{R}$, which in particular implies the desired statement since $\sup_{i \in I} f_i^-(y) \in \mathbb{R}$.

- Let $y \in [-\infty, \infty]$. By the assumed properties of f and the fifth item, we have $f(g(x)) \geq y$ if and only if $g(x) \geq f^-(y)$. Therefore

$$(f \circ g)^-(y) = \inf\{x \in \mathbb{R} \mid f(g(x)) \geq y\}$$
$$= \inf\{x \in \mathbb{R} \mid g(x) \geq f^-(y)\}$$
$$= g^-(f^-(y)) = (g^- \circ f^-)(y).$$

- We start with the inequality ($\leq$). Let $x \in \mathbb{R}$. Since $(f_n(x))_{n \in \mathbb{N}}$ is decreasing, we have $f(x) = \lim_{n \to \infty} f_n(x) = \inf_{n \in \mathbb{N}} f_n(x)$. Therefore, for all $n \in \mathbb{N}$, we have $f_n \geq f$. By the second item, we therefore have $f_n^- \leq f^-$. Taking supremum over n yields the result.

  For the inequality ($\geq$), let $y \in \mathbb{R}$, and suppose that $\sup_{n \in \mathbb{N}} f_n^-(y) < f^-(y)$. If $\sup_{n \in \mathbb{N}} f_n^-(y) = -\infty$, then for all $x \in \mathbb{R}$ and for all $n \in \mathbb{N}$, $f_n(x) \geq y$, which implies that for all $x \in \mathbb{R}$, $f(x) = \lim_{n \to \infty} f_n(x) \geq y$, and therefore $f^{-1}(y) = -\infty$, contradicting the strict inequality. Otherwise $x^* := \sup_{n \in \mathbb{N}} f_n^-(y) \in \mathbb{R}$, and either $f^-(y) = \infty$ or $f^-(y) \in \mathbb{R}$.

  If $f^-(y) = \infty$, then on the one hand, for all $x \in \mathbb{R}$, $\lim_{n \to \infty} f_n(x) = f(x) < y$. On the other hand, for all $n \in \mathbb{N}$, $f_n(x^*) \geq y$. Indeed, if $f_n^-(y) \in \mathbb{R}$, then by the fifth item $f_n(x^*) \geq f_n(f_n^-(y)) \geq y$. Otherwise, $f_n^-(y) = -\infty$ so that $f(x) \geq y$ for all $x \in \mathbb{R}$, and in particular $f_n(x^*) \geq y$. But then we get the contradiction $y > \lim_{n \to \infty} f_n(x^*) \geq y$.

  Finally, if $f^-(y) \in \mathbb{R}$, define $\varepsilon := f^-(y) - x^* > 0$. By definition of x^*, $f^-(y) - \varepsilon \geq f_n^-(y)$ for all $n \in \mathbb{N}$. We claim that for all $n \in \mathbb{N}$

$$f_n(f^-(y) - \varepsilon) \geq y.$$

  Indeed, if $f_n^-(y) \in \mathbb{R}$, then by the fifth item $f_n(f^-(y) - \varepsilon) \geq f_n(f_n^-(y)) \geq y$. Otherwise, $f_n^-(y) = -\infty$ so that $f(x) \geq y$ for all $x \in \mathbb{R}$, and in particular $f_n(f^-(y) - \varepsilon) \geq y$ since $f^-(y) - \varepsilon \in \mathbb{R}$. Taking the limit as $n \to \infty$ yields

$$\lim_{n \to \infty} f_n(f^-(y) - \varepsilon) = f(f^-(y) - \varepsilon) \geq y$$

  contradicting the minimality of $f^-(y)$.

  ∎

For a random variable X, we denote by $F_X : \mathbb{R} \to \mathbb{R}$ its cumulative distribution function $F_X(x) := \mathrm{P}(X \leq x)$, and by $Q_X : [-\infty, \infty] \to [-\infty, \infty]$ its quantile function $Q_X(p) := F_X^-(p)$. Since F_X is right-continuous, then by the fifth item of Lemma 20 we have

$$Q_X(p) = \min\{x \in \mathbb{R} \mid F_X(x) \geq p\}.$$

Furthermore, since $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$, it is easy to verify that $Q_X(p) \in \mathbb{R}$ for all $p \in (0, 1)$ and $Q_X(0) = -\infty$. If X, Y are two random variables, we define the random variable $F_{X|Y}(x) := \mathrm{P}(X \leq x \mid Y)$ and we note that $F_X(x) = \mathrm{E}[F_{X|Y}(x)]$ for all $x \in \mathbb{R}$.

Lemma 21 *Let $\varphi : \mathbb{R} \to \mathbb{R}$ be strictly increasing and left continuous. Then for all $p \in (0, 1)$*

$$Q_{\varphi(X)}(p) = \varphi(Q_X(p)).$$

where we define $\varphi(\infty) := \infty$ and $\varphi(-\infty) := -\infty$.

Proof Let $\varepsilon_- := \mathrm{P}(X = \infty)$ and $\varepsilon_+ = \mathrm{P}(X = \infty)$. By definition, $\varphi(X) = \infty \Leftrightarrow X = -\infty$, so the identity holds trivially for all $p \in (0, \varepsilon_-] \cup [1 - \varepsilon_+, 1)$. Now consider the case $p \in I := (\varepsilon_-, 1 - \varepsilon_+)$. First, since $\lim_{x \to -\infty} F_X(x) = \varepsilon_-$ and $\lim_{x \to \infty} F_X(x) = 1 - \varepsilon_+$ by continuity of probability measures, we have $Q_X(p) \in \mathbb{R}$. The same argument shows that $Q_{\varphi(X)}(p) \in \mathbb{R}$. Now, since φ is strictly increasing,

$$F_X(x) = \mathrm{P}(X \le x) = \mathrm{P}(\varphi(X) \le \varphi(x)) = F_{\varphi(X)}(\varphi(x)) = (F_{\varphi(X)} \circ \varphi)(x).$$

Therefore, by the penultimate item of Lemma 20, we have

$$Q_X = F_X^- = \varphi^- \circ F_{\varphi(X)}^- = \varphi^- \circ Q_{\varphi(X)}. \tag{5}$$

We claim that for all $p \in I$,

$$(\varphi \circ \varphi^- \circ Q_{\varphi(X)})(p) = Q_{\varphi(X)}(p). \tag{6}$$

By the fourth item of Lemma 20, it is enough to show that $Q_{\varphi(X)}(p) \in \mathrm{Im}(\varphi)$ for all $p \in I$. This will be the goal of the proof. Let $p \in I$, and define $S := \{x \in \mathbb{R} \mid \varphi(x) \le Q_{\varphi(X)}(p)\}$. We claim that S is non-empty and upper bounded. Indeed, suppose not. Then either $\varphi(x) > Q_{\varphi(X)}(p)$ for all $x \in \mathbb{R}$ or $\varphi(x) \le Q_{\varphi(X)}(p)$ for all $x \in \mathbb{R}$. In the former case, this implies that for all $x \in \mathbb{R}$

$$F_X(x) = \mathrm{P}(X \le x) = \mathrm{P}(\varphi(X) \le \varphi(x)) \ge \mathrm{P}(\varphi(X) \le Q_{\varphi(X)}(p)) \ge p > \varepsilon_-,$$

where the second inequality follows from the fifth item of Lemma 20 and the fact that $Q_{\varphi(X)}(p) \in \mathbb{R}$. This leads to the contradiction

$$\varepsilon_- = \lim_{x \to -\infty} F_X(x) \ge p > \varepsilon_-.$$

In the latter case, we get that for all $x \in \mathbb{R}$

$$F_X(x) = \mathrm{P}(X \le x) = \mathrm{P}(\varphi(X) \le \varphi(x)) \le \mathrm{P}(\varphi(X) \le Q_{\varphi(X)}(p)).$$

This leads to

$$1 - \varepsilon_+ = \lim_{x \to \infty} F_X(x) \le \mathrm{P}(\varphi(X) \le Q_{\varphi(X)}(p)) \le 1 - \varepsilon_+.$$

where the last inequality follows from the fact that $Q_{\varphi(X)}(p) \in \mathbb{R}$. Now, since

$$\mathrm{P}\left(\varphi(X) \le \lim_{x \to \infty} \varphi(x)\right) = 1 - \varepsilon_+$$

and $\varphi(x) \le Q_{\varphi(X)}(p)$ for all $x \in \mathbb{R}$, we get by the minimality property of $Q_{\varphi(X)}(p)$

$$Q_{\varphi(X)}(p) = \lim_{x \to \infty} \varphi(x).$$

But, on the one hand, we have by continuity of probability,

$$\lim_{n \to \infty} \mathrm{P}(\varphi(X) \le \varphi(n)) = \mathrm{P}\left(\varphi(X) \le \lim_{n \to \infty} \varphi(n)\right) = 1 - \varepsilon_+$$

21

yet on the other, since φ is strictly increasing, we have $\varphi(n) < \lim_{x \to \infty} \varphi(x) = Q_{\varphi(X)}(p)$ for all $n \in \mathbb{N}$, so by the minimality of $Q_{\varphi(X)}(p)$, $\mathrm{P}(\varphi(X) \leq \varphi(n)) < p$ for all $n \in \mathbb{N}$, from which we obtain the contradiction

$$1 - \varepsilon_+ = \lim_{n \to \infty} \mathrm{P}(\varphi(X) \leq \varphi(n)) \leq p$$

This proves that S is non-empty and upper bounded. Now define $x_0 := \sup S$, which is guaranteed to satisfy $x_0 \in \mathbb{R}$ by the upper boundedness of S and its non-emptiness. We claim that $\varphi(x_0) = Q_{\varphi(X)}(p)$. Indeed, by the left-continuity of φ, we have, for any sequence $(x_n)_{n \in \mathbb{N}}$ in S such that $x_n \to x_0$

$$\varphi(x_0) = \lim_{n \to \infty} \varphi(x_n) \leq Q_{\varphi(X)}(p) \tag{7}$$

where the last inequality follows from the definition of S and the fact that $x_n \in S$ for all $n \in \mathbb{N}$. On the other hand, by the maximality of x_0, we have for all $x > x_0$, $\varphi(x) > Q_{\varphi(X)}(p)$, which implies that

$$Q_{\varphi(X)}(p) \leq \lim_{x \to x_0^+} \varphi(x) \tag{8}$$

Combining (7) and (8), we obtain

$$Q_{\varphi(X)}(p) \in \left[\varphi(x_0), \lim_{x \to x_0^+} \varphi(x) \right]$$

But for any $y \in \left[\varphi(x_0), \lim_{x \to x_0^+} \varphi(x) \right]$, we have

$$\mathrm{P}(\varphi(X) \leq y) = \mathrm{P}(\varphi(X) \leq \varphi(x_0))$$

Indeed on the one hand

$$\mathrm{P}(\varphi(X) \leq y) \geq \mathrm{P}(\varphi(X) \leq \varphi(x_0))$$

On the other, if $X > x_0$, then since φ is strictly increasing, $\varphi(X) > \lim_{x \to x_0^+} \varphi(x) \geq y$. Therefore

$$\mathrm{P}(\varphi(X) \leq y) \leq \mathrm{P}\left(\varphi(X) \leq \lim_{x \to x_0^+} \varphi(x) \right) \leq \mathrm{P}(X \leq x_0) = \mathrm{P}(\varphi(X) \leq \varphi(x_0))$$

but then, by the minimality of $Q_{\varphi(X)}(p)$, we obtain $Q_{\varphi(X)}(p) = \varphi(x_0)$. This proves (6). Now applying φ to both sides of (5) and using (6) yields the result. $\blacksquare$

A.2. Convexity

Definition 22 *A subset $A \subset \mathbb{R}^d$ is*

- *convex if for all $x, y \in A$ and $t \in [0, 1]$, $(1 - t)x + ty \in A$.*

- *symmetric if for all $x \in A$, $-x \in A$.*

Lemma 23 *Let A be a non-empty convex symmetric set. Then for all $\lambda \geq 1$, $A \subseteq \lambda A$.*

Proof We start by proving that λA is convex. Indeed let $x, y \in \lambda A$ and $t \in [0, 1]$. Then by definition $x/\lambda, y/\lambda \in A$, so by convexity of A

$$(1 - t)\frac{x}{\lambda} + t\frac{y}{\lambda} \in A,$$

which implies

$$(1 - t)x + ty = \lambda \cdot \left((1 - t)\frac{x}{\lambda} + t\frac{y}{\lambda}\right) \in \lambda A.$$

Next we prove that $0 \in A$. Let $v \in A$. Since A is symmetric, $-v \in A$, and by convexity of A

$$0 = \frac{1}{2}x - \frac{1}{2}x = \frac{1}{2}x + \frac{1}{2}(-x) \in A.$$

Finally, we prove the main claim. Let $x \in A$. Then by definition $\lambda x \in \lambda A$. But then by convexity of λA and since $\lambda \geq 1$

$$x = \left(1 - \frac{1}{\lambda}\right)0 + \frac{1}{\lambda}\lambda x \in \lambda A$$

∎

Definition 24 *A function $f : \mathbb{R}^d \to \mathbb{R}$ is*

- *quasiconvex if for all $x, y \in \mathbb{R}^d$ and $t \in [0, 1]$, $f((1 - t)x + ty) \leq \max(f(x), f(y))$.*

- *symmetric if for all $v \in \mathbb{R}^d$, $f(v) = f(-v)$.*

Remark 25 *Every convex function is quasiconvex. The function $f(x) = \log(x)$ is quasiconvex but not convex. Every norm is quasiconvex (and, in fact, convex) and symmetric.*

Lemma 26 *The following holds.*

- *$f : \mathbb{R}^d \to \mathbb{R}$ is quasiconvex and symmetric if and only if for all $y \in \mathbb{R}$, $f^{-1}((-\infty, y])$ is convex and symmetric.*

- *If $f : \mathbb{R}^d \to \mathbb{R}$ is quasiconvex and symmetric then $0 \in \operatorname{argmin}_{x \in \mathbb{R}} f(x)$.*

A.3. Gaussian measures

Lemma 27 *Let $Z \sim \mathcal{N}(0, \sigma^2)$. Then*

$$\sqrt{1 - \exp\left(-\frac{r^2}{2\sigma^2}\right)} \leq F_{|Z|}(r) \leq \sqrt{1 - \exp\left(-\frac{2r^2}{\pi\sigma^2}\right)}.$$

Proof Consider first the case where $\sigma^2 = 1/2$. Then

$$\left(F_{|Z|}(r)\right)^2 = (\mathrm{P}(-r \leq Z \leq r))^2 = \left(\frac{2}{\sqrt{\pi}}\int_0^r e^{-t^2}dt\right)^2$$

$$= \frac{4}{\pi}\int_0^r\int_0^r e^{-(t^2+s^2)}dtds = \frac{4}{\pi}\int_S e^{-(t^2+s^2)}dtds$$

where $S := \{(x, y) \in \mathbb{R}^2 \mid 0 \leq x, y \leq r\}$ is the square of length r whose lower left corner is at 0. For a radius $\rho > 0$, define the quarter disks $D(\rho) := \{(x, y) \subset \mathbb{R}^2 \mid x, y \geq 0, \sqrt{x^2 + y^2} \leq \rho\}$. Clearly, $D(r) \subset S$, so that

$$\frac{4}{\pi} \int_S e^{-(t^2+s^2)} dt ds \leq \frac{4}{\pi} \int_{D(r)} e^{-(t^2+s^2)} dt ds = 1 - \exp(-r^2)$$

where the last equality is obtained by an explicit integration using polar coordinates. On the other hand, consider the quarter disk $D(2r/\sqrt{\pi})$, and define $A := D(2r/\sqrt{\pi}) \setminus S$ and $B := S \setminus D(2r/\sqrt{\pi})$. Since S and $D(2r/\sqrt{\pi})$ have the same area, so do A and B. But for all $(t, s) \in A$ and all $(x, y) \in B$, we have

$$t^2 + s^2 \leq \frac{4r^2}{\pi} \leq x^2 + y^2 \Rightarrow \exp(-(t^2 + s^2)) \geq \exp(-(x^2 + y^2))$$

Therefore

$$\begin{aligned}
\frac{4}{\pi} \int_S e^{-(t^2+s^2)} dt ds &= \frac{4}{\pi} \left(\int_{S \cap D(2r/\sqrt{\pi})} e^{-(t^2+s^2)} dt ds + \int_B e^{-(t^2+s^2)} dt ds \right) \\
&\leq \frac{4}{\pi} \left(\int_{S \cap D(2r/\sqrt{\pi})} e^{-(t^2+s^2)} dt ds + \int_A e^{-(t^2+s^2)} dt ds \right) \\
&= \frac{4}{\pi} \int_{D(2r/\sqrt{\pi})} e^{-(t^2+s^2)} dt ds = 1 - \exp\left(-\frac{4r^2}{\pi}\right)
\end{aligned}$$

This proves the statement for $\sigma^2 = 1/2$. For $\sigma^2 \in (0, \infty)$, note that $Z \overset{d}{=} \sqrt{2\sigma^2} \tilde{Z}$ where $\tilde{Z} \sim \mathcal{N}(0, 1/2)$, so

$$F_{|Z|}(r) = \mathrm{P}(-r \leq Z \leq r) = \mathrm{P}\left(-\frac{r}{\sqrt{2\sigma^2}} \leq \tilde{Z} \leq \frac{r}{\sqrt{2\sigma^2}}\right) = F_{|\tilde{Z}|}\left(\frac{r}{\sqrt{2\sigma^2}}\right),$$

and applying the result for $\sigma^2 = 1/2$ yields the general result. ∎

Lemma 28 *Let X, Y be random vectors such that $X \mid Y = y \sim \mathcal{N}(y, \Sigma)$ for some fixed $\Sigma \in \mathbb{S}_{++}^d$ and for all y in the image of Y. Then $X - Y \sim \mathcal{N}(0, \Sigma)$.*

Proof Let B be a Borel subset of $\mathbb{R}^d$, and let $Z \sim \mathcal{N}(0, \Sigma)$. We have

$$\mathrm{P}(X - Y \in B) = \mathrm{E}[\mathrm{P}(X - Y \in B \mid Y)] = \mathrm{E}[\mathrm{P}(Z \in B)] = \mathrm{P}(Z \in B).$$

∎

Lemma 29 (Anderson's Lemma) *Let $Z \sim \mathcal{N}(0, \Sigma)$ for some $\Sigma \in S_{++}^d$ and $d \in \mathbb{N}$. Let $A \subset \mathbb{R}^d$ be a convex symmetric set. Then for all $a \in \mathbb{R}^d$, we have*

$$\mathrm{P}(Z \in A) \geq \mathrm{P}(Z \in A + a).$$

Lemma 30 (Gaussian Poincaré/Concentration) *Let $d \in \mathbb{N}$, $Z \sim \mathcal{N}(0, I_{d \times d})$, and $f : \mathbb{R}^d \to \mathbb{R}$ be an L-Lipschitz function with respect to the Euclidean metric. Then $\mathrm{Var}[f(Z)] \leq L^2$ and*

$$\mathrm{P}(f(Z) - \mathrm{E}[f(Z)] \geq t) \leq \exp\left(-\frac{t^2}{2L^2}\right).$$

A.3.1. CONCENTRATION OF NORMS OF GAUSSIAN VECTORS

For this subsection, fix $d \in \mathbb{N}$, an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^d$, and a covariance matrix $\Sigma \in S_{++}^d$. Let $Z \sim \mathcal{N}(0, \Sigma)$, and define $M := \mathrm{E}[\|Z\|]$. Let S denote the unit sphere of the dual norm $\|\cdot\|_*$, and recall that $\|x\| = \|x\|_{**} = \sup_{v \in S} |\langle v, x \rangle|$. Define $R := \sup_{v \in S} v^T \Sigma v$, and $v_* = \max_{v \in S} v^T \Sigma v$ where the maximum is attained since S is compact and the function is continuous.

