An elementary analysis of ridge regression with random design

Jaouad Mourtada[*] and Lorenzo Rosasco[†]

April 22, 2022

Abstract

In this note, we provide an elementary analysis of the prediction error of ridge regression with random design. The proof is short and self-contained. In particular, it bypasses the use of Rudelson's deviation inequality for covariance matrices, through a combination of exchangeability arguments, matrix perturbation and operator convexity.

Keywords. Ridge regression; reproducing kernel Hilbert spaces; covariance matrices.

1 Introduction

Let (X, Y) be a random vector in $\mathbf{R}^d \times \mathbf{R}$ with distribution P. We consider the problem of random-design regression, namely prediction of Y by linear functions of X. (This setting also allows to consider nonlinear functions of general covariates X', taking values in a measurable space $\mathcal{X}'$, by letting $X = \Phi(X')$ for some feature map $\Phi : \mathcal{X}' \to \mathbf{R}^d$.) Specifically, the prediction error of a regression parameter $\theta \in \mathbf{R}^d$ is defined by its *risk* $L(\theta) = \mathbb{E}[(Y - \langle \theta, X\rangle)^2]$, where $\langle \theta, x\rangle = \theta^\top x$ is the standard scalar product on $\mathbf{R}^d$. In the statistical setting, the true distribution P, and in particular the (population) risk $L : \mathbf{R}^d \to \mathbf{R}^+$ and its minimizer $\theta^* = \arg\min_{\theta \in \mathbf{R}^d} L(\theta)$, are unknown. The aim is then, given a random independent and identically distributed (i.i.d.) sample $(X_1, Y_1), \dots, (X_n, Y_n)$ from P, to find a good parameter $\widehat{\theta}$, as measured by its *excess risk*

$$\mathcal{E}(\widehat{\theta}) = L(\widehat{\theta}) - L(\theta^*) \,. \tag{1}$$

A popular approach to this problem is the method of regularized least squares (also called empirical risk minimization), where the estimator $\widehat{\theta}$ minimizes the sum of an empirical error term and a regularization term favoring some structure on the parameter. In this note, we consider the classical *ridge* estimator [13, 34], defined for $\lambda > 0$ by

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

where $\|\theta\| := \langle \theta, \theta \rangle^{1/2}$ is the Euclidean norm and $\widehat{\Sigma}_n := n^{-1} \sum_{i=1}^n X_i X_i^\top$ is the empirical covariance matrix. While this estimator is well-studied (see Section 3), our aim here is to present a short and elementary analysis of its performance. In particular, the analysis presented here does not rely on matrix concentration [29, 1, 26, 35] or uniform deviation bounds for empirical processes [22, 16, 4], but rather on a combination of exchangeability and matrix convexity arguments. It draws inspiration from an analysis of [25] in the context of conditional density estimation. Our main error estimate is provided in Theorem 1 below.

[*] CREST, ENSAE, Palaiseau, France; jaouad.mourtada@ensae.fr

[†] Universita' di Genova & Istituto Italiano di Tecnologia, Genova, Italy; Center for Brains, Minds and Machines, MIT, Cambridge, United States; lorenzo.rosasco@unige.it

Notation. Given a $d \times d$ matrix A, we denote by $\mathrm{Tr}(A)$ its trace and $\|A\|_{\mathrm{op}}$ its operator norm. The $d \times d$ identity matrix is denoted I_d, or simply I; for $\lambda \in \mathbf{R}$, we denote $A + \lambda = A + \lambda I$. The symbol $\preccurlyeq$ denotes the standard order on symmetric matrices: $A \preccurlyeq B$ means that $\langle Av, v \rangle \leqslant \langle Bv, v \rangle$ for all $v \in \mathbf{R}^d$.

2 Risk analysis of ridge regression

Assumptions. The analysis requires two assumptions on the joint distribution P of (X, Y), the first one on the distribution of the error $Y - \langle \theta^*, X \rangle$, and the second one on the distribution of X.

