High-dimensional analysis of double descent for linear regression with random projections

Francis Bach

High-dimensional analysis of double descent for linear regression with random projections

Francis Bach
Inria, Ecole Normale Supérieure
PSL Research University
francis.bach@inria.fr

March 13, 2023

Abstract

We consider linear regression problems with a varying number of random projections, where we provably exhibit a double descent curve for a fixed prediction problem, with a high-dimensional analysis based on random matrix theory. We first consider the ridge regression estimator and review earlier results using classical notions from non-parametric statistics, namely degrees of freedom, also known as effective dimensionality. We then compute asymptotic equivalents of the generalization performance (in terms of squared bias and variance) of the minimum norm least-squares fit with random projections, providing simple expressions for the double descent phenomenon.

1 Introduction

Over-parameterized models estimated with some form of gradient descent come in various forms, such as linear regression with potentially non-linear features, neural networks, or kernel methods. The double descent phenomenon can be seen empirically in several of these models [6, 15]: Given a fixed prediction problem, when the number of parameters of the model is increasing from zero to the number of observations, the generalization performance traditionally goes down and then up, due to overfitting. Once the number of parameters exceeds the number of observations, the generalization error decreases again, as illustrated in Figure 1.

The phenomenon has been theoretically analyzed in several settings, such as random features based on neural networks [27], random Fourier features [24], or linear regression [7, 17]. While the analysis of [27, 24] for random features corresponds to a single prediction problem with a sequence of increasingly larger prediction models, most of the analysis of [17] for linear regression does not consider a single problem, but varying problems, which does not actually lead to a double descent curve. Random subsampling on a single prediction problem was analyzed with a simpler model with isotropic covariance matrices in [7] and [17, Section 5.2], but without a proper double descent as the model is too simple to account for a U-shaped curve in the under-parameterized regime. In work related to ours, principal component regression was analyzed by [37] with a double descent curve but with less general assumptions regarding the spectrum of the covariance matrix and the optimal predictor.

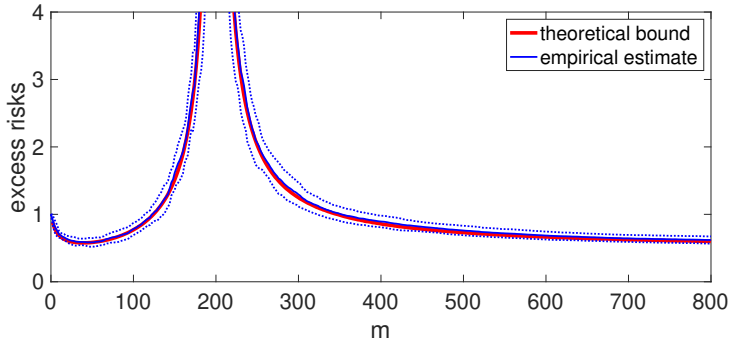

Figure 1: Example of a double descent curve, for linear regression with random projections with $n = 200$ observations, in dimension $d = 400$ and a non-isotropic covariance matrix. The data are normalized so that predicting zero leads to an excess risk of 1 and the noise so that the optimal expected risk is $1/4$. The empirical estimate is obtained by sampling 20 datasets and 20 different random projections from the same distribution and averaging the corresponding excess risks. We plot the empirical performance together with our asymptotic equivalents from Section 6.

In this paper, we consider linear regression problems and consider *random projections*, whose number increases, where we provably exhibit a double descent curve for a fixed prediction problem. Our analysis follows the high-dimensional analysis of [17, 14, 31, 23, 5] based on random matrix theory [2], and we give asymptotic expressions for the (squared) bias and the variance terms of the excess risk. These expressions and the trade-offs they lead to will be the same as what can be obtained with ridge regression [19], where a squared Euclidean penalty is added to the empirical risk.

The paper is organized as follows.

- We first present the asymptotic set-up we will follow in Section 2, and review in Section 3 the results from random matrix theory that we will need for our main result on random projections.
- We consider in Section 4 the ridge regression estimator and re-interpret the results of [14, 31, 11, 36, 5] using classical notions from non-parametric statistics, namely the degrees of freedom, a.k.a. effective dimensionality [38, 8]. When going from a fixed design analysis (where inputs are assumed deterministic) to a random design analysis (where inputs are random), the prediction performance in terms of bias and variance has the same expression, but with a larger regularization parameter, which corresponds to an additional regularization, which, following [28], we will refer to as "self-induced".
- With our new interpretation, we consider in Section 5 the minimum norm least-squares estimate and analyze its performance (which corresponds to $\lambda = 0$ above for ridge regression), thus recovering the results of [17, 4]. This corresponds to the end of the double descent curve.
- In Section 6, we compute asymptotic equivalents of the generalization performance (in terms of bias and variance) of the minimum norm least-squares fit with random projections, providing simple expressions for the double descent phenomenon. If n is the number of observations and m is the number of random projections, the variance term goes up and explodes at $m = n$ and then goes down. In contrast, the bias term may exhibit a U-shaped curve on its own in the under-parameterized regime

$(m < n)$, blows up at $m = n$, and then goes down. Our result relies on using a high-dimensional analysis both on the data and on the random projections.

