Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices

Jaouad Mourtada*

Abstract

We consider random-design linear prediction and related questions on the lower tail of random matrices. It is known that, under boundedness constraints, the minimax risk is of order d/n in dimension d with n samples. Here, we study the minimax expected excess risk over the full linear class, depending on the distribution of covariates. First, the least squares estimator is exactly minimax optimal in the well-specified case, for every distribution of covariates. We express the minimax risk in terms of the distribution of statistical leverage scores of individual samples, and deduce a minimax lower bound of $d/(n-d+1)$ for any covariate distribution, nearly matching the risk for Gaussian design. We then obtain sharp nonasymptotic upper bounds for covariates that satisfy a "small ball"-type regularity condition in both well-specified and misspecified cases.

Our main technical contribution is the study of the lower tail of the smallest singular value of empirical covariance matrices at small values. We establish a lower bound on this lower tail, valid for any distribution in dimension $d \geqslant 2$, together with a matching upper bound under a necessary regularity condition. Our proof relies on the PAC-Bayes technique for controlling empirical processes, and extends an analysis of Oliveira [Oli16] devoted to a different part of the lower tail.

1 Introduction

Linear least-squares regression, also called random-design linear regression or linear aggregation, is one of the basic statistical prediction problems. Specifically, given a random pair (X, Y) where X is a covariate vector in $\mathbf{R}^d$ and Y is a scalar response, the aim is to predict Y using a linear function $\langle \beta, X \rangle = \beta^\top X$ (with $\beta \in \mathbf{R}^d$) of X as well as possible, in a sense measured by the prediction risk with squared error $R(\beta) = \mathbb{E}[(Y - \langle \beta, X \rangle)^2]$. The best prediction is achieved by the population risk minimizer β^*, which equals:

$$\beta^* = \Sigma^{-1}\mathbb{E}[YX]$$

where $\Sigma := \mathbb{E}[XX^\top]$, assuming that both Σ and $\mathbb{E}[YX]$ are well-defined and that Σ is invertible. In the statistical setting considered here, the joint distribution P of the pair (X, Y) is unknown. The goal is then, given a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of n i.i.d. realizations of P, to find a predictor (also called *estimator*) $\widehat{\beta}_n$ with small *excess risk*

$$\mathcal{E}_P(\widehat{\beta}_n) := R(\widehat{\beta}_n) - R(\beta^*) = \|\widehat{\beta}_n - \beta^*\|_\Sigma^2 \,,$$

where we define $\|\beta\|_\Sigma^2 := \langle \Sigma\beta, \beta \rangle = \|\Sigma^{1/2}\beta\|^2$. Arguably the most common procedure is the *Ordinary Least Squares* (OLS) estimator (that is, the empirical risk minimizer), defined by

$$\widehat{\beta}_n^{\mathrm{LS}} := \operatorname*{arg\,min}_{\beta \in \mathbf{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^n (Y_i - \langle \beta, X_i \rangle)^2 \right\} = \widehat{\Sigma}_n^{-1} \cdot \frac{1}{n} \sum_{i=1}^n Y_i X_i \,,$$

*CREST, ENSAE, Institut Polytechnique de Paris, France; jaouad.mourtada@ensae.fr

with $\widehat{\Sigma}_n := n^{-1} \sum_{i=1}^n X_i X_i^\top$ the sample covariance matrix.

Linear classes are of particular importance to regression problems, both in themselves and since they naturally appear in the context of nonparametric estimation [GKKW02, Tsy09]. In this note, we analyze this problem from a decision-theoretic perspective, focusing on the minimax excess risk with respect to the full linear class $\mathcal{F} = \{x \mapsto \langle \beta, x \rangle : \beta \in \mathbf{R}^d\}$, and in particular on its dependence on the distribution of X. The minimax perspective is relevant when little is known or assumed on the optimal parameter β^*. Specifically, define the *minimax excess risk* (see, e.g., [LC98]) with respect to $\mathcal{F}$ under a set $\mathcal{P}$ of joint distributions P on (X, Y) as:

$$\inf_{\widehat{\beta}_n} \sup_{P \in \mathcal{P}} \mathbb{E}[\mathcal{E}_P(\widehat{\beta}_n)] = \inf_{\widehat{\beta}_n} \sup_{P \in \mathcal{P}} \Big(\mathbb{E}[R(\widehat{\beta}_n)] - \inf_{\beta \in \mathbf{R}^d} R(\beta) \Big) , \tag{1}$$

where the infimum in (1) spans over all estimators $\widehat{\beta}_n$ based on n samples, while the expectation and the risk R depend the underlying distribution P. Our aim is to characterize the influence of the distribution P_X of covariates on the hardness of the problem. Hence, our considered classes $\mathcal{P}$ of distributions are obtained by fixing the marginal distribution of X, and letting the optimal regression parameter β^* vary freely in $\mathbf{R}^d$ (see Section 2).

Some minimal regularity condition on the distribution P_X is required to ensure even finiteness of the minimax risk (1) in the random-design setting. Indeed, assume that the distribution P_X charges some positive mass on a hyperplane $H \subset \mathbf{R}^d$ (we call such a distribution *degenerate*, see Definition 1). Then, with positive probability, all points $X_1, \dots, X_n$ in the sample lie within H, so that the component of the optimal parameter β^* which is orthogonal to H cannot be estimated. However, this component matters for out-of-sample prediction, in case the point X for which one wishes to compute prediction does not belong to H. Such a degeneracy (or quantitative variants, where P_X puts too much mass at the neighborhood of a hyperplane) turns out to be the main obstruction to achieving controlled uniform excess risk over $\mathbf{R}^d$.

The second part of this note (Section 3) is devoted to the study of the *sample covariance matrix*

$$\widehat{\Sigma}_n := \frac{1}{n} \sum_{i=1}^n X_i X_i^\top , \tag{2}$$

where $X_1, \dots, X_n$ are i.i.d. samples from P_X. Indeed, upper bounds on the minimax risk require a control of relative deviations of the empirical covariance matrix $\widehat{\Sigma}_n$ with respect to its population counterpart Σ, in the form of *negative moments* of the rescaled covariance matrix $\widetilde{\Sigma}_n := \Sigma^{-1/2} \widehat{\Sigma}_n \Sigma^{-1/2}$, namely

$$\mathbb{E}[\lambda_{\min}(\widetilde{\Sigma}_n)^{-q}] \tag{3}$$

where $q \geqslant 1$ and $\lambda_{\min}(A)$ is the smallest eigenvalue of symmetric matrix A.

Control of lower relative deviations of $\widehat{\Sigma}_n$ with respect to Σ can be expressed in terms of lower-tail bounds, of the form

$$\mathbb{P}\big(\lambda_{\min}(\widetilde{\Sigma}_n) \leqslant t\big) \leqslant \delta , \tag{4}$$

where $t, \delta \in (0, 1)$. Sub-Gaussian tail bounds for $\lambda_{\min}(\widetilde{\Sigma}_n)$, of the form (4) with

$$\delta = \exp\Big(-cn \Big(1 - C\sqrt{\frac{d}{n}} - t \Big)_+^2 \Big)$$

for some constants c, C depending on P_X, as well as similar bounds for the largest eigenvalue $\lambda_{\max}(\widetilde{\Sigma}_n)$, can be obtained under the (strong) assumption that X is sub-Gaussian (see, e.g., [Ver12]). Remarkably, it is shown in [Oli16, KM15] that such bounds can be obtained for the *smallest* eigenvalue under much weaker assumptions on X, namely bounded fourth moments of linear marginals of X.

While sub-Gaussian bounds provide a precise control of deviations (4) for $t \in (c, 1 - C\sqrt{d/n})$ (for some constants c, C), they do not suffice to control moments of $\lambda_{\min}(\widetilde{\Sigma}_n)^{-1}$. Indeed, such bounds "saturate" in the sense that $\delta = \delta(t)$ does not tend to 0 as $t \to 0$; in other words, they provide no nonvacuous guarantee (4) with $t > 0$ as the confidence level $1 - \delta$ tends to 1. This prevents one from integrating such tail bounds and deduce a control of moments of the form (3). In fact, the covariance matrix of a sub-Gaussian matrix can be singular with positive probability (exponentially small in n), for instance for matrices with independent Bernoulli entries; in order to ensure invertibility at all confidence levels, different regularity assumptions are required. In Section 3, we complement the sub-Gaussian tail bounds by a study of non-asymptotic large deviation bounds (4) with $\delta = \exp(-n\psi(t))$ for small values of t, namely $t \in (0, c)$.

1.1 Summary of contributions

Below is an overview of our results on least squares regression, which appear in Section 2:

1. We determine the minimax excess risk in the well-specified case (where the true regression function $x \mapsto \mathbb{E}[Y|X = x]$ is linear) for every distribution P_X of features and noise level σ^2. For some "degenerate" distributions (Definition 1), the minimax risk is infinite (Proposition 1); while for non-degenerate ones, the OLS estimator is exactly minimax (Theorem 1) irrespective of P_X, σ^2.

2. We express the minimax risk in terms of the distribution of *statistical leverage scores* of samples drawn from P_X (Theorem 2). Quite intuitively, distributions of X for which leverage scores are uneven are seen to be harder from a minimax point of view. We deduce a precise minimax lower bound of $\sigma^2 d/(n - d + 1)$, valid for every distribution P_X of covariates. This lower bound nearly matches the $\sigma^2 d/(n-d-1)$ risk for centered Gaussian covariates, in both low ($d/n \to 0$) and moderate ($d/n \to \gamma \in (0, 1)$) dimensions; hence, Gaussian covariates are almost the "easiest" ones in terms of minimax risk. This provides a counterpart to results obtained in the moderate-dimensional regime for *independent* covariates from the Marchenko-Pastur law.

3. We then turn to upper bounds on the minimax risk. Under some quantitative variant of the non-degeneracy assumption (Assumption 1) together with a fourth-moment condition on P_X (Assumption 2 or 3), we show that the minimax risk is finite and scales as $(1 + o(1))\sigma^2 d/n$ when $d = o(n)$, both in the well-specified (Theorem 3) and misspecified (Proposition 3) cases. In particular, OLS is asymptotically minimax in the misspecified case as well, as $d/n \to 0$. To our knowledge, this gives the first bounds on the expected risk of the OLS estimator for general random design distribution.

The previous upper bounds rely on the study of the lower tail of the sample covariance matrix $\widehat{\Sigma}_n$, carried out in Section 3. Our contributions here are the following (assuming, to simplify notation, that $\mathbb{E}[XX^\top] = I_d$):

4. First, we establish a *lower bound* on the lower tail of $\lambda_{\min}(\widehat{\Sigma}_n)$, for $d \geqslant 2$ and *any* distribution P_X such that $\mathbb{E}[XX^\top] = I_d$, of the form: $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t) \geqslant (ct)^{n/2}$ for some numerical constant c and every $t \in (0, 1)$ (Proposition 4). We also exhibit a "small-ball" condition (Assumption 1) which is necessary to achieve similar upper bounds.

5. Under Assumption 1, we show a matching *upper bound* on the lower tail $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t)$, valid for all $t \in (0, 1)$, and in particular for small t. This result (Theorem 4) is the core technical contribution of this paper. Its proof relies the PAC-Bayesian technique for

controlling empirical processes, which was used by [Oli16] to control a different part of the lower tail; however, some non-trivial refinements (such as non-Gaussian smoothing) are needed to handle small values of t. This result can be equivalently stated as an upper bound on moments of $\lambda_{\min}(\widehat{\Sigma}_n)^{-1}$, namely $\|\lambda_{\min}(\widehat{\Sigma}_n)^{-1}\|_{L^q} = O(1)$ for $q \asymp n$ (Corollary 4).

6. Finally, we discuss in Section 3.3 the case of independent covariates. In this case, the "small-ball" condition (Assumption 1) holds naturally under mild regularity assumptions on the distribution of individual coordinates. A result of [RV14] establishes this for coordinates with bounded density; we complement it by a general anti-concentration result for linear combination of independent variables (Proposition 6), implying Assumption 1 for sufficiently "non-atomic" coordinates.

1.2 Related work

Linear least squares regression is a classical problem, and the literature on this topic is too vast to be surveyed here; we refer to [GKKW02, AC10, HKZ14] (and references therein) for a more thorough overview. In addition, while we focus on mean-squared prediction error, different criteria can be considered, as in the predictive inference literature [RWG19]. Analysis of least squares regression is most standard and straightforward in the *fixed design* setting, where the covariates $X_1, \dots, X_n$ are deterministic and the risk is evaluated within-sample; in this case, the expected excess risk of the OLS estimator is bounded by $\sigma^2 d/n$ (see, e.g., [HKZ14]).

In the random design setting considered here, a classical result [GKKW02, Theorem 11.3] states that, if $\mathrm{Var}(Y|X) \leqslant \sigma^2$ and the true regression function $g^*(x) = \mathbb{E}[Y|X = x]$ satisfies $|g^*(X)| \leqslant L$ almost surely, then the risk $R(g) = \mathbb{E}[(g(X) - Y)^2]$ of the (nonlinear) *truncated* ERM estimator, defined by $\widehat{g}_n^L(x) = \min(-L, \max(L, \langle \widehat{\beta}_n^{\mathrm{LS}}, x\rangle))$, is at most

$$\mathbb{E}[R(\widehat{g}_n^L)] - R(g^*) \leqslant 8\big(R(\beta^*) - R(g^*)\big) + C\max(\sigma^2, L^2)\frac{d(\log n + 1)}{n} \tag{5}$$

for some universal constant $C > 0$. This result is an *inexact oracle inequality*, where the risk is bounded by a constant times that of the best linear predictor β^*. Such guarantees are adequate in a nonparametric setting, where the approximation error $R(\beta^*) - R(g^*)$ of the linear model is itself of order $O(d/n)$ [GKKW02]. On the other hand, when no assumption is made on the magnitude of the approximation error, this bound does not ensure that the risk of the estimator approaches that of β^*. By contrast, in the *linear aggregation* problem as defined by [Nem00] (and studied by [Tsy03, Cat04, BTW07, AC11, HKZ14, LM16, Men15, Oli16]), one seeks to obtain excess risk bounds, also called *exact* oracle inequalities (where the constant 8 in the bound (5) is replaced by 1), with respect to the linear class. In this setting, Tsybakov [Tsy03] showed that the minimax rate of aggregation is of order $O(d/n)$, under boundedness assumptions on the regression function and on covariates. It is also worth noting that bounds on the regression function also implicitly constrain the optimal regression parameter to lie in some ball. This contrasts with the approach considered here, where minimax risk with respect to the full linear class is considered. Perhaps most different from the point of view adopted here is the approach from [Fos91, Vov01, AW01, Sha15, BKM+15], whose authors consider worst-case covariates (either in the individual sequences or in the agnostic learning setting) under boundedness assumptions on both covariates and outputs, and investigate achievable excess risk (or regret) bounds with respect to bounded balls in this case. By contrast, we take the distribution of covariates as given and allow the optimal regression parameter to be arbitrary, and study under which conditions on the covariates uniform bounds are achievable. Another type of non-uniform guarantees over linear classes is achieved by Ridge regression [Hoe62, Tik63] in the context of reproducing kernel Hilbert spaces [CS02a, CS02b, DVCR05, CDV07, SZ07, SHS09,

AC11, HKZ14], where the bounds do not depend explicitly on the dimension d, but rather on spectral properties of Σ and some norm of β^*.

This work is concerned with the expected risk. Risk bounds in probability are obtained, among others, by [AC11, HKZ14, HS16, Oli16, Men15, LM16]. While such bounds hold with high probability, the probability is upper bounded and cannot be arbitrarily close to 1, so that they cannot be integrated to control the expected risk. Indeed, some additional regularity conditions are required in order to have finite minimax risk, as will be seen below. To the best of our knowledge, the only available uniform expected risk bounds for random-design regression are obtained in the case of Gaussian covariates, where they rely on the knowledge of the closed-form distribution of inverse covariance matrices [Ste60, BF83, And03]. One reason for considering the expected risk is that it is a single scalar, which can be more tightly controlled (in terms of matching upper and lower bounds) and compared across distributions than quantiles. In addition, random-design linear regression is a classical statistical problem, which justifies its precise decision-theoretic analysis. On the other hand, expected risk only provides limited information on the tails of the risk in the high-confidence regime: in the case of heavy-tailed noise, the OLS estimator may perform poorly, and dedicated robust estimators may be required (see, e.g., [AC11] and the references in [LM19]).

Another line of work [EK13, Dic16, DM16, EK18, DW18] considers the limiting behavior of regression procedures in the high-dimensional asymptotic regime where d, n tend to infinity at a proportional rate, with their ratio kept constant [Hub73]. The results in this setting take the form of a convergence in probability of the risk to a limit depending on the ratio d/n as well as the properties of β^*. With the notable exception of [EK18], the previous results hold under the assumption that the covariates are either Gaussian, or have a joint independence structure that leads to the same limiting behavior in high dimension. In contrast, here we consider non-asymptotic bounds valid for fixed n, d, general design distribution and uniformly over $\beta^* \in \mathbf{R}^d$.

The study of spectral properties of sample covariance matrices has a rich history (see for instance [BS10, AGZ10, Tao12] and references therein); we refer to [RV10] for an overview of results (up to 2010) on the non-asymptotic control of the smallest eigenvalue of sample covariance matrices, which is the topic of Section 3. It is well-known [Ver12] that sub-Gaussian tail bounds on both the smallest and largest eigenvalues can be obtained under sub-Gaussian assumptions on covariates (see also [KL17] for operator norm concentration under general population covariance). A series of work obtained control on these quantities under weaker assumptions [ALPTJ10, MP14, SV13, Tik18]. A key observation, which has been exploited in a series of work [SV13, KM15, Oli16, Yas14, Yas15, vdGM14], is that the smallest eigenvalue can be controlled under much weaker tail assumptions than the largest one. Our study follows this line of work, but considers a different part of the lower tail, which poses additional technical difficulties; we also provide a general lower bound on the lower tail.

Notation. Throughout this text, the transpose of an $m \times n$ real matrix A is denoted $A^\top$, its trace (when $m = n$) $\mathrm{Tr}(A)$, and vectors in $\mathbf{R}^d$ are identified with $d \times 1$ column vectors. In addition, the coordinates of a vector $x \in \mathbf{R}^d$ are indicated as superscripts: $x = (x^j)_{1 \leqslant j \leqslant d}$. We also denote $\langle x, z \rangle = x^\top z = \sum_{j=1}^d (x^j) \cdot (z^j)$ the canonical scalar product of $x, z \in \mathbf{R}^d$, and $\|x\| = \langle x, x \rangle^{1/2}$ the associated Euclidean norm. In addition, for any symmetric and positive $d \times d$ matrix A, we define the scalar product $\langle x, z \rangle_A = \langle Ax, z \rangle$ and norm $\|x\|_A = \langle Ax, x \rangle^{1/2} = \|A^{1/2} x\|$. The $d \times d$ identity matrix is denoted I_d, while $S^{d-1} = \{x \in \mathbf{R}^d : \|x\| = 1\}$ refers to the unit sphere. The smallest and largest eigenvalues of a symmetric matrix A are denoted $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ respectively; if A is positive definite, then $\lambda_{\max}(A) = \|A\|_{\mathrm{op}}$ is the operator norm of A (with respect to $\|\cdot\|$), while $\lambda_{\min}(A) = \|A^{-1}\|_{\mathrm{op}}^{-1}$. We denote by $\mathrm{dist}(x, A) = \inf_{y \in A} \|x - y\|$ the

distance of $x \in \mathbf{R}^d$ to a subset $A \subset \mathbf{R}^d$.

2 Exact minimax analysis of least-squares regression

This section is devoted to the minimax analysis of the linear least-squares problem, and in particular on the dependence of its hardness on the distribution P_X of covariates. In Section 2.1, we indicate the exact minimax risk and estimator in the well-specified case, namely on the class $\mathcal{P}_{\text{well}}(P_X, \sigma^2)$. In Section 2.2, we express the minimax risk in terms of the distribution of statistical leverage scores, and deduce a general lower bound. Finally, Section 2.3 provides upper bounds on the minimax risk under some regularity condition on the distribution P_X, both in the well-specified and misspecified cases.