Assumption 1. There exist $\theta^* \in \mathbf{R}^d$ and $\sigma > 0$ such that

$$\mathbb{E}[Y|X] = \langle \theta^*, X \rangle, \quad \text{and} \quad \mathrm{Var}(Y|X) \leqslant \sigma^2. \tag{3}$$

The first condition in Assumption 1 states that the linear model is well-specified, in the sense that the true regression function $x \mapsto \mathbb{E}[Y|X = x]$ is linear. This condition is standard, although restrictive when the dimension d is low. On the other hand, the guarantees we consider do not explicit depend on the dimension d, and extend with minor changes in notation to the case where $\mathbf{R}^d$ is replaced by an infinite-dimensional Hilbert space. This allows to handle the case of reproducing kernel Hilbert spaces [2] (such as certain Sobolev spaces), for which ridge regression is a classical estimator [41, 32]. When considering a "universal" kernel, the corresponding Hilbert space is dense in the space $L^2(P_X)$ of square-integrable functions of X [32], in which case the well-specified assumption is considerably less restrictive. The main issue is then the dependence of the bound on θ^*. In this respect, the bound of Theorem 1 will only depend on θ^* through the approximation properties of balls of the Hilbert space. Finally, the second condition in Assumption 1 is a bound on the conditional variance of Y given X, which controls the amount of noise. It holds for instance if Y is bounded, or if the error $Y - \langle \theta^*, X \rangle$ is independent of X with finite variance.

Assumption 2. There exists a constant $R > 0$ such that $\|X\| \leqslant R$ almost surely.

The boundedness assumption 2 is classical in the context of ridge regression. This condition is automatically satisfied, for instance, in the case where X is the feature associated to a bounded reproducing kernel Hilbert space [32]. Assumption 2 implies in particular that the covariance matrix $\Sigma = \mathbb{E}[XX^\top]$ of X is well-defined and satisfies $\mathrm{Tr}(\Sigma) = \mathbb{E}\|X\|^2 \leqslant R^2$.

Risk analysis of ridge regression. The proof hinges on two main lemmas. Before presenting them, we start with a classical bias-variance decomposition.

Lemma 1 (Error decomposition). *Under Assumptions 1 and 2, we have for every $\lambda > 0$,*

$$\mathbb{E}[\mathcal{E}(\widehat{\theta}_\lambda)] \leqslant \lambda^2 \mathbb{E}[\langle (\widehat{\Sigma}_n + \lambda)^{-1} \Sigma (\widehat{\Sigma}_n + \lambda)^{-1} \theta^*, \theta^* \rangle] + \frac{\sigma^2}{n} \cdot \mathbb{E}[\mathrm{Tr}\{(\widehat{\Sigma}_n + \lambda)^{-1} \Sigma\}] \, . \tag{4}$$

Proof. Let $Y = \langle \theta^*, X \rangle + \varepsilon$, so that $\mathbb{E}[\varepsilon|X] = 0$ and $\mathbb{E}[\varepsilon^2|X] \leqslant \sigma^2$ almost surely. Then,

$$\widehat{\theta}_\lambda = (\widehat{\Sigma}_n + \lambda)^{-1} \frac{1}{n} \sum_{i=1}^n Y_i X_i = (\widehat{\Sigma}_n + \lambda)^{-1} \widehat{\Sigma}_n \theta^* + (\widehat{\Sigma}_n + \lambda)^{-1} \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \, ,$$

so that

$$\widehat{\theta}_\lambda - \theta^* = -\lambda (\widehat{\Sigma}_n + \lambda)^{-1} \theta^* + (\widehat{\Sigma}_n + \lambda)^{-1} \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \, .$$

Moreover, it is standard (by Pythagoras' theorem in L^2) that $\mathcal{E}(\widehat{\theta}_\lambda) = \|\widehat{\theta}_\lambda - \theta^*\|_\Sigma^2$, where $\|v\|_\Sigma^2 = \langle \Sigma v, v\rangle$ for all $v \in \mathbf{R}^d$. Since $\mathbb{E}[\varepsilon_i | X_1, \dots, X_n] = 0$ and $\mathbb{E}[\varepsilon_i^2 | X_1, \dots, X_n] \leqslant \sigma^2$, then

$$\begin{aligned}
\mathbb{E}[\mathcal{E}(\widehat{\theta}_\lambda)] &= \mathbb{E}[\|\widehat{\theta}_\lambda - \theta^*\|_\Sigma^2] \\
&= \lambda^2 \mathbb{E}\|(\widehat{\Sigma}_n + \lambda)^{-1}\theta^*\|_\Sigma^2 + \frac{1}{n^2}\mathbb{E}\Big[\sum_{i=1}^n \|(\widehat{\Sigma}_n + \lambda)^{-1}\varepsilon_i X_i\|_\Sigma^2\Big] \\
&\leqslant \lambda^2 \mathbb{E}\|(\widehat{\Sigma}_n + \lambda)^{-1}\theta^*\|_\Sigma^2 + \frac{\sigma^2}{n^2}\mathbb{E}\Big[\sum_{i=1}^n \mathrm{Tr}\{\Sigma(\widehat{\Sigma}_n + \lambda)^{-1} X_i X_i^\top (\widehat{\Sigma}_n + \lambda)^{-1}\}\Big] \\
&= \lambda^2 \mathbb{E}[\langle (\widehat{\Sigma}_n + \lambda)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda)^{-1}\theta^*, \theta^*\rangle] + \frac{\sigma^2}{n}\mathbb{E}[\mathrm{Tr}\{(\widehat{\Sigma}_n + \lambda)^{-1}\widehat{\Sigma}_n(\widehat{\Sigma}_n + \lambda)^{-1}\Sigma\}].
\end{aligned}$$