2 High-dimensional analysis of linear regression

We consider the traditional random design linear regression model, where $x_1, \dots, x_n \in \mathbb{R}^d$ are sampled independently and with identical distributions (i.i.d.) with covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$, and $y_i = x_i^\top \theta_* + \varepsilon_i$, with ε_i and x_i independent, and $\mathbb{E}[\varepsilon_i] = 0$ and $\mathrm{var}(\varepsilon_i) = \sigma^2$ for some $\theta_* \in \mathbb{R}^d$.

We denote $y \in \mathbb{R}^n$ the response vector, $X \in \mathbb{R}^{n \times d}$ the design matrix, and $\varepsilon \in \mathbb{R}^n$ the noise vector. We denote by $\widehat{\Sigma} = \frac{1}{n} X^\top X \in \mathbb{R}^{d \times d}$ the non-centered empirical covariance matrix, while $XX^\top \in \mathbb{R}^{n \times n}$ is the kernel matrix.

The excess risk for an estimator $\hat{\theta}$ is $\mathcal{R}(\hat{\theta}) = (\hat{\theta} - \theta_*)\Sigma(\hat{\theta} - \theta_*)$, and we will always consider expectations with respect to ε, thus conditioned on X and on the potential additional random projections. The expectation of the excess risk will be composed of two terms: a (squared) "bias" term $\mathcal{R}^{(\mathrm{bias})}(\hat{\theta})$ corresponding to $\sigma = 0$ (and thus independent of ε), and a "variance" term $\mathbb{E}_\varepsilon\big[\mathcal{R}^{(\mathrm{var})}(\hat{\theta})\big]$ corresponding to $\theta_* = 0$ (and after taking the expectation with respect to ε). All of our asymptotic results will then be almost surely in all other random quantities (e.g., X and the random projections S later).

We make similar high-dimensional assumptions as [14, 31], that is:

(A1) $X = Z\Sigma^{1/2}$ with $Z \in \mathbb{R}^{n \times d}$ with sub-Gaussian i.i.d. components with mean zero and unit variance.

(A2) The sample size n and the dimension d go to infinity, with $\frac{d}{n}$ tending to $\gamma > 0$.

(A3) The spectral measure $\frac{1}{d}\sum_{i=1}^d \delta_{\sigma_i}$ of Σ converges to a probability distribution μ on $\mathbb{R}_+$, where $\sigma_1, \dots, \sigma_d$ are the eigenvalues of Σ. Moreover, μ has compact support in $\mathbb{R}_+^*$, and Σ is invertible and bounded in operator norm.

(A4) The measure $\sum_{i=1}^d (v_i^\top \theta_*)^2 \delta_{\sigma_i}$ converges to a measure ν with bounded mass, where v_i is the unit-norm eigenvector of Σ associated to σ_i. The norm of θ_* is bounded.

Assumption **(A1)** does *not* assume Gaussian data but includes Z with standard Gaussian components or Rademacher random variables (uniform in $\{-1, 1\}$).

Assumption **(A2)** states that the ratio of dimensions tends to a constant, but could be relaxed by a uniform boundedness assumption [32]. See [10] for an analysis that goes beyond this assumption of n and d being of the same order.

Assumption **(A3)** implies that for any bounded function $r : \mathbb{R}_+ \to \mathbb{R}$, $\frac{1}{d}\mathrm{tr}[r(\Sigma)] \to \int_0^{+\infty} r(\sigma) d\mu(\sigma)$. Note that in **(A3)**, we assume that the support of the limiting μ is bounded away from zero (e.g., no vanishing eigenvalues).

Assumption **(A4)** is equivalent to: for any bounded function $r : \mathbb{R}_+ \to \mathbb{R}$, $\theta_*^\top r(\Sigma)\theta_* \to \int_0^{+\infty} r(\sigma) d\nu(\sigma)$. Moreover, it is often replaced by θ_* being random with mean zero and covariance matrix proportional to identity [14], or a spectral variant of Σ [31]. This corresponds to having ν having a density with respect to μ.

3 Random matrix theory tools

We consider the kernel matrix $XX^\top = Z\Sigma Z^\top \in \mathbb{R}^{n\times n}$ with all components of $Z \in \mathbb{R}^{n\times d}$ being i.i.d. sub-Gaussian with zero mean and unit variance, that is, following Assumption **(A1)**. We also assume **(A2)** and **(A3)** throughout this section. We denote by $\widehat{\Sigma} = \frac{1}{n}X^\top X \in \mathbb{R}^{d\times d}$ the empirical covariance matrix.