Lemma 31 *The function $f : \mathbb{R}^d \to \mathbb{R}$ given by $f(x) = \|\Sigma^{1/2} x\|$ is $\sqrt{R}$-Lipschitz in the Euclidean metric.*

Proof

$$|f(x) - f(y)| = \left| \|\Sigma^{1/2} x\| - \|\Sigma^{1/2} y\| \right| \leq \|\Sigma^{1/2}(x - y)\|$$

$$= \sup_{v \in S} \langle \Sigma^{1/2} v, x - y \rangle \leq \left(\max_{v \in S} \|\Sigma^{1/2} v\|_2 \right) \|x - y\|_2$$

∎

Lemma 32 *For all $t \geq 0$,*

$$M^2 \leq \mathrm{E}[\|Z\|^2] \leq \left(1 + \frac{\pi}{2} \right) M^2,$$

$$\mathrm{P}(M - \|Z\| \geq t) \leq \exp\left(-\frac{t^2}{\pi M^2} \right).$$

Proof Notice that

$$M = \mathrm{E}[\|Z\|] = \mathrm{E}\left[\sup_{v \in S} |\langle v, Z \rangle| \right] \geq \sup_{v \in S} \mathrm{E}[|\langle v, Z \rangle|] = \sup_{v \in S} \sqrt{\frac{2}{\pi} v^T \Sigma v} = \sqrt{\frac{2}{\pi} v_*^T \Sigma v_*}.$$

where the inequality follows by convexity of the supremum and Jensen's inequality, and the third equality by the fact that $\langle v, Z \rangle \sim \mathcal{N}(0, v^T \Sigma v)$ and an explicit calculation of the expectation. We now prove the first item. The first inequality follows from Jensen's inequality. For the second, notice that $\Sigma^{-1/2} Z \sim \mathcal{N}(0, I_{d \times d})$, so that an application of Lemmas 30 and 31 yields

$$\mathrm{E}[\|Z\|^2] - (\mathrm{E}[\|Z\|])^2 = \mathrm{Var}[\|Z\|] = \mathrm{Var}\left[f(\Sigma^{-1/2} Z) \right] \leq R = v_*^T \Sigma v_* \leq \frac{\pi}{2} (\mathrm{E}[\|Z\|])^2.$$

where f is as defined in Lemma 31. For the second item, notice that $-f$ is also $\sqrt{R}$-Lipschitz, so that again an application of Lemma 30 yields

$$\mathrm{P}(M - \|Z\| > t) = \mathrm{P}\left(-f(\Sigma^{-1/2} Z) - \mathrm{E}\left[-f(\Sigma^{-1/2} Z) \right] > t \right)$$

$$\leq \exp\left(-\frac{t^2}{2 v_*^T \Sigma v_*} \right) \leq \exp\left(-\frac{t^2}{\pi (\mathrm{E}[\|Z\|])^2} \right).$$

∎

Corollary 33 *For all $r \in \mathbb{R}$,*

$$l(r) \leq F_{\|Z\|}(r) \leq \min\{u_1(r), u_2(r)\}$$

where

$$l(r) := \left[1 - \exp\left(-\frac{(r-M)^2}{2R}\right)\right] \mathbb{1}_{[M,\infty)},$$

$$u_1(r) := \exp\left(\frac{-(M-r)^2}{\pi M^2}\right) \mathbb{1}_{[0,M)}(r) + \mathbb{1}_{[M,\infty)}(r), \quad u_2(r) := \sqrt{1 - \exp\left(-\frac{2r^2}{\pi R}\right)} \mathbb{1}_{[0,\infty)}(r).$$

Proof We start with the lower bound. Let $r \geq M$. Then

$$F_{\|Z\|}(r) = \mathrm{P}(\|Z\| \leq r) = 1 - \mathrm{P}(\|Z\| > r) = 1 - \mathrm{P}(\|Z\| - M > r - M) \geq 1 - \exp\left(-\frac{(r-M)^2}{2R}\right)$$

where the last inequality follows from Lemmas 30 and 31. For the lower bound, we have, by Lemma 32

$$F_{\|Z\|}(r) = \mathrm{P}(\|Z\| \leq r) = \mathrm{P}(\|Z\| - M \leq r - M) = \mathrm{P}(M - \|Z\| \geq M - r) \leq u_1(r).$$

Furthermore,

$$F_{\|Z\|}(r) = \mathrm{P}\left(\sup_{v \in S} |\langle v, Z \rangle| \leq r\right) \leq \mathrm{P}(|\langle v_*, Z \rangle| \leq r) \leq u_2(r),$$

where the second inequality follows from the fact that $\langle v_*, Z \rangle \sim \mathcal{N}(0, R)$ and Lemma 27. ∎

A.4. Inverse Gamma measure

Let $\alpha, \beta > 0$. The inverse gamma measure Inv-Gamma(α, β) on $(0, \infty)$ has density

$$f_{\alpha,\beta}(x) := \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right)$$

with respect to Lebesgue measure, where Γ is the gamma function.

Lemma 34 *Let $X \sim$ Inv-Gamma(α, β) and $Z \sim$ Inv-Gamma(α, α). Let $r > 0$, and define*

$$x_{\alpha,\beta}(r) := \frac{\beta[\exp(r) - \exp(-r)]}{2\alpha r} \qquad p_\alpha(r) := \mathrm{P}\left(\frac{1 - \exp(-2r)}{2r} \leq Z \leq \frac{\exp(2r) - 1}{2r}\right)$$

Then for all $x \in (0, \infty)$

$$p_\alpha(r) = \mathrm{P}(|\log(X/x_{\alpha,\beta}(r))| \leq r) \geq \mathrm{P}(|\log(X/x)| \leq r).$$

26

Proof Fix $r \in (0, \infty)$. Define $h_r(x) := P(|\log(X/x)| \leq r)$. Then we have

$$\begin{aligned}
\frac{d}{dx}(h_r(x)) &= \frac{d}{dx}(P(|\log(X/x)| \leq r)) \\
&= \frac{d}{dx}(P(x\exp(-r) \leq X \leq x\exp(r))) \\
&= \frac{d}{dx}\left(\int_{x\exp(-r)}^{x\exp(r)} f_{\alpha,\beta}(t)dt \right) \\
&= \exp(r)f_{\alpha,\beta}(x\exp(r)) - \exp(-r)f_{\alpha,\beta}(x\exp(-r))
\end{aligned}$$

where in the last line we used Leibniz integral rule. Setting the derivative to 0 and solving yields $x_{\alpha,\beta}(r)$. Examining its derivative, we notice that h_r is non-decreasing on $(0, x_{\alpha,\beta}(r)]$ and non-increasing on $[x_{\alpha,\beta}(r), \infty)$. Therefore $x_{\alpha,\beta}(r)$ is the global maximizer of h_r. Now

$$\begin{aligned}
P(|\log(X/x_{\alpha,\beta}(r))| \leq r) &= P(x_{\alpha,\beta}(r)\exp(-r) \leq X \leq x_{\alpha,\beta}(r)\exp(r)) \\
&= P\left(\frac{1-\exp(-2r)}{2r} \leq \frac{\alpha}{\beta}X \leq \frac{\exp(2r)-1}{2r} \right) \\
&= p_\alpha(r)
\end{aligned}$$

where in the last line we used that if $X \sim$ Inv-Gamma(α, β), then $cX \sim$ Inv-Gamma$(\alpha, c \cdot \beta)$ for all $c > 0$. $\blacksquare$

Lemma 35 *Let p_α be as defined in Lemma 34. The following holds.*

- $p_\alpha(r)$ *is non-decreasing in α for all $r > 0$.*

- $p_\alpha(r)$ *is strictly increasing in r for all $\alpha > 0$ and $\text{Im}(p_\alpha) = (0, 1)$*

Appendix B. Suprema of truncated empirical processes

B.1. Truncation function

Let $\alpha, \beta \in \mathbb{R}$ such that $\alpha \leq \beta$. Define

$$\phi_{\alpha,\beta}(x) := \begin{cases} \beta & \text{if} \quad x > \beta, \\ x & \text{if} \quad x \in [\alpha, \beta], \\ \alpha & \text{if} \quad x < \alpha. \end{cases} \tag{9}$$

Lemma 36 *The following holds for all $x \in \mathbb{R}$.*

- $c \cdot \phi_{\alpha,\beta}(x) = \phi_{c\alpha,c\beta}(cx)$ *for all $c \in [0, \infty)$.*

- $-\phi_{\alpha,\beta}(x) = \phi_{-\beta,-\alpha}(-x)$.

- $\phi_{\alpha,\beta}(x) + y = \phi_{\alpha+y,\beta+y}(x+y)$ *for all $y \in \mathbb{R}$.*

Proof Just check the three possible cases for each item. ∎

Fix $n \in \mathbb{N}$. For a real valued sequence $a := (a_i)_{i=1}^n$, define the sequence $a^* = (a_i^*)_{i=1}^n$ by $a_i^* := a_{\pi(i)}$ for all $i \in [n]$ and where $\pi : [n] \to [n]$ is a permutation that orders a non-decreasingly, i.e. $a_{\pi(1)} \leq \ldots \leq a_{\pi(n)}$. Note that this is well-defined since any such permutation gives the same a^*. Addition and scalar multiplication of sequences are as usual. For two sequences a, b, we say that $a \leq b$ if $a_i \leq b_i$ for all $i \in [n]$.

Lemma 37 *Let $a = (a_i)_{i=1}^n$ and $b = (b_i)_{i=1}^n$ be real valued sequences.*

$$a \leq b \Rightarrow a^* \leq b^*$$

Proof Let π and σ be permutations of $[n]$ that order a and b non-decreasingly, respectively. Let $i \in [n]$. We show that $a_{\pi(i)} \leq b_{\sigma(i)}$. We consider two cases. If $\pi(i) \in \{\sigma(1), \ldots, \sigma(i)\}$, then $a_{\pi(i)} \leq b_{\pi(i)} \leq b_{\sigma(i)}$. Otherwise, $\pi(i) \in \{\sigma(i+1), \ldots, \sigma(n)\}$. This implies that there exists a $j \in \{i+1, \ldots, n\}$ such that $\pi(j) \in \{\sigma(1), \ldots, \sigma(i)\}$, from which we conclude that $a_{\pi(i)} \leq a_{\pi(j)} \leq b_{\pi(j)} \leq b_{\sigma(i)}$. ∎

Let $k \in \{1, \ldots, \lfloor n/2 \rfloor\}$. Define

$$\varphi_k(a) := \sum_{i=1}^n \phi_{a_{1+k}^*, a_{n-k}^*}(a_i). \tag{10}$$

Lemma 38 *The following holds for all real-valued sequences $a = (a_i)_{i=1}^n$.*

- $c \cdot \varphi_k(a) = \varphi_k(c \cdot a)$ *for all $c \in \mathbb{R}$.*

- $\varphi_k(a) + n \cdot c = \varphi_k(a + c)$ *for all $c \in \mathbb{R}$.*

- $\varphi_k(a) \leq \varphi_k(b)$ *for all sequences $b = (b_i)_{i=1}^n$ such that $a \leq b$.*

Proof We start with the first item. Let $\pi : [n] \to [n]$ be a permutation that orders a non-decreasingly. The case $c = 0$ is trivial. Now consider the case $c > 0$. Then since $c > 0$, π also orders $c \cdot a$ non-decreasingly. Therefore $(c \cdot a)_i^* = c \cdot a_i^*$ and

$$c \cdot \varphi_k(a) = \sum_{i=1}^n c \cdot \phi_{a_k^*, a_{n-k}^*}(a_i) = \sum_{i=1}^n \phi_{c \cdot a_k^*, c \cdot a_{n-k}^*}(c \cdot a_i) = \sum_{i=1}^n \phi_{(c \cdot a)_k^*, (c \cdot a)_{n-k}^*}(c \cdot a_i) = \varphi_k(c \cdot a),$$

where the second equality follows from the first item of Lemma 36. Now consider the case $c = -1$. Then the permutation π orders $-a$ non-increasingly so that $(-a)_i^* = -a_{n-i}^*$ and

$$-\varphi_k(a) = \sum_{i=1}^n -\phi_{a_k^*, a_{n-k}^*}(a_i) = \sum_{i=1}^n \phi_{-a_{n-k}^*, -a_k^*}(-a_i) = \sum_{i=1}^n \phi_{(-a)_k^*, (-a)_{n-k}^*}(-a_i) = \varphi_k(-a),$$

where the second equality follows from the second item of Lemma 37. For the case $c < 0$, we have

$$c \cdot \varphi_k(a) = (-c) \cdot -\varphi_k(a) = (-c) \cdot \varphi_k(-a) = \varphi_k(c \cdot a).$$

For the second item, we have by Lemma 37 that $a^* \leq b^*$, from which we conclude

$$\varphi_k(a) = \sum_{i=1}^{n} \phi_{a_k^*, a_{n-k}^*}(a_i) = ka_k^* + \sum_{i=k+1}^{n-k-1} a_i^* + ka_{n-k}^* \leq kb_k^* + \sum_{i=k+1}^{n-k-1} b_i^* + kb_{n-k}^* = \varphi_k(b)$$

∎

Lemma 39 *Let $a = (a_i)_{i=1}^{n}$ and $b = (b_i)_{i=1}^{n}$ be real valued sequences such that $b \geq 0$. Then*

$$\varphi_k(a + b) \geq \varphi_k(a) + \sum_{i=1}^{n-2k} b_i^*$$

Proof Let π and σ be permutations that order $a + b$ and a non-decreasingly, respectively. By definition

$$\varphi_k(a + b) = k(a + b)_{1+k}^* + \sum_{i=k+1}^{n-k} (a + b)_i^* + k(a + b)_{n-k}^*.$$

We lower bound each of the three terms separately. For the first, define the sets $I_1 := \{\pi(1), \ldots, \pi(1 + k)\}$ and $J_1 := \{\sigma(1 + k), \ldots, \sigma(n)\}$, and notice that

$$|I_1 \cap J_1| = |I_1| + |J_1| - |I_1 \cup J_1| \geq (1 + k) + (n - k) - n = 1.$$

Therefore, we have

$$(a+b)_{1+k}^* = a_{\pi(1+k)} + b_{\pi(1+k)} = \max\{a_i + b_i \mid i \in I_1\} \geq \max\{a_i + b_i \mid i \in I_1 \cap J_1\} \geq a_{\sigma(1+k)},$$

where the last inequality uses the non-negativity of b. Similarly, for the third term, define the sets $I_2 := \{\pi(1), \ldots, \pi(n - k)\}$ and $J_2 := \{\sigma(n - k), \ldots, \sigma(n)\}$, and notice that

$$|I_2 \cap J_2| = |I_2| + |J_2| - |I_2 \cup J_2| \geq (n - k) + (1 + k) - n = 1.$$

Therefore, we get

$$(a+b)_{n-k}^* = a_{\pi(n-k)} + b_{\pi(n-k)} = \max\{a_i + b_i \mid i \in I_2\} \geq \max\{a_i + b_i \mid i \in I_2 \cap J_2\} \geq a_{\sigma(n-k)},$$

where again we used the non-negativity of b in the last inequality. It remains to lower bound the second term. Let $S^* \subset I_2$ such that $|S| = k$ and $\{\sigma(1), \ldots, \sigma(k)\} \cap I_2 \subset S$. Notice that

$$\begin{aligned}
\sum_{i=k+1}^{n-k} (a + b)_i^* &= \sum_{i=k+1}^{n-k} (a_{\pi(i)} + b_{\pi(i)}) \\
&= \max_{\substack{S \subset I_2 \\ |S|=k}} \sum_{i \in I_2 \setminus S} (a_i + b_i) \\
&\geq \sum_{i \in I_2 \setminus S^*} a_i + \sum_{i \in I_2 \setminus S^*} b_i
\end{aligned}$$

29

Let us further bound each term. For the first, notice that by definition of S^*, we have $(I_2 \setminus S^*) \subset J_1$ and $|I_2 \setminus S^*| = n - 2k$, therefore

$$\sum_{i \in I_2 \setminus S^*} a_i \geq \min_{\substack{T \subset J_1 \\ |T| = n - 2k}} \sum_{i \in T} a_i = \sum_{i=k+1}^{n-k} a_{\sigma(i)}.$$

For the second, we have

$$\sum_{i \in I_2 \setminus S^*} b_i \geq \min_{\substack{T \subset [n] \\ |T| = n - 2k}} \sum_{i \in T} b_i = \sum_{i=1}^{n-2k} b_i^*.$$

Combining the bounds yields the desired result. ∎

B.2. Suprema of truncated empirical processes

Let $\mathcal{T}$ be a countable index set, and let $(\{Z_{i,s}\}_{s \in \mathcal{T}})_{i=1}^n$ be independent real-valued $\mathcal{T}$-indexed stochastic processes. Define $Z := \sup_{s \in \mathcal{T}} \sum_{i=1}^n Z_{i,s}$. For $s \in \mathcal{T}$, define $Z_s := (Z_{i,s})_{i=1}^n$. For i.i.d. Rademacher random variables $(\varepsilon_i)_{i=1}^n$, define $\mu := \mathrm{E}[\sup_{s \in \mathcal{T}} \sum_{i=1}^n \varepsilon_i Z_{i,s}]$. We assume throughout that $\sigma^2 := \sup_{s \in \mathcal{T}} \sum_{i=1}^n \mathrm{E}\left[Z_{i,s}^2\right] < \infty$. We start by recalling the following result.