The bound (4) ensues by further bounding $(\widehat{\Sigma}_n + \lambda)^{-1}\widehat{\Sigma}_n(\widehat{\Sigma}_n + \lambda)^{-1} \preccurlyeq (\widehat{\Sigma}_n + \lambda)^{-1}$, and using that $\mathrm{Tr}(A\Sigma) \leqslant \mathrm{Tr}(B\Sigma)$ for symmetric matrices A, B since Σ is positive semi-definite. □

Lemma 1 shows the main quantities that need to be controlled in the random-design setting. The first term in (4) is a bias term, due to the use of a regularization favoring solutions with small norm. The second one is a variance term due to the presence of errors $\varepsilon_i = Y_i - \langle \theta^*, X_i\rangle$. Both terms depend on the (random) sample covariance matrix $\widehat{\Sigma}_n$, as well as the population covariance matrix Σ. The fact that both of these matrices appear comes from the fact that, in the random-design/statistical learning setting, one is interested in making a prediction at a new point X, rather than at the points $X_1, \dots, X_n$ in the sample.

In order to argue that the sample covariance matrix $\widehat{\Sigma}_n$ is "close" to Σ in a suitable sense and deduce an explicit error bound, some assumption on the distribution of X is needed. This is where Assumption 2 is used. Under this assumption, one can apply Rudelson's inequality for sample covariance matrices [29]; this is the approach adopted, for instance, in [7, 14]. Rudelson's inequality, a consequence of the non-commutative Khintchine inequality of [21], is however a non-trivial result, despite subsequent simplifications to its proof [1, 26, 35]. In addition, matrix concentration through Rudelson's inequality introduces an additional logarithmic term [35, 39].

In what follows, we present an alternative approach to controlling the random matrix terms in the right-hand side of (4), which only uses short and elementary arguments. The proof relies on a combination of exchangeability, matrix perturbation and operator convexity. It draws inspiration from an analysis of [25], where similar arguments were used in the context of conditional density estimation and logistic regression.

Lemma 2. *Under Assumption 2, we have that for every $\lambda > 0$,*

$$\mathrm{Tr}[(\Sigma + \lambda I)^{-1}\Sigma] \leqslant \mathbb{E}\mathrm{Tr}[(\widehat{\Sigma}_n + \lambda I)^{-1}\Sigma] \leqslant \Big(1 + \frac{R^2}{\lambda n}\Big) \cdot \mathrm{Tr}[(\Sigma + \lambda I)^{-1}\Sigma]\,. \tag{5}$$

Proof. The lower bound comes from convexity of $A \mapsto \mathrm{Tr}(A^{-1}\Sigma)$ over positive matrices (see Lemma 5 below) and Jensen's inequality. Let us now prove the upper bound. We start by writing:

$$\mathbb{E}\big[\mathrm{Tr}((\widehat{\Sigma}_n + \lambda I)^{-1}\Sigma)\big] = n\,\mathbb{E}\big[\langle (n\widehat{\Sigma}_n + \lambda n I)^{-1} X_{n+1}, X_{n+1}\rangle\big]\,,$$

where X_{n+1} is a random variable distributed as X and independent of $X_1, \dots, X_n$. Now, the Sherman-Morrison identity (10) with $S = n\widehat{\Sigma}_n + \lambda n I$ and $v = X_{n+1}$ shows that

$$\begin{aligned}
&\langle (n\widehat{\Sigma}_n + \lambda n I)^{-1} X_{n+1}, X_{n+1}\rangle \\
&= \big(1 + n^{-1}\langle (\widehat{\Sigma}_n + \lambda I)^{-1} X_{n+1}, X_{n+1}\rangle\big)\langle (n\widehat{\Sigma}_n + X_{n+1}X_{n+1}^\top + \lambda n I)^{-1} X_{n+1}, X_{n+1}\rangle \\
&\leqslant \Big(1 + \frac{R^2}{\lambda n}\Big)\big\langle \big((n+1)\widehat{\Sigma}_{n+1} + \lambda n I\big)^{-1} X_{n+1}, X_{n+1}\big\rangle
\end{aligned}$$