Throughout this note, we assume that the covariate vector X satisfies $\mathbb{E}[\|X\|^2] < +\infty$, and denote $\Sigma = \mathbb{E}[XX^\top]$ its covariance matrix (by a slight but common abuse of terminology, we refer to Σ as the covariance matrix of X even when X is not centered). In addition, we assume that Σ is invertible, or equivalently that the support of X is not contained in any hyperplane; this assumption is not restrictive (up to restricting to the span of the support of X, a linear subspace of $\mathbf{R}^d$) and only serves to simplify notations. Then, for every distribution of Y given X such that $\mathbb{E}[Y^2] < +\infty$, the risk $R(\beta) = \mathbb{E}[(\langle \beta, X\rangle - Y)^2]$ of any $\beta \in \mathbf{R}^d$ is finite; this risk is uniquely minimized by $\beta^* = \Sigma^{-1}\mathbb{E}[YX]$, where $\mathbb{E}[YX]$ is well-defined since $\mathbb{E}[\|YX\|] \leqslant \mathbb{E}[Y^2]^{1/2}\mathbb{E}[\|X\|^2]^{1/2} < +\infty$. The response Y then writes

$$Y = \langle \beta^*, X\rangle + \varepsilon\,, \tag{6}$$

where ε is the *error*, with $\mathbb{E}[\varepsilon X] = \mathbb{E}[YX] - \Sigma\beta^* = 0$. The distribution P of (X, Y) is then characterized by the distribution P_X of X, the coefficient $\beta^* \in \mathbf{R}^d$ as well as the conditional distribution of ε given X, which satisfies $\mathbb{E}[\varepsilon^2] \leqslant \mathbb{E}[Y^2] < +\infty$ and $\mathbb{E}[\varepsilon X] = 0$. Now, given a distribution P_X of covariates and a bound σ^2 on the conditional second moment of the error, define the following three classes, where Y is given by (6):

$$\begin{aligned}
\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2) &= \left\{P_{(X,Y)} : X \sim P_X,\, \beta^* \in \mathbf{R}^d,\, \varepsilon|X \sim \mathcal{N}(0, \sigma^2)\right\} \\
\mathcal{P}_{\text{well}}(P_X, \sigma^2) &= \left\{P_{(X,Y)} : X \sim P_X,\, \beta^* \in \mathbf{R}^d,\, \mathbb{E}[\varepsilon|X] = 0,\, \mathbb{E}[\varepsilon^2|X] \leqslant \sigma^2\right\} \\
\mathcal{P}_{\text{mis}}(P_X, \sigma^2) &= \left\{P_{(X,Y)} : X \sim P_X,\, \beta^* \in \mathbf{R}^d,\, \mathbb{E}[\varepsilon^2|X] \leqslant \sigma^2\right\}.
\end{aligned} \tag{7}$$

The class $\mathcal{P}_{\text{Gauss}}$ corresponds to the standard case of independent Gaussian noise, while $\mathcal{P}_{\text{well}}$ includes all *well-specified* distributions, such that the true regression function $x \mapsto \mathbb{E}[Y|X = x]$ is linear. Finally, $\mathcal{P}_{\text{mis}}$ corresponds to the general *misspecified* case, where the regression function $x \mapsto \mathbb{E}[Y|X = x]$ is not assumed to be linear.

2.1 Minimax analysis of linear least squares

We start with the following definition.

Definition 1. The distribution P_X on $\mathbf{R}^d$ is *degenerate* if there exists a linear hyperplane $H \subset \mathbf{R}^d$ such that $\mathbb{P}(X \in H) > 0$ (that is, if there exists some $\theta \in S^{d-1}$ such that $\mathbb{P}(\langle \theta, X\rangle = 0) > 0$).

Fact 1. *Let $n \geqslant d$. The following properties are equivalent:*

1. *The distribution P_X is non-degenerate;*
2. *The sample covariance matrix $\widehat{\Sigma}_n$ is invertible almost surely;*

3. The ordinary least-squares (OLS) estimator

$$\widehat{\beta}_n^{\mathrm{LS}} := \operatorname*{arg\,min}_{\beta \in \mathbf{R}^d} \sum_{i=1}^{n} (\langle \beta, X_i \rangle - Y_i)^2 \tag{8}$$

is uniquely defined almost surely, and equals $\widehat{\beta}_n^{\mathrm{LS}} = \widehat{\Sigma}_n^{-1} n^{-1} \sum_{i=1}^n Y_i X_i$.

Proof. The equivalence between the second and third points is standard: the empirical risk being convex, its global minimizers are the critical points β characterized by $\widehat{\Sigma}_n \beta = n^{-1} \sum_{i=1}^n Y_i X_i$.

We now prove that the second point implies the first, by contraposition. If $\mathbb{P}(\langle \theta, X \rangle = 0) = p > 0$ for some $\theta \in S^{d-1}$, then with probability p^n, $\langle \theta, X_i \rangle = 0$ for $i = 1, \dots, n$, so that $\widehat{\Sigma}_n \theta = n^{-1} \sum_{i=1}^n \langle \theta, X_i \rangle X_i = 0$ and thus $\widehat{\Sigma}_n$ is not invertible.

Conversely, let us show that the first point implies the second one. Note that the latter amounts to saying that $X_1, \dots, X_n$ span $\mathbf{R}^d$ almost surely. It suffices to show this for $n = d$, which we do by showing that, almost surely, $V_k = \operatorname{span}(X_1, \dots, X_k)$ is of dimension k for $0 \leqslant k \leqslant d$, by induction on k. The case $k = 0$ is clear. Now, assume that $k \leqslant d$ and that V_{k-1} is of dimension $k - 1 \leqslant d - 1$ almost surely. Then, V_{k-1} is contained in a hyperplane of $\mathbf{R}^d$, and since X_k is independent of V_{k-1}, the first point implies that $\mathbb{P}(X_k \in V_{k-1}) = 0$, so that V_k is of dimension k almost surely. □

Remark 1 (Intercept). Assume that $X = (X^j)_{1 \leqslant j \leqslant d}$, where $X^d \equiv 1$ is an intercept variable. Then, the distribution P_X is degenerate if and only if there exists $\theta = (\theta^j)_{1 \leqslant j < d} \in \mathbf{R}^{d-1} \setminus \{0\}$ and $c \in \mathbf{R}$ such that $\sum_{j=1}^{d-1} \theta^j X^j = c$ with positive probability. This amounts to say that $(X^1, \dots, X^{d-1})$ belongs to some fixed affine hyperplane of $\mathbf{R}^{d-1}$ with positive probability.

The following result shows that non-degeneracy of the design distribution is necessary to obtain finite minimax risk.

Proposition 1 (Degenerate case)**.** *Assume that either $n < d$, or that the distribution P_X of X is degenerate, in the sense of Definition 1. Then, the minimax excess risk with respect to the class $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$ is infinite.*

An infinite minimax excess risk means that some dependence on the true parameter β^* (for instance, through its norm) is unavoidable in the expected risk of any estimator $\widehat{\beta}_n$. From now on and until the rest of this section, we assume that the distribution P_X is non-degenerate and that $n \geqslant d$. In particular, the OLS estimator is well-defined, and the empirical covariance matrix $\widehat{\Sigma}_n$ is invertible almost surely. Theorem 1 below provides the exact minimax excess risk and estimator in the well-specified case.

Theorem 1. *Assume that P_X is non-degenerate and $n \geqslant d$. The minimax risks over classes $\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$ and $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$ coincide, and equal*

$$\inf_{\widehat{\beta}_n} \sup_{P \in \mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)} \mathbb{E}\big[\mathcal{E}_P(\widehat{\beta}_n)\big] = \frac{\sigma^2}{n} \cdot \mathbb{E}\big[\operatorname{Tr}(\widetilde{\Sigma}_n^{-1})\big] \tag{9}$$

where $\widetilde{\Sigma}_n = \Sigma^{-1/2} \widehat{\Sigma}_n \Sigma^{-1/2}$ is the rescaled empirical covariance matrix. In addition, the minimax risk is achieved by the OLS estimator (8) over the classes $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$ and $\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$ for every P_X and σ^2.

The proof of Theorem 1 and Proposition 1 is provided in Section 5.2, and relies on standard decision-theoretic arguments (see [Tsy09, Chapter 2] and [Joh19, Section 4.10]). First, an upper bound (in the non-degenerate case) over $\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$ is obtained for the OLS estimator. Then,

a matching lower bound on the minimax risk over the subclass $\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)$ is established by considering the Bayes risk under Gaussian prior on β^* and using a monotone convergence argument.

Remark 2 (Linear changes of covariates). The minimax risk is invariant under invertible linear transformations of the covariates x. This can be seen a priori, by noting that the class of linear functions of x is invariant under linear changes of variables. To recover it from Theorem 1, let $X' = AX$, where A is an invertible $d \times d$ matrix. Since $\Sigma' = \mathbb{E}[X'X'^\top]$ equals $A\Sigma A^\top$ and $\widehat{\Sigma}'_n = n^{-1}\sum_{i=1}^n X'_i X'^\top_i$ equals $A\widehat{\Sigma}_n A^\top$, we have

$$\widehat{\Sigma}'^{-1}_n \Sigma' = ((A^\top)^{-1}\widehat{\Sigma}_n^{-1}A^{-1})(A\Sigma A^\top) = (A^\top)^{-1}(\widehat{\Sigma}_n^{-1}\Sigma)A^\top\,,$$

which is conjugate to $\widehat{\Sigma}_n^{-1}\Sigma$ and hence has the same trace; this concludes by Theorem 1 (as $\mathrm{Tr}(\widetilde{\Sigma}_n^{-1}) = \mathrm{Tr}(\widehat{\Sigma}_n^{-1}\Sigma)$). In particular, the minimax risk for the design X is the same as the one for $\widetilde{X} = \Sigma^{-1/2}X$.

Note that the OLS estimator $\widehat{\beta}_n^{\text{LS}}$ is minimax optimal for every distribution of covariates P_X and noise level σ^2. This shows in particular that the knowledge of neither of those properties of the distribution is helpful to achieve improved risk uniformly over the linear class. On the other hand, when additional knowledge on the optimal parameter β^* is available, OLS may no longer be optimal, and knowledge of σ^2 may be helpful.

Another consequence of Theorem 1 is that independent Gaussian noise is the least favorable noise structure (in terms of minimax risk) in the well-specified case for a given noise level σ^2.

Finally, the convexity of the map $A \mapsto \mathrm{Tr}(A^{-1})$ on positive matrices [Bha09] implies (by Jensen's inequality combined with the identity $\mathbb{E}[\widetilde{\Sigma}_n] = I_d$) that the minimax risk (9) is always at least as large as $\sigma^2 d/n$, which is the minimax risk in the fixed-design case. We will however show in what follows that a strictly better lower bound can be obtained for $d \geqslant 2$.

2.2 Connection with statistical leverage and distribution-independent lower bound

In this section, we provide another expression for the minimax risk over the classes $\mathcal{P}_{\text{well}}(P_X, \sigma^2)$ and $\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)$, by relating it to the notion of *statistical leverage score* [HW78, CH88, Hub81].

Theorem 2 (Minimax risk and leverage score)**.** *Under the assumptions of Theorem 1, the minimax risk* (9) *over the classes* $\mathcal{P}_{\text{well}}(P_X, \sigma^2)$ *and* $\mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)$ *is equal to*

$$\inf_{\widehat{\beta}_n} \sup_{P \in \mathcal{P}_{\text{Gauss}}(P_X, \sigma^2)} \mathbb{E}\big[\mathcal{E}_P(\widehat{\beta}_n)\big] = \sigma^2 \cdot \mathbb{E}\left[\frac{\widehat{\ell}_{n+1}}{1 - \widehat{\ell}_{n+1}}\right] \tag{10}$$

where the expectation holds over an i.i.d. sample $X_1, \ldots, X_{n+1}$ *drawn from* P_X*, and where* $\widehat{\ell}_{n+1}$ *denotes the* statistical leverage score *of* X_{n+1} *among* $X_1, \ldots, X_{n+1}$*, defined by:*

$$\widehat{\ell}_{n+1} = \left\langle \left(\sum_{i=1}^{n+1} X_i X_i^\top \right)^{-1} X_{n+1}, X_{n+1} \right\rangle . \tag{11}$$

The leverage score $\widehat{\ell}_{n+1}$ of X_{n+1} among $X_1, \ldots, X_{n+1}$ measures the influence of the response Y_{n+1} on the associated fitted value $\widehat{Y}_{n+1} = \langle \widehat{\beta}_{n+1}^{\text{LS}}, X_{n+1} \rangle$: $\widehat{Y}_{n+1}$ is an affine function of Y_{n+1}, with slope $\widehat{\ell}_{n+1} = \partial \widehat{Y}_{n+1}/\partial Y_{n+1}$ [HW78, CH88]. Theorem 2 shows that the minimax predictive risk under the distribution P_X is characterized by the distribution of leverage scores of samples

drawn from this distribution. Intuitively, uneven leverage scores (with some points having higher leverage) imply that the estimator $\widehat{\beta}_n^{\mathrm{LS}}$ is determined by a smaller number of points, and therefore has higher variance. This is consistent with the message from robust statistics that points with high leverage (typically seen as outliers) can be detrimental to the performance of the least squares estimator [HW78, CH88, Hub81], see also [RM16].

Proof of Theorem 2. By Theorem 1, the minimax risk over $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$ and $\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$ equals, letting $X_{n+1} \sim P_X$ be independent from $X_1, \dots, X_n$:

$$\begin{aligned}
\frac{\sigma^2}{n} \cdot \mathbb{E}\big[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})\big] &= \frac{\sigma^2}{n} \cdot \mathbb{E}\big[\mathrm{Tr}(\widehat{\Sigma}_n^{-1}\Sigma)\big] \\
&= \sigma^2 \cdot \mathbb{E}\big[\mathrm{Tr}\big((n\widehat{\Sigma}_n)^{-1} X_{n+1}X_{n+1}^\top\big)\big] \\
&= \sigma^2 \cdot \mathbb{E}\big[\langle (n\widehat{\Sigma}_n)^{-1} X_{n+1}, X_{n+1}\rangle\big] \\
&= \sigma^2 \cdot \mathbb{E}\left[\frac{\langle (n\widehat{\Sigma}_n + X_{n+1}X_{n+1}^\top)^{-1} X_{n+1}, X_{n+1}\rangle}{1 - \langle (n\widehat{\Sigma}_n + X_{n+1}X_{n+1}^\top)^{-1} X_{n+1}, X_{n+1}\rangle}\right] \qquad (12) \\
&= \sigma^2 \cdot \mathbb{E}\left[\frac{\widehat{\ell}_{n+1}}{1 - \widehat{\ell}_{n+1}}\right],
\end{aligned}$$

where (12) follows from Lemma 1 below, with $S = n\widehat{\Sigma}_n$ and $v = X_{n+1}$. □

Lemma 1. *For any symmetric positive $d \times d$ matrix S and $v \in \mathbf{R}^d$,*

$$\langle S^{-1}v, v\rangle = \frac{\langle (S + vv^\top)^{-1}v, v\rangle}{1 - \langle (S + vv^\top)^{-1}v, v\rangle}\,. \qquad (13)$$

Proof. Since $S + vv^\top \succcurlyeq S$ is positive, it is invertible, and the Sherman-Morrison formula [HJ90] shows that

$$\begin{aligned}
(S + vv^\top)^{-1} &= S^{-1} - \frac{S^{-1}vv^\top S^{-1}}{1 + v^\top S^{-1} v}\,, \quad \text{so that} \\
\langle (S + vv^\top)^{-1}v, v\rangle &= \langle S^{-1}v, v\rangle - \frac{\langle S^{-1}v, v\rangle^2}{1 + \langle S^{-1}v, v\rangle} = \frac{\langle S^{-1}v, v\rangle}{1 + \langle S^{-1}v, v\rangle},
\end{aligned}$$

hence $\langle (S + vv^\top)^{-1}v, v\rangle \in [0, 1)$. Inverting this equality yields (13). □

We now deduce from Theorem 2 a precise lower bound on the minimax risk (9), valid for every distribution of covariates P_X. By Proposition 1, it suffices to consider the case when $n \geqslant d$ and P_X is nondegenerate (since otherwise the minimax risk is infinite).

Corollary 1 (Minimax lower bound)**.** *Under the assumptions of Theorem 1, the minimax risk (9) over $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$ satisfies*

$$\inf_{\widehat{\beta}_n} \sup_{P \in \mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)} \mathbb{E}\big[\mathcal{E}_P(\widehat{\beta}_n)\big] \geqslant \frac{\sigma^2 d}{n - d + 1}\,. \qquad (14)$$

Proof of Corollary 1. By Theorem 2, the minimax excess risk over $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$ writes:

$$\sigma^2 \cdot \mathbb{E}\left[\frac{\widehat{\ell}_{n+1}}{1 - \widehat{\ell}_{n+1}}\right] \geqslant \sigma^2 \cdot \frac{\mathbb{E}[\widehat{\ell}_{n+1}]}{1 - \mathbb{E}[\widehat{\ell}_{n+1}]}\,, \qquad (15)$$

where the inequality follows from the convexity of the map $x \mapsto x/(1-x) = 1 - 1/(1-x)$ on $[0,1)$. Now, by exchangeability of $(X_1, \dots, X_{n+1})$,

$$\begin{aligned}\mathbb{E}[\widehat{\ell}_{n+1}] &= \frac{1}{n+1}\sum_{i=1}^{n+1}\mathbb{E}\left[\left\langle\left(\sum_{i=1}^{n+1}X_iX_i^\top\right)^{-1}X_i, X_i\right\rangle\right] \\ &= \frac{1}{n+1}\mathbb{E}\left[\mathrm{Tr}\left\{\left(\sum_{i=1}^{n+1}X_iX_i^\top\right)^{-1}\left(\sum_{i=1}^{n+1}X_iX_i^\top\right)\right\}\right] = \frac{d}{n+1}\,. \end{aligned} \tag{16}$$

Plugging equation (16) into (15) yields the lower bound (14). □

Since $n-d+1 \leqslant n$, Corollary 1 implies a lower bound of $\sigma^2 d/n$. The minimax risk for linear regression has been determined under additional boundedness assumptions on Y (and thus on β^*) by [Tsy03], showing that it scales as $\Theta(d/n)$ up to numerical constants. The proof of the lower bound relies on information-theoretic arguments, and in particular on Fano's inequality [Tsy09]. Although widely applicable, such techniques often lead to loose constant factors. By contrast, the approach relying on Bayes risk leading to Corollary 1 recovers the optimal leading constant, owing to the analytical tractability of the problem.

In fact, the lower bound of Corollary 1 is more precise than the $\sigma^2 d/n$ lower bound, in particular when the dimension d is commensurate to n. Indeed, in the case of centered Gaussian design, namely when $X \sim \mathcal{N}(0, \Sigma)$ for some positive matrix Σ, the risk of the OLS estimator (and thus, by Theorem 1, the minimax risk) can be computed exactly [And03, BF83], and equals

$$\mathbb{E}\big[\mathcal{E}_P(\widehat{\beta}_n^{\mathrm{LS}})\big] = \frac{\sigma^2 d}{n-d-1}\,. \tag{17}$$

The distribution-independent lower bound of Corollary 1 is very close to the above whenever $n - d \gg 1$. Hence, it is almost the best possible distribution-independent lower bound on the minimax risk. This also shows that Gaussian design is almost the easiest design distribution in terms of minimax risk. This can be understood as follows: degeneracy (a large value of $\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})$) occurs whenever the rescaled sample covariance matrix $\widetilde{\Sigma}_n$ is small in some direction; this occurs if either the direction of $\widetilde{X} = \Sigma^{-1/2}X$ is far from uniform (so that the projection of $\widetilde{X}$ in some direction can be small), or if its norm can be small. If $\widetilde{X} \sim \mathcal{N}(0, I_d)$, then $\widetilde{X}/\|\widetilde{X}\|$ is uniformly distributed on the unit sphere, while $\|\widetilde{X}\| = \sqrt{\sum_{j=1}^d (\widetilde{X}^j)^2}$ is sharply concentrated around $\sqrt{d}$: with high probability, $\|\widetilde{X}\| = \sqrt{d} + O(1)$ (see e.g. [Ver18, Eq. 3.7]).

In particular, in the high-dimensional regime where d and n are large and commensurate, namely $d, n \to \infty$ and $d/n \to \gamma$, the lower bound of Corollary 1 matches the minimax risk (17) in the Gaussian case, which converges to $\sigma^2\gamma/(1-\gamma)$. The limit $\sigma^2\gamma/(1-\gamma)$ has a form of universality in the high-dimensional regime: indeed, it is connected to the Marchenko-Pastur law for the spectrum of random matrices [MP67], which extends to more general distributions with jointly independent coordinates. However, the "universality" of this limiting behavior is quite restrictive [EKK11, EK18], since it relies on the assumption of independent covariates, which induces in high dimension a very specific geometry due to the concentration of measure phenomenon [Led01, BLM13]. For instance, [EK18] obtains different limiting risks for robust regression in high dimension when considering non-independent coordinates. Corollary 1 shows that, if not universal, the limiting risk obtained in the independent case provides a *lower bound* for general design distributions.

Finally, the property of the design distribution that leads to the minimal excess risk in high dimension can be formulated succinctly in terms of leverage scores, using Theorem 2.

Corollary 2. *Let $(d_n)_{n\geqslant 1}$ be a sequence of positive integers such that $d_n/n \to \gamma \in (0,1)$, and $(P_X^{(n)})_{n\geqslant 1}$ a sequence of non-degenerate distributions on $\mathbf{R}^{d_n}$. Assume that the minimax excess risk (9) over $\mathcal{P}_{\mathrm{well}}(P_X^{(n)}, \sigma^2)$ converges to $\sigma^2\gamma/(1-\gamma)$. Then, the distribution of the leverage score $\widehat{\ell}_{n+1}^{(n)}$ of one sample among $n+1$ under $P_X^{(n)}$ converges in probability to γ.*

Proof. Let $\phi(x) = x/(1-x)$ for $x \in [0,1)$, and $\psi(x) := \phi(x) - \phi(\gamma) - \phi'(\gamma)(x-\gamma)$ (with $\psi(\gamma) = 0$). Since ϕ is strictly convex, $\psi(x) > 0$ for $x \neq \gamma$, and ψ is also strictly convex. Hence, ψ is decreasing on $[0,\gamma]$ and increasing on $[\gamma,1)$. In particular, for every $\varepsilon > 0$, $\eta_\varepsilon := \inf_{|x-\gamma|\geqslant\varepsilon} \psi(x) > 0$.