We now present the tools from random matrix theory that we will need. Most of them have already been used in the same context [14, 17, 31, 23], but more refined ones will be needed along the lines of [12, 23] (Section 3.3) and we will give explicit interpretations in terms of degrees of freedom (Section 3.1) and self-induced regularization (Section 3.2).

3.1 Summary and re-interpretation of existing results

We will need to relate the spectral properties of the empirical covariance matrix $\widehat{\Sigma}$ to the ones of the population covariance matrix Σ. This typically includes the distribution of eigenvalues, but in this paper, we will only need spectral functions of the form $\mathrm{tr}[r(\widehat{\Sigma})]$, or more general quantities, such as $\mathrm{tr}[Ar(\widehat{\Sigma})]$, $\mathrm{tr}[Ar(\widehat{\Sigma})Br(\widehat{\Sigma})]$, for matrices $A, B \in \mathbb{R}^{d\times d}$.

We summarize the relevant results from random matrix theory through the asymptotic equivalence,[1] for any $\lambda > 0$,

$$\mathrm{tr}\left[\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-1}\right] \sim \mathrm{tr}\left[\Sigma\big(\Sigma+\kappa(\lambda)I\big)^{-1}\right], \tag{1}$$

where $\kappa : \mathbb{R}_+ \to \mathbb{R}_+$ is an increasing function. Within the analysis of ridge regression, these are often referred to as the "degrees of freedom" [8, 18], and denoted[2]

$$\widehat{\mathrm{df}}_1(\lambda) = \mathrm{tr}\left[\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-1}\right] \quad \text{and} \quad \mathrm{df}_1(\kappa) = \mathrm{tr}\left[\Sigma(\Sigma+\kappa I)^{-1}\right].$$

In the limit when d tends to infinity, by definition of μ in Assumption **(A3)**, then $\frac{1}{d}\mathrm{df}_1(\kappa) \to \int_0^{+\infty} \frac{\sigma d\mu(\sigma)}{\sigma+\kappa}$, which is strictly decreasing in κ, with a value of 1 at $\kappa = 0$. Since $\mathrm{tr}\left[\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-1}\right] \leqslant d$, this asymptotically defines uniquely $\kappa(\lambda)$.

The extra knowledge from random matrix theory will be the self-consistency equation

$$\kappa(\lambda) - \lambda = \kappa(\lambda)\cdot\gamma\int_0^{+\infty}\frac{\sigma d\mu(\sigma)}{\sigma+\kappa},$$

that allows to define $\kappa(\lambda)$, which we will write equivalently

$$\kappa(\lambda) - \lambda \sim \kappa(\lambda)\cdot\frac{1}{n}\mathrm{df}_1(\kappa(\lambda)).$$

As shown below, for λ large, then $\kappa(\lambda) \sim \lambda$. When λ tends to zero (which will be the case in classical scenarios where we regularize less as we observe more data), $\kappa(\lambda)$ will tend to zero only for under-parameterized models ($\gamma < 1$), while for over-parameterized model ($\gamma > 1$), it will tend to a constant.

[1] In this paper, we use the asymptotic equivalent notation $u \sim v$, to mean that the ratio u/v tends to one when the dimensions n, d go to infinity. This allows to provide results for diverging quantities which are more easily interpretable, such as degrees of freedom.

[2] We use the notation df_1 as we will introduce a related notion df_2 later.

In statistical terms, the degrees of freedom for the empirical covariance matrix correspond to the degrees of freedom of the population covariance matrix with a larger regularization parameter, leading to an additional regularization.

Beyond Eq. (1), we will need asymptotic equivalents for the quantities $\operatorname{tr}\left[A\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-1}\right]$ and $\operatorname{tr}\left[A\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-1}B\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-1}\right]$ for matrices $A, B \in \mathbb{R}^{d\times d}$. They will be valid when certain quantities for the matrices A and B converge (see Prop. 1 and Prop. 2 below).

These results recover existing work with $A, B = I$ or Σ [22, 14], and lead to the same formulas as [12, 23] obtained with similar assumptions. They are needed for the ridge regression results in Section 4 and for the random projection results in Section 6, where they will, for example, be used with $A = \theta_*\theta_*^\top$.