where $\widehat{\Sigma}_{n+1} = (n+1)^{-1}\sum_{i=1}^{n+1} X_i X_i^\top$, and where we used that, by Assumption 2, $\langle(\widehat{\Sigma}_n + \lambda I)^{-1}X_{n+1}, X_{n+1}\rangle \leqslant \|X_{n+1}\|^2/\lambda \leqslant R^2/\lambda$. It follows that

$$\begin{aligned}
\mathbb{E}\big[\mathrm{Tr}((\widehat{\Sigma}_n + \lambda I)^{-1}\Sigma)\big] &\leqslant n\Big(1 + \frac{R^2}{\lambda n}\Big)\mathbb{E}[\langle((n+1)\widehat{\Sigma}_{n+1} + \lambda n I)^{-1}X_{n+1}, X_{n+1}\rangle] \\
&= n\Big(1 + \frac{R^2}{\lambda n}\Big) \cdot \frac{1}{n+1}\sum_{i=1}^{n+1}\mathbb{E}[\mathrm{Tr}\{((n+1)\widehat{\Sigma}_{n+1} + \lambda n I)^{-1}X_i X_i^\top\}] \qquad (6) \\
&= \Big(1 + \frac{R^2}{\lambda n}\Big) \cdot \mathbb{E}[\mathrm{Tr}\{((1+1/n)\widehat{\Sigma}_{n+1} + \lambda I)^{-1}\widehat{\Sigma}_{n+1}\}] \\
&\leqslant \Big(1 + \frac{R^2}{\lambda n}\Big)\mathbb{E}[\mathrm{Tr}\{(\widehat{\Sigma}_{n+1} + \lambda I)^{-1}\widehat{\Sigma}_{n+1}\}] \\
&\leqslant \Big(1 + \frac{R^2}{\lambda n}\Big)\mathrm{Tr}[(\Sigma + \lambda I)^{-1}\Sigma] \qquad (7)
\end{aligned}$$

where (6) follows from exchangeability of $(X_1, \dots, X_{n+1})$, while (7) follows from concavity of the map $A \mapsto \mathrm{Tr}[(A + \lambda I)^{-1}A]$ over positive matrices (by Lemma 5). □

Next, we turn to controlling the bias term in Lemma 3 below. The proof follows a similar recipe as that of Lemma 2.

Lemma 3. *Under Assumption 2, we have that for every $\lambda > 0$,*

$$\mathbb{E}[(\widehat{\Sigma}_n + \lambda)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda)^{-1}] \preccurlyeq \Big(1 + \frac{R^2}{\lambda n}\Big)^2 \lambda^{-1}(\Sigma + \lambda)^{-1}\Sigma\,. \qquad (8)$$

Proof. Similarly to the proof of Lemma 2, we start by writing:

$$\mathbb{E}[(\widehat{\Sigma}_n + \lambda)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda)^{-1}] = n^2\,\mathbb{E}[(n\widehat{\Sigma}_n + \lambda n)^{-1}X_{n+1}X_{n+1}^\top(n\widehat{\Sigma}_n + \lambda n)^{-1}]\,.$$

Next, the Sherman-Morrison identity (10) applied to $S = n\widehat{\Sigma}_n + \lambda n$ and $v = X_{n+1}$ implies that

$$\begin{aligned}
&(n\widehat{\Sigma}_n + \lambda n)^{-1}X_{n+1}X_{n+1}^\top(n\widehat{\Sigma}_n + \lambda n)^{-1} \\
&= (1 + \langle(n\widehat{\Sigma}_n + \lambda n)^{-1}X_{n+1}, X_{n+1}\rangle)^2(n\widehat{\Sigma}_n + \lambda n + X_{n+1}X_{n+1}^\top)^{-1}X_{n+1}X_{n+1}^\top(n\widehat{\Sigma}_n + \lambda n + X_{n+1}X_{n+1}^\top)^{-1} \\
&\preccurlyeq \Big(1 + \frac{R^2}{\lambda n}\Big)^2(n+1)^{-2}(\widehat{\Sigma}_{n+1} + \lambda')^{-1}X_{n+1}X_{n+1}^\top(\widehat{\Sigma}_{n+1} + \lambda')^{-1}
\end{aligned}$$