Lemma 40 ([Klein and Rio](2005)) *Assume that for all $s \in \mathcal{T}$ and $i \in [n]$, $\mathrm{E}[Z_{i,s}] = 0$, and that $R := \sup_{(s,i) \in \mathcal{T} \times [n]} \|Z_{i,s}\|_\infty < \infty$. Define $v := 2R\,\mathrm{E}[Z] + \sigma^2$. Then*

$$\mathrm{P}(Z \geq \mathrm{E}[Z] + t) \leq \exp\left(-\frac{4v}{9R^2} h\left(\frac{3Rt}{2v}\right)\right),$$

where $h(t) := 1 + t - \sqrt{1 + 2t}$ with inverse $h^{-1}(t) = t + \sqrt{2t}$. Consequently, with probability at least $1 - \delta$

$$Z < \mathrm{E}[Z] + \frac{3R \log(1/\delta)}{2} + \sqrt{2v \log(1/\delta)}.$$

The following result is due to [Lugosi and Mendelson](2021).

Lemma 41 *Let $T > 0$. Then with probability at least $1 - \delta$*

$$\sup_{s \in \mathcal{T}} |\{i \in [n] \mid |Z_{i,s}| > T\}|$$

$$< \inf_{\varepsilon \in (0,1)} \left\{ \frac{2\mu}{\varepsilon T} + \frac{\sigma^2}{(1-\varepsilon)^2 T^2} + \sqrt{\left(\frac{8\mu}{\varepsilon T} + \frac{2\sigma^2}{(1-\varepsilon)^2 T^2}\right) \log(1/\delta)} + \frac{3 \log(1/\delta)}{2} \right\}.$$

Proof Let $T > 0$ and $\varepsilon \in (0,1)$, and define the function $\chi_{T,\varepsilon} : \mathbb{R} \to [0,1]$ by

$$\chi_{T,\varepsilon}(x) := \begin{cases} 0 & \text{if } x \leq (1-\varepsilon)T \\ \dfrac{x}{\varepsilon T} - \dfrac{1-\varepsilon}{\varepsilon} & \text{if } x \in ((1-\varepsilon)T, T] \\ 1 & \text{if } x > T. \end{cases}$$

Note that $\mathbb{1}_{(T,\infty)} \le \chi_{T,\varepsilon} \le \mathbb{1}_{((1-\varepsilon)T,\infty)}$ and $\chi_{T,\varepsilon}$ is $(1/\varepsilon T)$-Lipschitz. Now we have

$$\sup_{s\in\mathcal{T}}\sum_{i=1}^{n}\mathbb{1}_{(T,\infty)}(|Z_{i,s}|) \le \sup_{s\in\mathcal{T}}\sum_{i=1}^{n}\chi_{T,\varepsilon}(|Z_{i,s}|)$$

$$\le \sup_{s\in\mathcal{T}}\sum_{i=1}^{n}\underbrace{\chi_{T,\varepsilon}(|Z_{i,s}|) - \mathrm{E}[\chi_{T,\varepsilon}(|Z_{i,s}|)]}_{W_{i,s}:=} + \sup_{s\in\mathcal{T}}\sum_{i=1}^{n}\mathrm{E}[\chi_{T,\varepsilon}(|Z_{i,s}|)] \quad (11)$$

The second term of (11) is bounded by

$$\sup_{s\in\mathcal{T}}\sum_{i=1}^{n}\mathrm{E}[\chi_{T,\varepsilon}(|Z_{i,s}|)] \le \sup_{s\in\mathcal{T}}\sum_{i=1}^{n}\mathrm{P}(|Z_{i,s}| > (1-\varepsilon)T) \le \frac{\sigma^2}{(1-\varepsilon)^2 T^2}. \quad (12)$$

We now turn to the first term of (11) which we denote by W. We note that $\mathrm{E}[W_{i,s}] = 0$, $|W_{i,s}| \le 1$, so by Lemma 28 we have with probability at least $1 - \delta$

$$W < \mathrm{E}[W] + \frac{3\log(1/\delta)}{2} + \sqrt{2(2\,\mathrm{E}[W] + \alpha^2)\log(1/\delta)}, \quad (13)$$

where $\alpha^2 := \sup_{s\in\mathcal{T}}\sum_{i=1}^{n}\mathrm{E}[W_{i,s}^2]$. It remains to bound $\mathrm{E}[W]$ and α^2. The former is bounded by

$$\mathrm{E}[W] = \mathrm{E}\left[\sup_{s\in\mathcal{T}}\sum_{i=1}^{n}\chi_{T,\varepsilon}(|Z_{i,s}|) - \mathrm{E}[\chi_{T,\varepsilon}(|Z_{i,s}|)]\right]$$

$$\le 2\,\mathrm{E}\left[\sup_{s\in\mathcal{T}}\sum_{i=1}^{n}\varepsilon_i\chi_{T,\varepsilon}(|Z_{i,s}|)\right] \le \frac{2}{\varepsilon T}\,\mathrm{E}\left[\sup_{s\in\mathcal{T}}\sum_{i=1}^{n}\varepsilon_i Z_{i,s}\right], \quad (14)$$

where the first inequality is by symmetrization and the second by the contraction principle and the $(1/\varepsilon T)$-Lipschitzness of $\chi_{T,\varepsilon}\circ|\cdot|$. The latter is bounded by

$$\alpha^2 = \sup_{s\in\mathcal{T}}\sum_{i=1}^{n}\mathrm{E}\big[W_{i,s}^2\big] \le \sup_{s\in\mathcal{T}}\sum_{i=1}^{n}\mathrm{E}\big[\chi_{T,\varepsilon}^2(|Z_{i,s}|)\big] \le \sup_{s\in\mathcal{T}}\sum_{i=1}^{n}\mathrm{P}(|Z_{i,s}| > (1-\varepsilon)T) \le \frac{\sigma^2}{(1-\varepsilon)^2 T^2}. \quad (15)$$

Combining (13), (14), and (15) yields that with probability at least $1 - \delta$

$$W < \frac{2\mu}{\varepsilon T} + \sqrt{\left(\frac{8\mu}{\varepsilon T} + \frac{2\sigma^2}{(1-\varepsilon)^2 T^2}\right)\log(1/\delta)} + \frac{3\log(1/\delta)}{2} \quad (16)$$

Combining (11), (12), (16), and optimizing over $\varepsilon \in (0,1)$ yields the result. ∎

Corollary 42 *Using the same notation as in Lemma 29, we have with probability at least $1 - \delta$*

$$\sup_{s\in\mathcal{T}}|\{i \in [n] \mid |Z_{i,s}| > T_0\}| < 8\log(1/\delta),$$

where

$$T_0 := 2\max\left\{\frac{\mu}{\log(1/\delta)}, \sqrt{\frac{\sigma^2}{\log(1/\delta)}}\right\}.$$

Proof The result follows from taking $\varepsilon = 1/2$ in the bound of Lemma 29, replacing T by T_0, and straightforwardly bounding the resulting expression. ∎

Lemma 43 *Let $\delta \in (0,1)$ be such that $k := 8\log(2/\delta)$ is an integer satisfying $1 \le k \le \lfloor n/2 \rfloor$. Assume that for all $s \in \mathcal{T}$ and $i \in [n]$, $\mathrm{E}[Z_{i,s}] = 0$. Then with probability at least $1 - \delta$*

$$\sup_{s \in \mathcal{T}} |\{i \in [n] \mid |Z_{i,s}| > T_0\}| < k,$$

and

$$\sup_{s \in \mathcal{T}} \varphi_k(Z_s) \le 50 \max\Big\{\mu, \sqrt{\sigma^2 \log(2/\delta)}\Big\}.$$

where

$$T_0 := 2 \max\left\{\frac{\mu}{\log(1/\delta)}, \sqrt{\frac{\sigma^2}{\log(1/\delta)}}\right\}.$$

Proof By Lemma 29, with probability at least $1 - \delta/2$, we have for all $s \in \mathcal{T}$

$$-T_0 \le Z^*_{1+k,s} \le Z^*_{n-k,s} \le T_0.$$

Therefore

$$\begin{aligned}
\sup_{s \in \mathcal{T}} \varphi_k(Z_s) &= \sup_{s \in \mathcal{T}} \sum_{i=1}^n \phi_{Z^*_{1+k,s}, Z^*_{n-k,s}}(Z_{i,s}) \\
&= \sup_{s \in \mathcal{T}} \sum_{i=1}^n \phi_{-T_0, T_0}(Z_{i,s}) + k(Z_{1+k,s} + T_0) + \underbrace{k(Z^*_{n-k,s} - T_0)}_{\le 0} \\
&\le \sup_{s \in \mathcal{T}} \sum_{i=1}^n \underbrace{\phi_{-T_0, T_0}(Z_{i,s}) - \mathrm{E}[\phi_{-T_0, T_0}(Z_{i,s})]}_{W_{i,s} :=} + \sup_{s \in \mathcal{T}} \sum_{i=1}^n \mathrm{E}[\phi_{-T_0, T_0}(Z_{i,s})] + 2kT_0.
\end{aligned}$$

$$(17)$$

We now bound the second term of (17) by

$$\sup_{s \in \mathcal{T}} \sum_{i=1}^n \mathrm{E}[\phi_{-T_0, T_0}(Z_{i,s})] = \sup_{s \in \mathcal{T}} \sum_{i=1}^n \underbrace{\mathrm{E}[Z_{i,s}]}_{= 0} + \underbrace{\mathrm{E}\big[(T_0 - Z_{i,s})\mathbb{1}_{(T_0, \infty)}(Z_{i,s})\big]}_{\le 0} \qquad (18)$$

$$+ \mathrm{E}\big[(-T_0 - Z_{i,s})\mathbb{1}_{(-\infty, -T_0)}(Z_{i,s})\big]$$

$$\le \sup_{s \in \mathcal{T}} \sum_{i=1}^n \mathrm{E}\big[Z_{i,s}^2\big]^{1/2} \cdot \mathrm{P}(Z_{i,s} < -T_0)^{1/2} \le \frac{\sigma^2}{T_0}, \qquad (19)$$

where we used the Cauchy-Schwarz inequality and Markov's inequality respectively. Denote the first term of (17) by W, and note that $\mathrm{E}[W_{i,s}] = 0$ and $|W_{i,s}| \le 2T_0$, so by Lemma 28 we have with probability at least $1 - \delta/2$

$$W < \mathrm{E}[W] + 2T_0 \log(2/\delta) + \sqrt{2(4T_0 \mathrm{E}[W] + \alpha^2) \log(2/\delta)}, \qquad (20)$$

where $\alpha := \sup_{s \in \mathcal{T}} \sum_{i=1}^{n} \mathrm{E}\left[W_{i,s}^2\right]$. It remains to bound $\mathrm{E}[W]$ and α^2. The former is bounded by

$$
\mathrm{E}[W] = \mathrm{E}\left[\sup_{s \in \mathcal{T}} \sum_{i=1}^{n} \phi_{-T_0,T_0}(Z_{i,s}) - \mathrm{E}[\phi_{-T_0,T_0}(Z_{i,s})]\right]
$$
$$
\leq 2\,\mathrm{E}\left[\sup_{s \in \mathcal{T}} \sum_{i=1}^{n} \varepsilon_i \phi_{-T_0,T_0}(Z_{i,s})\right] \leq 2\,\mathrm{E}\left[\sup_{s \in \mathcal{T}} \sum_{i=1}^{n} \varepsilon_i Z_{i,s}\right], \quad (21)
$$

where we used symmetrization and the contraction principle along with the 1-Lipschitzness of ϕ_{-T_0,T_0} respectively. The latter is bounded by

$$
\alpha^2 \leq \sup_{s \in \mathcal{T}} \sum_{i=1}^{n} \mathrm{E}\left[W_{i,s}^2\right] \leq \sup_{s \in \mathcal{T}} \sum_{i=1}^{n} \mathrm{E}\left[\phi_{-T_0,T_0}^2(Z_{i,s})\right] \leq \sup_{s \in \mathcal{T}} \sum_{i=1}^{n} \mathrm{E}\left[Z_{i,s}^2\right] \quad (22)
$$

Combining (20), (21), (22), and using the definition of T_0, we obtain with probability at least $1-\delta/2$

$$
W < 16 \max\left\{\mu, \sqrt{\sigma^2 \log(1/\delta)}\right\} \quad (23)
$$

Combining (17), (19), (16), and the definition of T_0 yields the result. ∎

Appendix C. Proofs of Section 4

C.1. Proof of Theorem 13

By definition of the minimax risk, we have

$$
R_\delta^*(\ell) = \inf_d R_\delta(\ell, d) = \inf_d \sup_{P \in \mathcal{P}} R_\delta(\ell, P, d) = \inf_d \sup_{P \in \mathcal{P}} Q_{\ell(P,d(O))}(1-\delta) = \inf_d \sup_{P \in \mathcal{P}} F_{\ell(P,d(O))}^{-}(1-\delta).
$$

Applying the sixth item of Lemma 20 to the last expression yields

$$
R_\delta^*(\ell) = \inf_d \left(\inf_{P \in \mathcal{P}} F_{\ell(P,d(O))}\right)^{-}(1-\delta).
$$

Now let $k \in \mathbb{N}$. Since $\inf_{P \in \mathcal{P}} F_{\ell(P,d(O))} \leq \mathrm{E}_{P \sim \pi_k}\left[F_{\ell(P,d(O))|P}\right] = F_{\ell(P,d(O))}^{\pi_k}$, where $O \mid P \sim P$ inside the expectation, we have by the second item of Lemma 20

$$
R_\delta^*(\ell) \geq \inf_d \left(F_{\ell(P,d(O))}^{\pi_k}\right)^{-1}(1-\delta) \geq \left(\sup_d F_{\ell(P,d(O))}^{\pi_k}\right)^{-}(1-\delta) = p_{\ell,k}^{-}(1-\delta).
$$

where the second inequality follows from the third item of Lemma 20, and the last by definition of $p_{\ell,k}$. Taking supremum over k, and combining our assumptions on the sequence $(p_{\ell,k})_{k \in \mathbb{N}}$ with the last item of Lemma 20 yields the result.

C.2. Proof of Proposition 14

The first statement follows from the assumption on φ and Lemma 21. For the second statement, define $S := \{R_\delta(\ell, P, d) \mid P \in \mathcal{P}\} \subset [-\infty, \infty)$ and $x_0 := \sup S$. If $x_0 = -\infty$, then $\varphi(x_0) = -\infty$, and $\varphi(\ell(P, d(O)))) = \ell(P, d(O)) = -\infty$ with probability at least $1 - \delta$ for all P, so the statement holds. Otherwise $x_0 \in \mathbb{R}$. Now for any $x \in S$, we have $x \leq x_0$, so $\varphi(x) \leq \varphi(x_0)$, and hence $\sup_{x \in S} \varphi(x) \leq \varphi(x_0)$. On the other hand, let $(x_k)_{k \in \mathbb{N}}$ be an increasing sequence in S such that $x_k \to x_0$ as $k \to \infty$. Then by the left-continuity of φ, we obtain $\sup_{x \in S} \varphi(x) \geq \lim_{k \to \infty} \varphi(x_k) = \varphi(x_0)$, which proves the statement. For the last statement, suppose that $d^* \in \operatorname{argmin}_d R_\delta(\ell, d)$, then by assumption $R_\delta(\ell, d^*) < \infty$, so that by the second statement $R_\delta(\varphi \circ \ell, d^*) = \varphi(R_\delta(\ell, d^*))$. Now let d be any other decision rule. If $R_\delta(\ell, d) < \infty$, then by the minimality of d^* we get $R_\delta(\ell, d^*) \leq R_\delta(\ell, d)$, and since φ is increasing and using the second statement again, $R_\delta(\varphi \circ \ell, d^*) = \varphi(R_\delta(\ell, d^*)) \leq \varphi(R_\delta(\ell, d)) = R_\delta(\varphi \circ \ell, d)$. If $R_\delta(\ell, d) = \infty$ then there exists $P_0 \in \mathcal{P}$ such that $R_\delta(\ell, P_0, d) \geq R_\delta(\ell, d^*)$, but then since φ is increasing, $R_\delta(\varphi \circ \ell, d) = \sup_{P \in \mathcal{P}} R_\delta(\varphi \circ \ell, P, d) \geq \varphi(R_\delta(\ell, P_0, d)) \geq \varphi(R_\delta(\ell, d^*)) = R_\delta(\varphi \circ \ell, d^*)$. This proves the last statement.