By Theorem 2, the assumption of Corollary 2 means that $\mathbb{E}[\phi(\widehat{\ell}_{n+1}^{(n)})] \to \phi(\gamma)$. Since in addition $\mathbb{E}[\widehat{\ell}_{n+1}^{(n)}] = d_n/(n+1) \to \gamma$ (the first equality, used in the proof of Corollary 1, holds for $d_n \leqslant n+1$, hence for n large enough since $\gamma < 1$), we have $\mathbb{E}[\psi(\widehat{\ell}_{n+1}^{(n)})] \to 0$. Now, for every $\varepsilon > 0$, $\psi(x) \geqslant \eta_\varepsilon \cdot \mathbf{1}(|x-\gamma| \geqslant \varepsilon)$, so that $\mathbb{P}(|\widehat{\ell}_{n+1}^{(n)} - \gamma| \geqslant \varepsilon) \leqslant \eta_\varepsilon^{-1}\mathbb{E}[\psi(\widehat{\ell}_{n+1}^{(n)})] \to 0$. □

2.3 Upper bounds on the minimax risk

In this section, we complement the lower bound of Corollary 1 by providing matching *upper bounds* on the minimax risk. Since by Proposition 1 the minimax risk is infinite when the design distribution is degenerate, we introduce the following quantitative version of the non-degeneracy condition:

Assumption 1 (Small-ball condition)**.** The whitened design $\widetilde{X} = \Sigma^{-1/2}X$ satisfies the following: there exist constants $C \geqslant 1$ and $\alpha \in (0,1]$ such that, for every linear hyperplane H of $\mathbf{R}^d$ and $t > 0$,

$$\mathbb{P}\big(\mathrm{dist}(\widetilde{X}, H) \leqslant t\big) \leqslant (Ct)^\alpha . \tag{18}$$

Equivalently, for every $\theta \in \mathbf{R}^d \setminus \{0\}$ and $t > 0$,

$$\mathbb{P}\big(|\langle \theta, X\rangle| \leqslant t\|\theta\|_\Sigma\big) \leqslant (Ct)^\alpha . \tag{19}$$

The equivalence between (18) and (19) comes from the fact that the distance $\mathrm{dist}(\widetilde{X}, H)$ of $\widetilde{X}$ to the hyperplane H equals $|\langle \theta', \widetilde{X}\rangle|$, where $\theta' \in S^{d-1}$ is a normal vector to H. Condition (19) is then recovered by letting $\theta = \Sigma^{-1/2}\theta'$ (such that $\|\theta\|_\Sigma = \|\theta'\| = 1$) and by homogeneity.

Assumption 1 states that $\widetilde{X}$ does not lie too close to any fixed hyperplane. This assumption is a strengthened variant of the "small ball" condition introduced by [KM15, Men15, LM16] in the analysis of sample covariance matrices and least squares regression, which amounts to assuming (19) for a single value of $t < C^{-1}$. This latter condition amounts to a uniform equivalence between the L^1 and L^2 norms of one-dimensional marginals $\langle \theta, X\rangle$ ($\theta \in \mathbf{R}^d$) of X [KM15]. Here, we require that the condition holds for arbitrarily small t; the reason for this is that in order to control the minimax excess risk (9) (and thus $\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})]$), we are led to control the lower tail of the rescaled covariance matrix $\widetilde{\Sigma}_n$ at all confidence levels. The study of the lower tail of $\widetilde{\Sigma}_n$ (on which the results of this section rely) is deferred to Section 3. We also illustrate Assumption 1 in Section 3.3, by discussing conditions under which it holds in the case of independent coordinates.

First, Assumption 1 itself suffices to obtain an upper bound on the minimax risk of $O(\sigma^2 d/n)$, without additional assumptions on the upper tail of $XX^\top$ (apart from integrability).

Proposition 2. *If Assumption 1 holds, then for every $P \in \mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$, letting $C' = 3C^4 e^{1+9/\alpha}$ we have:*

$$\mathbb{E}[\mathcal{E}(\widehat{\beta}_n^{\mathrm{LS}})] \leqslant 2C' \cdot \frac{\sigma^2 d}{n} . \tag{20}$$

Proposition 2 (a consequence of Corollary 4 from Section 3.2) is optimal in terms of the rate of convergence; however, it exhibits the suboptimal $2C'$ factor in the leading term. As we show next, it is possible to obtain an optimal constant in the first-order term (as well as a second-order term of the correct order) under a modest additional assumption.

Assumption 2 (Norm kurtosis)**.** $\mathbb{E}[\|\Sigma^{-1/2}X\|^4] \leqslant \kappa d^2$ for some $\kappa > 0$.

Remark 3. Since $\mathbb{E}[\|\Sigma^{-1/2}X\|^2] = d$, Assumption 2 is a bound on the kurtosis of the variable $\|\Sigma^{-1/2}X\|$. This condition is implied by the following L^2-L^4 equivalence for one-dimensional marginals of X: for every $\theta \in \mathbf{R}^d$, $\mathbb{E}[\langle\theta, X\rangle^4]^{1/4} \leqslant \kappa^{1/4} \cdot \mathbb{E}[\langle\theta, X\rangle^2]^{1/2}$ (Assumption 3 below). Indeed, assuming that the latter holds, then taking $\theta = \Sigma^{-1/2}e_j$ (where $(e_j)_{1\leqslant j\leqslant d}$ denotes the canonical basis of $\mathbf{R}^d$), so that $\langle\theta, X\rangle$ is the j-th coordinate $\widetilde{X}^j$ of $\widetilde{X}$, we get $\mathbb{E}[(\widetilde{X}^j)^4] \leqslant \kappa\widetilde{\mathbb{E}}[(\widetilde{X}^j)^2]^2 = \kappa$ (since $\mathbb{E}[\widetilde{X}\widetilde{X}^\top] = I_d$). This implies that

$$\begin{aligned}\mathbb{E}\big[\|\widetilde{X}\|^4\big] &= \mathbb{E}\bigg[\bigg(\sum_{j=1}^d (\widetilde{X}^j)^2\bigg)^2\bigg] = \sum_{1\leqslant j,k\leqslant d} \mathbb{E}\big[(\widetilde{X}^j)^2(\widetilde{X}^k)^2\big] \\ &\leqslant \sum_{1\leqslant j,k\leqslant d} \mathbb{E}\big[(\widetilde{X}^j)^4\big]^{1/2}\mathbb{E}\big[(\widetilde{X}^k)^4\big]^{1/2} \leqslant \sum_{1\leqslant j,k\leqslant d} \kappa^{1/2}\cdot\kappa^{1/2} = \kappa\cdot d^2\,,\end{aligned}$$

where the first inequality above comes from the Cauchy-Schwarz inequality. The converse is false: if $\widetilde{X}$ is uniform on $\{\sqrt{d}e_j : 1 \leqslant j \leqslant d\}$, then the first condition holds with $\kappa = 1$, while the second only holds for $\kappa \geqslant d$ (taking $\theta = e_1$). Hence, Assumption 2 on the upper tail of X is weaker than an L^2-L^4 equivalence of the one-dimensional marginals of X; on the other hand, we do require a small-ball condition (Assumption 1) on the lower tail of X.

Theorem 3 (Upper bound in the well-specified case)**.** *Grant Assumptions 1 and 2. Let* $C' = 3C^4e^{1+9/\alpha}$ *(which only depends on* α, C*). If* $n \geqslant \min(6\alpha^{-1}d, 12\alpha^{-1}\log(12\alpha^{-1}))$*, then*

$$\frac{1}{n}\mathbb{E}\big[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})\big] \leqslant \frac{d}{n} + 8C'\kappa\Big(\frac{d}{n}\Big)^2. \tag{21}$$

In particular, the minimax excess risk over the class $\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$ *satisfies:*

$$\frac{\sigma^2 d}{n} \leqslant \inf_{\widehat{\beta}_n} \sup_{P\in\mathcal{P}_{\mathrm{well}}(P_X,\sigma^2)} \mathbb{E}\big[\mathcal{E}_P(\widehat{\beta}_n)\big] \leqslant \frac{\sigma^2 d}{n}\Big(1 + 8C'\frac{\kappa d}{n}\Big). \tag{22}$$

The proof of Theorem 3 is given in Section 5.3; it relies in particular on Lemma 7 herein and on Theorem 4 from Section 3. From a technical point of view, some care is required since the assumptions of Theorem 3 provide control on lower, rather than upper, relative deviations of $\widehat{\Sigma}_n$ with respect to Σ. As shown by the lower bound (established in Corollary 1), the constant in the first-order term in (22) is tight; in addition, one could see from a higher-order expansion (under additional moment assumptions) that the second-order term is also tight, up to the constant $8C'$ factor.

Consider now the general misspecified case, namely the class $\mathcal{P}_{\mathrm{mis}}(P_X, \sigma^2)$. Here, we will need the slightly stronger Assumption 3.

Assumption 3 (L^2-L^4 norm equivalence)**.** There exists a constant $\kappa > 0$ such that, for every $\theta \in \mathbf{R}^d$, $\mathbb{E}[\langle\theta, X\rangle^4] \leqslant \kappa \cdot \mathbb{E}[\langle\theta, X\rangle^2]^2$.

Proposition 3 (Upper bound in the misspecified case)**.** *Assume that* P_X *satisfies Assumptions 1 and 3, and that*

$$\chi := \mathbb{E}\big[\mathbb{E}[\varepsilon^2|X]^2\|\Sigma^{-1/2}X\|^4\big]/d^2 < +\infty$$

(*note that* $\chi \leqslant \mathbb{E}[(Y - \langle \beta^*, X\rangle)^4 \|\Sigma^{-1/2}X\|^4]/d^2$). *Then, for* $n \geqslant \max(96, 6d)/\alpha$, *the risk of the OLS estimator satisfies*

$$\mathbb{E}\big[\mathcal{E}(\widehat{\beta}_n^{\mathrm{LS}})\big] \leqslant \frac{1}{n}\mathbb{E}\big[(Y - \langle \beta^*, X\rangle)^2 \|\Sigma^{-1/2}X\|^2\big] + 276C'^2\sqrt{\kappa\chi}\Big(\frac{d}{n}\Big)^{3/2}. \tag{23}$$

In particular, we have

$$\frac{\sigma^2 d}{n} \leqslant \inf_{\widehat{\beta}_n} \sup_{P \in \mathcal{P}_{\mathrm{mis}}(P_X, \sigma^2)} \mathbb{E}\big[\mathcal{E}(\widehat{\beta}_n)\big] \leqslant \frac{\sigma^2 d}{n}\left(1 + 276C'^2\kappa\sqrt{\frac{d}{n}}\right). \tag{24}$$

The proof of Proposition 3 is provided in Section 5.4; it combines results from Section 3 with a tail bound from [Oli16]. Proposition 3 shows that, under Assumptions 1 and 3, the minimax excess risk over the class $\mathcal{P}_{\mathrm{mis}}(P_X, \sigma^2)$ scales as $(1 + o(1))\sigma^2 d/n$ as $d/n \to 0$. This implies that the OLS estimator is asymptotically minimax on the misspecified class $\mathcal{P}_{\mathrm{mis}}(P_X, \sigma^2)$ when $d = o(n)$, and that independent Gaussian noise is asymptotically the least favorable structure for the error ε.

2.4 Parameter estimation

Let us briefly discuss how the results of this section obtained for prediction can be adapted to the problem of parameter estimation, where the loss of an estimate $\widehat{\beta}_n$ given β^* is $\|\widehat{\beta}_n - \beta^*\|^2$.

By the same proof as that of Theorem 1 (replacing the norm $\|\cdot\|_\Sigma$ by $\|\cdot\|$), the minimax excess risk over the classes $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$ and $\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$ is $(\sigma^2/n)\mathbb{E}[\mathrm{Tr}(\widehat{\Sigma}_n^{-1})]$, achieved by the OLS estimator. By convexity of $A \mapsto \mathrm{Tr}(A^{-1})$ over positive matrices [Löw34], this quantity is larger than $\sigma^2\mathrm{Tr}(\Sigma^{-1})/n$.

In the case of centered Gaussian covariates, $\mathbb{E}[\mathrm{Tr}(\widehat{\Sigma}_n^{-1})] = \mathrm{Tr}(\Sigma^{-1}\mathbb{E}[\widetilde{\Sigma}_n^{-1}]) = \mathrm{Tr}(\Sigma^{-1})n/(n - d - 1)$ [And03], so the minimax risk is $\sigma^2\mathrm{Tr}(\Sigma^{-1})/(n - d - 1)$. On the other hand, the improved lower bound for general design of Corollary 1 for prediction does not appear to extend to estimation. The reason for this is that the map $A \mapsto A/(1 - \mathrm{Tr}(A))$ is not convex over positive matrices for $d \geqslant 2$ (where convexity is defined with respect to the positive definite order, see e.g. [BV04, Section 3.6.2]), although its trace is.

Finally, the results of Section 3 on the lower tail of $\widetilde{\Sigma}_n$ can be used to obtain upper bounds in a similar fashion as for prediction. For instance, an analogue of Proposition 2 can be directly obtained by bounding $\mathrm{Tr}(\widehat{\Sigma}_n^{-1}) \leqslant \lambda_{\min}(\widetilde{\Sigma}_n)^{-1} \cdot \mathrm{Tr}(\Sigma^{-1})$. Since this work is primarily focused on prediction, we do not elaborate further in this direction.

3 Bounding the lower tail of a sample covariance matrix at all probability levels

Throughout this section, up to replacing X by $\Sigma^{-1/2}X$, we assume unless otherwise stated that $\mathbb{E}[XX^\top] = I_d$. Our aim is to obtain non-asymptotic large deviation inequalities of the form:

$$\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t) \leqslant e^{-n\psi(t)}$$

where $\psi(t) \to \infty$ as $t \to 0^+$. Existing bounds [Ver12, SV13, KM15, Oli16] are typically sub-Gaussian bounds with $\psi(t) = c(1 - C\sqrt{d/n} - t)_+^2$ for some constants $c, C > 0$, which "saturate" for small t. In this section, we study the behavior of the large deviations for small values of t, namely $t \in (0, c)$, where $c < 1$ is a fixed constant. In Section 3.1, we provide a lower bound on these tail probabilities, namely an upper bound on ψ, valid for every distribution of X when $d \geqslant 2$. In Section 3.2, we show that Assumption 1 is necessary and sufficient to obtain tail bounds of the optimal order. Finally, in Section 3.3 we show that Assumption 1 is naturally satisfied in the case of independent coordinates, under a mild regularity condition on their distributions.

3.1 A general lower bound on the lower tail

First, Proposition 4 below shows that in dimension $d \geqslant 2$, the probability of deviations of $\lambda_{\min}(\widehat{\Sigma}_n)$ cannot be arbitrarily small.

Proposition 4. *Assume that $d \geqslant 2$. Let X be a random vector in $\mathbf{R}^d$ such that $\mathbb{E}[XX^\top] = I_d$. Then, for every $t \leqslant 1$,*

$$\sup_{\theta \in S^{d-1}} \mathbb{P}(|\langle \theta, X \rangle| \leqslant t) \geqslant 0.16 \cdot t\,, \tag{25}$$

and therefore

$$\mathbb{P}\big(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t\big) \geqslant (0.025 \cdot t)^{n/2}\,. \tag{26}$$

The assumption that $d \geqslant 2$ is necessary since for $d = 1$, if $X = 1$ almost surely, then $\lambda_{\min}(\widehat{\Sigma}_n) = 1$ almost surely. Proposition 4 is proved in Section 6.1 through a probabilistic argument, namely by considering a random vector θ drawn uniformly on the sphere S^{d-1}.

Proposition 4 shows that $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t)$ is at least $(Ct)^{cn}$, where $C = 0.025$ and $c = 1/2$ are absolute constants; this bound writes $e^{-n\psi(t)}$, where $\psi(t) \asymp \log(1/t)$ as $t \to 0^+$. In the following section, we study matching upper bounds on this lower tail.

3.2 Optimal control of the lower tail

In this section, we study conditions under which an upper bound matching the lower bound from Proposition 4 can be obtained. We start by noting that Assumption 1 is necessary to obtain such bounds:

Remark 4 (Necessity of small ball condition). Assume that there exists $c_1, c_2 > 0$ such that $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t) \leqslant (c_1 t)^{c_2 n}$ for all $t \in (0,1)$. Then, Lemma 2 below implies that $\sup_{\theta \in S^{d-1}} \mathbb{P}(|\langle \theta, X \rangle| \leqslant t) \leqslant (c_1 t^2)^{c_2}$ for all $t \in (0,1)$. Hence, P_X satisfies Assumption 1 with $C = \sqrt{c_1}$ and $\alpha = 2c_2$.

Lemma 2. *For $t \in (0,1)$, let $p_t = \sup_{\theta \in S^{d-1}} \mathbb{P}(|\langle \theta, X \rangle| \leqslant t)$. Then, $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t) \geqslant p_{\sqrt{t}}^n$.*

Proof of Lemma 2. Let $p < p_{\sqrt{t}}$. By definition of $p_{\sqrt{t}}$, there exists $\theta \in S^{d-1}$ such that $\mathbb{P}(\langle \theta, X \rangle^2 \leqslant t) \geqslant p$. Hence, by independence, with probability at least p^n, $\langle \theta, X_i \rangle^2 \leqslant t$ for $i = 1, \dots, n$, so that $\lambda_{\min}(\widehat{\Sigma}_n) \leqslant \langle \widehat{\Sigma}_n \theta, \theta \rangle \leqslant t$. Taking $p \to p_{\sqrt{t}}$ concludes the proof. □

As Theorem 4 shows, Assumption 1 is also sufficient to obtain an optimal control on the lower tail.

Theorem 4. *Let X be a random vector in $\mathbf{R}^d$. Assume that $\mathbb{E}[XX^\top] = I_d$ and that X satisfies Assumption 1. If $n \geqslant 6d/\alpha$, then for every $t \in (0,1)$:*

$$\mathbb{P}\big(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t\big) \leqslant (C't)^{\alpha n/6} \tag{27}$$

where $C' = 3C^4 e^{1+9/\alpha}$.

Theorem 4 can be stated in the non-isotropic case, where $\Sigma = \mathbb{E}[XX^\top]$ is arbitrary:

Corollary 3. *Let X be a random vector in $\mathbf{R}^d$ such that $\mathbb{E}[\|X\|^2] < +\infty$, and let $\Sigma = \mathbb{E}[XX^\top]$. Assume that X satisfies Assumption 1. Then, if $d/n \leqslant \alpha/6$, for every $t \in (0,1)$, the empirical covariance matrix $\widehat{\Sigma}_n$ formed with an i.i.d. sample of size n satisfies*

$$\widehat{\Sigma}_n \succcurlyeq t\Sigma \tag{28}$$

with probability at least $1 - (C't)^{\alpha n/6}$, where C' is as in Theorem 4.

Proof of Corollary 3. We may assume that Σ is invertible: otherwise, we can just consider the span of the support of X, a subspace of $\mathbf{R}^d$ of dimension $d' \leqslant d \leqslant \alpha n/6$. Now, let $\widetilde{X} = \Sigma^{-1/2}X$; by definition, $\mathbb{E}[\widetilde{X}\widetilde{X}^\top] = I_d$, and $\widetilde{X}$ satisfies Assumption 1 since X does. By Theorem 4, with probability at least $1-(C't)^{\alpha n/6}$, $\lambda_{\min}(\Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2}) \geqslant t$, which amounts to $\Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2} \succcurlyeq tI_d$, and thus $\widehat{\Sigma}_n \succcurlyeq t\Sigma$. □

It is worth noting that Theorem 4 does not require any condition on the upper tail of $XX^\top$, aside from the assumption $\mathbb{E}[XX^\top] = I_d$. Indeed, as noted in Remark 4, it only requires the necessary Assumption 1. In particular, it does not require any sub-Gaussian assumption on X, similarly to the results from [KM15, Oli16, vdGM14, Yas14, Yas15]; this owes to the fact that lower bounds for sums of non-negative random variables hold under weak assumptions.

Remark 5 (Extension to random quadratic forms). Theorem 4 extends (up to straightforward changes in notations) to random quadratic forms $v \mapsto \langle A_i v, v\rangle$ where $A_1, \dots, A_n$ are positive semi-definite and i.i.d., with $\mathbb{E}[A_i] = I_d$ (Theorem 4 corresponds to the rank 1 case where $A_i = X_iX_i^\top$). On the other hand, the lower bound of Proposition 4 is specific to rank 1 matrices, as can be seen by considering the counterexample where $A_i = I_d$ almost surely.