with $\lambda' = \lambda n/(n+1)$, where we used that $\langle(\widehat{\Sigma}_n + \lambda)^{-1}X_{n+1}, X_{n+1}\rangle \leqslant R^2/\lambda$. It follows that, by exchangeability of $(X_1, \dots, X_{n+1})$,

$$\begin{aligned}
\mathbb{E}[(\widehat{\Sigma}_n + \lambda)^{-1}\Sigma(\widehat{\Sigma}_n + \lambda)^{-1}] &= n^2\mathbb{E}[(n\widehat{\Sigma}_n + \lambda n)^{-1}X_{n+1}X_{n+1}^\top(n\widehat{\Sigma}_n + \lambda n)^{-1}] \\
&\preccurlyeq \Big(1 + \frac{R^2}{\lambda n}\Big)^2\frac{n^2}{(n+1)^2}\mathbb{E}\Big[(\widehat{\Sigma}_{n+1} + \lambda')^{-1}X_{n+1}X_{n+1}^\top(\widehat{\Sigma}_{n+1} + \lambda')^{-1}\Big] \\
&= \Big(1 + \frac{R^2}{\lambda n}\Big)^2\frac{n^2}{(n+1)^2}\frac{1}{n+1}\sum_{j=1}^{n+1}\mathbb{E}\Big[(\widehat{\Sigma}_{n+1} + \lambda')^{-1}X_j X_j^\top(\widehat{\Sigma}_{n+1} + \lambda')^{-1}\Big] \\
&= \Big(1 + \frac{R^2}{\lambda n}\Big)^2\frac{n^2}{(n+1)^2}\mathbb{E}\big[(\widehat{\Sigma}_{n+1} + \lambda')^{-1}\widehat{\Sigma}_{n+1}(\widehat{\Sigma}_{n+1} + \lambda')^{-1}\big] \\
&\preccurlyeq \Big(1 + \frac{R^2}{\lambda n}\Big)^2\frac{n^2}{(n+1)^2}\lambda'^{-1}\mathbb{E}\big[(\widehat{\Sigma}_{n+1} + \lambda')^{-1}\widehat{\Sigma}_{n+1}\big] \\
&\preccurlyeq \Big(1 + \frac{R^2}{\lambda n}\Big)^2\frac{n^2}{(n+1)^2}\lambda'^{-1}(\Sigma + \lambda')^{-1}\Sigma
\end{aligned}$$

where the last inequality follows from operator concavity of $x \mapsto x(x+\lambda')^{-1}$ over $\mathbf{R}^+$ (a consequence of Lemma 5). Inequality (8) is then obtained after substituting $\lambda' = \lambda n/(n+1)$. □

Before deriving the excess risk bound, we express the bias term in a more relatable form.

Lemma 4. *For every* $\lambda > 0$*,*

$$\lambda \|(\Sigma+\lambda)^{-1/2}\Sigma^{1/2}\theta^*\|^2 = \inf_{\theta \in \mathbf{R}^d} \left\{ L(\theta) + \lambda \|\theta\|^2 \right\} - L(\theta^*) \,.$$

Proof. Letting $\theta_\lambda = (\Sigma+\lambda)^{-1}\Sigma\theta^*$, direct computations show that:

$$\begin{aligned}
\inf_{\theta \in \mathbf{R}^d} \left\{ L(\theta) + \lambda \|\theta\|^2 \right\} - L(\theta^*) &= \|\theta_\lambda - \theta^*\|_\Sigma^2 + \lambda \|\theta_\lambda\|^2 \\
&= \lambda^2 \|\Sigma^{1/2}(\Sigma+\lambda)^{-1}\theta^*\|^2 + \lambda \|(\Sigma+\lambda)^{-1}\Sigma\theta^*\|^2 \\
&= \big\langle \big(\lambda^2(\Sigma+\lambda)^{-2}\Sigma + \lambda(\Sigma+\lambda)^{-2}\Sigma^2\big)\theta^*, \theta^* \big\rangle \\
&= \big\langle \lambda(\Sigma+\lambda)^{-1}\Sigma\theta^*, \theta^* \big\rangle \\
&= \lambda \|(\Sigma+\lambda)^{-1/2}\Sigma^{1/2}\theta^*\|^2 \,.
\end{aligned}$$

□

Finally, plugging Lemmas 2, 3 and 4 (the first for the variance term, the last two for the bias term) into the decomposition of Lemma 1, we obtain the following excess risk bound.