C.3. Proof of Proposition 15

Proof We present here the proof for the case $\varphi(x) = x$. The general statement follows from Proposition 14. Our aim is to apply Theorem 13. We select $\pi_k := \mathcal{N}(0, \Sigma/\lambda_k)$ for a decreasing strictly positive sequence $(\lambda_k)_{k \in \mathbb{N}}$ satisfying $\lambda_k \to 0$ as $k \to \infty$. We want to compute, for all $t \in \mathbb{R}$,

$$p_{\ell,k}(t) = \sup_{\hat{\mu}} \mathrm{P}(e(\hat{\mu}((X_i)_{i=1}^n - \mu)) \leq t),$$

where $\mu \sim \pi_k$ and $X_i \mid \mu \sim \mathcal{N}(\mu, \Sigma)$ for all $i \in [n]$ independently. A classical Bayesian calculation shows that $\mu \mid (X_i)_{i=1}^n \sim \mathcal{N}(\overline{X}_k, \Sigma_k)$ where $\overline{X}_k := \frac{n}{n+\lambda_k}\overline{X}$ and $\Sigma_k := \frac{1}{n+\lambda_k}\Sigma$, where $\overline{X} := n^{-1} \sum_{i=1}^n X_i$ is the sample mean. Now we compute, for $Z_k \sim \mathcal{N}(0, \Sigma_k)$,

$$
\begin{aligned}
p_{\ell,k}(t) &= \sup_{\hat{\mu}} \mathrm{P}(e(\hat{\mu}((X_i)_{i=1}^n - \mu)) \leq t) \\
&= \mathrm{E}\left[\sup_{a \in \mathbb{R}^d} \mathrm{P}(e(\mu - a) \leq t \mid (X_i)_{i=1}^n)\right] \\
&= \mathrm{E}\left[\sup_{a \in \mathbb{R}^d} \mathrm{P}\big(\mu - a \in e^{-1}((-\infty, t]) \mid (X_i)_{i=1}^n\big)\right] \\
&= \mathrm{E}\big[\mathrm{P}\big(\mu - \overline{X}_k \in e^{-1}((-\infty, t] \mid (X_i)_{i=1}^n)\big)\big] \\
&= \mathrm{P}(e(Z_k) \leq t) = F_{e(Z_k)}(t)
\end{aligned}
$$

The second line follows from conditioning on $(X_i)_{i=1}^n$ and the symmetry of e. The fourth line follows from combining the assumptions on e with the first item of Lemma 26, as well as an application of Lemma 29, known as Anderson's Lemma. The last line follows from the fact that $\mu - \overline{X}_k \mid (X_i)_{i=1}^n \overset{d}{=} Z_k$. To conclude it remains to prove the needed properties for the sequence $(p_{\ell,k})_{k \in N}$. The right-continuity follows directly from the fact that $F_{e(Z_k)}$ is a CDF. To see that the

sequence is decreasing, define $n_k := n + \lambda_k$ and note that $n_k \geq n_{k+1}$. Then, for all $t \in \mathbb{R}$

$$F_{e(Z_k)}(t) = \mathrm{P}(e(Z_k) \leq t) = \mathrm{P}\big(Z_k \in e^{-1}((-\infty, t])\big) = \mathrm{P}\bigg(\sqrt{\frac{n_{k+1}}{n_k}} \cdot Z_{k+1} \in e^{-1}((-\infty, t])\bigg)$$

$$= \mathrm{P}\bigg(Z_{k+1} \in \sqrt{\frac{n_k}{n_{k+1}}} \cdot e^{-1}((-\infty, t])\bigg) \geq \mathrm{P}\big(Z_{k+1} \in e^{-1}((-\infty, t])\big) = F_{e(Z_{k+1})}(t),$$

where the inequality follows from the fact that $\sqrt{n_k/n_{k+1}} \geq 1$, $e^{-1}((-\infty, t]$ is convex and symmetric, and Lemma 23. Finally, let $Z \sim \mathcal{N}(0, \Sigma/n)$. We compute

$$\lim_{k \to \infty} F_{e(Z_k)}(t) = \lim_{k \to \infty} \mathrm{P}\big(Z_k \in e^{-1}((-\infty, t])\big) = \lim_{k \to \infty} \mathrm{P}\bigg(Z \in \sqrt{\frac{n_k}{n}} \cdot e^{-1}((-\infty, t])\bigg)$$

$$= \mathrm{P}\bigg(Z \in \bigcap_{k=1}^{\infty} \bigg\{\sqrt{\frac{n_k}{n}} \cdot e^{-1}((-\infty, t])\bigg\}\bigg) = \mathrm{P}\big(Z \in e^{-1}((-\infty, t])\big) = F_{e(Z)}(t),$$

Finally, the worst-case risk of the sample mean is given by $Q_{e(Z)}(1 - \delta)$ as can be checked with a simple explicit calculation. An application of Theorem 13 concludes the proof. ∎

C.4. Proof of Proposition 16

Proof We aim at applying Theorem 13. We select $\pi_k := \text{Inv-Gamma}(\lambda_k, \lambda_k)$ for a decreasing strictly positive sequence $(\lambda_k)_{k=1}^{\infty}$ satisfying $\lambda_k \to 0$ as $k \to \infty$. We need to compute

$$p_{\ell,k}(t) = \sup_{\hat{\sigma}} \mathrm{P}\bigg(\log\bigg(\frac{\sigma^2}{\hat{\sigma}^2((X_i)_{i=1}^n)}\bigg) \leq t\bigg),$$

where $\sigma^2 \sim \pi_k$ and $X_i \mid \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$ for all $i \in [n]$ independently. A classical Bayesian calculation shows that $\sigma^2 \mid (X_i)_{i=1}^n \sim \text{Inv-Gamma}(\alpha_k, \beta_k)$, where $\alpha_k := n/2 + \lambda_k$ and $\beta_k := \lambda_k + \sum_{i=1}^n (X_i - \mu)^2/2$. Recalling the definition of the fucntion p_α from the statement, we obtain

$$\sup_{\hat{\sigma}} \mathrm{P}\bigg(\log\bigg(\frac{\sigma^2}{\hat{\sigma}^2((X_i)_{i=1}^n)}\bigg) \leq t\bigg) = \mathrm{E}\bigg[\sup_{b \in (0,\infty)} \mathrm{P}\bigg(\log\bigg(\frac{\sigma^2}{b}\bigg) \leq t \,\bigg|\, (X_i)_{i=1}^n\bigg)\bigg] = \mathrm{E}[p_{\alpha_k}(t)] = p_{\alpha_k}(t)$$

where the last equality follows from Lemma 34. It is straightforward to check that p_α is continuous for all values of $\alpha \in (0, \infty)$. Furthermore, by Lemma 35, the sequence $(p_{\alpha_k})_{k \in \mathbb{N}}$ is decreasing with limit $p_{n/2}$. This provides us with the first part needed for Theorem 13. Now note that, for any $\sigma^2 \in (0, \infty)$ and $X_i \sim \mathcal{N}(\mu, \sigma^2)$ for all $i \in [n]$ and independently, we have $(n \cdot \sigma^2)/\sum_{i=1}^n (X_i - \mu)^2 \sim \text{Inv-Gamma}(n/2, n/2)$, so that for the estimator $\hat{\sigma}^2$ defined in the theorem, we have

$$\mathrm{P}\Big(\big|\log(\sigma^2/\hat{\sigma}^2((X_i)_{i=1}^n))\big| \leq p_{n/2}^{-1}(1 - \delta)\Big)$$

$$= \mathrm{P}\bigg(\exp(-p_{n/2}^{-1}(1 - \delta)) \leq \frac{\sigma^2}{\hat{\sigma}^2((X_i)_{i=1}^n)} \leq \exp(p_{n/2}^{-1}(1 - \delta))\bigg)$$

$$= \mathrm{P}\bigg(\frac{1 - \exp(-2p_{n/2}^{-1}(1 - \delta))}{2p_{n/2}^{-1}(1 - \delta)} \leq \frac{n \cdot \sigma^2}{\sum_{i=1}^n (X_i - \mu)^2} \leq \frac{\exp(2p_{n/2}^{-1}(1 - \delta)) - 1}{2p_{n/2}^{-1}(1 - \delta)}\bigg)$$

$$= p_{n/2}(p_{n/2}^{-1}(1 - \delta))$$

$$= 1 - \delta$$

and therefore the worst-case risk of this estimator is equal to $p_{n/2}^{-1}(1 - \delta)$. Applying Theorem 13 proves the minimaxity of this estimator. An explicit calculation of the worst-case risk of the sample variance combined with the uniqueness of the minimizer in Lemma 34 shows that it is not minimax. ∎

Appendix D. Proofs of Section 2

D.1. Proof of Theorem 3

Before we proceed with the proof, we start with a simple lemma.

Lemma 44 *Under the setup of Theorem 3, the functions $\widetilde{\mathcal{E}}, \widetilde{E}$ are strictly convex and symmetric with unique minimizer 0. Furthermore, if $(X, Y) \sim P \in \mathcal{P}_{Gauss}(P_X, \sigma^2)$ so that $Y = \langle w^*, X \rangle + \eta$, then $E(v) = \widetilde{E}(v - w^*)$ for all v, and w^* is the unique minimizer of $E(w)$.*

Proof We prove the convexity and symmetry of $\widetilde{E}$ first. We start with the symmetry.

$$\widetilde{E}(-\Delta) = \mathrm{E}[e(-\langle \Delta, X \rangle + \eta)] = \mathrm{E}[e(\langle \Delta, X \rangle - \eta)] = \mathrm{E}[e(\langle \Delta, X \rangle + \eta)] = \widetilde{E}(\Delta),$$

where the second equality follows from the symmetry of e, and the fourth equality follows from $\eta \overset{d}{=} -\eta$. For the strict convexity, let $t \in (0, 1)$ and $\Delta, \Delta' \in \mathbb{R}^d$. Then

$$\widetilde{E}((1 - t)\Delta + t\Delta') = \mathrm{E}\big[e\big((1 - t)\{\langle \Delta, X \rangle + \eta\} + t\{\langle \Delta', X \rangle + \eta\}\big)\big] < (1 - t)\widetilde{E}(\Delta) + t\widetilde{E}(\Delta'),$$

where the inequality follows from the strict convexity of e. Therefore $\widetilde{E}$ is strictly convex and symmetric, and since $\widetilde{\mathcal{E}}$ and $\widetilde{E}$ differ by a constant, the same holds for $\widetilde{\mathcal{E}}$.

For the second statement, notice that, by symmetry of η,

$$E(v) = \mathrm{E}[e(\langle v, X \rangle - Y)] = \mathrm{E}[e(\langle v - w^*, X \rangle - \eta)] = \mathrm{E}[e(\langle v - w^*, X \rangle + \eta)] = \widetilde{E}(v - w^*).$$

After routine calculations and an application of the chain rule, this also shows that E is strictly convex, symmetric, and differentiable at w^* with $\nabla E(w^*) = \nabla \widetilde{E}(0)$. We compute

$$\nabla E(w^*) = \nabla \widetilde{E}(0) = \mathrm{E}[\nabla e(\eta)] = \mathrm{E}\big[e'(\eta)X\big] = \mathrm{E}\big[e'(\eta)\big]\mathrm{E}[X],$$

where $\eta \sim \mathcal{N}(0, \sigma^2)$ and the last equality follows from the independence of η and X. Now

$$\mathrm{E}\big[e'(\eta)\big] = \frac{1}{\sigma^2}\mathrm{E}[e(\eta)\eta] = \frac{1}{\sigma^2}\mathrm{E}[e(-\eta) \cdot (-\eta)] = -\frac{1}{\sigma^2}\mathrm{E}[e(\eta)\eta] = -\mathrm{E}\big[e'(\eta)\big]$$

where the first and last equalities are by Stein's lemma, the second since $\eta \overset{d}{=} -\eta$, and the third by the symmetry of e. This proves that $\mathrm{E}[e'(\eta)] = 0$, and hence that w^* is the unique minimizer of E by strong convexity. ∎

We now present the main proof of the theorem.

Proof [Proof of Theorem 3] Our strategy is to use Theorem 13 with a properly chosen sequence of distributions $(\pi_k)_{k \in \mathbb{N}}$. Notice that, associated to each $P \in \mathcal{P}_{Gauss}(P_X, \sigma^2)$ is a unique minimizer $w^* \in \mathbb{R}^d$ of the expected error $E(w)$. So putting a distribution on the set of the latter, $\mathbb{R}^d$, induces a

distribution on the set of the former, $\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)$. Specifically, let $(\lambda_k)_{k \in \mathbb{N}}$ be a strictly positive sequence converging to 0, and define $\pi_k := \mathcal{N}(0, \lambda_k^{-1} \cdot (\sigma^2/n) \cdot I_{d \times d})$. With the goal of applying Theorem 13, we need to compute

$$p_k(t) := \sup_{\hat{w}} \mathrm{P}(\mathcal{E}(\hat{w}((X_i, Y_i)_{i=1}^n)) \leq t),$$

where $w^* \sim \pi_k$, $(X_i)_{i=1}^n \sim P_X^n$, and $Y_i \mid (w^*, X_i) \sim \mathcal{N}(\langle w^*, X_i \rangle, \sigma^2)$ for all $i \in [n]$, and independently. A basic calculation shows that $w^* \mid (X_i, Y_i)_{i=1}^n \sim \mathcal{N}(w_k, (\sigma^2/n)\Sigma_k^{-1})$, where

$$w_k := \Sigma_k^{-1}\left(\frac{1}{n} \sum_{i=1}^n Y_i X_i\right), \quad \Sigma_k := \widehat{\Sigma}_n + \lambda_k I_d, \quad \widehat{\Sigma}_n := \frac{1}{n} \sum_{i=1}^n X_i X_i^T$$

Therefore, using Lemma 44,

$$\begin{aligned}
p_k(t) &= \sup_{\hat{w}} \mathrm{P}(\mathcal{E}(\hat{w}) \leq t) = \sup_{\hat{w}} \mathrm{P}\left(\widetilde{\mathcal{E}}(\hat{w} - w^*) \leq t\right) \\
&= \sup_{\hat{w}} \mathrm{E}\left[\mathrm{P}\left(\widetilde{\mathcal{E}}(\hat{w} - w^*) \leq t \,\Big|\, (X_i, Y_i)_{i=1}^n\right)\right] \\
&= \mathrm{E}\left[\sup_{v \in \mathbb{R}^d} \mathrm{P}\left(w^* - v \in \widetilde{\mathcal{E}}^{-1}((-\infty, t]) \,\Big|\, (X_i, Y_i)_{i=1}^n\right)\right] \\
&= \mathrm{E}\left[\mathrm{P}\left(w^* - w_k \in \widetilde{\mathcal{E}}^{-1}((-\infty, t]) \,\Big|\, (X_i, Y_i)_{i=1}^n\right)\right] \\
&= \mathrm{P}\left(\widetilde{\mathcal{E}}(Z_k) \leq t\right) = F_{\widetilde{\mathcal{E}}(Z_k)}(t)
\end{aligned}$$

where $Z_k \mid (X_i)_{i=1}^n \sim \mathcal{N}(0, (\sigma^2/n)\Sigma_k^{-1})$. The fifth equality is obtained by combining the first item of Lemma 44 with the first item of Lemma 26, and an application of Lemma 29. With the goal of applying Theorem 13, we verify the needed properties on the sequence $(p_k)_{k \in \mathbb{N}}$.

First, since each p_k is a CDF, it is right-continuous. To show that $(p_k)_{k \in \mathbb{N}}$ is decreasing, let $k \in \mathbb{N}$. Since $\lambda_k \geq \lambda_{k+1}$ by assumption, $\Sigma_k \succeq \Sigma_{k+1}$, and therefore $\Sigma_k^{-1} \preceq \Sigma_{k+1}^{-1}$. We conclude that $Z_{k+1} \overset{d}{=} Z_k + Y_k$ where $Y_k \perp\!\!\!\perp Z_k \mid (X_i)_{i=1}^n$ and $Y_k \mid (X_i)_{i=1}^n \sim \mathcal{N}(0, \frac{\sigma^2}{n}\{\Sigma_{k+1}^{-1} - \Sigma_k^{-1}\})$. Now

$$\begin{aligned}
F_{\widetilde{\mathcal{E}}(Z_{k+1})|(X_i)_{i=1}^n}(t) &= \mathrm{P}\left(\widetilde{\mathcal{E}}(Z_{k+1}) \leq t \,\Big|\, (X_i)_{i=1}^n\right) \\
&= \mathrm{P}\left(Z_{k+1} \in \widetilde{\mathcal{E}}^{-1}((-\infty, t]) \,\Big|\, (X_i)_{i=1}^n\right) \\
&= \mathrm{P}\left(Z_k + Y_k \in \widetilde{\mathcal{E}}^{-1}((-\infty, t]) \,\Big|\, (X_i)_{i=1}^n\right) \\
&= \mathrm{E}\left[\mathrm{P}\left(Z_k + Y_k \in \widetilde{\mathcal{E}}^{-1}((-\infty, t]) \,\Big|\, (X_i)_{i=1}^n, Y_k\right)\right] \\
&\leq \mathrm{E}\left[\sup_{a \in \mathbb{R}^d} \mathrm{P}\left(Z_k + a \in \widetilde{\mathcal{E}}^{-1}((-\infty, t]) \,\Big|\, (X_i)_{i=1}^n, Y_k\right)\right] \\
&= \mathrm{E}\left[\mathrm{P}\left(Z_k \in \widetilde{\mathcal{E}}^{-1}((-\infty, t]) \,\Big|\, (X_i)_{i=1}^n, Y_k\right)\right] \\
&= F_{\widetilde{\mathcal{E}}(Z_k)|(X_i)_{i=1}^n}(t),
\end{aligned}$$

where the penultimate equality follows from Lemma 28 and the fact that, given $((X_i)_{i=1}^n, Y_k)$, Z_k is a centred Gaussian vector. Taking expectation of both sides with respect to $(X_i)_{i=1}^n$ proves that the sequence $(p_k)_{k\in\mathbb{N}}$ is decreasing. It remains to compute its limit.