Remark 6 (Gaussian case). It may be worth comparing the bound (27) to known estimates in the special case of the Gaussian distribution, namely $X \sim \mathcal{N}(0, I_d)$. In this case, the joint density of eigenvalues of $\widehat{\Sigma}_n$ admits a closed-form expression, which provides by marginalization the density of $\lambda_{\min}(\widehat{\Sigma}_n)$ [Ede88, p. 533]. From this expression, the following bound is deduced in [WV12, eq. (99)]:

$$\mathbb{P}\Big(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t\Big) \leqslant \frac{2(n/2)^{(n-d+1)/2}}{n-d+1} \frac{\sqrt{\pi}\Gamma(\frac{n+1}{2})}{\Gamma(\frac{d}{2})\Gamma(\frac{n-d+1}{2})\Gamma(\frac{n-d+2}{2})} t^{n-d+1}\,.$$

Letting $d = d_n$ such that $d_n/n \to \alpha \in (0,1)$ and applying Stirling's approximation, this implies the following large deviation estimate [WV12, Lemma 1]: for any fixed $t \in (0,1)$,

$$\mathbb{P}\Big(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t\Big) \leqslant \Big(\frac{n}{d}\Big)^{d/2} \Big(\frac{\sqrt{e}t}{1-d/n}\Big)^{n-d+o(n)}\,.$$

The bound (27) is of this form; it holds for general distributions of X, at the cost of worst constants in the Gaussian case.

Idea of the proof. The proof of Theorem 4 is provided in Section 4. It builds on the analysis of [Oli16], who obtains sub-Gaussian deviation bounds under fourth moment assumptions (Assumption 3), although some refinements are needed to handle our considered regime (with t small). We now discuss some general ideas about the proof technique.

The proof starts with the representation of $\lambda_{\min}(\widehat{\Sigma}_n)$ as the infimum of an empirical process:

$$\lambda_{\min}(\widehat{\Sigma}_n) = \inf_{\theta \in S^{d-1}} \langle \widehat{\Sigma}_n\theta, \theta\rangle = \inf_{\theta \in S^{d-1}} \Big\{ Z(\theta) := \frac{1}{n}\sum_{i=1}^n \langle \theta, X_i\rangle^2 \Big\}\,. \tag{29}$$

In order to control this infimum, a natural approach is to first control $Z(\theta)$ on a suitable finite ε-covering of S^{d-1} using Assumption 1, independence, and a union bound, and then to extend this control to S^{d-1} by approximation. However, this approach (see e.g. [Ver12, Theorem 5.39] for a use of this argument) fails here, since the control of the approximation term would require an exponential upper bound on $\|\widehat{\Sigma}_n\|_{\mathrm{op}}$, which does not hold for heavy-tailed distributions. Instead, as in [Oli16], we use the so-called PAC-Bayesian inequality for empirical processes [McA99, McA03, LST03, Cat07, AC11], which is based on a variational representation of the

relative entropy. This technique enables one to control a smoothed version of the process $Z(\theta)$, namely

$$Z(\rho) := \int_{\mathbf{R}^d} Z(\theta)\rho(\mathrm{d}\theta)\,,$$

indexed by probability distributions ρ on Θ.

Specifically, let π be a probability distribution on some subset $\Theta \subset \mathbf{R}^d$ containing S^{d-1}. In addition, let $\psi : \mathbf{R}_+^* \to \mathbf{R}$ be a bound on the moment generating function of $-\langle\theta, X\rangle^2$, such that for all $\lambda > 0$ and $\theta \in \Theta$,

$$\mathbb{E}\exp\big(-\lambda\langle\theta, X\rangle^2\big) \leqslant e^{-\psi(\lambda)}, \quad \text{so that} \quad \mathbb{E}\exp\big(-\lambda n Z(\theta) - n\psi(\lambda)\big) \leqslant 1.$$

The PAC-Bayes variational inequality (see Lemma 4 for a general statement) allows to turn this (pointwise, for every θ) bound on the moment generating function into a uniform bound for the smoothed process: for every $t > 0$,

$$\mathbb{P}\left(\forall\rho,\ -\lambda n\big[Z(\rho) + \psi(\lambda)\big] \leqslant \mathrm{KL}(\rho,\pi) + t\right) \geqslant 1 - e^{-t}\,,$$

where ρ spans all distributions over Θ and $\mathrm{KL}(\rho,\pi) = \int \log\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\mathrm{d}\rho$ is the relative entropy between ρ and π. One then deduce from these inequalities the following decomposition. To each $\theta \in S^{d-1}$, we associate a smoothing distribution ρ_θ around θ; then, with probability at least $1 - e^{-t}$, for every $\theta \in S^{d-1}$,

$$\begin{aligned} Z(\theta) &= Z(\theta) - \int_\Theta Z(\theta')\rho_\theta(\mathrm{d}\theta') + \int_\Theta Z(\theta')\rho_\theta(\mathrm{d}\theta') \\ &\geqslant \underbrace{Z(\theta) - \int_\Theta Z(\theta')\rho_\theta(\mathrm{d}\theta')}_{\text{approximation term}} - \underbrace{\frac{\mathrm{KL}(\rho_\theta,\pi)}{\lambda n}}_{\text{entropy term}} - \frac{\psi(\lambda)+t}{\lambda n}\,. \end{aligned}$$

The proof then involves controlling (i) the Laplace transform of the process; (ii) the approximation term; and (iii) the entropy term. In order to control the last two, a careful choice of smoothing distribution (and prior) is needed.

Remark 7 (PAC-Bayes vs. ε-net argument). As indicated above, the use of an ε-net argument would fail here, since it would lead to an approximation term depending on $\|\widehat{\Sigma}_n\|_{\mathrm{op}}$. On the other hand, the use of a smoothing distribution which is "isotropic" and centered at a point θ enables one to obtain an approximation term in terms of $\mathrm{Tr}(\widehat{\Sigma}_n)/d$, which can be bounded after proper truncation of X (in a way that does not overly degrade Assumption 1).

Remark 8 (Choice of prior and posteriors: entropy term). The PAC-Bayesian technique is classically employed in conjunction with Gaussian prior and smoothing distribution [LST03, AC11, Oli16]. This choice is convenient, since both the approximation and entropy term have closed-form expressions (in addition, a Gaussian distribution centered at θ yields the desired "isotropic" approximation term).

However, in order to obtain non-vacuous bounds for small t, we need the approximation term (and thus the "radius" γ of the smoothing distribution) to be small. But as $\gamma \to 0$, the entropy term for Gaussian distributions grows rapidly (as d/γ^2, instead of the $d\log(1/\gamma)$ rate suggested by covering numbers), which ultimately leads to vacuous bounds. In order to bypass this difficulty, we employ a more refined choice of prior and smoothing distributions, leading to an optimal entropy term of $d\log(1/\gamma)$. In addition, symmetry arguments show that this choice of smoothing also leads to an "isotropic" approximation term controlled by $\mathrm{Tr}(\widehat{\Sigma}_n)/d$ instead of $\|\widehat{\Sigma}_n\|_{\mathrm{op}}$.

Formulation in terms of moments. The statements of this section on the lower tail of $\lambda_{\min}(\widehat{\Sigma}_n)$ can equivalently be rephrased in terms of its negative moments. For $q \geqslant 1$, we denote $\|Z\|_{L^q} := \mathbb{E}[|Z|^q]^{1/q} \in [0, +\infty]$ the L^q norm of a real random variable Z.

Corollary 4. *Under the assumptions of Theorem 4 and for $n \geqslant 12/\alpha$, for any $1 \leqslant q \leqslant \alpha n/12$,*

$$\| \max(1, \lambda_{\min}(\widehat{\Sigma}_n)^{-1}) \|_{L^q} \leqslant 2^{1/q} \cdot C' \,. \tag{30}$$

Conversely, the previous inequality implies that $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n) \leqslant t) \leqslant (2C't)^{\alpha n/12}$ for all $t \in (0,1)$.

Finally, for any random vector X in $\mathbf{R}^d$, $d \geqslant 2$, such that $\mathbb{E}[XX^\top] = I_d$, we have for any $q \geqslant n/2$:

$$\|\lambda_{\min}(\widehat{\Sigma}_n)^{-1}\|_{L^q} = +\infty \,.$$

The proof of Corollary 4 is provided in Section 6.2.

3.3 The small-ball condition for independent covariates

We now discuss conditions under which Assumption 1 holds in the case of independent coordinates. In this section, we assume that the coordinates X^j, $1 \leqslant j \leqslant d$, of $X = \widetilde{X}$ are independent and centered. Note that the condition $\mathbb{E}[XX^\top] = I_d$ means that the X^j have unit variance.

Let us introduce the *Lévy concentration function* $Q_Z : \mathbf{R}^+ \to [0,1]$ of a real random variable Z defined by, for $t \geqslant 0$,

$$Q_Z(t) := \sup_{a \in \mathbf{R}} \mathbb{P}(|Z - a| \leqslant t) \,.$$

Anti-concentration (or small ball) estimates [NV13] refer to nonvacuous upper bounds on this function. Here, in order to establish Assumption 1, it suffices to show that $Q_{\langle \theta, X\rangle}(t) \leqslant (Ct)^\alpha$ for all $t > 0$ and $\theta \in S^{d-1}$. This amounts to establishing anti-concentration of linear combinations of independent variables $\langle \theta, X\rangle = \sum_{j=1}^d \theta^j X^j$, uniformly over $\theta \in S^{d-1}$, namely to provide upper bounds on:

$$Q_X(t) := \sup_{\theta \in S^{d-1}} Q_{\langle \theta, X\rangle}(t) \,.$$

Small-ball probabilities naturally appear in the study of the smallest singular value of a random matrix (see [RV10]). [TV09a, TV09b, RV08, RV09] studied anti-concentration for variables of the form $\langle \theta, X\rangle$, and deduced estimates of the smallest singular value of random matrices. These bounds are however slightly different from the one we need: indeed, they hold for "unstructured" vectors θ (which do not have additive structure, see [RV10]), rather than uniformly over $\theta \in S^{d-1}$. Here, in order to show that Assumption 1 holds, we need bounds over Q_X, which requires some assumption on the distribution of the coordinates X^j.

Clearly, $Q_X \geqslant \max_{1 \leqslant j \leqslant d} Q_{X^j}$, and in particular the coordinates X^j themselves must be anti-concentrated. Remarkably, a result of [RV14] (building on a reduction by [Rog87] to uniform variables) shows that, if the X^j have bounded densities, a reverse inequality holds:

Proposition 5 ([RV14], Theorem 1.2)**.** *Assume that $X^1, \dots, X^d$ are independent and have density bounded by $C_0 > 0$. Then, for every $\theta \in S^{d-1}$, $\sum_{j=1}^d \theta^j X^j$ has density bounded by $\sqrt{2}\, C_0$. In other words, $Q_X(t) \leqslant 2\sqrt{2}\, C_0 t$ for every $t > 0$, i.e., Assumption 1 holds with $\alpha = 1$ and $C = 2\sqrt{2}\, C_0$.*

Equivalently, if $\max_{1 \leqslant j \leqslant d} Q_{X^j}(t) \leqslant Ct$ for all $t > 0$, then $Q_X(t) \leqslant \sqrt{2}Ct$ for all $t > 0$, and the constant $\sqrt{2}$ is optimal [RV14]. Whether a general bound of Q_X in terms of $\max_{1 \leqslant j \leqslant d} Q_{X_j}$ holds

is unclear (for instance, the inequality $Q_X \leqslant \sqrt{2}\max_{1\leqslant j\leqslant d} Q_{X_j}$ does not hold, as shown by considering X^1, X^2 independent Bernoulli $1/2$ variables, and $\theta = (1/\sqrt{2}, 1/\sqrt{2})$: then $Q_{X^j}(3/8) = 1/2$ but $Q_{\langle\theta,X\rangle}(3/8) = 3/4$). While independence gives

$$Q_{\langle\theta,X\rangle}(t) \leqslant \min_{1\leqslant j\leqslant d} Q_{\theta^j X^j}(t) = \min_{1\leqslant j\leqslant d} Q_{X^j}(t/|\theta^j|) \leqslant \max_{1\leqslant j\leqslant d} Q_{X^j}(\sqrt{d}\cdot t)\,,$$

this bound features an undesirable dependence on the dimension d.

Another way to express the "non-atomicity" of the distributions of coordinates X^j, which is stable through linear combinations of independent variables, is the rate of decay of their Fourier transform. Indeed, if X^j is atomic, then its characteristic function does not vanish at infinity. Proposition 6 below (proved in Section 6.3), which follows from an inequality by Esséen, provides uniform anti-concentration for one-dimensional marginals $\langle\theta, X\rangle$ in terms of the Fourier transform of the X^j, establishing Assumption 1 beyond bounded densities. We let Φ_Z be the characteristic function of a real random variable Z, defined by $\Phi_Z(\xi) = \mathbb{E}[e^{i\xi Z}]$ for $\xi \in \mathbf{R}$.

Proposition 6. *Assume that $X^1, \dots, X^d$ are independent and that there are constants $C_0 > 0$ and $\alpha \in (0,1)$ such that, for $1 \leqslant j \leqslant d$ and $\xi \in \mathbf{R}$,*

$$|\Phi_{X^j}(\xi)| \leqslant (1 + |\xi|/C_0)^{-\alpha}\,. \tag{31}$$

Then, $X = (X^1, \dots, X^d)$ satisfies Assumption 1 with $C = 2^{1/\alpha}(2\pi)^{1/\alpha - 1}(1-\alpha)^{-1/\alpha}C_0$.

4 Proof of Theorem 4

4.1 Truncation and small-ball condition

The first step of the proof is to replace X by the truncated vector $X' := \big(1 \wedge \frac{\sqrt{d}}{\|X\|}\big)X$; likewise, let $X_i' = \big(1 \wedge \frac{\sqrt{d}}{\|X_i\|}\big)X_i$ for $1 \leqslant i \leqslant n$, and $\widehat{\Sigma}_n' := n^{-1}\sum_{i=1}^n X_i'(X_i')^\top$. Note that $X'(X')^\top \preccurlyeq XX^\top$ and $\|X'\| = \sqrt{d} \wedge \|X\|$, so that $\widehat{\Sigma}_n' \preccurlyeq \widehat{\Sigma}_n$ and $\mathbb{E}[\|X'\|^2] \leqslant \mathbb{E}[\|X\|^2] = d$. It follows that $\lambda_{\min}(\widehat{\Sigma}_n') \leqslant \lambda_{\min}(\widehat{\Sigma}_n)$, hence it suffices to establish a lower bound for $\lambda_{\min}(\widehat{\Sigma}_n')$.

In addition, for every $\theta \in S^{d-1}$, $t \in (0, C^{-1})$ and $a \geqslant 1$,

$$\begin{aligned}
\mathbb{P}(|\langle X', \theta\rangle| \leqslant t) &\leqslant \mathbb{P}\left(|\langle X, \theta\rangle| \leqslant at\right) + \mathbb{P}\left(\frac{\sqrt{d}}{\|X\|} \leqslant \frac{1}{a}\right) \\
&\leqslant (Cat)^\alpha + \mathbb{P}(\|X\| \geqslant a\sqrt{d}) \\
&\leqslant (Cat)^\alpha + \frac{\mathbb{E}[\|X\|^2]}{a^2 d} \qquad (32) \\
&= (Ct)^\alpha a^\alpha + \frac{1}{a^2} \qquad (33)
\end{aligned}$$

where we applied Markov's inequality in (32). In particular, letting $a = (Ct)^{-\alpha/(2+\alpha)}$, inequality (33) becomes

$$\mathbb{P}(|\langle X', \theta\rangle| \leqslant t) \leqslant 2(Ct)^{2\alpha/(2+\alpha)}\,. \tag{34}$$

4.2 Concentration and PAC-Bayesian inequalities

The smallest eigenvalue $\lambda_{\min}(\widehat{\Sigma}_n')$ of $\widehat{\Sigma}_n'$ may be written as the infimum of an empirical process indexed by the unit sphere $S^{d-1} = \{v \in \mathbf{R}^d : \|v\| = 1\}$:

$$\lambda_{\min}(\widehat{\Sigma}_n') = \inf_{v\in S^{d-1}} \langle \widehat{\Sigma}_n' v, v\rangle = \inf_{v\in S^{d-1}} \frac{1}{n}\sum_{i=1}^n \langle X_i', v\rangle^2\,.$$

Now, recall that the variables $\langle X_i', \theta\rangle^2$ are i.i.d. and distributed as $\langle X', \theta\rangle^2$ for every $\theta \in S^{d-1}$. The inequality (34) on the left tail of this variable can be expressed in terms of its Laplace transform, through the following lemma:

Lemma 3. *Let Z be a nonnegative random variable. Assume that there exists $\alpha \in (0,1]$ and $C > 0$ such that, for every $t \geqslant 0$, $\mathbb{P}(Z \leqslant t) \leqslant (Ct)^\alpha$. Then, for every $\lambda > 0$,*

$$\mathbb{E}[\exp(-\lambda Z)] \leqslant (C/\lambda)^\alpha \,. \tag{35}$$

Proof of Lemma 3. Since $0 \leqslant \exp(-\lambda Z) \leqslant 1$, we have

$$\mathbb{E}[e^{-\lambda Z}] = \int_0^1 \mathbb{P}(e^{-\lambda Z} \geqslant t)\mathrm{d}t = \int_0^1 \mathbb{P}\left(Z \leqslant \frac{\log(1/t)}{\lambda}\right)\mathrm{d}t \leqslant \int_0^1 \left(C\frac{\log(1/t)}{\lambda}\right)^\alpha \mathrm{d}t.$$

Now, for $u > 0$, the map $\alpha \mapsto u^\alpha = e^{\alpha \log u}$ is convex on $\mathbf{R}$, so that $u^\alpha \leqslant \alpha u + (1-\alpha)$ for $0 \leqslant \alpha \leqslant 1$. It follows that

$$\int_0^1 \log^\alpha(1/t)\mathrm{d}t \leqslant \alpha \int_0^1 (-\log t)\mathrm{d}t + (1-\alpha) = \alpha\big[-t\log t + t\big]_0^1 + (1-\alpha) = 1,$$

which establishes inequality (35). □

Here, inequality (34) implies that, for every $\theta \in S^{d-1}$,

$$\mathbb{P}(\langle X', \theta\rangle^2 \leqslant t) = \mathbb{P}(|\langle X', \theta\rangle| \leqslant \sqrt{t}) \leqslant 2(C\sqrt{t})^{2\alpha/(2+\alpha)} = 2(C^2 t)^{\alpha/(2+\alpha)} \,.$$

Hence, Lemma 3 with $Z = \langle X', \theta\rangle^2$ implies that, for every $\lambda > 0$,

$$\mathbb{E}[\exp(-\lambda\langle X', \theta\rangle^2)] \leqslant 2(C^2/\lambda)^{\alpha/(2+\alpha)} \,.$$

In other words, for $i = 1, \dots, n$, $\mathbb{E}[\exp(Z_i(\theta))] \leqslant 1$, where, letting $\alpha' = \alpha/(2+\alpha)$, we define

$$Z_i(\theta) = -\lambda\langle X_i', \theta\rangle^2 + \alpha' \log\left(\frac{\lambda}{C^2}\right) - \log 2$$

with $\lambda > 0$ a fixed parameter that will be optimized later. In particular, letting

$$Z(\theta) = Z_1(\theta) + \cdots + Z_n(\theta) = n\left[-\lambda\langle \widehat{\Sigma}_n'\theta, \theta\rangle + \alpha' \log\left(\frac{\lambda}{C^2}\right) - \log 2\right] ,$$

the independence of $Z_1(\theta), \dots, Z_n(\theta)$ implies that, for every $\theta \in S^{d-1}$,

$$\mathbb{E}[\exp(Z(\theta))] = \mathbb{E}[\exp(Z_1(\theta))] \cdots \mathbb{E}[\exp(Z_n(\theta))] \leqslant 1 \,. \tag{36}$$

The bound (36) controls the upper tail of $Z(\theta)$ for fixed $\theta \in \Theta$. In order to obtain a uniform control over θ, similarly to [AC11, Oli16] we will use the PAC-Bayesian technique for bounding empirical processes [McA99, McA03, Cat07]. For completeness, we include a proof of Lemma 4 (which is standard) below.

Lemma 4 (PAC-Bayesian deviation bound)**.** *Let Θ be a measurable space, and $Z(\theta)$, $\theta \in \Theta$, be a real-valued measurable process. Assume that $\mathbb{E}[\exp Z(\theta)] \leqslant 1$ for every $\theta \in \Theta$. Let π be a probability distribution on Θ. Then,*

$$\mathbb{P}\left(\forall \rho, \int_\Theta Z(\theta)\rho(\mathrm{d}\theta) \leqslant \mathrm{KL}(\rho, \pi) + t\right) \geqslant 1 - e^{-t} \,, \tag{37}$$

where ρ spans all probability measures on Θ, and $\mathrm{KL}(\rho, \pi) := \int_\Theta \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)\mathrm{d}\rho \in [0, +\infty]$ is the Kullback-Leibler divergence between ρ and π, and where we define the integral in (37) to be $-\infty$ when the negative part is not integrable.