Theorem 1. *Under Assumptions 1 and 2, we have for every* $\lambda > 0$*,*

$$\mathbb{E}[\mathcal{E}(\widehat{\theta}_\lambda)] \leqslant \Big(1 + \frac{R^2}{\lambda n}\Big)^2 \inf_{\theta \in \mathbf{R}^d} \left\{ L(\theta) + \lambda \|\theta\|^2 - L(\theta^*) \right\} + \Big(1 + \frac{R^2}{\lambda n}\Big) \frac{\sigma^2 \mathrm{Tr}[(\Sigma+\lambda)^{-1}\Sigma]}{n} \,. \tag{9}$$

3 Discussion

Comments on the bound (9). For $\lambda \geqslant cR^2/n$, the bound (9) is at most

$$C \cdot \Big(\inf_{\theta \in \mathbf{R}^d} \left\{ L(\theta) + \lambda \|\theta\|^2 - L(\theta^*) \right\} + \frac{\sigma^2 \mathrm{Tr}[(\Sigma+\lambda)^{-1}\Sigma]}{n} \Big) \,,$$

where c, C are constants. In addition, the first term above is at most $\lambda \|\theta^*\|^2$ (take $\theta = \theta^*$ instead of $\inf_\theta$). A bound in finite dimension can be deduced by letting $\lambda \asymp R^2/n$ and bounding $d_\lambda := \mathrm{Tr}[(\Sigma+\lambda)^{-1}\Sigma] \leqslant d$, which yields a $O((\sigma^2 d + R^2\|\theta^*\|^2)/n)$ bound.

Both the variance and the bias term in (9) are distribution-dependent; they depend, respectively, on the spectrum of Σ (through the effective dimension $d_\lambda = \mathrm{Tr}[(\Sigma+\lambda)^{-1}\Sigma]$) and on the approximation properties of Euclidean balls of $\mathbf{R}^d$.

As shown in the lower bound of Lemma 2, the upper bound on the variance term from the decomposition of Lemma 1 is sharp up to universal constants in the regime $n \gtrsim R^2/\lambda$. We note that the variance term from Lemma 2 is itself only an upper bound on the actual variance (after bounding $(\widehat{\Sigma}_n + \lambda)^{-1}\widehat{\Sigma}_n \preccurlyeq I$), though it is generally of the correct order (an exception is the "interpolation" regime [5], where λ is very small or equal to 0 and $n \ll d$). For instance, under the "nonparametric" regime $d \gg n$ and under a polynomial decay of eigenvalues of Σ, namely if $\mathrm{Tr}(\Sigma^{1/b}) \leqslant B$ for some $b > 1$ and $B > 0$, then $d_\lambda \leqslant 2B\lambda^{-1/b}$ and this gives the optimal variance under this assumption [7].

The bound on the bias term is somewhat less accurate (in particular, there is no matching lower bound in Lemma 3), though still of correct order in some relevant regimes. Ideally, one may wish to replace $\widehat{\Sigma}_n$ by Σ (at least for n large enough) in the bias term of Lemma 1, leading to a term of

$$\lambda^2 \langle (\Sigma+\lambda)^{-1}\Sigma(\Sigma+\lambda)^{-1}\theta^*, \theta^* \rangle = L(\theta_\lambda) - L(\theta^*) \,,$$

where $\theta_\lambda = \arg\min_{\theta \in \mathbf{R}^d} \{L(\theta) + \lambda \|\theta\|^2\} = (\Sigma + \lambda)^{-1} \Sigma \theta^*$. Instead, the bound (9) gives a term $L(\theta_\lambda) - L(\theta^*) + \lambda \|\theta_\lambda\|^2$, with an additional $\lambda \|\theta_\lambda\|^2$ component. Roughly speaking, this extra term dominates $L(\theta_\lambda) - L(\theta^*)$ when θ^* is highly aligned with the leading eigenvectors of Σ. Similarly to the variance term, a simple way to assess this bound is to consider the stylized nonparametric regime, with d large or infinite, and polynomial decay (this time, of coefficients of θ^* in the basis of eigenvectors of Σ). Specifically, assume as in [7] that $\|\Sigma^{(1-r)/2} \theta^*\| \leqslant \rho$ for some $r > 0$ and $\rho > 0$; the parameter $r > 0$ controls the rate of decay of components of θ^*. One can bound the bias term as $C(\rho) \lambda^{\min(r,1)}$, while it is known (e.g., from [7]) that the actual bias of ridge can be bounded as $\widetilde{C}(\rho) \lambda^{\min(r,2)}$ under these assumptions[1]. The bound is therefore of the correct order for $0 < r \leqslant 1$, but suboptimal in the regime $r > 1$.