By the monotone convergence theorem, we have

$$
\begin{aligned}
\lim_{k\to\infty} F_{\widetilde{\mathcal{E}}(Z_k)}(t) &= \lim_{k\to\infty} \mathrm{P}\Big(\widetilde{\mathcal{E}}(Z_k) \le t\Big) \\
&= \lim_{k\to\infty} \mathrm{E}\Big[\mathrm{P}\Big(\widetilde{\mathcal{E}}(Z_k) \le t \,\Big|\, (X_i)_{i=1}^n\Big)\Big] \\
&= \mathrm{E}\Big[\lim_{k\to\infty} \mathrm{P}\Big(\widetilde{\mathcal{E}}(Z_k) \le t \,\Big|\, (X_i)_{i=1}^n\Big)\Big].
\end{aligned}
\tag{24}
$$

Furthermore, letting $Z \sim \mathcal{N}(0, I_{d\times d})$, we have

$$
\begin{aligned}
\lim_{k\to\infty} \mathrm{P}\Big(\widetilde{\mathcal{E}}(Z_k) \le t \,\Big|\, (X_i)_{i=1}^n\Big) &= \lim_{k\to\infty} \mathrm{P}\Big(Z_k \in \widetilde{\mathcal{E}}^{-1}((-\infty, t]) \,\Big|\, (X_i)_{i=1}^n\Big) \\
&= \lim_{k\to\infty} \mathrm{P}\Big(Z \in \frac{\sqrt{n}}{\sigma}\Sigma_k^{1/2}\widetilde{\mathcal{E}}^{-1}((-\infty, t]) \,\Big|\, (X_i)_{i=1}^n\Big) \\
&= \mathrm{P}\Big(Z \in \bigcap_{k=1}^\infty \Big\{\frac{\sqrt{n}}{\sigma}\Sigma_k^{1/2}\widetilde{\mathcal{E}}^{-1}((-\infty, t])\Big\} \,\Big|\, (X_i)_{i=1}^n\Big) \\
&= \mathrm{P}\Big(Z \in \frac{\sqrt{n}}{\sigma}\widehat{\Sigma}_n^{1/2}\widetilde{\mathcal{E}}^{-1}((-\infty, t]) \,\Big|\, (X_i)_{i=1}^n\Big),
\end{aligned}
\tag{25}
$$

where the second line follows from the fact that $Z_k \stackrel{d}{=} \frac{\sigma}{\sqrt{n}}\Sigma_k^{1/2}Z$ and the third line from the continuity of probability and the fact that for all $k \in \mathbb{N}$,

$$
\frac{\sqrt{n}}{\sigma}\Sigma_{k+1}^{1/2}\widetilde{\mathcal{E}}^{-1}((-\infty, t]) \subset \frac{\sqrt{n}}{\sigma}\Sigma_k^{1/2}\widetilde{\mathcal{E}}^{-1}((-\infty, t]).
$$

Indeed, by the spectral theorem, there exists an orthogonal matrix Q and a diagonal matrix Λ such that $\widehat{\Sigma}_n = Q\Lambda Q^T$, so $\Sigma_k^{1/2} = Q(\Lambda^{1/2} + \lambda_k^{1/2}I)Q^T$. Now since $\lambda_{k+1} \le \lambda_k$, we have by Lemma 23

$$
(\Lambda^{1/2} + \lambda_{k+1}^{1/2})Q^T\widetilde{\mathcal{E}}^{-1}((-\infty, t]) \subset (\Lambda^{1/2} + \lambda_k^{1/2})Q^T\widetilde{\mathcal{E}}^{-1}((-\infty, t]),
$$

Mapping the above sets through Q yields the desired statement. Now if $\mathrm{rank}(\widehat{\Sigma}_n) < d$, then $\dim(\mathrm{Im}(\widehat{\Sigma}_n^{1/2})) < d$ and

$$
0 \le \mathrm{P}\Big(Z \in \frac{\sqrt{n}}{\sigma}\widehat{\Sigma}_n^{1/2}\widetilde{\mathcal{E}}^{-1}((-\infty, t]) \,\Big|\, (X_i)_{i=1}^n\Big) \le \mathrm{P}\Big(Z \in \mathrm{Im}(\widehat{\Sigma}_n^{1/2})\Big) = 0
\tag{26}
$$

where the last equality follows since Z is a standard normal vector, so its distribution is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^d$, and Lebesgue measure assigns zero measure to all hyperplanes. Otherwise, $\mathrm{rank}(\widehat{\Sigma}_n) = d$, and we get

$$
\mathrm{P}\Big(Z \in \frac{\sqrt{n}}{\sigma}\widehat{\Sigma}_n^{1/2}\widetilde{\mathcal{E}}^{-1}((-\infty, t]) \,\Big|\, (X_i)_{i=1}^n\Big) = \mathrm{P}\Big(\widetilde{\mathcal{E}}\Big(\frac{\sigma}{\sqrt{n}}\widehat{\Sigma}_n^{-1/2}Z\Big) \le t \,\Big|\, (X_i)_{i=1}^n\Big)
\tag{27}
$$

Combining (24), (25), (26), and (27) proves that

$$\lim_{k \to \infty} p_k(t) = \mathrm{E}\left[\mathrm{P}\left(\widetilde{\mathcal{E}}\left(\frac{\sigma}{\sqrt{n}} \widehat{\Sigma}_n^{-1/2} Z \right) \le t \,\Big|\, (X_i)_{i=1}^n \right) \mathbb{1}_{\left\{ \mathrm{rank}(\widehat{\Sigma}_n) = d \right\}} \big((X_i)_{i=1}^n \big) \right]$$

which can be interpreted as the CDF of the random variable

$$A((X_i)_{i=1}^n, Z) := \begin{cases} \widetilde{\mathcal{E}}\left(\dfrac{\sigma}{\sqrt{n}} \widehat{\Sigma}_n^{-1/2} Z \right) & \text{if } \mathrm{rank}(\widehat{\Sigma}_n) = d \\ \infty & \text{otherwise} \end{cases}$$

so we write $\lim_{k \to \infty} p_k(t) = F_A(t)$. It remains to show that the worst case risk of the procedures defined in the theorem is $Q_A(1 - \delta)$. Let $\hat{w}$ be a procedure satisfying the condition stated in the theorem and fix $w^* \in \mathbb{R}^d$. Then, on the event that $\mathrm{rank}(\widehat{\Sigma}_n) = d$, and through an elementary explicit calculation, we have $\hat{w} - w^* = \widehat{\Sigma}_n^{-1}(\frac{1}{n} \sum_{i=1}^n \eta_i X_i)$ where $\eta_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d.. Therefore, $\frac{1}{n} \sum_{i=1}^n \eta_i X_i \mid (X_i)_{i=1}^n \sim \mathcal{N}(0, \sigma^2/n \cdot \widehat{\Sigma}_n)$, and hence $\hat{w} - w^* \mid (X_i)_{i=1}^n \sim \mathcal{N}(0, \sigma^2/n \cdot \widehat{\Sigma}_n^{-1})$, so the worst case risk of this procedure is upper bounded by $Q_A(1 - \delta)$. Applying Theorem 13 concludes the proof. ∎

D.2. Proof of Proposition 4

The proof is a simple application of the second-order delta method. Let $(Z_n, (X_i)_{i=1}^n)$ be such that $(X_i)_{i=1}^n \sim P_X^n$ and $Z_n \mid (X_i)_{i=1}^n \sim \mathcal{N}(0, \frac{\sigma^2}{n} \widehat{\Sigma}_n^{-1})$ whenever $\widehat{\Sigma}_n$ is invertible and set $Z_n = 0$ otherwise. The conclusion of Theorem 3 can then be rewritten as

$$R_{n,\delta}^*(\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)) = Q_{\widetilde{\mathcal{E}}(Z_n)}(1 - \delta),$$

with the additional specification that $\widetilde{\mathcal{E}}(Z_n) := \infty$ whenever $(X_i)_{i=1}^n$ is such that $\widehat{\Sigma}_n$ is singular. Recall that $Z \sim \mathcal{N}(0, I_{d \times d})$. By a property of Gaussian vectors, we have that on the event that $\widehat{\Sigma}_n$ is invertible, $Z_n \overset{d}{=} \frac{\sigma}{\sqrt{n}} \widehat{\Sigma}_n^{-1/2} Z$. By the weak law of large numbers and the continuous mapping theorem, we have $\widehat{\Sigma}_n^{-1/2} \overset{p}{\to} \Sigma^{-1/2}$, so that an application of Slutsky's theorem yields $\sqrt{n} \cdot Z_n \overset{d}{\to} \mathcal{N}(0, \sigma^2 \Sigma^{-1})$. Now by assumption, $\widetilde{\mathcal{E}}$ is twice differentiable at 0 where its gradient vanishes by Lemma 44, and where its Hessian is given by $\nabla^2 \widetilde{\mathcal{E}}(0) = \mathrm{E}\big[e''(\eta) X X^T \big] = 2\alpha \Sigma$ by independence of η and X. Therefore, by an application of the delta method, we obtain

$$\lim_{n \to \infty} n \cdot \widetilde{\mathcal{E}}(Z_n) \overset{d}{\to} \sigma^2 \alpha \|Z\|_2^2$$

Since convergence in distribution implies the pointwise convergence of quantiles, we obtain the first equality in the proposition. The second statement follows from Lemma 27.

D.3. Proof of Lemma 5

We start with the first statement. Let $\delta \in (\varepsilon_n, 1)$. By the monotone convergence theorem and the fact that $\widetilde{\mathcal{E}}(w) < \infty$ for all $w \in \mathbb{R}^d$ by assumption on P_X, we have

$$\lim_{t \to \infty} F_{\widetilde{\mathcal{E}}(Z)}(t) = \lim_{t \to \infty} \mathrm{E}\left[\mathrm{P}\left(\widetilde{\mathcal{E}}(Z) \le t \,\Big|\, (X_i)_{i=1}^n \right) \mathbb{1}_{\left\{ \mathrm{rank}(\widehat{\Sigma}_n) = d \right\}} \big((X_i)_{i=1}^n \big) \right] = 1 - \varepsilon_n$$

Therefore, since $1 - \delta < 1 - \varepsilon_n$ there exists a $t \in \mathbb{R}$, such that $F_{\widetilde{\mathcal{E}}(Z)}(t) \geq 1 - \delta$, so $Q_{\widetilde{\mathcal{E}}(Z)}(1-\delta) < \infty$. On the other hand, for all $t \in \mathbb{R}$, $F_{\widetilde{\mathcal{E}}(Z)}(t) < 1 - \varepsilon_n$, so for any $\delta \in [0, \varepsilon_n]$, $Q_{\widetilde{\mathcal{E}}(Z)}(1 - \delta) = \infty$. As for the lower bound on ε_n, El Hanchi and Erdogdu (2023, Lemma 1) proved that there exists a $w_0 \in S^{d-1}$ such that $\rho(P_X) = \sup_{w \in \mathbb{R}^d \setminus \{0\}} \mathrm{P}(\langle w, X \rangle = 0) = \mathrm{P}(\langle w_0, X \rangle = 0) < 1$. Therefore

$$\varepsilon_n = \mathrm{P}\Big(\lambda_{\min}(\widehat{\Sigma}_n) = 0\Big) \geq \mathrm{P}\bigg(\bigcap_{i=1}^{n}\{\langle w_0, X_i \rangle = 0\}\bigg) = \rho(P_X)^n.$$

The upper bound on ε_n follows from the proof of (El Hanchi and Erdogdu, 2023, Theorem 4).

D.4. Proof of Proposition 7

Proof In this proof we will let $Z \sim \mathcal{N}(0, \frac{\sigma^2}{n}\widetilde{\Sigma}_n^{-1})$, so that the minimax risk is given by $Q_{\|Z\|_2^2}(1 - \varepsilon_n - \delta)/2$. Define the random variables

$$M((X_i)_{i=1}^{n}) := \begin{cases} \mathrm{E}[\|Z\|_2 \,|\, (X_i)_{i=1}^{n}] & \text{if } \mathrm{rank}(\widehat{\Sigma}_n) = d \\ \infty & \text{otherwise} \end{cases}$$

$$R((X_i)_{i=1}^{n}) := \begin{cases} \lambda_{\max}\left(\dfrac{\sigma^2}{n}\widetilde{\Sigma}_n^{-1}\right) & \text{if } \mathrm{rank}(\widehat{\Sigma}_n) = d \\ \infty & \text{otherwise} \end{cases}$$

To simplify notation, we will write M and R only, and leave the dependence on $(X_i)_{i=1}^{n}$ implicit.

Upper bound. We have, for all $r \in \mathbb{R}$,

$$\begin{aligned}
F_{\|Z\|_2^2}(r^2) &= \mathrm{E}\Big[\mathrm{P}(\|Z\|_2 \leq r \,|\, (X_i)_{i=1}^{n})\mathbb{1}_{\{\mathrm{rank}(\widehat{\Sigma}_n)=d\}}((X_i)_{i=1}^{n})\Big] \\
&= \mathrm{E}\Big[\{1 - \mathrm{P}(\|Z\|_2 > r \,|\, (X_i)_{i=1}^{n})\}\mathbb{1}_{\{\mathrm{rank}(\widehat{\Sigma}_n)=d\}}((X_i)_{i=1}^{n})\Big] \\
&\geq \mathrm{E}\bigg[\bigg\{1 - \exp\bigg(-\frac{|r - M|^2}{2R}\bigg)\bigg\}\mathbb{1}_{[M,\infty)}(r)\mathbb{1}_{\{\mathrm{rank}(\widehat{\Sigma}_n)=d\}}((X_i)_{i=1}^{n})\bigg] \\
&= \mathrm{E}\bigg[\bigg\{1 - \exp\bigg(-\frac{|r - M|^2}{2R}\bigg)\bigg\}\mathbb{1}_{[0,r]}(M)\bigg] \\
&= \mathrm{P}(M \leq r) - \mathrm{E}\bigg[\exp\bigg(-\frac{|r - M|^2}{2R}\bigg)\mathbb{1}_{[0,r]}(M)\bigg] =: L(r)
\end{aligned}$$

where the inequality follows from the Gaussian concentration (Lemma 32), and where the expression inside the expectation is defined to be 0 whenever $\mathrm{rank}(\widehat{\Sigma}) < d$. The penultimate equality follows from the fact that $\mathbb{1}_{[M,\infty)}(r)\mathbb{1}_{\mathrm{rank}(\widehat{\Sigma})=d}((X_i)_{i=1}^{n}) = \mathbb{1}_{[0,r]}(M)$. Now let $0 < c \leq 1$ and

define $q := Q_M(1 - \varepsilon_n - c\delta)$. Then, recalling the definition of W from the statement,

$$L(r + q) \geq \mathrm{P}(M \leq q) + \mathrm{P}(M \in (q, r + q])$$

$$- \mathrm{E}\left[\exp\left(-\frac{r^2}{2R}\right)\mathbb{1}_{[0,q]}(M)\right] - \mathrm{E}\left[\underbrace{\exp\left(-\frac{|r - M|^2}{2R}\right)}_{\leq 1}\mathbb{1}_{(q,r+q]}(M)\right]$$

$$\geq \mathrm{P}(M \leq q) - \mathrm{E}\left[\exp\left(-\frac{r^2}{2R}\right)\mathbb{1}_{[0,q]}(M)\right]$$

$$\geq 1 - \varepsilon_n - c\delta - \mathrm{E}\left[\exp\left(-\frac{r^2}{2R}\right)\mathbb{1}_{\mathrm{rank}(\widehat{\Sigma})=d}((X_i)_{i=1}^n)\right]$$

$$= \mathrm{E}\left[\left\{1 - \exp\left(-\frac{r^2}{2R}\right)\right\}\mathbb{1}_{\mathrm{rank}(\widehat{\Sigma})=d}((X_i)_{i=1}^n)\right] - c\delta$$

$$= \mathrm{P}\left(\sqrt{\frac{2\sigma^2}{n}}W \leq r\right) - c\delta$$

hence taking $r = \sqrt{\frac{2\sigma^2}{n}Q_W(1 - \varepsilon_n - c\delta)}$ and $c = 1/2$ in the last display yields

$$L\left(Q_M(1 - \varepsilon_n - \delta/2) + \sqrt{\frac{2\sigma^2}{n}Q_W(1 - \varepsilon_n - \delta/2)}\right) \geq 1 - \varepsilon_n - \delta$$

And since $F_{\|Z\|_2^2} \circ \varphi \geq L$ where $\varphi(r) = r^2$, we get by the second item of Lemma 20 that $\varphi^{-1} \circ Q_{\|Z\|_2^2} \leq L^-$. Applying φ to both sides yields and using Lemma 21 we obtain,

$$Q_{\|Z\|_2^2}(1 - \varepsilon_n - \delta) \leq (L^-(1 - \varepsilon_n - \delta))^2$$

$$\leq \left(Q_M(1 - \varepsilon_n - \delta/2) + \sqrt{\frac{2\sigma^2}{n}Q_W(1 - \varepsilon_n - \delta/2)}\right)^2$$

$$\leq 2\left[Q_{M^2}(1 - \varepsilon_n - \delta/2) + \frac{2\sigma^2}{n}Q_W(1 - \varepsilon_n - \delta/2)\right]$$

$$\leq \frac{2\sigma^2}{n}\left(Q_{\mathrm{Tr}(\widetilde{\Sigma}^{-1})}(1 - \varepsilon_n - \delta/2) + 2Q_W(1 - \varepsilon_n - \delta/2)\right)$$

$$\leq 4 \cdot \frac{\sigma^2}{n}\left(Q_{\mathrm{Tr}(\widetilde{\Sigma}^{-1})}(1 - \varepsilon_n - \delta/2) + Q_W(1 - \varepsilon_n - \delta/2)\right),$$

where in the penultimate inequality, we used the fact that $M^2 \leq \mathrm{Tr}\left(\widetilde{\Sigma}_n^{-1}\right)$ by Jensen's inequality.

Lower bound. For any $(X_i)_{i=1}^n$, define $v((X_i)_{i=1}^n)$ to be the eigenvector corresponding to the smallest eigenvalue of $\widetilde{\Sigma}_n$. Then we have.

$$
\begin{aligned}
F_{\|Z\|_2^2}(r^2) &= \mathrm{E}\Big[\mathrm{P}(\|Z\|_2 \le r \,|\, (X_i)_{i=1}^n)\mathbb{1}_{\mathrm{rank}(\widehat{\Sigma})=d}((X_i)_{i=1}^n)\Big]\\
&\le \mathrm{E}\Big[\mathrm{P}(|\langle v((X_i)_{i=1}^n), Z\rangle| \le r \,|\, (X_i)_{i=1}^n)\mathbb{1}_{\mathrm{rank}(\widehat{\Sigma})=d}((X_i)_{i=1}^n)\Big]\\
&\le \mathrm{E}\left[\sqrt{1 - \exp\left(-\frac{2r^2}{\pi R}\right)}\,\mathbb{1}_{\mathrm{rank}(\widehat{\Sigma})=d}((X_i)_{i=1}^n)\right]\\
&\le \mathrm{E}\left[\left\{1 - \frac{1}{2}\exp\left(-\frac{2r^2}{\pi R}\right)\right\}\mathbb{1}_{\mathrm{rank}(\widehat{\Sigma})=d}((X_i)_{i=1}^n)\right]\\
&= \frac{1}{2}(1 - \varepsilon_n) + \frac{1}{2}\mathrm{E}\left[\left\{1 - \exp\left(-\frac{2r^2}{\pi R}\right)\right\}\mathbb{1}_{\mathrm{rank}(\widehat{\Sigma})=d}((X_i)_{i=1}^n)\right]\\
&= \frac{1}{2}\left(1 - \varepsilon_n + \mathrm{P}\left(\sqrt{\frac{\pi\sigma^2}{2n}}W \le r\right)\right) =: U_1(r)
\end{aligned}
$$

Where the third line follows from Lemma 27. Now let $\varepsilon > 0$ and define

$$
r(\varepsilon) := \sqrt{\frac{\pi\sigma^2}{2n}Q_W(1 - \varepsilon_n - 2\delta)} - \varepsilon
$$

Then

$$
U_1(r(\varepsilon)) < \frac{1}{2}(1 - \varepsilon_n + 1 - \varepsilon_n - 2\delta) = 1 - \varepsilon_n - \delta.
$$

Since this holds for all $\varepsilon > 0$, we obtain $U_1^-(1 - \varepsilon_n - \delta) \ge r(0)$. Therefore,

$$
Q_{\|Z\|_2^2}(1 - \varepsilon_n - \delta) \ge (U_1^-(1 - \varepsilon_n - \delta))^2 \ge r^2(0) = \frac{\pi\sigma^2}{2n}Q_W(1 - \varepsilon_n - 2\delta)
$$