Proof of Lemma 4. By integrating the inequality $\mathbb{E}[\exp Z(\theta)] \leqslant 1$ with respect to π and using the Fubini-Tonelli theorem, we obtain

$$\mathbb{E}\left[\int_\Theta \exp Z(\theta)\pi(\mathrm{d}\theta)\right] \leqslant 1\,. \tag{38}$$

In addition, using the duality between the log-Laplace transform and the Kullback-Leibler divergence (see, e.g., [Cat04, p. 159]):

$$\log \int_\Theta \exp(Z(\theta))\pi(\mathrm{d}\theta) = \sup_\rho \left\{ \int_\Theta Z(\theta)\rho(\mathrm{d}\theta) - \mathrm{KL}(\rho,\pi) \right\}$$

where the supremum spans over all probability distributions ρ over Θ, the inequality (38) writes

$$\mathbb{E}\left[\exp \sup_\rho \left\{ \int_\Theta Z(\theta)\rho(\mathrm{d}\theta) - \mathrm{KL}(\rho,\pi) \right\}\right] \leqslant 1\,. \tag{39}$$

Applying Markov's inequality to (39) yields the desired bound (37). □

Here, we let $\Theta = S^{d-1}$ and $Z(\theta)$ as defined above. In addition, we take π to be the uniform distribution on S^{d-1}, and for $v \in S^{d-1}$ and $\gamma > 0$ we define $\Theta(v,\gamma) := \{\theta \in S^{d-1} : \|\theta - v\| \leqslant \gamma\}$ and let $\pi_{v,\gamma} = \pi(\Theta(v,\gamma))^{-1}\mathbf{1}(\Theta(v,\gamma)) \cdot \pi$ be the uniform distribution over $\Theta(v,\gamma)$. In this case, the PAC-Bayesian bound of Lemma 4 writes: for every $t > 0$, with probability at least $1 - e^{-t}$, for every $v \in S^{d-1}$ and $\gamma > 0$,

$$n\left[-\lambda F_{v,\gamma}(\widehat{\Sigma}'_n) + \alpha' \log\left(\frac{\lambda}{C^2}\right) - \log 2\right] \leqslant \mathrm{KL}(\pi_{v,\gamma},\pi) + t\,, \tag{40}$$

where we define for every symmetric matrix Σ:

$$F_{v,\gamma}(\Sigma) := \int_\Theta \langle \Sigma\theta,\theta\rangle \pi_{v,\gamma}(\mathrm{d}\theta)\,. \tag{41}$$

4.3 Control of the approximation term

Now, using the symmetries of the smoothing distributions $\pi_{v,\gamma}$, we will show that, for every $\gamma > 0$, $v \in S^{d-1}$ and symmetric matrix Σ,

$$F_{v,\gamma}(\Sigma) = \big(1 - \phi(\gamma)\big)\langle \Sigma v, v\rangle + \phi(\gamma) \cdot \frac{1}{d}\mathrm{Tr}(\Sigma)\,, \tag{42}$$

where for $\gamma > 0$,

$$\phi(\gamma) := \frac{d}{d-1}\int_\Theta \big(1 - \langle\theta, v\rangle^2\big)\pi_{v,\gamma}(\mathrm{d}\theta) \in [0, d/(d-1)\gamma^2]\,. \tag{43}$$

First, note that

$$F_{v,\gamma}(\Sigma) = \mathrm{Tr}(\Sigma A_{v,\gamma})\,, \qquad \text{where} \qquad A_{v,\gamma} := \int_\Theta \theta\theta^\top \pi_{v,\gamma}(\mathrm{d}\theta)\,.$$

In addition, for every isometry $U \in O(d)$ of $\mathbf{R}^d$ and $v \in S^{d-1}$, $\gamma > 0$, the image measure $U_*\pi_{v,\gamma}$ of $\pi_{v,\gamma}$ under U is $\pi_{Uv,\gamma}$ (since U sends $\Theta(v,\gamma)$ to $\Theta(Uv,\gamma)$ and preserves the uniform distribution π on S^{d-1}). It follows that

$$UA_{v,\gamma}U^{-1} = \int_\Theta (U\theta)(U\theta)^\top \pi_{v,\gamma}(\mathrm{d}\theta) = \int_\Theta \theta\theta^\top \pi_{Uv,\gamma}(\mathrm{d}\theta) = A_{Uv,\gamma}\,. \tag{44}$$

In particular, $A_{v,\gamma}$ commutes with every isometry $U \in O(d)$ such that $Uv = v$. Taking U to be the orthogonal reflection with respect to $H_v := (\mathbf{R}v)^\perp$, $A_{v,\gamma}$ preserves $\ker(U - I_d) = \mathbf{R}v$ and is therefore of the form $\phi_1(v,\gamma)vv^\top + C_{v,\gamma}$ where $\phi_1(v,\gamma) \in \mathbf{R}$ and $C_{v,\gamma}$ is a symmetric operator with $C_{v,\gamma}H_v \subset H_v$ and $C_{v,\gamma}v = v$. Next, taking $U = vv^\top + U_v$ where U_v is an arbitrary isometry of H_v, it follows that $C_{v,\gamma}$ commutes on H_v with all isometries U_v, and is therefore of the form $\phi_2(v,\gamma)P_v$, where $P_v = I_d - vv^\top$ is the orthogonal projection on H_v and $\phi_2(v,\gamma) \in \mathbf{R}$. To summarize, we have:

$$A_{v,\gamma} = \phi_1(v,\gamma)vv^\top + \phi_2(v,\gamma)(I_d - vv^\top)\,.$$

Now, the identity (44) shows that, for every $U \in O(d)$ and v, γ, $\phi_1(Uv,\gamma) = \phi_1(v,\gamma)$ and $\phi_2(Uv,\gamma) = \phi_2(v,\gamma)$; hence, these constants do not depend on v and are simply denoted $\phi_1(\gamma), \phi_2(\gamma)$. Defining $\phi(\gamma) := d \cdot \phi_2(\gamma)$ and $\widetilde{\phi}(\gamma) := \phi_1(\gamma) - \phi_2(\gamma)$, we therefore have:

$$A_{v,\gamma} = \widetilde{\phi}(\gamma)vv^\top + \phi(\gamma) \cdot \frac{1}{d} I_d\,. \tag{45}$$

Next, observe that

$$\int_{S^{d-1}} \pi_{v,\gamma}\pi(\mathrm{d}v) = \pi\,; \tag{46}$$

this follows from the fact that the measure π' on the left-hand side of (46) is a probability distribution on S^{d-1} invariant under any $U \in O(d)$, since

$$U_*\pi' = \int_{S^{d-1}} U_*\pi_{v,\gamma}\pi(\mathrm{d}v) = \int_{S^{d-1}} \pi_{Uv,\gamma}\pi(\mathrm{d}v) = \int_{S^{d-1}} \pi_{v,\gamma}\pi(\mathrm{d}v) = \pi'\,.$$

Equation (46), together with Fubini's theorem, implies that

$$\int_{S^{d-1}} A_{v,\gamma}\pi(\mathrm{d}v) = \int_{S^{d-1}}\int_{S^{d-1}} \theta\theta^\top \pi_{v,\gamma}(\mathrm{d}\theta)\pi(\mathrm{d}v) = \int_{S^{d-1}} \theta\theta^\top \pi(\mathrm{d}\theta) =: A\,.$$

Since A commutes with isometries (by invariance of π), it is of the form cI_d with $c = \mathrm{Tr}(A)/d = (1/d)\int_{S^{d-1}} \|\theta\|^2 \pi(\mathrm{d}\theta) = 1/d$. Plugging (45) into the previous equality, we obtain

$$\frac{1}{d} I_d = \int_{S^{d-1}} \Big[\widetilde{\phi}(\gamma)vv^\top + \phi(\gamma) \cdot \frac{1}{d} I_d\Big]\pi(\mathrm{d}v) = \frac{1}{d}\widetilde{\phi}(\gamma) I_d + \frac{1}{d}\phi(\gamma) I_d\,,$$

so that $\widetilde{\phi}(\gamma) = 1 - \phi(\gamma)$. The decomposition (45) then writes:

$$A_{v,\gamma} = \big(1 - \phi(\gamma)\big)vv^\top + \phi(\gamma) \cdot \frac{1}{d} I_d\,.$$

Recalling that $F_{v,\gamma}(\Sigma) = \mathrm{Tr}(\Sigma A_{v,\gamma})$, we obtain the desired expression (42) for $F_{v,\gamma}$.

Finally, note that on the one hand,

$$\langle A_{v,\gamma}v, v\rangle = (1 - \phi(\gamma))\|v\|^2 + \phi(\gamma) \cdot \frac{1}{d}\|v\|^2 = 1 - \frac{d-1}{d}\phi(\gamma)\,,$$

while on the other hand:

$$\langle A_{v,\gamma}v, v\rangle = \int_{S^{d-1}} \langle \theta, v\rangle^2 \pi_{v,\gamma}(\mathrm{d}\theta)\,,$$

so that

$$\phi(\gamma) = \frac{d}{d-1}\int_{S^{d-1}} \big(1 - \langle \theta, v\rangle^2\big)\pi_{v,\gamma}(\mathrm{d}\theta) \geqslant 0\,,$$

where we used that $\langle \theta, v\rangle^2 \leqslant 1$ by the Cauchy-Schwarz inequality.

Now, let α denote the angle between θ and v. We have $\langle \theta, v\rangle = \cos\alpha$ and $\|\theta - v\|^2 = (1-\cos\alpha)^2 + \sin^2\alpha = 2(1-\cos\alpha)$, so that $\langle \theta, v\rangle = 1 - \frac{1}{2}\|\theta - v\|^2$. Since $\pi_{v,\gamma}(\mathrm{d}\theta)$-almost surely, $\|\theta - v\| \leqslant \gamma$, this implies

$$1 - \langle \theta, v\rangle^2 = 1 - \Big(1 - \frac{1}{2}\|\theta - v\|^2\Big)^2 = \|\theta - v\|^2 - \frac{1}{4}\|\theta - v\|^4 \leqslant \gamma^2 \,.$$

Integrating this inequality over $\pi_{v,\gamma}$ yields $\phi(\gamma) \leqslant d/(d-1)\gamma^2$; this establishes (43).

4.4 Control of the entropy term

We now turn to the control of the entropy term in (40). Specifically, we will show that, for every $v \in S^{d-1}$ and $\gamma > 0$,

$$\mathrm{KL}(\pi_{v,\gamma}, \pi) \leqslant d \log\Big(1 + \frac{2}{\gamma}\Big) \,. \tag{47}$$

First, since $\mathrm{d}\pi_{v,\gamma}/\mathrm{d}\pi = \pi[\Theta(v,\gamma)]^{-1}$ $\pi_{v,\gamma}$-almost surely, $\mathrm{KL}(\pi_{v,\gamma}, \pi) = \log \pi[\Theta(v,\gamma)]^{-1}$. Now, let $N = N_c(\gamma, S^{d-1})$ denote the γ-covering number of S^{d-1}, namely the smallest $N \geqslant 1$ such that there exists $\theta_1, \dots, \theta_N \in S^{d-1}$ with

$$S^{d-1} = \bigcup_{i=1}^{N} \Theta(\theta_i, \gamma) \,. \tag{48}$$

Applying a union bound to (48) and using the fact that $\pi[\Theta(\theta_i,\gamma)] = \pi[\Theta(v,\gamma)]$ yields $1 \leqslant N\pi[\Theta(v,\gamma)]$, namely

$$\mathrm{KL}(\pi_{v,\gamma}, \pi) \leqslant \log N \,. \tag{49}$$

Now, let $N_p(\gamma, S^{d-1})$ denote the γ-packing number of S^{d-1}, which is the largest number of points in S^{d-1} with pairwise distances at least γ. We have, denoting $B^d = \{x \in \mathbf{R}^d : \|x\| \leqslant 1\}$,

$$N \leqslant N_p(\gamma, S^{d-1}) \leqslant N_p(\gamma, B^d) \leqslant \Big(1 + \frac{2}{\gamma}\Big)^d \,, \tag{50}$$

where the first inequality follows from a comparison of covering and packing numbers [Ver18, Lemma 4.2.8], the second one from the inclusion $S^{d-1} \subset B^d$ and the last one from a volumetric argument [Ver18, Lemma 4.2.13]. Combining (49) and (50) establishes (47).

4.5 Conclusion of the proof

First note that, since $\|X_i'\|^2 = \|X_i\|^2 \wedge d \leqslant d$ for $1 \leqslant i \leqslant n$,

$$\mathrm{Tr}(\widehat{\Sigma}_n') = \frac{1}{n}\sum_{i=1}^{n} \|X_i'\|^2 \leqslant d \,. \tag{51}$$

Putting together the previous bounds (40), (42), (47) and (51), we get with probability $1 - e^{-nu}$, for every $v \in S^{d-1}$, $\gamma \in (0, 1/2]$,

$$\begin{aligned}
\alpha' \log\Big(\frac{\lambda}{C^2}\Big) - \log 2 - \frac{d}{n}\log\Big(1 + \frac{2}{\gamma}\Big) - u &\leqslant \lambda F_{v,\gamma}(\widehat{\Sigma}_n') \\
&= \lambda\Big((1-\phi(\gamma))\langle \widehat{\Sigma}_n' v, v\rangle + \phi(\gamma)\cdot\frac{1}{d}\mathrm{Tr}(\widehat{\Sigma}_n')\Big) \\
&\leqslant \lambda\Big[(1-\phi(\gamma))\langle \widehat{\Sigma}_n' v, v\rangle + \phi(\gamma)\Big]
\end{aligned}$$

In particular, rearranging, and using the fact that $\phi(\gamma) \leqslant 1/2$ for $\gamma \leqslant 1/2$, as well as $\phi(\gamma) \leqslant \gamma^2$ and $\lambda_{\min}(\widehat{\Sigma}'_n) = \inf_v \langle \widehat{\Sigma}'_n v, v\rangle$, we get with probability $1 - e^{-nu}$,

$$\lambda_{\min}(\widehat{\Sigma}'_n) \geqslant \frac{2}{\lambda}\left[\alpha' \log\left(\frac{\lambda}{C^2}\right) - \log 2 - \frac{d}{n}\log\left(1 + \frac{2}{\gamma}\right) - u\right] - 2\gamma^2 \tag{52}$$

We first approximately maximize the above lower bound in γ, given λ. Since $\gamma \leqslant 1/2$, $1 + 2/\gamma \leqslant 1 + 1/\gamma^2 \leqslant 5/(4\gamma^2)$. We are therefore led to minimize

$$\frac{2d}{\lambda n}\log\left(\frac{5}{4\gamma^2}\right) + 2\gamma^2$$

over $\gamma^2 \leqslant 1/4$. Now, let $\gamma^2 = d/(2\lambda n)$, which belongs to the prescribed range if

$$\lambda \geqslant \frac{2d}{n}\,. \tag{53}$$

For this choice of γ, the lower bound (52) becomes

$$\begin{aligned}\lambda_{\min}(\widehat{\Sigma}'_n) &\geqslant \frac{2}{\lambda}\left[\alpha' \log\left(\frac{\lambda}{C^2}\right) - \log 2 - \frac{d}{n}\log\left(\frac{5\lambda n}{2d}\right) - u\right] - \frac{d}{\lambda n} \\ &= \frac{2}{\lambda}\left[\left(\alpha' - \frac{d}{n}\right)\log\lambda - \alpha' \log C^2 - \left\{\log 2 + \frac{d}{n}\log\left(\frac{5n}{2d}\right) + \frac{d}{2n}\right\} - u\right]\end{aligned}$$

Now, recall that by assumption, $d/n \leqslant \alpha/6 \leqslant 1/6$, so that (by monotonicity of $x \mapsto -x\log x$ on $(0, e^{-1}]$, replacing d/n by $1/6$) the term inside braces is smaller than $c_0 = 1.3$. In addition, assume that $\lambda \geqslant C^4$, so that $\log(\lambda/C^4) \geqslant 0$; in this case, condition (53) is automatically satisfied, since $2d/n \leqslant 1/3 \leqslant C^4$. Finally, since $\alpha' = \alpha/(2+\alpha) \geqslant \alpha/3$ and $d/n \leqslant \alpha/6$, $\alpha' \leqslant 2(\alpha' - d/n)$ and $\alpha' - d/n \geqslant \alpha/6$, so that

$$\left(\alpha' - \frac{d}{n}\right)\log\lambda - \alpha'\log C^2 \geqslant \left(\alpha' - \frac{d}{n}\right)\log\left(\frac{\lambda}{C^4}\right) \geqslant \frac{\alpha}{6}\log\left(\frac{\lambda}{C^4}\right),$$

the previous inequalities implies that, for every $\lambda \geqslant C^4$ and $u > 0$, with probability at least $1 - e^{-nu}$,

$$\lambda_{\min}(\widehat{\Sigma}'_n) \geqslant \frac{2}{\lambda}\left[\frac{\alpha}{6}\log\left(\frac{\lambda}{C^4}\right) - c_0 - u\right] = \frac{\alpha}{3C^4}\,\frac{\log\lambda' - 6\alpha^{-1}(c_0+u)}{\lambda'}$$

where $\lambda' = \lambda/C^4 \geqslant 1$. A simple analysis shows that for $c \in \mathbf{R}$, the function $\lambda' \mapsto (\log\lambda' - c)/\lambda'$ admits a maximum on $(0, +\infty)$ of e^{-c-1}, reached at $\lambda' = e^{c+1}$. Here $c = 6\alpha^{-1}(c_0 + u) > 0$, so that $\lambda' > e > 1$. Hence, for every $u > 0$, with probability at least $1 - e^{-nu}$,

$$\lambda_{\min}(\widehat{\Sigma}'_n) \geqslant \frac{\alpha}{3C^4}\exp\left(-1 - \frac{6(c_0+u)}{\alpha}\right) \geqslant C'^{-1}e^{-6u/\alpha} =: t\,, \tag{54}$$

where we let $C' := 3C^4 e^{1+9/\alpha}$ (using the fact that $6c_0 \leqslant 8$ and $1/\alpha \leqslant e^{1/\alpha}$). Inverting the bound (54), we obtain that for every $t < C'^{-1}$,

$$\mathbb{P}\big(\lambda_{\min}(\widehat{\Sigma}'_n) \leqslant t\big) \leqslant (C't)^{\alpha n/6}\,.$$

Since $\lambda_{\min}(\widehat{\Sigma}_n) \geqslant \lambda_{\min}(\widehat{\Sigma}'_n)$, and since the bound trivially holds for $t \geqslant C'^{-1}$, this concludes the proof.

5 Proofs from Section 2

In this section, we gather the remaining proofs of results from Section 2 on least squares regression, namely those of Proposition 1, Theorem 1, Proposition 2, Theorem 3 and Proposition 3.

5.1 Preliminary: risk of Ridge and OLS estimators

We start with general expressions for the risk, which will be used several times in the proofs. Here, we assume that (X, Y) is as in Section 2, namely $\mathbb{E}[Y^2] < +\infty$, $\mathbb{E}[\|X\|^2] < +\infty$ and $\Sigma := \mathbb{E}[XX^\top]$ is invertible. Letting $\varepsilon := Y - \langle \beta^*, X\rangle$ denote the error, where $\beta^* := \Sigma^{-1}\mathbb{E}[YX]$ is the risk minimizer, we let $m(X) := \mathbb{E}[\varepsilon|X] = \mathbb{E}[Y|X] - \langle \beta^*, X\rangle$ denote the misspecification (or approximation) error of the linear model, and $\sigma^2(X) := \mathrm{Var}(\varepsilon|X) = \mathrm{Var}(Y|X)$ denote the conditional variance of the noise.