Additional comments and references. A possible approach to analyzing ridge regression is based on viewing it as empirical risk minimization and using tools from empirical process theory together with localization and fixed-point arguments [37, 22, 4, 16], see for instance [17, Example 2 p. 86], [23] and [38] for analyses in this spirit.

A direct approach is based on matrix concentration [29, 21, 1, 26, 35]. Results in this direction were first derived in [10] and then refined in a series of works [40, 30, 31, 7]. In particular, optimal bounds depending on the effective dimension d_λ were first derived in [7], see also [14, 33, 6]. In fact, two-sided matrix concentration is not necessary to control the error, and one-sided lower bounds on the sample covariance matrix suffice, see [27, 20, 9, 24, 43] for results of this type. In a different direction, a risk analysis of ridge regression (assuming that X is a sub-Gaussian random vector, see [18] for matrix concentration results in this case) covering more regimes of choices of $\theta^*, \Sigma, \lambda, n$ can be found in [36].

The elementary analysis of ridge regression presented here does not explicitly rely on matrix concentration (or lower tail) results. The main arguments, namely exchangeability, matrix perturbation and matrix convexity, were also used in [25], but for different estimators and for conditional density estimation rather than regression. For ridge regression, an analysis related to the one presented in this note, based on average stability arguments, is proposed in [38]. Additional relevant works include [19, 12], that use average stability to analyze empirical minimization in exp-concave statistical learning (with bounds depending on the dimension d). It is worth noting that, while exchangeability/leave-one-out/average stability arguments typically lead to simple and direct proofs, a shortcoming of these approaches is that they generally give in-expectation rather than deviation bounds.

Finally, a precise in-expectation finite-sample analysis of estimators based on stochastic gradient descent (SGD) can be found in [11] (extending results in [3]), see also [42, 28, 15] (and references therein) for more information on SGD for least squares. The arguments of [11] are direct and do not rely on matrix concentration either. This analysis is however specific to iterative methods such as stochastic gradient descent, and it is unclear whether it applies to ridge regression.

A Operator convexity and Sherman-Morrison's identity

In this appendix, we provide for the sake of completeness statements and short proofs of facts used in the proofs of Lemmas 2 and 3.

We start with (operator) convexity of the matrix inverse.

[1] Different estimators can give better bias terms, see e.g. [6] and references therein.

Lemma 5 (Lemma 2.7 in [8])**.** *Let S be a symmetric, positive semi-definite matrix. Then, the map*

$$A \mapsto \mathrm{Tr}(A^{-1}S)\,,$$

defined on the cone of positive-definite matrices, is convex.

Proof. By continuity of the map $A \mapsto \mathrm{Tr}(A^{-1}S)$ on its domain, it suffices to prove that it is midpoint-convex. Hence, it suffices to show that, for any positive-definite matrices A, B,

$$\Big(\frac{A+B}{2}\Big)^{-1} \preccurlyeq \frac{A^{-1}+B^{-1}}{2}\,.$$

Now, letting $C = A^{-1/2}BA^{-1/2}$, since

$$\Big(\frac{A+B}{2}\Big)^{-1} = A^{-1/2}\Big(\frac{I+C}{2}\Big)^{-1}A^{-1/2}$$

and

$$\frac{A^{-1}+B^{-1}}{2} = A^{-1/2}\Big(\frac{I+C^{-1}}{2}\Big)A^{-1/2}\,,$$

it suffices to show that $(I+C)^{-1}/2 \preccurlyeq (I+C^{-1})/2$. Up to conjugating with a rotation, one may assume that C is diagonal. In this case, the desired inequality follows from the convexity of the (scalar) inverse on $\mathbf{R}_+^*$, applied to each entry of the diagonal. □

We also use the following identity involving the inverse of rank-one perturbations of matrices, which follows from the Sherman-Morrison identity.

Lemma 6. *Let S be a positive-definite $d \times d$ matrix, and $v \in \mathbf{R}^d$. Then, one has*

$$S^{-1}v = \big(1 + \langle S^{-1}v, v\rangle\big)(S + vv^\top)^{-1}v\,. \tag{10}$$

Proof. We recall the Sherman-Morrison identity:

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

which can be checked by multiplying both sides by $S + vv^\top$, and deduce that

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

Acknowledgements. We would like to thank Francis Bach for encouraging us to share this note and for many helpful comments. L.R. acknowledges the financial support of the European Research Council (grant SLING 819789), the AFOSR projects FA9550-18-1-7009, FA9550-17-1-0390 and BAA-AFRL-AFOSR-2016-0007 (European Office of Aerospace Research and Development), and the EU H2020-MSCA-RISE project NoMADS - DLV-777826.