This finishes the proof of the first part of the lower bound. For the second part of the lower bound, we also have by Gaussian concentration, and in particular Lemma 32,

$$
\begin{aligned}
F_{\|Z\|_2^2}(r^2) &= \mathrm{E}\Big[\mathrm{P}(\|Z\|_2 \le r \,|\, (X_i)_{i=1}^n)\mathbb{1}_{\mathrm{rank}(\widehat{\Sigma})=d}((X_i)_{i=1}^n)\Big]\\
&\le \mathrm{E}\left[\exp\left(-\frac{|M - r|^2}{\pi M^2}\right)\mathbb{1}_{[0,M]}(r)\mathbb{1}_{\mathrm{rank}(\widehat{\Sigma})=d}((X_i)_{i=1}^n) + \mathbb{1}_{(M,\infty)}(r)\mathbb{1}_{\mathrm{rank}(\widehat{\Sigma})=d}((X_i)_{i=1}^n)\right]\\
&= \mathrm{E}\left[\exp\left(-\frac{|M - r|^2}{\pi M^2}\right)\mathbb{1}_{[r,\infty)}(M) + \mathbb{1}_{[0,r)}(M)\right]\\
&= \mathrm{P}(M < r) + \mathrm{E}\left[\exp\left(-\frac{|M - r|^2}{\pi M^2}\right)\mathbb{1}_{[r,\infty)}(M)\right] =: U_2(r)
\end{aligned}
$$

Let $a \in (0,1)$, $c > 1$, and $q := Q_M(1 - \varepsilon_n - c\delta)$. Then we have,

$$U((1-a)q) = \mathrm{E}\Bigg[\underbrace{\exp\left(-\frac{|M - (1-a)q|^2}{\pi M^2}\right) \mathbb{1}_{[(1-a)q,q)}(M)}_{\leq 1}\Bigg] + \mathrm{E}\left[\exp\left(-\frac{|M - (1-a)q|^2}{\pi M^2}\right) \mathbb{1}_{[q,\infty)}(M)\right]$$

$$+ \mathrm{P}(M < q) - \mathrm{P}((1-a)q \leq M < q)$$

$$\leq \mathrm{P}(M < q) + \mathrm{E}\left[\exp\left(-\frac{|M - (1-a)q|^2}{\pi M^2}\right) \mathbb{1}_{[q,\infty)}(M)\right]$$

$$\leq \mathrm{P}(M < q) + \exp\left(-\frac{a^2}{\pi}\right) \mathrm{P}(q \leq M < \infty)$$

$$\leq \mathrm{P}(M < q) + \exp\left(-\frac{a^2}{\pi}\right)(1 - \varepsilon_n - \mathrm{P}(M < q))$$

$$\leq \left(1 - \exp\left(-\frac{a^2}{\pi}\right)\right)\mathrm{P}(M < q) + \exp\left(-\frac{a^2}{\pi}\right)(1 - \varepsilon_n)$$

$$\leq \left(1 - \exp\left(-\frac{a^2}{\pi}\right)\right)(1 - \varepsilon_n - c\delta) + \exp\left(-\frac{a^2}{\pi}\right)(1 - \varepsilon_n)$$

$$= 1 - \varepsilon_n - \left(1 - \exp\left(-\frac{a^2}{\pi}\right)\right)c\delta$$

$$< 1 - \varepsilon_n - \delta$$

where the last line follows from taking $a = 0.96$, and $c = 4$, and noticing that with these choices $c\left(1 - \exp\left(-\frac{a^2}{\pi}\right)\right) > 1$. Now since $F_{\|Z\|_2^2} \circ \varphi \leq U_2$ where $\varphi(r) = r^2$, we get by the second item of Lemma 20 and an application of Lemma 21,

$$Q_{\|Z\|_2^2}(1 - \varepsilon_n - \delta) \geq (U^-(1 - \varepsilon_n - \delta))^2$$

$$\geq \left(\frac{1}{25} Q_M(1 - \varepsilon_n - 4\delta)\right)^2$$

$$= \frac{1}{625} Q_{M^2}(1 - \varepsilon_n - 4\delta)$$

$$\geq \frac{1}{625(1 + \pi/2)} \frac{\sigma^2}{n} Q_{\mathrm{Tr}(\widetilde{\Sigma}^{-1})}(1 - \varepsilon_n - 4\delta).$$

Averaging the two lower bounds yields the result. ∎

D.5. Proof of Lemma 8

We start with the bounds on $Q_{\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})}(1 - \delta)$, and in particular the lower bound. If $(a_i)_{i=1}^d$ is a finite sequence of non-negative real numbers, then twice applying the AM-GM inequality we obtain

$$\frac{d}{\sum_{i=1}^d \frac{1}{a_i}} \leq \left(\prod_{i=1}^d a_i\right)^{1/d} \leq \frac{\sum_{i=1}^n a_i}{d} \implies \sum_{i=1}^d \frac{1}{a_i} \geq \frac{d^2}{\sum_{i=1}^d a_i}.$$

Using this, we have

$$\mathrm{Tr}\left(\widetilde{\Sigma}_n^{-1}\right) = \sum_{i=1}^{d} \lambda_i(\widetilde{\Sigma}_n^{-1}) = \sum_{i=1}^{d} \frac{1}{\lambda_i(\widetilde{\Sigma}_n)} \geq \frac{d^2}{\mathrm{Tr}(\widetilde{\Sigma}_n)}.$$

Now, since $\mathrm{E}[\mathrm{Tr}(\widetilde{\Sigma}_n)] = d$, we have

$$\mathrm{P}\left(\frac{d^2}{\mathrm{Tr}(\widetilde{\Sigma}_n)} \leq t\right) = \mathrm{P}\left(\mathrm{Tr}(\widetilde{\Sigma}_n) \geq \frac{d^2}{t}\right) \leq \frac{\mathrm{E}[\mathrm{Tr}(\widetilde{\Sigma}_n)]}{d^2/t} = \frac{t}{d}.$$

Applying the second item of Lemma 20, we obtain the desired lower bound

$$Q_{\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})}(1-\delta) \geq Q_{d^2/\mathrm{Tr}(\widetilde{\Sigma}_n)}(1-\delta) \geq d \cdot (1-\delta).$$

The upper bound follows from the simple observation $\mathrm{Tr}(\widetilde{\Sigma}_n^{-1}) \leq d \cdot \lambda_{\max}(\widetilde{\Sigma}_n^{-1})$. We now move to bounds on $Q_W(1-\delta)$, and we start with the lower bound. By definition, we have

$$1 - \delta \leq \mathrm{P}(W \leq Q_W(1-\delta)) = 1 - \mathrm{E}\left[\exp(-Q_W(1-\delta) \cdot \lambda_{\min}(\widetilde{\Sigma}_n))\right],$$

hence, by Jensen's inequality

$$\delta \geq \mathrm{E}\left[\exp(-Q_W(1-\delta) \cdot \lambda_{\min}(\widetilde{\Sigma}_n))\right] \geq \exp\left(-Q_W(1-\delta) \cdot \mathrm{E}\left[\lambda_{\min}(\widetilde{\Sigma}_n)\right]\right),$$

and using the variational characterization of the smallest eigenvalue we get, for any $v \in S^{d-1}$,

$$\mathrm{E}\left[\lambda_{\min}(\widetilde{\Sigma}_n)\right] = \mathrm{E}\left[\inf_{v \in S^{d-1}} \frac{1}{n} \sum_{i=1}^{n} \langle v, \Sigma^{-1/2} X_i \rangle^2\right] \leq \mathrm{E}\left[\langle v, \Sigma^{-1/2} X \rangle^2\right] = 1.$$

Therefore $Q_W(1-\delta) \geq \log(1/\delta)$ as desired. For the upper bound, let $q := Q_{\lambda_{\max}(\widetilde{\Sigma}_n^{-1})}(1-\delta/2)$ and define the event $A := \left\{\lambda_{\max}(\widetilde{\Sigma}_n^{-1}) \leq q\right\}$ which satisfies $\mathrm{P}(A) \geq 1 - \delta/2$. Notice that

$$\mathrm{P}(W \leq t) \geq \mathrm{E}\left[\left\{1 - \exp\left(-\frac{t}{\lambda_{\max}(\widetilde{\Sigma}_n^{-1})}\right)\right\} \mathbb{1}_A((X_i)_{i=1}^{n})\right] \geq (1-\delta/2)(1 - \exp(t/q)).$$

Taking $t \geq q \cdot \log(2/\delta)$ ensures that the above probability is at least $1 - \delta$. By the minimality of the quantile, we get that $Q_W(1-\delta) \leq Q_{\lambda_{\max}(\widetilde{\Sigma}_n^{-1})}(1-\delta/2) \cdot \log(2/\delta)$, which is the desired upper bound.

D.6. Proof of Corollary 9

Proof We claim that for all the allowed sample sizes,

$$Q_{\lambda_{\max}(\widetilde{\Sigma}_n))}(1-\delta/2) \leq 2.$$

Indeed, the restriction on the sample size n is chosen in such a way that by the upper bound in Proposition 17, we have

$$Q_{1-\lambda_{\min}(\widetilde{\Sigma}_n)}(1-\delta/2) \leq \frac{1}{2}$$

Now if $1 - \lambda_{\min}(\widetilde{\Sigma}_n) \leq 1/2$, then $\lambda_{\min}(\widetilde{\Sigma}_n) \geq 1/2$ and $\lambda_{\max}(\widetilde{\Sigma}_n^{-1}) = \lambda_{\min}^{-1}(\widetilde{\Sigma}_n) \leq 2$. Therefore

$$P\left(\lambda_{\max}(\widetilde{\Sigma}_n^{-1}) \leq 2\right) \geq P\left(\lambda_{\min}(\widetilde{\Sigma}_n) \leq 1/2\right) \geq P\left(\lambda_{\min}(\widetilde{\Sigma}_n) \leq Q_{1-\lambda_{\min}(\widetilde{\Sigma}_n)}(1 - \delta/2)\right) \geq 1 - \delta/2.$$

which finishes the proof of the bound on $Q_{\lambda_{\max}(\widetilde{\Sigma}_n))}(1 - \delta/2)$. Now appealing to Lemma 8 proves the result. ∎

D.7. Proof of Proposition 10

Using Lemma 2.5 in Adil et al. (2023), we have for the p-th power error $e(t) = |t|^p/[p(p-1)]$,

$$\widetilde{\mathcal{E}}(\Delta) = \mathrm{E}[e(\langle \Delta, X \rangle + \eta)] - \mathrm{E}[e(\eta)] \geq \frac{1}{8(p-1)} \Delta^T \mathrm{E}\big[e''(\eta) X X^T\big] \Delta.$$

Since $e''(t) = |t|^{p-2}$, and η and X are independent,

$$\widetilde{\mathcal{E}}(\Delta) \geq \frac{1}{8(p-1)} \cdot m(p-2)\sigma^{p-2}\Delta^T \Sigma \Delta.$$

Therefore, by Theorem 3,

$$R_{n,\delta}(\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)) = Q_{\widetilde{\mathcal{E}}(Z)}(1 - \delta) \geq \frac{m(p-2)\sigma^{p-2}}{8(p-1)} \cdot \frac{\sigma^2}{n} Q_{\|A\|_2^2}(1 - \delta)$$

where $A \sim \mathcal{N}(0, \widetilde{\Sigma}_n^{-1})$. Now noting that $\frac{\sigma^2}{n} Q_{\|A\|_2^2}(1 - \delta)$ is the minimax risk under the square error, applying Proposition 7 and Lemma 8, and using the constraint on δ, we obtain the desired lower bound.

Appendix E. Proofs of Section 3

E.1. Proof of Theorem 11

Fix a distribution $P \in \mathcal{P}_2(P_X, \sigma^2)$. We will prove an upper bound on the risk of the proposed procedure under P. We follow the approach developed by Lugosi and Mendelson (2019a). Define

$$\phi(w) = \max_{v \in \mathbb{R}^d} \psi_k(w, v),$$

and note that by definition of $\hat{w}_{n,\delta}$, we have

$$\psi_k(\hat{w}_{n,k}, w^*) \leq \phi(\hat{w}_{n,k}) \leq \phi(w^*). \tag{28}$$

The key idea of the proof is to show that $\|\hat{w}_{n,k} - w^*\|_\Sigma$ is small, by simultaneously showing that

- For all $w \in \mathbb{R}^d$, if $\|w - w^*\|_\Sigma$ is large, then so is $\psi_k(w, w^*)$,

- $\phi(w^*)$ is small.

The combination of these statements combined with (28) will show that $\|\hat{w}_{n,k} - w^*\|_\Sigma$ is indeed small. Define

$$\Delta(\delta) := 50\sqrt{\frac{\sigma^2[d + \log(4/\delta)]}{n}}$$

All the following lemmas are stated under the conditions of Theorem 11. The first step of the proof is a simple application of Lemma 43.

Lemma 45

$$\mathrm{P}\left(\sup_{\|v\|_\Sigma \leq 1} n^{-1}\varphi_k[(\langle\nabla e(\langle w^*, X_i\rangle - Y_i), v\rangle)_{i=1}^n] > \Delta(\delta)\right) \leq \delta/2$$

Proof For $v \in \mathbb{R}^d$ such that $\|v\|_\Sigma \leq 1$ and $i \in [n]$, define

$$Z_{i,v} := \frac{1}{n}\langle\nabla e(\langle w^*, X_i\rangle - Y_i), v\rangle = \frac{1}{n}\xi_i\langle X_i, v\rangle$$

Our aim is to apply Lemma 43, so we make the necessary computations here. We have

$$\mathrm{E}[Z_{i,v}] = \frac{1}{n}\langle\mathrm{E}[\nabla e(\langle w^*, X_i\rangle - Y_i)], v\rangle = \frac{1}{n}\langle\nabla E(w^*), v\rangle = 0,$$

$$\sup_{\|v\|_\Sigma \leq 1}\sum_{i=1}^n \mathrm{E}\big[Z_{i,v}^2\big] = \frac{1}{n}\sup_{\|v\|_\Sigma = 1}\mathrm{E}\big[\xi^2\langle X, v\rangle^2\big] \leq \frac{\sigma^2}{n}.$$

where the last inequality follows from the assumption $\mathrm{E}\big[\xi^2 \mid X\big] \leq \sigma^2$. Now, for independent Rademacher variables $(\varepsilon_i)_{i=1}^n$, we have

$$\mathrm{E}\left[\sup_{\|v\|_\Sigma = 1}\sum_{i=1}^n \varepsilon_i Z_{i,v}\right] = \mathrm{E}\left[\sup_{\|v\|_\Sigma = 1}\left\langle\frac{1}{n}\sum_{i=1}^n \varepsilon_i \xi_i X_i, v\right\rangle\right]$$

$$= \mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \xi_i X_i\right\|_{\Sigma^{-1}}\right]$$

$$\leq \mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \xi_i X_i\right\|_{\Sigma^{-1}}^2\right]^{1/2}$$

$$= \sqrt{\frac{\mathrm{E}[\xi^2\|X\|_{\Sigma^{-1}}^2]}{n}} \leq \sqrt{\frac{\sigma^2 d}{n}}.$$

Where again we have used the assumption $\mathrm{E}\big[\xi^2 \mid X\big] \leq \sigma^2$. Recalling that $k = 8\log(2/(\delta/2))$ from the statement of the theorem, and applying Lemma 43 with the above constants yields the result. ■

From this result, we can deduce the following estimate, which will help us bound $\phi(w^*)$ later on.

Corollary 46 *For any $r \in (0, \infty)$,*

$$\mathrm{P}\left(\sup_{\|v - w^*\|_\Sigma < r}\psi_k(w^*, v) > r \cdot \Delta(\delta)\right) \leq \delta/2.$$

Proof We have

$$
\sup_{\|v-w^*\|_\Sigma < r} \psi_k(w^*, v) = \sup_{\|v-w^*\|_\Sigma < r} \varphi_k[(e(\langle w^*, X_i\rangle - Y_i) - e(\langle v, X_i\rangle - Y_i))_{i=1}^n]
$$

$$
= \sup_{\|v-w^*\|_\Sigma < r} \varphi_k\left[\left(-\langle\nabla e(\langle w^*, X_i\rangle - Y_i), v - w^*\rangle - \frac{1}{2}\langle X_i, v - w^*\rangle^2\right)_{i=1}^n\right]
$$

$$
\leq \sup_{\|v-w^*\|_\Sigma < r} \varphi_k[(-\langle\nabla e(\langle w^*, X_i\rangle - Y_i), v - w^*\rangle)_{i=1}^n]
$$

$$
= \sup_{\|v-w^*\|_\Sigma < r} r \cdot \varphi_k\left[\left(\left\langle\nabla e(\langle w^*, X_i\rangle - Y_i), \frac{v - w^*}{r}\right\rangle\right)_{i=1}^n\right]
$$

$$
= r \cdot \sup_{\|v\|_\Sigma < 1} \varphi_k[(\langle\nabla e(\langle w^*, X_i\rangle - Y_i), v\rangle)_{i=1}^n]
$$

where the first line is by definition, the second holds since e is quadratic so its second order Taylor expansion is exact, the third by the third item of Lemma 38, and the fourth by the first item of Lemma 38. Applying Lemma 45 to the last line yields the result. ∎

The key technical novelty of this proof is the following lemma, which uses our new results Proposition 18 and Lemma 39.