Lemma 5 (Risk of the Ridge estimator)**.** *Assume that (X, Y) is of the previous form. Let $\lambda \geqslant 0$, and assume that either $\lambda > 0$ or that P_X is non-degenerate and $n \geqslant d$. The risk of the Ridge estimator $\widehat{\beta}_{\lambda,n}$, defined by*

$$\widehat{\beta}_{\lambda,n} := \operatorname*{arg\,min}_{\beta\in\mathbf{R}^d} \left\{ \frac{1}{n}\sum_{i=1}^n (Y_i - \langle \beta, X_i\rangle)^2 + \lambda\|\beta\|^2 \right\} = \big(\widehat{\Sigma}_n + \lambda I_d\big)^{-1} \cdot \frac{1}{n}\sum_{i=1}^n Y_i X_i \,, \tag{55}$$

equals

$$\begin{aligned}\mathbb{E}\big[\mathcal{E}(\widehat{\beta}_{\lambda,n})\big] = \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n m(X_i)X_i - \lambda\beta^*\right\|^2_{(\widehat{\Sigma}_n+\lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n+\lambda I_d)^{-1}}\right] + \\ + \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^n \sigma^2(X_i)\|X_i\|^2_{(\widehat{\Sigma}_n+\lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n+\lambda I_d)^{-1}}\right].\end{aligned} \tag{56}$$

Proof. Since $Y_i = \langle \beta^*, X_i\rangle + \varepsilon_i$ for $i = 1, \ldots, n$, and since $\langle \beta^*, X_i\rangle X_i = X_i X_i^\top \beta^*$, we have

$$\frac{1}{n}\sum_{i=1}^n Y_i X_i = \widehat{\Sigma}_n\beta^* + \frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i \,. \tag{57}$$

Hence, the excess risk of $\widehat{\beta}_{\lambda,n}$ (which is well-defined by the assumptions) is

$$\begin{aligned}\mathbb{E}\big[\mathcal{E}(\widehat{\beta}_{\lambda,n})\big] &= \mathbb{E}\left[\left\|(\widehat{\Sigma}_n + \lambda I_d)^{-1}\left(\widehat{\Sigma}_n\beta^* + \frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i\right) - \beta^*\right\|^2_\Sigma\right] \\ &= \mathbb{E}\left[\left\|(\widehat{\Sigma}_n + \lambda I_d)^{-1} \cdot \frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i - \lambda(\widehat{\Sigma}_n + \lambda I_d)^{-1}\beta^*\right\|^2_\Sigma\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i X_i - \lambda\beta^*\right\|^2_{(\widehat{\Sigma}_n+\lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n+\lambda I_d)^{-1}}\,\middle|\, X_1, \ldots, X_n\right]\right] \\ &= \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n m(X_i)X_i - \lambda\beta^*\right\|^2_{(\widehat{\Sigma}_n+\lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n+\lambda I_d)^{-1}}\right] + \\ &\quad + \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^n \sigma^2(X_i)\|X_i\|^2_{(\widehat{\Sigma}_n+\lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n+\lambda I_d)^{-1}}\right]\end{aligned} \tag{58}$$

where (58) is obtained by expanding and using the fact that, for $i \neq j$,

$$\begin{aligned}\mathbb{E}\big[\varepsilon_i\varepsilon_j|X_1, \ldots, X_n\big] &= m(X_i)m(X_j)\,, \\ \mathbb{E}\big[\varepsilon_i^2|X_1, \ldots, X_n\big] &= m(X_i)^2 + \sigma^2(X_i)\,. \square\end{aligned}$$

In the special case where $\lambda = 0$, the previous risk decomposition becomes:

Lemma 6 (Risk of the OLS estimator)**.** *Assume that P_X is non-degenerate and $n \geqslant d$. Then,*

$$\mathbb{E}\big[\mathcal{E}(\widehat{\beta}_n^{\mathrm{LS}})\big] = \mathbb{E}\bigg[\bigg\|\frac{1}{n}\sum_{i=1}^n m(X_i)\widetilde{X}_i\bigg\|^2_{\widetilde{\Sigma}_n^{-2}}\bigg] + \frac{1}{n^2}\mathbb{E}\bigg[\sum_{i=1}^n \sigma^2(X_i)\|\widetilde{X}_i\|^2_{\widetilde{\Sigma}_n^{-2}}\bigg]\,, \tag{59}$$

where we let $\widetilde{X}_i = \Sigma^{-1/2}X_i$ and $\widetilde{\Sigma}_n = \Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2}$.

Proof. This follows from Lemma 5 and the fact that, when $\lambda = 0$, for every $x \in \mathbf{R}^d$,

$$\|x\|_{(\widehat{\Sigma}_n+\lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n+\lambda I_d)^{-1}} = \|\Sigma^{-1/2}x\|_{\Sigma^{1/2}\widehat{\Sigma}_n^{-1}\Sigma\widehat{\Sigma}_n^{-1}\Sigma^{1/2}} = \|\Sigma^{-1/2}x\|_{\widetilde{\Sigma}_n^{-2}}\,. \qquad \square$$

5.2 Proof of Theorem 1 and Proposition 1

Upper bound on the minimax risk. We start with an upper bound on the risk the least-squares estimator over the class $\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$. As in Theorem 1, we assume that $n \geqslant d$ and that P_X is non-degenerate. Let $(X, Y) \sim P \in \mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$, so that $m(X) = 0$ and $\sigma^2(X) \leqslant \sigma^2$. It follows from Lemma 6 that

$$\begin{aligned}\mathbb{E}\big[\mathcal{E}(\widehat{\beta}_n^{\mathrm{LS}})\big] &\leqslant \frac{\sigma^2}{n^2}\mathbb{E}\bigg[\sum_{i=1}^n \sigma^2(X_i)\|\widetilde{X}_i\|^2_{\widetilde{\Sigma}_n^{-2}}\bigg] = \frac{\sigma^2}{n^2}\mathbb{E}\bigg[\mathrm{Tr}\bigg(\widetilde{\Sigma}_n^{-2}\sum_{i=1}^n \widetilde{X}_i\widetilde{X}_i^\top\bigg)\bigg] \\ &= \frac{\sigma^2}{n}\mathbb{E}\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})\,.\end{aligned}$$

Hence, the maximum risk of the OLS estimator $\widehat{\beta}_n^{\mathrm{LS}}$ over the class $\mathcal{P}_{\mathrm{well}}(P_X, \sigma^2)$ (and thus the minimax risk over this class) is at most $\sigma^2\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})]/n$.

Lower bound on the minimax risk. We now provide a lower bound on the minimax risk over $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$. We will in fact establish the lower bound both in the setting of Theorem 1 (namely, P_X is non-degenerate and $n \geqslant d$) and that of Proposition 1 (the remaining cases). In particular, we do not assume for now that P_X is non-degenerate or that $n \geqslant d$.

For $\beta^* \in \mathbf{R}^d$, let P_{β^*} denote the joint distribution of (X, Y) where $X \sim P_X$ and $Y = \langle\beta^*, X\rangle + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ independent of X. Now, consider the decision problem with model $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2) = \{P_{\beta^*} : \beta^* \in \mathbf{R}^d\}$, decision space $\mathbf{R}^d$ and loss function $\mathcal{L}(\beta^*, \beta) = \mathcal{E}_{P_{\beta^*}}(\beta) = \|\beta - \beta^*\|^2_\Sigma$. Let $\mathcal{R}(\beta^*, \widehat{\beta}_n) = \mathbb{E}_{\beta^*}[\mathcal{L}(\beta^*, \widehat{\beta}_n)]$ denote the risk under P_{β^*} of a decision rule $\widehat{\beta}_n$ (that is, an estimator of β^* using an i.i.d. sample of size n from P_{β^*}), namely its expected excess risk. Consider the prior $\Pi_\lambda = \mathcal{N}(0, \sigma^2/(\lambda n)I_d)$ on $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$. A standard computation (see, e.g., [GCS+13]) shows that the posterior $\Pi_\lambda(\cdot|(X_1, Y_1), \ldots, (X_n, Y_n))$ is $\mathcal{N}(\widehat{\beta}_{\lambda,n}, (\sigma^2/n)\cdot(\widehat{\Sigma}_n+\lambda I_d)^{-1})$. Since the loss function $\mathcal{L}$ is quadratic, the Bayes estimator under Π_λ is the expectation of the posterior, which is $\widehat{\beta}_{\lambda,n}$. Hence, using the comparison between minimax and Bayes risks:

$$\inf_{\widehat{\beta}_n}\sup_{P_{\beta^*}\in\mathcal{P}_{\mathrm{Gauss}}(P_X,\sigma^2)} \mathcal{R}(\beta^*, \widehat{\beta}_n) \geqslant \inf_{\widehat{\beta}_n} \mathbb{E}_{\beta^*\sim\Pi_\lambda}\big[\mathcal{R}(\beta^*, \widehat{\beta}_n)\big] = \mathbb{E}_{\beta^*\sim\Pi_\lambda}\big[\mathcal{R}(\beta^*, \widehat{\beta}_{\lambda,n})\big]\,, \tag{60}$$

where the infimum is over all estimators $\widehat{\beta}_n$. Note that the left-hand side of (60) is simply the minimax excess risk over $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$. On the other hand, applying Lemma 5 with $m(X) = 0$ and $\sigma^2(X) = \sigma^2$ and noting that

$$\begin{aligned}\mathbb{E}\bigg[\sum_{i=1}^n \|X_i\|^2_{(\widehat{\Sigma}_n+\lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n+\lambda I_d)^{-1}}\bigg] &= \mathbb{E}\bigg[\mathrm{Tr}\bigg\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}\sum_{i=1}^n X_iX_i^\top\bigg\}\bigg] \\ &= n\,\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}\widehat{\Sigma}_n\big\}\big]\,,\end{aligned}$$

we obtain

$$\mathcal{R}(\beta^*, \widehat{\beta}_{\lambda,n}) = \lambda^2 \, \mathbb{E}\Big[\|\beta^*\|^2_{(\widehat{\Sigma}_n+\lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n+\lambda I_d)^{-1}}\Big] + \frac{\sigma^2}{n}\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}\widehat{\Sigma}_n\big\}\big] \, .$$

This implies that

$$\begin{aligned}\mathbb{E}_{\beta^*\sim\Pi_\lambda}\big[\mathcal{R}(\beta^*, \widehat{\beta}_{\lambda,n})\big] &= \mathbb{E}_{\beta^*\sim\Pi_\lambda}\Big[\lambda^2 \, \mathbb{E}\Big[\|\beta^*\|^2_{(\widehat{\Sigma}_n+\lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n+\lambda I_d)^{-1}}\Big]\Big] + \\ &\quad + \frac{\sigma^2}{n}\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}\widehat{\Sigma}_n\big\}\big] \qquad (61)\end{aligned}$$

where $\mathbb{E}$ simply denotes the expectation with respect to $(X_1, \dots, X_n) \sim P_X^n$. Now, by Fubini's theorem, and since $\mathbb{E}_{\beta^*\sim\Pi_\lambda}[\beta^*(\beta^*)^\top] = \sigma^2/(\lambda n) I_d$, we have

$$\begin{aligned}&\mathbb{E}_{\beta^*\sim\Pi_\lambda}\Big[\lambda^2 \, \mathbb{E}\Big[\|\beta^*\|^2_{(\widehat{\Sigma}_n+\lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n+\lambda I_d)^{-1}}\Big]\Big] \\ &= \lambda^2 \cdot \mathbb{E}\Big[\mathbb{E}_{\beta^*\sim\Pi_\lambda}\Big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}\beta^*(\beta^*)^\top\big\}\Big]\Big] \\ &= \frac{\sigma^2}{n}\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}\lambda I_d\big\}\big] \, . \qquad (62)\end{aligned}$$

Plugging (62) into (61) shows that the Bayes risk under Π_λ equals

$$\frac{\sigma^2}{n}\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda I_d)^{-1}(\widehat{\Sigma}_n + \lambda I_d)\big\}\big] = \frac{\sigma^2}{n}\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma\big\}\big] \, . \qquad (63)$$

Hence, by (60) the minimax risk is larger than $(\sigma^2/n) \cdot \mathbb{E}[\mathrm{Tr}\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma\}]$ for every $\lambda > 0$. We now distinguish the settings of Theorem 1 and Proposition 1.

Degenerate case. First, assume that P_X is degenerate or that $n < d$. By Fact 1, with probability $p > 0$, the matrix $\widehat{\Sigma}_n$ is non-invertible. When this occurs, let $\theta \in \mathbf{R}^d$ be such that $\|\theta\| = 1$ and $\widehat{\Sigma}_n(\Sigma^{-1/2}\theta) = 0$. We then have, for every $\lambda > 0$,

$$\langle \Sigma^{-1/2}(\widehat{\Sigma}_n + \lambda I_d)\Sigma^{-1/2}\theta, \theta\rangle = 0 + \lambda\|\Sigma^{-1/2}\theta\|^2 \leqslant \lambda \cdot \lambda_{\min}^{-1},$$

where $\lambda_{\min} = \lambda_{\min}(\Sigma)$ denotes the smallest eigenvalue of Σ. This implies that

$$\begin{aligned}\mathrm{Tr}\{\Sigma^{1/2}(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma^{1/2}\} &\geqslant \lambda_{\max}(\Sigma^{1/2}(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma^{1/2}) \\ &= \lambda_{\min}^{-1}(\Sigma^{-1/2}(\widehat{\Sigma}_n + \lambda I_d)\Sigma^{-1/2}) \geqslant \frac{\lambda_{\min}}{\lambda}\end{aligned}$$

so that

$$\frac{\sigma^2}{n}\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma\big\}\big] \geqslant \frac{\sigma^2}{n} \cdot p \cdot \frac{\lambda_{\min}}{\lambda} \, . \qquad (64)$$

Recalling that the left-hand side of equation (64) is a lower bound on the minimax risk for every $\lambda > 0$, and noting that the right-hand side tends to $+\infty$ as $\lambda \to 0$, the minimax risk is infinite as claimed in Proposition 1.

Non-degenerate case. Now, assume that P_X is non-degenerate and that $n \geqslant d$. By Fact 1, $\widehat{\Sigma}_n$ is invertible almost surely. In addition, $\mathrm{Tr}\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma\} = \mathrm{Tr}\{(\Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2} + \lambda\Sigma^{-1})^{-1}\}$ is decreasing in λ (since $\lambda \mapsto \Sigma^{-1/2}\widehat{\Sigma}_n\Sigma^{-1/2} + \lambda\Sigma^{-1}$ is increasing in λ), positive, and converges as $\lambda \to 0^+$ to $\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})$. By the monotone convergence theorem, it follows that

$$\lim_{\lambda\to 0^+} \frac{\sigma^2}{n}\mathbb{E}\big[\mathrm{Tr}\big\{(\widehat{\Sigma}_n + \lambda I_d)^{-1}\Sigma\big\}\big] = \frac{\sigma^2}{n}\mathbb{E}\big[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})\big] \, , \qquad (65)$$

where the limit in the right-hand side belongs to $(0, +\infty]$. Since the left-hand side is a lower bound on the minimax risk, the minimax risk over $\mathcal{P}_{\mathrm{Gauss}}(P_X, \sigma^2)$ is larger than $(\sigma^2/n)\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})]$.

Conclusion. Since $\mathcal{P}_{\mathrm{Gauss}}(P_X,\sigma^2) \subset \mathcal{P}_{\mathrm{well}}(P_X,\sigma^2)$, the minimax risk over $\mathcal{P}_{\mathrm{well}}(P_X,\sigma^2)$ is at least as large as that over $\mathcal{P}_{\mathrm{Gauss}}(P_X,\sigma^2)$. When P_X is degenerate or $n < d$, we showed that the minimax risk over $\mathcal{P}_{\mathrm{Gauss}}(P_X,\sigma^2)$ is infinite, establishing Proposition 1. When P_X is non-degenerate and $n \geqslant d$, the minimax risk over $\mathcal{P}_{\mathrm{well}}(P_X,\sigma^2)$ is smaller, and the minimax risk over $\mathcal{P}_{\mathrm{Gauss}}(P_X,\sigma^2)$ larger, than $(\sigma^2/n)\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})]$, so that these quantities agree and equal $(\sigma^2/n)\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})]$, as claimed in Theorem 1.

5.3 Proof of Theorem 3

The proof starts with the following lemma.

Lemma 7. *For any positive symmetric $d \times d$ matrix A and $p \in [1,2]$,*

$$\mathrm{Tr}(A^{-1}) + \mathrm{Tr}(A) - 2d \leqslant \max(1, \lambda_{\min}(A)^{-1}) \cdot \mathrm{Tr}\big(|A - I_d|^{2/p}\big)\,. \tag{66}$$

Proof of Lemma 7. Let us start by showing that, for every $a > 0$,

$$a^{-1} + a - 2 \leqslant \max(1, a^{-1}) \cdot |a-1|^{2/p}\,. \tag{67}$$

Multiplying both sides of (67) by $a > 0$, it amounts to

$$(a-1)^2 = 1 + a^2 - 2a \leqslant \max(a,1) \cdot |a-1|^{2/p}\,,$$

namely to $|a-1|^{2-2/p} \leqslant \max(a,1)$. For $a \in (0,2]$, this inequality holds since $|a-1| \leqslant 1$ and $2 - 2/p \geqslant 0$, so that $|a-1|^{2-2/p} \leqslant 1 \leqslant \max(a,1)$. For $a \geqslant 2$, the inequalities $|a-1| \geqslant 2$ and $2 - 2/p \leqslant 1$ imply that $|a-1|^{2-2/p} \leqslant |a-1| \leqslant a \leqslant \max(a,1)$. This establishes (67).

Now, let $a_1, \ldots, a_d > 0$ be the eigenvalues of A. Without loss of generality, assume that $a_d = \min_j(a_j) = \lambda_{\min}(A)$. Then, by inequality (67) and the bound $\max(1, a_j^{-1}) \leqslant \max(1, a_d^{-1})$, we have

$$\mathrm{Tr}(A^{-1}) + \mathrm{Tr}(A) - 2d = \sum_{j=1}^d (a_j^{-1} + a_j - 2) \leqslant \max(1, a_d^{-1}) \sum_{j=1}^d |a_j - 1|^{2/p}\,,$$

which is precisely the desired inequality (66). □

Proof of Theorem 3. Let $p \in (1,2]$ which will be determined later, and denote $q := p/(p-1)$ its complement. Applying Lemma 7 to $A = \widetilde{\Sigma}_n$ yields:

$$\mathrm{Tr}(\widetilde{\Sigma}_n^{-1}) + \mathrm{Tr}(\widetilde{\Sigma}_n) - 2d \leqslant \max(1, \lambda_{\min}(\widetilde{\Sigma}_n)^{-1}) \cdot \mathrm{Tr}\big(|\widetilde{\Sigma}_n - I_d|^{2/p}\big)\,.$$

Since $\mathbb{E}[\mathrm{Tr}(\widetilde{\Sigma}_n)] = d$, taking the expectation in the above bound and dividing by d yields:

$$\begin{aligned}
\frac{1}{d}\mathbb{E}\big[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})\big] - 1 &\leqslant \mathbb{E}\Big[\max(1, \lambda_{\min}(\widetilde{\Sigma}_n)^{-1}) \cdot \frac{1}{d}\mathrm{Tr}\big(|\widetilde{\Sigma}_n - I_d|^{2/p}\big)\Big] \\
&\leqslant \mathbb{E}\big[\max(1, \lambda_{\min}(\widetilde{\Sigma}_n)^{-1})^q\big]^{1/q} \cdot \mathbb{E}\Big[\Big(\frac{1}{d}\mathrm{Tr}\big(|\widetilde{\Sigma}_n - I_d|^{2/p}\big)\Big)^p\Big]^{1/p} \qquad (68) \\
&\leqslant \mathbb{E}\big[\max(1, \lambda_{\min}(\widetilde{\Sigma}_n)^{-q})\big]^{1/q} \cdot \mathbb{E}\Big[\frac{1}{d}\mathrm{Tr}\big((\widetilde{\Sigma}_n - I_d)^2\big)\Big]^{1/p} \qquad (69)
\end{aligned}$$

where (68) comes from Hölder's inequality, while (69) is obtained by noting that $x \mapsto x^p$ is convex and that $(1/d)\mathrm{Tr}(A)$ is the average of the eigenvalues of the symmetric matrix A. Next,

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{d}\mathrm{Tr}\big((\widetilde{\Sigma}_n - I_d)^2\big)\Big] &= \frac{1}{d}\mathrm{Tr}\Big\{\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^n(\widetilde{X}_i\widetilde{X}_i^\top - I_d)\Big)^2\Big]\Big\} \\ &= \frac{1}{n^2 d}\mathrm{Tr}\Big\{\sum_{1\leqslant i,j\leqslant n}\mathbb{E}\big[(\widetilde{X}_i\widetilde{X}_i^\top - I_d)(\widetilde{X}_j\widetilde{X}_j^\top - I_d)\big]\Big\} \\ &= \frac{1}{nd}\mathrm{Tr}\left\{\mathbb{E}\big[(\widetilde{X}\widetilde{X}^\top - I_d)^2\big]\right\}, \end{aligned} \tag{70}$$

where we used in (70) the fact that, for $i \neq j$, $\mathbb{E}\big[(\widetilde{X}_i\widetilde{X}_i^\top - I_d)(\widetilde{X}_j\widetilde{X}_j^\top - I_d)\big] = \mathbb{E}[\widetilde{X}_i\widetilde{X}_i^\top - I_d]\mathbb{E}[\widetilde{X}_j\widetilde{X}_j^\top - I_d] = 0$. Now, for $x \in \mathbf{R}^d$,

$$\mathrm{Tr}\{(xx^\top - I_d)^2\} = \mathrm{Tr}\{\|x\|^2 xx^\top - 2xx^\top + I_d\} = \|x\|^4 - 2\|x\|^2 + d\,,$$

so that (70) becomes, as $\mathbb{E}[\|\widetilde{X}\|^2] = d$ and $\mathbb{E}[\|\widetilde{X}\|^4] \leqslant \kappa d^2$ (Assumption 2),

$$\mathbb{E}\Big[\frac{1}{d}\mathrm{Tr}\big((\widetilde{\Sigma}_n - I_d)^2\big)\Big] = \frac{1}{nd}\Big(\mathbb{E}\|\widetilde{X}\|^4 - 2\mathbb{E}\|\widetilde{X}\|^2 + d\Big) = \frac{1}{n}\Big(\frac{1}{d}\mathbb{E}\|\widetilde{X}\|^4 - 1\Big) \leqslant \frac{\kappa d}{n}. \tag{71}$$

In addition, recall that $\widetilde{X}$ satisfies Assumption 1 and that $n \geqslant \max(6d/\alpha, 12/\alpha)$. Hence, letting $C' \geqslant 1$ be the constant in Theorem 4, we have by Corollary 4:

$$\mathbb{E}\big[\max(1, \lambda_{\min}(\widetilde{\Sigma}_n)^{-q})\big] \leqslant 2C'^q\,. \tag{72}$$