References

[1] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory*, 48(3):569–579, 2002.

[2] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68(3):337–404, 1950.

[3] F. Bach and É. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems 26*, pages 773–781, 2013.

[4] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *Ann. Statist.*, 33(4):1497–1537, 2005.

[5] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proc. Natl. Acad. Sci. USA*, 117(48):30063–30070, 2020.

[6] G. Blanchard and N. Mücke. Optimal rates for regularization of statistical inverse learning problems. *Found. Comput. Math.*, 18(4):971–1013, 2018.

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

[8] E. Carlen. Trace inequalities and quantum entropy: an introductory course. *Entropy and the quantum*, 529:73–140, 2010.

[9] O. Catoni. PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. *Preprint arXiv:1603.05229*, 2016.

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

[11] A. Dieuleveut and F. Bach. Nonparametric stochastic approximation with large step-sizes. *Ann. Statist.*, 44(4):1363–1399, 2016.

[12] A. Gonen and S. Shalev-Shwartz. Average stability is invariant to data preconditioning. Implications to exp-concave empirical risk minimization. *J. Mach. Learn. Res.*, 18(222):1–13, 2018.

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

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

[15] P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Parallelizing stochastic gradient descent for least squares regression: Mini-batching, averaging, and model misspecification. *J. Mach. Learn. Res.*, 18(223):1–42, 2018.

[16] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2593–2656, 2006.

[17] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, volume 2033 of *École d'Été de Probabilités de Saint-Flour*. Springer, 2011.

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

[19] T. Koren and K. Levy. Fast rates for exp-concave empirical risk minimization. In *Advances in Neural Information Processing Systems 28*, pages 1477–1485, 2015.

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

[21] F. Lust-Piquard and G. Pisier. Non commutative Khintchine and Paley inequalities. *Arkiv för Matematik*, 29(1):241–260, 1991.

[22] P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des sciences de Toulouse: Mathématiques*, 9(2):245–303, 2000.

[23] S. Mendelson. On the performance of kernel classes. *J. Mach. Learn. Res.*, 4:759–771, 2003.

[24] J. Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *Ann. Statist. (to appear), arXiv:1912.10754*, 2022.

[25] J. Mourtada and S. Gaïffas. An improper estimator with optimal excess risk in misspecified density estimation and logistic regression. *J. Mach. Learn. Res.*, 23(31):1–49, 2022.

[26] R. I. Oliveira. Sums of random Hermitian matrices and an inequality by Rudelson. *Electron. Commun. Probab.*, 15:203–212, 2010.

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

[28] L. Rosasco and S. Villa. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems 28*, pages 1630–1638, 2015.

[29] M. Rudelson. Random vectors in the isotropic position. *J. Funct. Anal.*, 164(1):60–72, 1999.

[30] S. Smale and D.-X. Zhou. Shannon sampling ii: Connections to learning theory. *Appl. Comput. Harmon. Anal.*, 19(3):285–302, 2005.

[31] S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, 26(2):153–172, 2007.

[32] I. Steinwart and A. Christmann. *Support Vector Machines.* Springer-Verlag New York, 2008.

[33] I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *Proc. 22nd Conference on Learning Theory*, pages 79–93, 2009.

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

[35] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012.

[36] A. Tsigler and P. L. Bartlett. Benign overfitting in ridge regression. *Preprint arXiv:2009.14286*, 2020.

[37] S. van de Geer. *Empirical Processes in M-estimation.* Cambridge University Press, Cambridge, 1999.

[38] T. Vaškevičius and N. Zhivotovskiy. Suboptimality of constrained least squares and improvements via non-linear predictors. *Bernoulli (to appear), arXiv:2009.09304*, 2022.

[39] R. Vershynin. *Introduction to the non-asymptotic analysis of random matrices*, pages 210–268. Cambridge University Press, Cambridge, 2012.

[40] E. D. Vito, L. Rosasco, A. Caponnetto, U. D. Giovannini, and F. Odone. Learning from examples as an inverse problem. *J. Mach. Learn. Res.*, 6(5):883–904, 2005.

[41] G. Wahba. *Spline Models for Observational Data*, volume 59. SIAM, 1990.

[42] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constr. Approx.*, 26(2):289–315, 2007.

[43] N. Zhivotovskiy. Dimension-free bounds for sums of independent matrices and simple tensors via the variational principle. *Preprint arXiv:2108.08198*, 2021.