Lemma 47 *Let $r \in [8\Delta(\delta), \infty)$. Then*

$$
P\left(\inf_{\|v-w^*\|\geq r} \psi_k(v, w^*) < \frac{r^2}{8} - r\Delta(\delta)\right) \leq \delta
$$

Proof We start with the case $\|v - w^*\|_\Sigma = r$. We have

$$
\inf_{\|v-w^*\|_\Sigma = r} \psi_k(v, w^*) = \inf_{\|v-w^*\|_\Sigma = r} n^{-1}\varphi_k[(e(\langle v, X_i\rangle - Y_i) - e(\langle w^*, X_i\rangle - Y_i))_{i=1}^n]
$$

$$
= \inf_{\|v-w^*\|_\Sigma = r} n^{-1}\varphi_k\left[\left(\langle\nabla e(\langle w^*, X_i\rangle - Y_i), v - w^*\rangle + \frac{1}{2}\langle v - w^*, X_i\rangle^2\right)_{i=1}^n\right]
$$

$$
= \inf_{v\in S^{d-1}} n^{-1}\varphi_k\left[\left(r \cdot \langle\nabla e(\langle w^*, X_i\rangle - Y_i), \Sigma^{-1/2}v\rangle + \frac{r^2}{2}\langle v, \Sigma^{-1/2}X_i\rangle^2\right)_{i=1}^n\right].
$$

Define $\widetilde{X}_i = \Sigma^{-1/2}X_i$, and $Z_{i,v} := \langle v, \widetilde{X}_i\rangle^2$ for $(i, v) \in [n] \times S^{d-1}$. Then we have by Lemma 39,

$$
\inf_{\|v-w^*\|_\Sigma = r} \psi_k(v, w^*)
$$

$$
\geq r \cdot \inf_{v\in S^{d-1}} n^{-1}\varphi_k\left[\left(\langle\nabla e(\langle w^*, X_i\rangle - Y_i), \Sigma^{-1/2}v\rangle\right)_{i=1}^n\right] + r^2 \cdot \inf_{v\in S^{d-1}} n^{-1}\sum_{i=1}^{n-2k} Z^*_{i,v}
$$

$$
= \frac{r^2}{2} \cdot \inf_{v\in S^{d-1}} n^{-1}\sum_{i=1}^{n-2k} Z^*_{i,v} - r \cdot \sup_{\|v\|_\Sigma = 1} n^{-1}\varphi_k[(\langle\nabla e(\langle w^*, X_i\rangle - Y_i), v\rangle)_{i=1}^n]
$$

The second term is bounded with probability $1-\delta/2$ by $r \cdot \Delta(\delta)$ by Lemma 45. For the first term, the restriction on the sample size in Theorem 11 is chosen such that by Proposition 17, with probability at least $1 - \delta^2/2 \geq 1 - \delta/2$

$$
\inf_{v\in S^{d-1}} n^{-1}\sum_{i=1}^{n-2k} Z^*_{i,v} \geq \frac{1}{4}
$$

Therefore, with probability at least $1 - \delta$

$$\inf_{\|v-w^*\|_\Sigma=r} \psi_k(v, w^*) \geq \frac{r^2}{8} - r\Delta(\delta).$$

We now extend this to all vectors $w \in \mathbb{R}^d$ such that $\|w - w^*\|_\Sigma \geq r$. On the same event, if $\|w - w^*\|_\Sigma = R > r$, then $v := w^* + \frac{r}{R}(w - w^*)$ satisfies $\|v - w^*\|_\Sigma = r$, and

$$\begin{aligned}
\psi_k(w, w^*) &= n^{-1}\varphi_k\left((\langle\nabla e(\langle w^*, X_i\rangle - Y_i), w - w^*\rangle + \tfrac{1}{2}\langle w - w^*, X_i\rangle^2)_{i=1}^n \right) \\
&= n^{-1}\varphi_k\left[\left(\frac{R}{r}\langle\nabla e(\langle w^*, X_i\rangle - Y_i), v - w^*\rangle + \frac{R^2}{r^2}\frac{1}{2}\langle v - w^*, X_i\rangle^2 \right)_{i=1}^n \right] \\
&\geq n^{-1}\varphi_k\left[\left(\frac{R}{r}\langle\nabla e(\langle w^*, X_i\rangle - Y_i), v - w^*\rangle + \frac{R}{r}\frac{1}{2}\langle v - w^*, X_i\rangle^2)_{i=1}^n \right)_{i=1}^n \right] \\
&= \frac{R}{r} \cdot \psi_k(v, w^*) \\
&\geq \psi_k(v, w^*)
\end{aligned}$$

where the first inequality follows from the fact that $R/r > 1$, $\langle v - w^*, X_i\rangle^2 > 0$, and Lemma 38, and the second inequality follows from the fact that by the condition on r, we have $\psi_k(v, w^*) \geq 0$ on the event we are considering. ∎

We are now ready to state the proof of Theorem 11. Set $r := 20\Delta(\delta)$, and recall from (28) that

$$\begin{aligned}
\psi_k(\hat{w}_{n,k}, w^*) &\leq \phi(w^*) = \sup_{v \in \mathbb{R}^d} \psi_k(w^*, v) \\
&= \max\left\{ \sup_{\|v-w^*\|_\Sigma\geq r} \psi_k(w^*, v), \sup_{\|v-w^*\|_\Sigma< r} \psi_k(w^*, v) \right\} \\
&= \max\left\{ -\inf_{\|v-w^*\|\geq r} \psi_k(v, w^*), \sup_{\|v-w^*\|_\Sigma< r} \psi_k(w^*, v) \right\},
\end{aligned}$$

where the last line uses the fact that $\psi_k(w, v) = -\psi_k(v, w)$. Now by combining Corollary 46 and Lemma 47, we have with probability $1 - \delta$ that the first term in the above maximum is negative, while the second is bounded by $r \cdot \Delta(\delta) = 20\Delta^2(\delta)$. On the other hand, we have on the same event by Lemma 47 that

$$\inf_{\|v-w^*\|_\Sigma\geq r} \psi_k(v, w^*) \geq \frac{r^2}{8} - r\Delta(\delta) = 30\Delta^2(\delta)$$

Therefore, we conclude that with probability at least $1 - \delta$

$$\|\hat{w}_{n,k} - w^*\|_\Sigma \leq 20\Delta(\delta).$$

finally, noticing that this implies, with probability at least $1 - \delta$

$$\mathcal{E}(\hat{w}_{n,k}) = \frac{1}{2}\|\hat{w}_{n,k} - w^*\|_\Sigma^2 \leq 20^2\Delta^2(\delta),$$

finishes the proof.

E.2. Proof of Theorem 12

The high-level idea behind the proof of Theorem 12 is similar to that of Theorem 11, but with a few more challenges. Fix $P \in \mathcal{P}_p(P_X, \sigma^2, \mu)$. We prove an upper bound on the risk of $\hat{w}_{n,k}$ under this fixed P. Define $H := \nabla^2 E(w^*)$, $c := \operatorname{ess\,inf}(\mathrm{E}\big[|\xi|^{p-2} \mid X\big])$, $C := \operatorname{ess\,sup}(\mathrm{E}\big[|\xi|^{2(p-1)} \mid X\big])$. Note that

$$H = \mathrm{E}\big[|\xi|^{p-2} X X^T\big] \succeq c \cdot \Sigma \tag{29}$$

Define

$$\Delta_p(\delta) := 50 \sqrt{\frac{m(2p-2)}{m(p-2)} \cdot \frac{\sigma^p[d + \log(4/\delta)]}{n}}$$

Our first statement is an analogue to Lemma 45.

Lemma 48

$$\mathrm{P}\left(\sup_{\|v\|_H \leq 1} n^{-1} \varphi_k[(\langle \nabla e(\langle w^*, X_i \rangle - Y_i), v \rangle)_{i=1}^n] > \Delta_p(\delta) \right) \leq \delta/2$$

Proof For $v \in \mathbb{R}^d$ such that $\|v\|_H \leq 1$ and $i \in [n]$, define

$$Z_{i,v} := \frac{1}{n} \langle \nabla e(\langle w^*, X_i \rangle - Y_i), v \rangle$$

Our aim is to apply Lemma 43, so we make the necessary computations here. We have

$$\mathrm{E}[Z_{i,v}] = \frac{1}{n} \langle \mathrm{E}[\nabla e(\langle w^*, X_i \rangle - Y_i)], v \rangle = \frac{1}{n} \langle \nabla E(w^*), v \rangle = 0$$

and

$$
\begin{aligned}
\sup_{\|v\|_H \leq 1} \sum_{i=1}^n \mathrm{E}\big[Z_{i,v}^2\big] &= \frac{1}{n} \sup_{\|v\|_H \leq 1} \mathrm{E}\Big[\xi^{2(p-1)} \langle X, v \rangle^2\Big] \\
&\leq \frac{1}{n} \sup_{\|v\|_\Sigma = 1} \frac{\operatorname{ess\,sup}\big(\mathrm{E}\big[\xi^{2(p-1)} \mid X\big]\big)}{\operatorname{ess\,inf}(\mathrm{E}[\xi^{p-2} \mid X]))} \mathrm{E}\big[\langle X, v \rangle^2\big] \\
&\leq \frac{m(2p-2)}{m(p-2)} \frac{\sigma^p}{n}
\end{aligned}
$$

where the first inequality follows from (29), and the second from the assumption on the class of distributions. Now, for independent Rademacher variables $(\varepsilon_i)_{i=1}^n$, we have

$$
\begin{aligned}
\mathrm{E}\left[\sup_{\|v\|_H=1}\sum_{i=1}^n\varepsilon_i Z_{i,v}\right] &= \mathrm{E}\left[\sup_{\|v\|_H=1}\left\langle\frac{1}{n}\sum_{i=1}^n\varepsilon_i\nabla e(\langle w^*,X_i\rangle-Y_i),v\right\rangle\right] \\
&= \mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\nabla e(\langle w^*,X_i\rangle-Y_i)\right\|_{H^{-1}}\right] \\
&\leq \mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\nabla e(\langle w^*,X_i\rangle-Y_i)\right\|_{H^{-1}}^2\right]^{1/2} \\
&= \sqrt{\frac{\mathrm{E}[\xi^{2(p-1)}\|X\|_{H^{-1}}^2]}{n}} \\
&\leq \sqrt{\frac{\operatorname{ess\,sup}\left(\mathrm{E}\left[\xi^{2(p-1)}\mid X\right]\right)}{\operatorname{ess\,inf}(\mathrm{E}[\xi^{p-2}\mid X]))}\cdot\frac{\mathrm{E}\left[\|X\|_{\Sigma^{-1}}^2\right]}{n}} \\
&\leq \sqrt{\frac{m(2p-2)}{m(p-2)}\cdot\frac{\sigma^p d}{n}}
\end{aligned}
$$

Where again we have used the assumption on the class of distributions, and where we used (29) in the penultimate line. Recalling that $k = 8\log(2/(\delta/2))$ from the statement of the theorem, and applying Lemma 43 with the above constants yields the result. ∎

The second statement is also similar to Corollary 46. The additional challenge here is that second order Taylor expansion is not exact.

Corollary 49

$$
\mathrm{P}\left(\sup_{\|v-w^*\|_H<r}\psi_k(w^*,v)>r\cdot\Delta_p(\delta)\right)\leq\delta/2.
$$

Proof By Lemma 2.5 in Adil et al. (2023), we have that, for all $t,s\in\mathbb{R}$,

$$
e(t)-e(s)-e'(s)(t-s)\geq\frac{1}{8(p-1)}e''(s)(t-s)^2
$$

Therefore,

$$
\begin{aligned}
\sup_{\|v-w^*\|_H<r}\psi_k(w^*,v) &= \sup_{\|v-w^*\|_H<r}\varphi_k[(e(\langle w^*,X_i\rangle-Y_i)-e(\langle v,X_i\rangle-Y_i))_{i=1}^n] \\
&\leq \sup_{\|v-w^*\|_H<r}\varphi_k\left[\left(-\langle\nabla e(\langle w^*,X_i\rangle-Y_i),v-w^*\rangle-\frac{|\xi_i|^{p-2}}{8(p-1)}\langle X_i,v-w^*\rangle^2\right)_{i=1}^n\right] \\
&\leq \sup_{\|v-w^*\|_H<r}\varphi_k[(-\langle\nabla e(\langle w^*,X_i\rangle-Y_i),v-w^*\rangle)_{i=1}^n] \\
&= \sup_{\|v-w^*\|_H<r}r\cdot\varphi_k\left[\left(\left\langle\nabla e(\langle w^*,X_i\rangle-Y_i),\frac{v-w^*}{r}\right\rangle\right)_{i=1}^n\right] \\
&= r\cdot\sup_{\|v\|_H<1}\varphi_k[(\langle\nabla e(\langle w^*,X_i\rangle-Y_i),v\rangle)_{i=1}^n]
\end{aligned}
$$

where the second line is by the inequality cited above, and the third by dropping negative terms. Applying Lemma 48 to the last line finishes the proof. ∎

It remains to show the analogue of Lemma 47. This is the most technical part of the proof.

Lemma 50 *Let $r \in [32(p-1)\Delta_p(\delta), \infty)$. Then*

$$\mathrm{P}\left(\inf_{\|v-w^*\|_H \geq r} \psi_k(v, w^*) < \frac{r^2}{32(p-1)} - r\Delta_p(\delta)\right) \leq \delta$$

Proof We start with the case $\|v - w^*\|_H = r$. We have, using the quoted lemma in the proof of Corollary 49,

$$\inf_{\|v-w^*\|_H=r} \psi_k(v, w^*)$$

$$= \inf_{\|v-w^*\|_H=r} n^{-1}\varphi_k[(e(\langle v, X_i\rangle - Y_i) - e(\langle w^*, X_i\rangle - Y_i))_{i=1}^n]$$

$$\geq \inf_{\|v-w^*\|_H=r} n^{-1}\varphi_k\left[\left(\langle \nabla e(\langle w^*, X_i\rangle - Y_i), v - w^*\rangle + \frac{|\xi_i|^{p-2}}{8(p-1)}\langle v - w^*, X_i\rangle^2\right)_{i=1}^n\right]$$

$$= \inf_{v \in S^{d-1}} n^{-1}\varphi_k\left[\left(r \cdot \langle \nabla e(\langle w^*, X_i\rangle - Y_i), H^{-1/2}v\rangle + r^2 \cdot \frac{|\xi_i|^{p-2}}{8(p-1)}\langle v, H^{-1/2}X_i\rangle^2\right)_{i=1}^n\right].$$

Now define the random vector $W := |\xi|^{(p-2)/2} \cdot X$, whose (uncentered) covariance matrix is H. Further define $\widetilde{W}_i := H^{-1/2}W_i$, and $Z_{i,v} := \langle v, \widetilde{W}_i\rangle^2$ for $(i, v) \in [n] \times S^{d-1}$. Then we have by Lemma 39,

$$\inf_{\|v-w^*\|_H=r} \psi_k(v, w^*)$$

$$\geq r \cdot \inf_{v \in S^{d-1}} n^{-1}\varphi_k\left[\left(\langle \nabla e(\langle w^*, X_i\rangle - Y_i), H^{-1/2}v\rangle\right)_{i=1}^n\right] + \frac{r^2}{8(p-1)} \cdot \inf_{v \in S^{d-1}} n^{-1}\sum_{i=1}^{n-2k} Z_{i,v}^*$$

$$= \frac{r^2}{8(p-1)} \cdot \inf_{v \in S^{d-1}} n^{-1}\sum_{i=1}^{n-2k} Z_{i,v}^* - r \cdot \sup_{\|v\|_H=1} n^{-1}\varphi_k[(\langle \nabla e(\langle w^*, X_i\rangle - Y_i), v\rangle)_{i=1}^n]$$

The second term is bounded with probability $1 - \delta/2$ by $r \cdot \Delta_p(\delta)$ by Lemma 48. For the first term, we claim that the restriction on the sample size in Theorem 12 is chosen such that by Proposition 17, with probability at least $1 - \delta^2/2 \geq 1 - \delta/2$

$$\inf_{v \in S^{d-1}} n^{-1}\sum_{i=1}^{n-2k} Z_{i,v}^* \geq \frac{1}{4} \tag{30}$$

51

Let us show why this is true. Let P_W be the distribution of W, and notice that

$$\lambda_{\max}(S(P_W)) + 1 = \sup_{v \in S^{d-1}} \mathrm{E}\Big[\|W\|^2_{H^{-1}} \langle v, H^{-1/2}W \rangle^2\Big]$$

$$= \sup_{v \in S^{d-1}} \mathrm{E}\Big[|\xi|^{2(p-2)} \cdot \|X\|^2_{H^{-1}} \langle H^{-1/2}v, X \rangle^2\Big]$$

$$\leq \mathrm{ess\,sup}(\mathrm{E}[|\xi|^{2(p-2)} \mid X]) \cdot \sup_{\|v\|_H = 1} \mathrm{E}\Big[\|X\|^2_{H^{-1}} \langle v, X \rangle^2\Big]$$

$$\leq \frac{C^{(p-2)/(p-1)}}{c^2} \cdot \sup_{\|v\|_\Sigma = 1} \mathrm{E}\Big[\|X\|_{\Sigma^{-1}} \langle v, X \rangle^2\Big]$$

$$\leq \left(\frac{m(2p-2)\sigma^p}{m(p-2)}\right)^{\frac{p-2}{p-1}} \frac{1}{\mu^{p/(p-1)}} \cdot [\lambda_{\max}(S(P_X)) + 1]$$

where the fourth line follows by Jensen's inequality, and the last line by the properties of the class of distributions. Note that this upper bound holds uniformly over all members of $\mathcal{P}_p(P_X, \sigma^2, \mu)$. Through a very similar argument, one may show

$$R(P_W) + 1 \leq \left(\frac{m(2p-2)\sigma^p}{m(p-2)}\right)^{\frac{p-2}{p-1}} \frac{1}{\mu^{p/(p-1)}} \cdot [R(P_X) + 1]$$

It is then straightforward to apply Proposition 18 under the above bounds and the sample size restriction and conclude that the claim (30) is true.