Finally, plugging the bounds (71) and (72) into (69) and letting $q = \alpha' n/2$, so that $1/p = 1 - 1/q = 1 - 2/(\alpha' n)$, we obtain

$$\frac{1}{d}\cdot\mathbb{E}\big[\mathrm{Tr}(\widetilde{\Sigma}_n^{-1})\big] - 1 \leqslant (2C'^q)^{1/q}\cdot\Big(\frac{\kappa d}{n}\Big)^{1/p} \leqslant 2C'\cdot\frac{\kappa d}{n}\cdot\Big(\frac{n}{\kappa d}\Big)^{2/(\alpha' n)}. \tag{73}$$

Now, since $\kappa = \mathbb{E}[\|\widetilde{X}\|^4]/\mathbb{E}[\|\widetilde{X}\|^2]^2 \geqslant 1$ and $d \geqslant 1$,

$$\Big(\frac{n}{\kappa d}\Big)^{2/(\alpha' n)} \leqslant n^{2/(\alpha' n)} = \exp\Big(\frac{2\log n}{\alpha' n}\Big).$$

An elementary analysis shows that the function $g : x \mapsto \log x/x$ is increasing on $(0, e]$ and decreasing on $[e, +\infty)$. Hence, if $x, y > 1$ satisfy $x \geqslant y\log y \geqslant e$, then

$$\frac{\log x}{x} \leqslant \frac{\log y + \log\log y}{y\log y} \leqslant \frac{1 + e^{-1}}{y}$$

where we used $\log\log y/\log y \leqslant g(e) = e^{-1}$. Here by assumption $n \geqslant 12\alpha^{-1}\log(12\alpha^{-1}) = 2\alpha'^{-1}\log(2\alpha'^{-1})$, and thus $\log n/n \leqslant (1 + e^{-1})/(2/\alpha')$, so that

$$\Big(\frac{n}{\kappa d}\Big)^{2/(\alpha' n)} \leqslant \exp\Big(\frac{2}{\alpha'}\cdot\frac{1 + e^{-1}}{2/\alpha'}\Big) = \exp\big(1 + e^{-1}\big) \leqslant 4\,.$$

Plugging this inequality into (73) yields the desired bound (21). Equation (22) then follows by Theorem 1. □

5.4 Proof of Proposition 3

Recall that, by Lemma 6, we have

$$\mathbb{E}\big[\mathcal{E}(\widehat{\beta}_n^{\mathrm{LS}})\big] = \mathbb{E}\bigg[\Big\|\frac{1}{n}\sum_{i=1}^n m(X_i)\Sigma^{-1/2}X_i\Big\|^2_{\widetilde{\Sigma}_n^{-2}}\bigg] + \frac{1}{n^2}\mathbb{E}\bigg[\sum_{i=1}^n \sigma^2(X_i)\big\|\Sigma^{-1/2}X_i\big\|^2_{\widetilde{\Sigma}_n^{-2}}\bigg]\,. \tag{74}$$

Now, since $\widetilde{\Sigma}_n^{-2} \leqslant \lambda_{\min}(\widetilde{\Sigma}_n)^{-2} I_d$, we have for every random variable V_n:

$$\begin{aligned}\mathbb{E}\big[\|V_n\|^2_{\widetilde{\Sigma}_n^{-2}}\big] &\leqslant \mathbb{E}\big[\|V_n\|^2\big] + \mathbb{E}\big[\big\{\lambda_{\min}(\widetilde{\Sigma}_n)^{-2} - 1\big\}_+ \cdot \|V_n\|^2\big] \\ &\leqslant \mathbb{E}\big[\|V_n\|^2\big] + \mathbb{E}\big[\big\{\lambda_{\min}(\widetilde{\Sigma}_n)^{-2} - 1\big\}_+^2\big]^{1/2} \cdot \mathbb{E}\big[\|V_n\|^4\big]^{1/2}\,, \end{aligned} \tag{75}$$

where (75) follows from the Cauchy-Schwarz inequality. Letting $V_n = \sigma(X_i)\Sigma^{-1/2}X_i$, we obtain from (75)

$$\begin{aligned}&\frac{1}{n^2}\mathbb{E}\bigg[\sum_{i=1}^n \sigma^2(X_i)\big\|\Sigma^{-1/2}X_i\big\|^2_{\widetilde{\Sigma}_n^{-2}}\bigg] \\ &\leqslant \frac{1}{n}\mathbb{E}\big[\sigma^2(X)\|\Sigma^{-1/2}X\|^2\big] + \frac{1}{n}\mathbb{E}\big[\{\lambda_{\min}(\widetilde{\Sigma}_n)^{-2} - 1\}_+^2\big]^{1/2}\mathbb{E}\big[\sigma^4(X)\|\Sigma^{-1/2}X\|^4\big]^{1/2}\,. \end{aligned} \tag{76}$$

On the other hand, let $V_n = n^{-1}\sum_{i=1}^n m(X_i)\Sigma^{-1/2}X_i$; we have, since $\mathbb{E}[m(X_i)X_i] = \mathbb{E}[\varepsilon_i X_i] = 0$,

$$\begin{aligned}\mathbb{E}\big[\|V_n\|^2\big] &= \mathbb{E}\bigg[\Big\|\frac{1}{n}\sum_{i=1}^n m(X_i)X_i\Big\|^2_{\Sigma^{-1}}\bigg] \\ &= \frac{1}{n^2}\sum_{1\leqslant i,j\leqslant n}\mathbb{E}\big[\,\langle m(X_i)X_i, m(X_j)X_j\rangle_{\Sigma^{-1}}\,\big] \\ &= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\big[m(X_i)^2\|\Sigma^{-1/2}X_i\|^2\big] + \frac{1}{n^2}\sum_{i\neq j}\big\langle\mathbb{E}[m(X_i)X_i], \mathbb{E}[m(X_j)X_j]\big\rangle_{\Sigma^{-1}} \\ &= \frac{1}{n}\mathbb{E}\big[m(X)^2\|\Sigma^{-1/2}X\|^2\big]\,. \end{aligned} \tag{77}$$

In addition,

$$\mathbb{E}\big[\|V_n\|^4\big] = \frac{1}{n^4}\sum_{1\leqslant i,j,k,l\leqslant n}\mathbb{E}\big[\,\langle m(X_i)X_i, m(X_j)X_j\rangle_{\Sigma^{-1}}\,\langle m(X_k)X_k, m(X_l)X_l\rangle_{\Sigma^{-1}}\,\big].$$

Now, by independence and since $\mathbb{E}[m(X)X] = 0$, each term in the sum above where one index among i, j, k, l is distinct from the others cancels. We therefore have

$$\begin{aligned}\mathbb{E}\big[\big\|V_n\big\|^4\big] &= \frac{1}{n^4}\sum_{i=1}^n \mathbb{E}\big[\|m(X_i)X_i\|^4_{\Sigma^{-1}}\big] + \frac{2}{n^4}\sum_{i<j}\mathbb{E}\big[\|m(X_i)X_i\|^2_{\Sigma^{-1}}\|m(X_j)X_j\|^2_{\Sigma^{-1}}\big] + \\ &\quad + \frac{4}{n^4}\sum_{1\leqslant i<j\leqslant n}\mathbb{E}\big[\langle m(X_i)X_i, m(X_j)X_j\rangle^2_{\Sigma^{-1}}\big] \\ &\leqslant \frac{1}{n^4}\sum_{i=1}^n \mathbb{E}\big[\|m(X_i)X_i\|^4_{\Sigma^{-1}}\big] + \frac{6}{n^4}\sum_{i<j}\mathbb{E}\big[\|m(X_i)X_i\|^2_{\Sigma^{-1}}\|m(X_j)X_j\|^2_{\Sigma^{-1}}\big] \qquad (78) \\ &= \frac{1}{n^3}\cdot\mathbb{E}\big[m(X)^4\|\Sigma^{-1/2}X\|^4\big] + \frac{6}{n^4}\cdot\frac{n(n-1)}{2}\cdot\mathbb{E}\big[m(X)^2\|X\|^2_{\Sigma^{-1}}\big]^2 \\ &\leqslant \frac{1}{n^3}\cdot\mathbb{E}\big[m(X)^4\|\Sigma^{-1/2}X\|^4\big] + \frac{3}{n^2}\cdot\mathbb{E}\big[m(X)^2\|\Sigma^{-1/2}X\|^2\big]^2 \\ &\leqslant \frac{4}{n^2}\cdot\mathbb{E}\big[m(X)^4\|\Sigma^{-1/2}X\|^4\big] \qquad (79)\end{aligned}$$

where (78) and (79) rely on the Cauchy-Schwarz inequality. Hence, it follows from (75), (77) and (79) that

$$\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{n}\sum_{i=1}^n m(X_i)\Sigma^{-1/2}X_i\Big\|^2_{\widetilde{\Sigma}_n^{-2}}\Big] \\
&\leqslant \frac{1}{n}\mathbb{E}\big[m(X)^2\|\Sigma^{-1/2}X\|^2\big] + \mathbb{E}\big[\{\lambda_{\min}(\widetilde{\Sigma}_n)^{-2}-1\}_+^2\big]^{1/2}\cdot\Big(\frac{4}{n^2}\cdot\mathbb{E}\big[m(X)^4\|\Sigma^{-1/2}X\|^4\big]\Big)^{1/2} \\
&\leqslant \frac{1}{n}\mathbb{E}\big[m(X)^2\|\Sigma^{-1/2}X\|^2\big] + \frac{2}{n}\mathbb{E}\big[\{\lambda_{\min}(\widetilde{\Sigma}_n)^{-2}-1\}_+^2\big]^{1/2}\mathbb{E}\big[m(X)^4\|\Sigma^{-1/2}X\|^4\big]^{1/2}. \qquad (80)
\end{aligned}$$

Plugging (76) and (80) into the decomposition (74) yields:

$$\begin{aligned}
\mathbb{E}\big[\mathcal{E}(\widehat{\beta}_n^{\mathrm{LS}})\big] \leqslant \frac{1}{n}\mathbb{E}\big[\big(m(X)^2+\sigma^2(X)\big)\|\Sigma^{-1/2}X\|^2\big] + \frac{1}{n}\mathbb{E}\big[\{\lambda_{\min}(\widetilde{\Sigma}_n)^{-2}-1\}_+^2\big]^{1/2}\times \\
\times\Big(\mathbb{E}\big[\sigma^4(X)\|\Sigma^{-1/2}X\|^4\big]^{1/2} + 2\mathbb{E}\big[m(X)^4\|\Sigma^{-1/2}X\|^4\big]^{1/2}\Big) \qquad (81)
\end{aligned}$$

Oliveira's lower tail bound. [Oli16] showed that, under Assumption 3, we have

$$\mathbb{P}\big(\lambda_{\min}(\widehat{\Sigma}_n)\geqslant 1-\varepsilon\big)\geqslant 1-\delta$$

provided that

$$n\geqslant\frac{81\kappa(d+2\log(2/\delta))}{\varepsilon^2}.$$

This can be rewritten as:

$$\mathbb{P}\Big(\lambda_{\min}(\widehat{\Sigma}_n) < 1-9\kappa^{1/2}\sqrt{\frac{d+2\log(2/\delta)}{n}}\Big)\leqslant\delta. \qquad (82)$$

Bound on the remaining term. Since the function $x\mapsto x^2$ is 2-Lipschitz on $[0,1]$, we have $(x^{-2}-1)_+ = (1-x^2)_+/x^2 \leqslant 2(1-x)_+/x^2$ for $x>0$, so that by Cauchy-Schwarz,

$$\begin{aligned}
\mathbb{E}\big[\{\lambda_{\min}(\widehat{\Sigma}_n)^{-2}-1\}_+^2\big]^{1/2} &\leqslant \mathbb{E}\Big[\frac{4\{1-\lambda_{\min}(\widehat{\Sigma}_n)\}_+^2}{\lambda_{\min}(\widehat{\Sigma}_n)^4}\Big]^{1/2} \\
&\leqslant 2\mathbb{E}\big[\{1-\lambda_{\min}(\widehat{\Sigma}_n)\}_+^4\big]^{1/4}\mathbb{E}\big[\lambda_{\min}(\widehat{\Sigma}_n)^{-8}\big]^{1/4}. \qquad (83)
\end{aligned}$$

First, note that

$$\begin{aligned}
\mathbb{E}\big[\{1-\lambda_{\min}(\widehat{\Sigma}_n)\}_+^4\big] &= \int_0^\infty \mathbb{P}\big(\{1-\lambda_{\min}(\widehat{\Sigma}_n)\}_+^4\geqslant u\big)\mathrm{d}u \\
&= \int_0^1 \mathbb{P}\big(\lambda_{\min}(\widehat{\Sigma}_n)\leqslant 1-u^{1/4}\big)\mathrm{d}u \\
&= \int_0^1 \mathbb{P}\big(\lambda_{\min}(\widehat{\Sigma}_n)\leqslant 1-v^{1/2}\big)2v\mathrm{d}v. \qquad (84)
\end{aligned}$$

Now, let $v^{1/2} = 9\kappa^{1/2}\sqrt{[d+2\log(2/\delta)]/n}$, so that the bound (82) yields $\mathbb{P}(\lambda_{\min}(\widehat{\Sigma}_n)\leqslant 1-v^{1/2})\leqslant\delta$. We have, equivalently,

$$\delta = 2\exp\Big(-\frac{n}{162\kappa}\Big(v-\frac{81\kappa d}{n}\Big)\Big)\leqslant 2\exp\Big(-\frac{n}{324\kappa}v\Big)$$

as long as $v \geqslant 162\kappa d/n$. Plugging this inequality into (84) yields

$$\begin{aligned}\mathbb{E}\big[\{1-\lambda_{\min}(\widehat{\Sigma}_n)\}_+^4\big] &\leqslant \int_0^{\min(162\kappa d/n,1)} 2v\mathrm{d}v + \int_{\min(162\kappa d/n,1)}^1 2\exp\Big(-\frac{n}{324\kappa}v\Big)2v\mathrm{d}v \\ &\leqslant \Big(\frac{162\kappa d}{n}\Big)^2 + \Big(\frac{324\kappa}{n}\Big)^2\int_0^\infty 4\exp(-w)w\mathrm{d}w \\ &= \Big(\frac{162\kappa d}{n}\Big)^2 + 4\Big(\frac{324\kappa}{n}\Big)^2\end{aligned}$$

so that, using the inequality $(x+y)^{1/4} \leqslant x^{1/4}+y^{1/4}$,

$$\mathbb{E}\big[\{1-\lambda_{\min}(\widehat{\Sigma}_n)\}_+^4\big]^{1/4} \leqslant 9\sqrt{\frac{2\kappa d}{n}} + 18\sqrt{\frac{2\kappa}{n}} \leqslant 27\sqrt{\frac{2\kappa d}{n}}\,. \tag{85}$$

Also, by Corollary 4 and the fact that $\alpha n/12 \geqslant 8$, $\mathbb{E}[\lambda_{\min}(\widehat{\Sigma}_n)^{-8}] \leqslant 2C'^8$, so that inequality (83) becomes

$$\mathbb{E}\big[\{\lambda_{\min}(\widehat{\Sigma}_n)^{-2}-1\}_+^2\big]^{1/2} \leqslant 2\times 27\sqrt{\frac{2\kappa d}{n}} \times 2^{1/4}C'^2 \leqslant 92C'^2\sqrt{\frac{\kappa d}{n}}\,. \tag{86}$$

Final bound. Now, let $\chi > 0$ as in Proposition 3. Since

$$\mathbb{E}[\varepsilon^2|X] = m(X)^2+\sigma^2(X) \geqslant \max(m(X)^2,\sigma^2(X))\,,$$

we have

$$\begin{aligned}&\max\Big(\mathbb{E}\big[m(X)^4\|\Sigma^{-1/2}X\|^4\big], \mathbb{E}\big[\sigma^4(X)\|\Sigma^{-1/2}X\|^4\big]\Big) \\ &\leqslant \mathbb{E}[\mathbb{E}[\varepsilon^2|X]^2\|\Sigma^{-1/2}X\|^4] = \chi d^2\,.\end{aligned} \tag{87}$$

Putting the bounds (86) and (87) inside (81) yields

$$\begin{aligned}\mathbb{E}\big[\mathcal{E}(\widehat{\beta}_n^{\mathrm{LS}})\big] &\leqslant \frac{1}{n}\mathbb{E}\big[\big(m(X)^2+\sigma^2(X)\big)\|\Sigma^{-1/2}X\|^2\big] + \frac{1}{n}\cdot 92C'^2\sqrt{\frac{\kappa d}{n}}\cdot 3\sqrt{\chi}d \\ &= \frac{1}{n}\mathbb{E}\big[(Y-\langle\beta^*,X\rangle)^2\|\Sigma^{-1/2}X\|^2\big] + 276C'^2\sqrt{\kappa\chi}\Big(\frac{d}{n}\Big)^{3/2}\,,\end{aligned} \tag{88}$$

where we used the fact that $\mathbb{E}[(Y-\langle\beta^*,X\rangle)^2|X] = m(X)^2+\sigma^2(X)$. This establishes (23). Finally, if $P \in \mathcal{P}_{\mathrm{mis}}(P_X,\sigma^2)$, then $\mathbb{E}[\varepsilon^2|X] \leqslant \sigma^2$, so that

$$\chi = \mathbb{E}[\mathbb{E}[\varepsilon^2|X]^2\|\Sigma^{-1/2}X\|^4]/d^2 \leqslant \sigma^4\mathbb{E}[\|\Sigma^{-1/2}X\|^4]/d^2 \leqslant \sigma^4\kappa\,,$$

where we used the fact that $\mathbb{E}[\|\Sigma^{-1/2}X\|^4] \leqslant \kappa d^2$ by Assumption 3 (see Remark 3). Plugging this inequality, together with $\mathbb{E}[(Y-\langle\beta^*,X\rangle)^2\|\Sigma^{-1/2}X\|^2] \leqslant \sigma^2 d$, inside (88), yields the upper bound (24). This concludes the proof.

6 Remaining proofs from Section 3

In this section, we gather the proofs of remaining results from Section 3, namely Proposition 4 and Corollary 4.

6.1 Proof of Proposition 4

Let Θ be a random variable distributed uniformly on the unit sphere S^{d-1} and independent of X. We have

$$\sup_{\theta\in S^{d-1}} \mathbb{P}(|\langle\theta, X\rangle| \leqslant t) \geqslant \mathbb{E}\big[\mathbb{P}(|\langle\Theta, X\rangle| \leqslant t|\Theta)\big] = \mathbb{E}\big[\mathbb{P}(|\langle\Theta, X\rangle| \leqslant t|X)\big]\,.$$

Next, note that for every $x \in \mathbf{R}^d$, $\langle\Theta, x\rangle$ is distributed as $\|x\| \cdot \Theta_1$, where Θ_1 denotes the first coordinate of Θ. Since X is independent of Θ, the above inequality becomes

$$\sup_{\theta\in S^{d-1}} \mathbb{P}(|\langle\theta, X\rangle| \leqslant t) \geqslant \mathbb{E}\left[\mathbb{P}\left(|\Theta_1| \leqslant \frac{t}{\|X\|}\Big|X\right)\right]. \tag{89}$$

Now, since $\mathbb{E}[\|X\|^2] = \mathrm{Tr}(\mathbb{E}[XX^\top]) = d$, Markov's inequality implies that $\mathbb{P}(\|X\| \geqslant 2\sqrt{d}) \leqslant \mathbb{E}[\|X\|^2]/(4d) \leqslant 1/4$. Since $r \mapsto \mathbb{P}_\theta(|\theta_1| \leqslant t/r)$ is non-increasing, plugging this into (89) yields

$$\sup_{\theta\in S^{d-1}} \mathbb{P}(|\langle\theta, X\rangle| \leqslant t) \geqslant \frac{3}{4} \cdot \mathbb{P}\left(|\Theta_1| \leqslant \frac{t}{2\sqrt{d}}\right). \tag{90}$$

Let us now derive the distribution of $|\Theta_1|$. Let $\phi : S^{d-1} \to \mathbf{R}$ be the projection on the first coordinate: $\phi(\theta) = \theta_1$ for $\theta \in S^{d-1}$. Note that for $u \in [-1, 1]$, $\phi^{-1}(u) = \{u\} \times (\sqrt{1-u^2} \cdot S^{d-2})$ which is isometric to $\sqrt{1-u^2} \cdot S^{d-2}$ and hence has $(d-2)$-dimensional Hausdorff measure $C_d(1-u^2)^{(d-2)/2}$ for some constant C_d. In addition, since $\phi(\theta) = \langle e_1, \theta\rangle$ (where $e_1 = (1, 0, \ldots, 0)$), $\nabla\phi(\theta) \in (\mathbf{R}\theta)^\perp$ is the orthogonal projection of e_1 on $(\mathbf{R}\theta)^\perp$, namely $e_1 - \theta_1\theta$, with norm $\|\nabla\phi(\theta)\| = \sqrt{1-\theta_1^2}$. Fix $t \in (0, 1]$ and define $g(\theta) = \mathbf{1}(|\theta_1| \leqslant t)/\sqrt{1-\theta_1^2}$, which equals $\mathbf{1}(|u| \leqslant t)/\sqrt{1-u^2}$ on $\phi^{-1}(u)$ (for $u \in (-1, 1)$), and such that $g(\theta) \cdot \|\nabla\phi(\theta)\| = \mathbf{1}(|\theta_1| \leqslant t)$. Hence, the coarea formula [Fed96, Theorem 3.2.2] implies that, for every $t \in (0, 1]$,

$$\begin{aligned}\mathbb{P}(|\Theta_1| \leqslant t) &= \int_{S^{d-1}} g(\theta)\|\nabla\phi(\theta)\|\pi(\mathrm{d}\theta) = \int_{-1}^{1} \frac{\mathbf{1}(|u| \leqslant t)}{\sqrt{1-u^2}} \times C_d(1-u^2)^{(d-2)/2}\mathrm{d}u \\ &= 2C_d \int_0^t (1-u^2)^{(d-3)/2}\mathrm{d}u\,. \end{aligned} \tag{91}$$