Therefore, with probability at least $1 - \delta$

$$\inf_{\|v-w^*\|_H = r} \psi_k(v, w^*) \geq \frac{r^2}{32(p-1)} - r\Delta_p(\delta).$$

We now extend this to all vectors $w \in \mathbb{R}^d$ such that $\|w - w^*\|_H \geq r$. On the same event, if $\|w - w^*\|_H = R > r$, then $v := w^* + \frac{r}{R}(w - w^*)$ satisfies $\|v - w^*\|_H = r$, and

$$\psi_k(w, w^*) \geq n^{-1} \varphi_k \left[\left(\langle \nabla e(\langle w^*, X_i \rangle - Y_i), w - w^* \rangle + \frac{|\xi_i|^{p-2}}{8(p-1)} \langle w - w^*, X_i \rangle^2\right)_{i=1}^n\right]$$

$$= n^{-1} \varphi_k \left[\left(\frac{R}{r} \langle \nabla e(\langle w^*, X_i \rangle - Y_i), v - w^* \rangle + \frac{R^2}{r^2} \frac{|\xi_i|^{p-2}}{8(p-1)} \langle v - w^*, X_i \rangle^2\right)_{i=1}^n\right]$$

$$\geq n^{-1} \varphi_k \left[\left(\frac{R}{r} \langle \nabla e(\langle w^*, X_i \rangle - Y_i), v - w^* \rangle + \frac{R}{r} \frac{|\xi_i|^{p-2}}{8(p-1)} \langle v - w^*, X_i \rangle^2\right)_{i=1}^n\right]$$

$$= \frac{R}{r} \cdot \left(\frac{r^2}{32(p-1)} - r\Delta_p(\delta)\right)$$

$$\geq \frac{r^2}{32(p-1)} - r\Delta_p(\delta)$$

where the first inequality follows from the fact that $R/r > 1$, $\langle v-w^*, X_i \rangle^2 > 0$, and Lemma 38, and the second inequality follows from the fact that by the condition on r, we have $\frac{r^2}{32(p-1)} - r\Delta_p(\delta) \geq 0$ on the event we are considering. ∎

Finally, we present the main proof. For the first step, we localize $\hat{w}_{n,k}$ using the lemmas we just proved. In particular, let $r := 96(p-1)\Delta_p(\delta)$. Then following the same argument as in the proof of Theorem 11, we obtain that with probability at least $1 - \delta$

$$\psi_k(\hat{w}_{n,k}, w^*) \leq r \cdot \Delta_p(\delta) = 96(p-1)\Delta_p^2(\delta)$$

On the other hand, and on the same event,

$$\inf_{\|v - w^*\|_H \geq r} \psi_k(v, w^*) \geq \frac{r^2}{32(p-1)} - r\Delta_p(\delta) = 192(p-1)^2\Delta_p^2(\delta).$$

Therefore,

$$\|\hat{w}_{n,k} - w^*\|_H \leq 96(p-1)\Delta_p(\delta)$$

It remains to bound the excess expected error. By Lemma 2.5 in Adil et al. (2023), we have the upper bound

$$e(t) - e(s) - e'(s)(t - s) \leq 4e''(s)(t - s)^2 + 2p^{p-2}|t - s|^p$$

Integrating this bound we obtain

$$\mathcal{E}(\hat{w}_{n,k}) \leq 4\|\hat{w}_{n,k} - w^*\|_H^2 + 2p^{p-2}\,\mathrm{E}[|\langle \hat{w}_{n,k} - w^*, X\rangle|^p].$$

We have control over the first term. We need to control the second, in a noise-independent way. We have, for any $w \in \mathbb{R}^d$, by (29)

$$\|w\|_H \geq \sqrt{c} \cdot \|w\|_\Sigma \geq \sqrt{\mu} \cdot \mathrm{E}\big[\langle w, X\rangle^2\big]^{1/2}$$

Therefore

$$\sup_{w \in \mathbb{R}^d \setminus \{0\}} \frac{\mathrm{E}[|\langle w, X\rangle|^p]^{1/p}}{\|w\|_H} \leq \frac{1}{\sqrt{\mu}} \sup_{w \in \mathbb{R}^d \setminus \{0\}} \frac{\mathrm{E}[|\langle w, X\rangle|^p]^{1/p}}{\mathrm{E}[\langle w, X\rangle^2]^{1/2}} = \frac{N(P_X, p)}{\sqrt{\mu}}.$$

Using this we obtain

$$\mathcal{E}(\hat{w}_{n,k}) \leq 4\|\hat{w}_{n,k} - w^*\|_H^2 + 2p^{p-2}\frac{N^p(P_X, p)}{\mu^{p/2}} \cdot \|\hat{w}_{n,k} - w^*\|_H^p$$

Under the restriction on the sample size stated in the theorem, in particular the second term, we have on the same event

$$2p^{p-2}\frac{N^p(P_X, p)}{\mu^{p/2}} \cdot \|\hat{w}_{n,k} - w^*\|_H^p \leq 4\|\hat{w}_{n,k} - w^*\|_H^2.$$

Hence with probability at least $1 - \delta$

$$\mathcal{E}(\hat{w}_{n,k}) \leq 8 \cdot [96(p-1)\Delta_p(\delta)]^2$$

Replacing $\Delta_p(\delta)$ with its value we recover the desired bound.

Appendix F. Proofs of Section 5

F.1. Proof of Proposition 17

Proof Asymptotic lower bound. By the central limit theorem, as $n \to \infty$, and by the finiteness of the fourth moments of P_X,

$$\sqrt{n}(\widetilde{\Sigma}_n - I) \xrightarrow{d} G,$$

where G is a centred symmetric Gaussian matrix with covariance

$$\mathrm{E}[g_{ij}g_{st}] = \mathrm{E}\Big[(\widetilde{X}_i\widetilde{X}_j - I_{i,j})(\widetilde{X}_s\widetilde{X}_t - I_{s,t})\Big],$$

for $i, j, s, t \in [d]$. Now since G is Gaussian and centred, we have $G \overset{d}{=} -G$. On the one hand, by the continuous mapping theorem, this implies

$$\sqrt{n}(1 - \lambda_{\min}(\widetilde{\Sigma}_n)) = \sqrt{n}\lambda_{\max}(I - \widetilde{\Sigma}_n) \xrightarrow{d} \lambda_{\max}(G).$$

On the other, $\lambda_{\max}(G) \overset{d}{=} \lambda_{\max}(-G) = -\lambda_{\min}(G)$, and therefore for $t \geq 0$,

$$\mathrm{P}(\|G\|_{\mathrm{op}} > t) = \mathrm{P}(\lambda_{\max}(G) > t \text{ or } \lambda_{\min}(G) < -t)$$
$$\leq \mathrm{P}(\lambda_{\max}(G) > t) + \mathrm{P}(\lambda_{\min}(G) < -t) = 2\,\mathrm{P}(\lambda_{\max}(G) > t).$$

so we conclude that

$$\lim_{n\to\infty} \mathrm{P}\left(1 - \lambda_{\min}(\widetilde{\Sigma}_n) \leq \frac{t}{\sqrt{n}}\right) = \mathrm{P}(\lambda_{\max}(G) \leq t) \leq \frac{1}{2}(1 + \mathrm{P}(\|G\|_{\mathrm{op}} \leq t)).$$

Now since convergence in distribution implies the pointwise convergence of quantiles, we obtain

$$\lim_{n\to\infty} \sqrt{n} \cdot Q_{1-\lambda_{\min}(\widetilde{\Sigma}_n)}(1 - \delta) = Q_{\lambda_{\max}(G)}(1 - \delta) \geq Q_{\|G\|_{\mathrm{op}}}(1 - 2\delta)$$

It remains to lower bound this last quantile. We do this by deriving two upper bounds on the CDF of $\|G\|_{\mathrm{op}}$. Let $v_* := \mathrm{argmax}_{v \in S^{d-1}} \mathrm{E}\Big[\big(\langle v, \widetilde{X}\rangle^2 - 1\big)^2\Big]$ and note that $v_*^T G v_* \sim \mathcal{N}(0, R(P_X))$. Therefore by Lemma 27,

$$\mathrm{P}(\|G\|_{\mathrm{op}} \leq t) \leq \mathrm{P}\big(\big|v_*^T G v_*^T\big| \leq t\big) \leq \sqrt{1 - \exp\left(-\frac{2t^2}{\pi R(P_X)}\right)}$$

On the other hand, it can be shown that $\|G\|_{\mathrm{op}}$ is a Lipschitz function of a standard normal vector (see e.g. Van Handel (2017)), with Lipschitz constant $\sqrt{R(P_X)}$. Therefore by Gaussian concentration (Lemma 30)

$$\mathrm{P}(\|G\|_{op} \leq t) = \mathrm{P}(\mathrm{E}[\|G\|_{op}] - \|G\|_{op} > \mathrm{E}[\|G\|_{op}] - t) \leq \exp\left(-\frac{(\mathrm{E}[\|G\|_{op}] - t)^2}{2R(P_X)}\right).$$

Now note that since $v^T G v$ is a Gaussian random variable for any $v \in \mathbb{R}^d$,

$$\mathrm{E}\big[\|G\|_{\mathrm{op}}\big] \geq \sup_{v \in \mathbb{S}^{d-1}} \mathrm{E}\big[\big|v^T G v\big|\big] = \sqrt{\frac{2}{\pi}} \mathrm{E}\big[(v^T G v)^2\big]^{1/2} = \sqrt{\frac{2}{\pi}} \sqrt{R(P_X)}$$

where the first equality is an explicit calculation of the first absolute moment of a Gaussian random variable. Bounding the right-most term in the previous display, we obtain

$$
\mathrm{P}\big(\|G\|_{\mathrm{op}} \leq t\big) \leq \exp\left(-\frac{(\mathrm{E}[\|G\|_{op}] - t)^2}{\pi \, \mathrm{E}\big[\|G\|_{\mathrm{op}}\big]^2}\right)
$$

Using the two bounds on the CDF of $\|G\|_{\mathrm{op}}$ and the second item of Lemma 20, we obtain the following lower bound

$$
Q_{\|G\|_{op}}(1 - 2\delta) \geq \frac{1}{2}\,\mathrm{E}\big[\|G\|_{\mathrm{op}}\big]\left(1 - \sqrt{\pi \log\left(\frac{1}{1 - 2\delta}\right)}\right) + \frac{1}{2}\sqrt{\frac{\pi}{2}}\sqrt{R(P_X)\log\left(\frac{1}{4\delta}\right)}
$$

using the restriction on $\delta \in (0, 0.1)$, we obtain

$$
Q_{\|G\|_{op}}(1 - 2\delta) \geq \frac{1}{20}\,\mathrm{E}\big[\|G\|_{\mathrm{op}}\big] + \frac{1}{2}\sqrt{R(P_X)\log(1/4\delta)}
$$

Finally by the Gaussian Poincaré inequality (Lemma 30)

$$
\mathrm{E}\big[\|G\|_{\mathrm{op}}^2\big] - (\mathrm{E}\big[\|G\|_{\mathrm{op}}\big])^2 \leq R(P_X) \leq \frac{\pi}{2}(\mathrm{E}\big[\|G\|_{\mathrm{op}}\big])^2
$$

rearranging yields

$$
\mathrm{E}\big[\|G\|_{\mathrm{op}}\big] \geq \frac{1}{\sqrt{1 + \pi/2}}\,\mathrm{E}\big[\|G\|_{\mathrm{op}}^2\big]^{1/2} \geq \frac{\big\|\mathrm{E}\big[G^2\big]\big\|_{\mathrm{op}}^{1/2}}{\sqrt{1 + \pi/2}} = \sqrt{\frac{\lambda_{\max}(S)}{1 + \pi/2}}
$$

and therefore

$$
Q_{\|G\|_{\mathrm{op}}}(1 - 2\delta) \geq \frac{\sqrt{\lambda_{\max}(S)}}{40} + \frac{1}{2}\sqrt{R(P_X)\log(1/4\delta)}
$$

This concludes the proof of the lower bound.

Upper bound. We have the variational representation

$$
1 - \lambda_{\min}(\widetilde{\Sigma}_n) = \lambda_{\max}(I - \widetilde{\Sigma}_n) = \sup_{v \in S^{d-1}} \sum_{i=1}^{n} \underbrace{\frac{1}{n}(\mathrm{E}\big[\langle v, \widetilde{X}\rangle^2\big] - \langle v, \widetilde{X}_i\rangle^2)}_{Z_{i,v} :=}. \tag{31}
$$

Now the processes $(\{Z_{i,v}\}_{v \in S^{d-1}})_{i=1}^{d}$ are i.i.d., $\mathrm{E}[Z_{i,v}] = 0$, and $Z_{i,v} \leq n^{-1}$ for all $(i, v) \in [n] \times S^{d-1}$, so that by Bousquet's inequality (Bousquet, 2002), with probability at least $1 - \delta$

$$
\sup_{v \in S^{d-1}} \sum_{i=1}^{n} Z_{i,v} < 2\,\mathrm{E}\left[\sup_{v \in S^{d-1}} \sum_{i=1}^{n} Z_{i,v}\right] + \sqrt{\frac{2R(P_X)\log(1/\delta)}{n}} + \frac{4\log(1/\delta)}{3n} \tag{32}
$$

It remains to bound the expectation in (32). We may rewrite it as

$$
\mathrm{E}\left[\sup_{v \in S^{d-1}} \sum_{i=1}^{n} Z_{i,v}\right] = \mathrm{E}\left[\sup_{v \in S^{d-1}} v^T \left\{\sum_{i=1}^{n} \frac{1}{n}(I - \widetilde{X}_i \widetilde{X}_i^T)\right\} v\right] = \mathrm{E}\left[\lambda_{\max}\left(\sum_{i=1}^{n} \frac{1}{n}(I - \widetilde{X}_i \widetilde{X}_i^T)\right)\right]
$$

Define the matrices $Y_i := \frac{1}{n}(I - \widetilde{X}_i \widetilde{X}_i^T)$ and notice that they are i.i.d. and satisfy $\lambda_{\max}(Y_i) = n^{-1}$, so that by the Matrix Bernstein inequality (Tropp, 2015b, Theorem 6.6.1) we obtain

$$\mathrm{E}\left[\lambda_{\max}\left(\sum_{i=1}^n Y_i\right)\right] \leq \sqrt{\frac{2\lambda_{\max}(S)\log(3d)}{n}} + \frac{\log(3d)}{3n}. \tag{33}$$

Combining (31), (32), and (33) yields the desired result. ∎

F.2. Proof of Proposition 18

Proof Define $\tilde{S} := S(P_X) + I = \mathrm{E}\left[\|\widetilde{X}\|_2^2 \widetilde{X} \widetilde{X}^T\right]$ and $\tilde{R} := R(P_X) + 1 = \sup_{v \in S^{d-1}} \mathrm{E}\left[\langle v, \widetilde{X}\rangle^4\right]$. Let

$$B := \sqrt{\frac{n\lambda_{\max}(\tilde{S})}{4(1 + 2\lceil \log(d)\rceil)}},$$

and define $X_B := \widetilde{X} \cdot \mathbb{1}_{[0,B)}(\|\widetilde{X}\|_2^2)$, $\Sigma_B := \mathrm{E}\left[X_B X_B^T\right]$, $\tilde{S}_B := \mathrm{E}\left[(X_B X_B^T)^2\right]$, and $\tilde{R}_B := \sup_{v \in S^{d-1}} \mathrm{E}\left[\langle v, X_B\rangle^4\right]$. Note that $\lambda_{\max}(\tilde{S}_B) \leq \lambda_{\max}(\tilde{S})$ and $\tilde{R}_B \leq \tilde{R}$. For $(i,v) \in [n] \times S^{d-1}$, define $Z_{i,v} := \langle v, X_{B,i}\rangle^2$, and note that $(Z_{i,v})_{i=1}^n$ are i.i.d. with mean $m(v) := \mathrm{E}[\langle v, X_B\rangle^2]$ and $Y_{i,v} \geq Z_{i,v}$. Now we have by Lemma 37

$$\sup_{v \in S^{d-1}} \sum_{i=k+1}^{n-k} \mathrm{E}\left[\langle v, \widetilde{X}\rangle^2\right] - Y_{i,v}^* \leq (n-2k)\sup_{v \in S^{d-1}} \mathrm{E}\left[\langle v, \widetilde{X}\rangle^2\right] - \mathrm{E}\left[\langle v, X_B\rangle^2\right] + \sup_{v \in S^{d-1}} \sum_{i=k+1}^{n-k} \mathrm{E}\left[\langle v, X_B\rangle^2\right] - Z_{i,v}^*$$

The first term is bounded by

$$\begin{aligned}
\sup_{v \in S^{d-1}} \mathrm{E}\left[\langle v, \widetilde{X}\rangle^2\right] - \mathrm{E}\left[\langle v, X_B\rangle^2\right] &= \sup_{v \in S^{d-1}} \mathrm{E}\left[\langle v, \widetilde{X}\rangle^2 \mathbb{1}_{[B,\infty)}(\|\widetilde{X}\|_2^2)\right] \\
&= \sup_{v \in S^{d-1}} \mathrm{E}\left[\langle v, \widetilde{X}\rangle^2 \|\widetilde{X}\|_2^2 \frac{1}{\|\widetilde{X}\|_2^2} \mathbb{1}_{[B,\infty)}(\|\widetilde{X}\|_2^2)\right] \\
&\leq \frac{\lambda_{\max}(\tilde{S})}{B} = \sqrt{4(1 + 2\lceil\log(d)\rceil)}\sqrt{\frac{\lambda_{\max}(\tilde{S})}{n}}
\end{aligned}$$

For the second term, define, for $(i, v) \in [n] \times S^{d-1}$, $W_{i,v} := \mathrm{E}\big[\langle v, X_B \rangle^2\big] - Z_{i,v}$, and note that $\mathrm{E}[W_{i,v}] = 0$, $\mathrm{E}\Big[W_{i,v}^2\Big] \leq \tilde{R}$. Furthermore

$$
2\,\mathrm{E}\left[\sup_{v \in S^{d-1}} \sum_{i=1}^{n} \varepsilon_i W_{i,v}\right]
$$

$$
= 2\,\mathrm{E}\left[\sup_{v \in S^{d-1}} v^T\left\{\sum_{i=1}^{n} \varepsilon_i(\Sigma_B - X_{B,i} X_{B,i}^T)\right\} v\right]
$$

$$
\leq 2\,\mathrm{E}\left[\left\|\sum_{i=1}^{n} \varepsilon_i(\Sigma_B - X_{B,i} X_{B,i}^T)\right\|_{op}\right]
$$

$$
\leq \sqrt{4(1 + 2\lceil \log(d) \rceil)}\sqrt{n\lambda_{\max}(\tilde{S}_B)} + 4(1 + 2\lceil \log(d) \rceil)\,\mathrm{E}\left[\max_{i \in [n]}\|X_{B,i} X_{B,i}^T - \Sigma_B\|_{op}^2\right]^{1/2}
$$

$$
\leq \sqrt{4(1 + 2\lceil \log(d) \rceil)}\sqrt{n\lambda_{\max}(\tilde{S})} + 4(1 + 2\lceil \log(d) \rceil)\,\mathrm{E}\left[\max\left\{\max_{i \in [n]}\|X_{B,i}\|_2^2, \lambda_{\max}(\Sigma_B)\right\}^2\right]^{1/2}
$$

$$
\leq \sqrt{4(1 + 2\lceil \log(d) \rceil)}\sqrt{n\lambda_{\max}(\tilde{S})} + 4(1 + 2\lceil \log(d) \rceil)\max\{B, 1\}
$$

$$
= 4\sqrt{1 + 2\lceil \log(d) \rceil}\sqrt{n\lambda_{\max}(\tilde{S})}
$$

where the fourth line follows from the proof of the second item of Theorem 5.1 in Tropp (2015a), the sixth line follows from the fact that $\|X_B\|_2^2 \leq B$, and the last line follows from the condition on n and the fact that $\lambda_{\max}(\tilde{S}) \geq 1$, which itself follows from the positive semi-definiteness of $S(P_X)$. Defining $W_v = (W_{i,v})_{i=1}^n$ for $v \in S^{d-1}$, and appealing to Lemma 43, we may therefore bound the second term as follows, with probability at least $1 - \delta$

$$
\sup_{v \in S^{d-1}} \sum_{i=k+1}^{n-k} \mathrm{E}\big[\langle v, X_B \rangle^2\big] - Z_{i,v}^* = \sup_{v \in S^{d-1}} \sum_{i=k+1}^{n-k} W_{i,v}^*
$$

$$
\leq \sup_{v \in S^{d-1}} \varphi_k(W_v) + \sup_{v \in S^{d-1}} k(|W_{1+k,v}| + |W_{n-k,v}|)
$$

$$
\leq 96 \cdot \left(\sqrt{4(1 + 2\lceil \log(d) \rceil)}\sqrt{n\lambda_{\max}(\tilde{S})} + \sqrt{n\tilde{R}\log(2/\delta)}\right)
$$

Combining the bounds, and bounding $\sqrt{4(1 + \lceil \log(d) \rceil)} \leq 8\log(6d)$ yields the result. ∎