If $d = 2$, (91) implies that $|\Theta_1|$ has density $(2/\pi)/\sqrt{1-t^2} \geqslant 2/\pi$ on $[0, 1]$, and hence for $t \in [0, 1]$:

$$\mathbb{P}\left(|\Theta_1| \leqslant \frac{t}{2\sqrt{d}}\right) \geqslant \frac{2}{\pi} \times \frac{t}{2\sqrt{2}}\,. \tag{92}$$

If $d = 3$, (91) implies that $|\Theta_1|$ is uniformly distributed on $[0, 1]$, so that for $t \in [0, 1]$

$$\mathbb{P}\left(|\Theta_1| \leqslant \frac{t}{2\sqrt{d}}\right) = \frac{t}{2\sqrt{3}}\,. \tag{93}$$

Now, assume that $d \geqslant 4$. Letting $t = 1$ in (91) yields the value of the constant C_d, which normalizes the right-hand side: since $1 - u^2 \leqslant e^{-u^2}$,

$$\begin{aligned}(2C_d)^{-1} &= \int_0^1 (1-u^2)^{(d-3)/2}\mathrm{d}u \leqslant \int_0^1 e^{-(d-3)u^2/2}\mathrm{d}u \\ &\leqslant \frac{1}{\sqrt{d-3}} \int_0^{\sqrt{d-3}} e^{-u^2/2}\mathrm{d}u \leqslant \frac{1}{\sqrt{d-3}} \times \sqrt{\frac{\pi}{2}}\,,\end{aligned}$$

so that $2C_d \geqslant \sqrt{2(d-3)/\pi}$. Finally, if $u \leqslant 1/(2\sqrt{d})$, then

$$\left(1-u^2\right)^{(d-3)/2} \geqslant \left(1-\frac{1}{4d}\right)^{d/2} \geqslant \left(1-\frac{1}{16}\right)^2,$$

using the fact that $4d \geqslant 16$ and that the function $x \mapsto (1-1/x)^{x/8}$ is increasing on $(1,+\infty)$. Plugging the above lower bounds in (91) shows that, for $t \leqslant 1$,

$$\mathbb{P}\Big(|\Theta_1| \leqslant \frac{t}{2\sqrt{d}}\Big) = 2C_d \int_0^{t/(2\sqrt{d})} (1-u^2)^{(d-3)/2} \mathrm{d}u \geqslant \sqrt{\frac{2(d-3)}{\pi}} \times \Big(\frac{15}{16}\Big)^2 \frac{t}{2\sqrt{d}} \geqslant \frac{t}{3} \tag{94}$$

where the last inequality is obtained by noting that $(d-3)/d \geqslant 1/4$ for $d \geqslant 4$ and lower bounding the resulting constant. The bounds (92), (93) and (94) imply that, for every $d \geqslant 2$ and $t \leqslant 1$,

$$\mathbb{P}\Big(|\Theta_1| \leqslant \frac{t}{2\sqrt{d}}\Big) \geqslant \frac{t}{\pi\sqrt{2}}\,. \tag{95}$$

The first inequality of Proposition 4 follows by combining inequalities (90) and (95). The second inequality (26) is a consequence of the first by Lemma 2.

6.2 Proof of Corollary 4

Corollary 4 directly follows from Theorem 4, Proposition 4 and Lemma 8 below.

Lemma 8. *Let Z be a nonnegative real variable.*

1. *If there exist some constants $C \geqslant 1$ and $a \geqslant 2$ such that $\mathbb{P}(Z \leqslant t) \leqslant (Ct)^a$ for all $t > 0$, then $\|Z^{-1}\|_{L^q} \leqslant \|\max(1, Z^{-1})\|_{L^q} \leqslant 2^{1/q} C \leqslant 2C$ for all $1 \leqslant q \leqslant a/2$.*
2. *Conversely, if $\|Z^{-1}\|_{L^q} \leqslant C$ for some constants $q \geqslant 1$ and $C > 0$, then $\mathbb{P}(Z \leqslant t) \leqslant (Ct)^q$ for all $t > 0$.*
3. *Finally, if there exist constants $c, a > 0$ such that $\mathbb{P}(Z \leqslant t) \geqslant (ct)^a$ for all $t \in (0,1)$, then $\|Z^{-1}\|_{L^q} = +\infty$ for $q \geqslant a$.*

Proof. For the first point, since $\max(1, Z^{-q})$ is nonnegative, we have

$$\mathbb{E}[\max(1, Z^{-q})] = \int_0^\infty \mathbb{P}(\max(1, Z^{-q}) \geqslant u)\, \mathrm{d}u = \int_0^\infty \mathbb{P}(\min(1, Z) \leqslant u^{-1/q})\, \mathrm{d}u\,.$$

For $u \leqslant C^q$, we bound $\mathbb{P}(\min(1, Z) \leqslant u^{-1/q}) \leqslant 1$, while for $u \geqslant C^q$ (so that $u^{-1/q} \leqslant C^{-1} \leqslant 1$), we bound $\mathbb{P}(\min(1, Z) \leqslant u^{-1/q}) = \mathbb{P}(Z \leqslant u^{-1/q}) \leqslant (Cu^{-1/q})^a$. We then conclude that

$$\|\max(1, Z^{-1})\|_{L^q}^q \leqslant C^q + \int_{C^q}^\infty (C^{-q}u)^{-a/q} \mathrm{d}u = C^q \left[1 + \int_1^\infty v^{-a/q} \mathrm{d}v\right] \leqslant 2C^q\,,$$

where we let $v = C^{-q}u$ and used the fact that $\int_1^\infty v^{-a/q} \mathrm{d}v \leqslant \int_1^\infty v^{-2} \mathrm{d}v = 1$ since $q \leqslant a/2$. The second point follows from Markov's inequality: for every $t > 0$,

$$\mathbb{P}(Z \leqslant t) = \mathbb{P}(Z^{-q} \geqslant t^{-q}) \leqslant t^q \cdot \mathbb{E}[Z^{-q}] \leqslant (Ct)^q\,.$$

Finally, for the third point, since $\mathbb{P}(Z \leqslant u^{-1/q}) \geqslant (cu^{-1/q})^a$ for $u > 1$, we have for $q \geqslant a$:

$$\mathbb{E}[Z^{-q}] = \int_0^\infty \mathbb{P}(Z \leqslant u^{-1/q}) \mathrm{d}u \geqslant \int_1^\infty c^a u^{-a/q} \mathrm{d}u \geqslant c^a \int_1^\infty u^{-1} \mathrm{d}u = +\infty\,. \qquad \square$$

6.3 Proof of Proposition 6

The proof relies on the following lemma.

Lemma 9. *Let $X^1, \dots, X^d$ be independent real random variables. Assume that there exists a sub-additive function $g : \mathbf{R}^+ \to \mathbf{R}$ such that, for every $j = 1, \dots, d$ and $\xi \in \mathbf{R}$,*

$$|\Phi_{X^j}(\xi)| \leqslant \exp(-g(\xi^2)) \,.$$

Then, for every $t \in \mathbf{R}$,

$$Q_X(t) \leqslant t \cdot \int_{-2\pi/t}^{2\pi/t} \exp(-g(\xi^2)) \, \mathrm{d}\xi \,. \tag{96}$$

Proof of Lemma 9. For every $\theta \in S^{d-1}$ and $\xi \in \mathbf{R}$, we have, by independence of the X^j,

$$\begin{aligned} |\Phi_{\langle \theta, X \rangle}(\xi)| &= \left|\mathbb{E}\left[e^{i\xi(\theta_1 X^1 + \cdots + \theta_d X^d)}\right]\right| = \left|\mathbb{E}\left[e^{i\xi\theta_1 X^1}\right]\right| \cdots \left|\mathbb{E}\left[e^{i\xi\theta_d X^d}\right]\right| \\ &\leqslant \exp\left[-\left(g(\theta_1^2\xi^2) + \cdots + g(\theta_d^2\xi^2)\right)\right] \leqslant \exp(-g(\xi^2)) \,, \end{aligned}$$

where the last inequality uses the sub-additivity of g and the fact that $\theta_1^2 + \cdots + \theta_d^2 = \|\theta\|^2 = 1$. Lemma 9 then follows from Esséen's inequality [Ess66], which states that for any real random variable Z,

$$Q_Z(t) \leqslant t \cdot \int_{-2\pi/t}^{2\pi/t} |\Phi_Z(\xi)| \, \mathrm{d}\xi \,. \qquad \square$$

Proof of Proposition 6. The functions $g_1 : u \mapsto \alpha \log(1+u)$ and $g_2 : u \mapsto C_0^{-1}\sqrt{u}$ are concave functions on $\mathbf{R}^+$ taking the value 0 at 0, and therefore sub-additive. Since g_1 is also increasing, the function $g : u \mapsto g_1 \circ g_2(u) = \alpha \log(1 + C_0^{-1}\sqrt{u})$ is also sub-additive. Condition (31) simply writes $\Phi_{X^j}(\xi) \leqslant \exp(-g(\xi^2))$, so that by Lemma 9

$$Q_X(t) \leqslant t \int_{-2\pi/t}^{2\pi/t} \frac{1}{(1 + |\xi|/C_0)^\alpha} \mathrm{d}\xi \leqslant 2t \int_0^{2\pi/t} \frac{\mathrm{d}\xi}{(\xi/C_0)^\alpha} = \frac{2tC_0^\alpha}{1-\alpha} \left(\frac{2\pi}{t}\right)^{1-\alpha} ,$$

which implies that $Q_X(t) \leqslant (Ct)^\alpha$, concluding the proof. $\square$

7 Conclusion

We analyzed random-design linear prediction from a minimax perspective, by obtaining matching upper and lower bounds on the risk under weak conditions. This revealed that the hardness of the problem is characterized by the distribution of leverage scores, and that Gaussian design is almost the most favorable one in high dimension.

The upper bounds relied on a study of the lower tail and negative moments of empirical covariance matrices. We showed a general lower bound on this lower tail in dimension $d \geqslant 2$, as well as a matching upper bound under a necessary regularity condition on the design. The proof of this result relied on the use of PAC-Bayesian smoothing of empirical processes, with refined non-Gaussian smoothing distributions.

It is worth noting that the upper bound of Theorem 4 on the lower tail of $\lambda_{\min}(\widehat{\Sigma}_n)$ requires $n \geqslant 6d$; the approach used here is not sufficient to obtain meaningful bounds for (nearly) square matrices, whose aspect ratio d/n is close to 1. It could be interesting to see if the bound of Theorem 4 can be extended to this case (for instance in the case of centered, variance 1 independent coordinates with bounded density, as in Section 3.3), by leveraging the techniques from [RV08, RV09, TV09b, TV09a].

Acknowledgements. The author would like to thank two anonymous referees and an associate editor for very helpful comments that improved the quality of this paper.

Funding. Part of this work was carried at Centre de Mathématiques Appliquées, École polytechnique, France, and supported by a public grant as part of the Investissement d'avenir project, reference ANR-11-LABX-0056-LMH, LabEx LMH. Part of this work was carried out at the Machine Learning Genoa center, Università di Genova, Italy.

References

[AC10] Jean-Yves Audibert and Olivier Catoni. Linear regression through PAC-Bayesian truncation. *arXiv preprint 1010.0072*, 2010.

[AC11] Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794, 2011.

[AGZ10] Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*. Cambridge University Press, 2010.

[ALPTJ10] Radosław Adamczak, Alexander Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *J. Amer. Math. Soc.*, 23(2):535–561, 2010.

[And03] Theodore W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley New York, 2003.

[AW01] Katy S. Azoury and Manfred K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Mach. Learn.*, 43(3):211–246, 2001.

[BF83] Leo Breiman and David Freedman. How many variables should be entered in a regression equation? *J. Amer. Statist. Assoc.*, 78(381):131–136, 1983.

[Bha09] Rajendra Bhatia. *Positive Definite Matrices*. Princeton University Press, 2009.

[BKM+15] Peter L. Bartlett, Wouter M. Koolen, Alan Malek, Eiji Takimoto, and Manfred K. Warmuth. Minimax fixed-design linear regression. In *Proc. 28th Conference on Learning Theory*, pages 226–239, 2015.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013.

[BS10] Zhidong Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics. Springer-Verlag, 2010.

[BTW07] Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Aggregation for gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[Cat04] Olivier Catoni. *Statistical Learning Theory and Stochastic Optimization: Ecole d'Eté de Probabilités de Saint-Flour XXXI - 2001*. Lecture Notes in Mathematics. Springer-Verlag, 2004.

[Cat07] Olivier Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *IMS Lecture Notes Monograph Series*. Institute of Mathematical Statistics, 2007.

[CDV07] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7(3):331–368, 2007.

[CH88] Samprit Chatterjee and Ali S. Hadi. *Sensitivity analysis in linear regression*, volume 327 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, New York, 1988.

[CS02a] Felipe Cucker and Steve Smale. Best choices for regularization parameters in learning theory: on the bias-variance problem. *Found. Comput. Math.*, 2(4):413–428, 2002.

[CS02b] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc.*, 39(1):1–49, 2002.

[Dic16] Lee H. Dicker. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1):1–37, 2016.

[DM16] David Donoho and Andrea Montanari. High dimensional robust m-estimation: asymptotic variance via approximate message passing. *Probab. Theory Related Fields*, 166(3):935–969, 2016.

[DVCR05] Ernesto De Vito, Andrea Caponnetto, and Lorenzo Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Found. Comput. Math.*, 5(1):59–85, 2005.

[DW18] Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *Ann. Statist.*, 46(1):247–279, 2018.

[Ede88] Alan Edelman. Eigenvalues and condition numbers of random matrices. *SIAM J. Matrix Anal. Appl.*, 9(4):543–560, 1988.

[EK13] Noureddine El Karoui. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv:1311.2445*, 2013.

[EK18] Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probab. Theory Related Fields*, 170(1):95–175, 2018.

[EKK11] Noureddine El Karoui and Holger Kösters. Geometric sensitivity of random matrix results: consequences for shrinkage estimators of covariance and related statistical methods. *arXiv preprint 1105.1404*, 2011.

[Ess66] Carl G. Esseen. On the Kolmogorov-Rogozin inequality for the concentration function. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 5(3):210–216, 1966.

[Fed96] Herbert Federer. *Geometric measure theory*. Springer, 1996.

[Fos91] Dean P. Foster. Prediction in the worst case. *Ann. Statist.*, 19:1084–1090, 1991.

[GCS+13] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2013.

[GKKW02] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2002.

[HJ90] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[HKZ14] Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. *Found. Comput. Math.*, 14(3):569–600, 2014.

[Hoe62] Arthur E. Hoerl. Application of ridge analysis to regression problems. *Chemical Engineering Progress*, 58:54–59, 1962.

[HS16] Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.*, 17(18):1–40, 2016.

[Hub73] Peter J. Huber. Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.*, 1(5):799–821, 1973.

[Hub81] Peter J. Huber. *Robust statistics*. John Wiley and Sons, 1981.

[HW78] David C. Hoaglin and Roy E. Welsch. The hat matrix in regression and ANOVA. *Amer. Statist.*, 32(1):17–22, 1978.

[Joh19] Iain M. Johnstone. Gaussian estimation: Sequence and wavelet models. Draft version, September 16, 2019, 2019.

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

[KM15] Vladimir Koltchinskii and Shahar Mendelson. Bounding the smallest singular value of a random matrix without concentration. *Int. Math. Res. Not. IMRN*, 2015(23):12991–13008, 2015.

[LC98] Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, 1998.

[Led01] Michel Ledoux. *The Concentration of Measure Phenomenon*. American Mathematical Society, 2001.

[LM16] Guillaume Lecué and Shahar Mendelson. Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3):1520–1534, 2016.

[LM19] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: a survey. *Found. Comput. Math.*, 19:1145–1190, 2019.

[Löw34] Karl Löwner. Über monotone matrixfunktionen. *Math. Z.*, 38(1):177–216, 1934.

[LST03] John Langford and John Shawe-Taylor. PAC-Bayes & margins. In *Advances in Neural Information Processing Systems 15*, pages 439–446, 2003.

[McA99] David A. McAllester. Some PAC-Bayesian theorems. *Mach. Learn.*, 37(3):355–363, 1999.

[McA03] David A. McAllester. PAC-Bayesian stochastic model selection. *Mach. Learn.*, 51(1):5–21, 2003.

[Men15] Shahar Mendelson. Learning without concentration. *J. ACM*, 62(3):21, 2015.

[MP67] Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

[MP14] Shahar Mendelson and Grigoris Paouris. On the singular values of random matrices. *J. Eur. Math. Soc.*, 16(4):823–834, 2014.

[Nem00] Arkadi Nemirovski. Topics in non-parametric statistics. *Ecole d'Ete de Probabilites de Saint-Flour XXVIII-1998*, 28:85–277, 2000.

[NV13] Hoi H. Nguyen and Van H. Vu. Small ball probability, inverse theorems, and applications. In *Erdös Centennial*, pages 409–463. Springer, 2013.

[Oli16] Roberto I. Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probab. Theory Related Fields*, 166(3):1175–1194, 2016.

[RM16] Garvesh Raskutti and Michael W. Mahoney. A statistical perspective on randomized sketching for ordinary least-squares. *J. Mach. Learn. Res.*, 17(1):7508–7538, 2016.

[Rog87] Boris A. Rogozin. The estimate of the maximum of the convolution of bounded densities. *Teor. Veroyatn. Primen.*, 32(1):53–61, 1987.

[RV08] Mark Rudelson and Roman Vershynin. The Littlewood–Offord problem and invertibility of random matrices. *Adv. Math.*, 218(2):600–633, 2008.

[RV09] Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Comm. Pure Appl. Math.*, 62(12):1707–1739, 2009.

[RV10] Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proc. International Congress of Mathematicians*, volume 3, pages 1576–1602, 2010.

[RV14] Mark Rudelson and Roman Vershynin. Small ball probabilities for linear images of high-dimensional distributions. *Int. Math. Res. Not. IMRN*, 2015(19):9594–9617, 2014.

[RWG19] Alessandro Rinaldo, Larry Wasserman, and Max G'Sell. Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *Ann. Statist.*, 47(6):3438–3469, 2019.

[Sha15] Ohad Shamir. The sample complexity of learning linear predictors with the squared loss. *J. Mach. Learn. Res.*, 16(108):3475–3486, 2015.

[SHS09] Ingo Steinwart, Don Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *Proc. 22nd Conference on Learning Theory*, pages 79–93, 2009.

[Ste60] Charles Stein. Multiple regression. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press, 1960.

[SV13] Nikhil Srivastava and Roman Vershynin. Covariance estimation for distributions with $2 + \varepsilon$ moments. *Ann. Probab.*, 41(5):3081–3111, 2013.

[SZ07] Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, 26(2):153–172, 2007.

[Tao12] Terence Tao. *Topics in random matrix theory*. American Mathematical Society, 2012.

[Tik63] Andrey N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics Doklady*, 4:1035–1038, 1963.

[Tik18] Konstantin Tikhomirov. Sample covariance matrices of heavy-tailed distributions. *Int. Math. Res. Not. IMRN*, 2018(20):6254–6289, 2018.

[Tsy03] Alexandre B. Tsybakov. Optimal rates of aggregation. In *Learning Theory and Kernel Machines*, Lecture Notes in Artificial Intelligence, pages 303–313. Springer, 2003.

[Tsy09] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer, 2009.

[TV09a] Terence Tao and Van Vu. From the Littlewood-Offord problem to the circular law: universality of the spectral distribution of random matrices. *Bull. Amer. Math. Soc.*, 46(3):377–396, 2009.

[TV09b] Terence Tao and Van H. Vu. Inverse Littlewood-Offord theorems and the condition number of random discrete matrices. *Ann. of Math.*, 169(2):595–632, 2009.

[vdGM14] Sara van de Geer and Alan Muro. On higher order isotropy conditions and lower bounds for sparse quadratic forms. *Electron. J. Stat.*, 8(2):3031–3061, 2014.

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

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018.

[Vov01] Volodya Vovk. Competitive on-line statistics. *Int. Stat. Rev.*, 69(2):213–248, 2001.

[WV12] Yihong Wu and Sergio Verdú. Optimal phase transitions in compressed sensing. *IEEE Trans. Inform. Theory*, 58(10):6241–6263, 2012.

[Yas14] Pavel Yaskov. Lower bounds on the smallest eigenvalue of a sample covariance matrix. *Electron. Commun. Probab.*, 19, 2014.

[Yas15] Pavel Yaskov. Sharp lower bounds on the least singular value of a random matrix without the fourth moment condition. *Electron. Commun. Probab.*, 20, 2015.