3.2 Self-induced regularization

We consider the Stieltjes transform of the spectral measure of the kernel matrix $XX^\top \in \mathbb{R}^{n\times n}$, with $z \in \mathbb{C}\backslash\mathbb{R}_+$:

$$\widehat{\varphi}(z) = \frac{1}{n}\operatorname{tr}\left[\left(\frac{1}{n}XX^\top - zI\right)^{-1}\right] = \operatorname{tr}\left[(XX^\top - nzI)^{-1}\right].$$

This transform is known to fully characterize the spectral distribution of $XX^\top$ (see, e.g., [2] and references therein). Then for all $z \in \mathbb{C}\backslash\mathbb{R}_+$, assuming **(A1)**, **(A2)**, and **(A3)**, $\widehat{\varphi}(z)$ is known to converge almost surely, and its limit $\varphi(z)$ satisfies the following equation (see Appendix A.1 for a simple argument leading to it) [2, 22]:

$$\frac{1}{\varphi(z)} + z = \gamma\int_0^{+\infty}\frac{\sigma d\mu(\sigma)}{1+\sigma\varphi(z)}. \tag{2}$$

When $\Sigma = \sigma I$, this allows to compute $\varphi(z)$ and, by inversion of the Stieltjes transform, to recover the Marchenko-Pastur distribution. In this paper, we will not need to know the limiting density (which is anyway uneasy to describe for general Σ) and only access it through its Stieltjes transform.

Indeed, for $z = -\lambda$ for $\lambda > 0$, we get $\widehat{\varphi}(-\lambda) = \operatorname{tr}\left[(XX^\top + n\lambda I)^{-1}\right] \to \varphi(-\lambda)$ almost surely, with

$$\frac{1}{\varphi(-\lambda)} - \lambda = \gamma\int_0^{+\infty}\frac{\sigma d\mu(\sigma)}{1+\sigma\varphi(-\lambda)}. \tag{3}$$

In the ridge regression context, as mentioned above, the quantity $\operatorname{df}_1(\kappa) = \operatorname{tr}[\Sigma(\Sigma+\kappa I)^{-1}] \in [0, d]$ is referred to as the "degrees of freedom". It is a strictly decreasing function of κ, with $\operatorname{df}_1(0) = \operatorname{rank}(\Sigma)$. It is asymptotically equivalent to $\sum_{i=1}^d \frac{\sigma_i}{\sigma_i+\kappa} \sim d\int_0^{+\infty}\frac{\sigma d\mu(\sigma)}{\sigma+\kappa}$. Thus, we can rewrite Eq. (3) as

$$\frac{1}{\varphi(-\lambda)} - \lambda \sim \frac{1}{\varphi(-\lambda)}\cdot\frac{1}{n}\operatorname{df}_1\Big(\frac{1}{\varphi(-\lambda)}\Big).$$

Therefore, we can define our equivalent regularization parameter $\kappa(\lambda) = \frac{1}{\varphi(-\lambda)} \in \mathbb{R}_+$ which is the almost sure limit of $1/\operatorname{tr}\left[(XX^\top + n\lambda I)^{-1}\right]$, and such that

$$\kappa(\lambda) - \lambda \sim \kappa(\lambda)\cdot\frac{1}{n}\operatorname{df}_1(\kappa(\lambda)) \quad\Leftrightarrow\quad \lambda \sim \kappa(\lambda)\Big(1 - \frac{1}{n}\operatorname{df}_1(\kappa(\lambda))\Big). \tag{4}$$

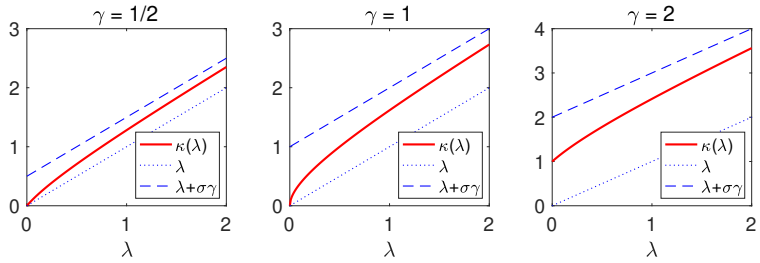

Figure 2: Implicit regularization parameter $\kappa(\lambda)$ in the three regimes for isotropic covariance matrices, with $\sigma = 1$. See text for details.

Depending on the relationship between d and n (that is, $d < n$ or $d > n$), we have different behaviors for the function κ (see below), but $\kappa(\lambda)$ is always larger than λ. This additional regularization has been explored in a number of works [23, 21, 10], and we refer to it as self-induced.

Note that in order to compute $\kappa(\lambda)$, we can either solve Eq. (4) if we can compute $\mathrm{df}_1(\kappa(\lambda))$, or simply use that $\kappa(\lambda)^{-1}$ is the almost sure limit of $\mathrm{tr}\left[(XX^\top + n\lambda I)^{-1}\right]$, when n, d go to infinity. We now provide properties of the function κ.

Isotropic covariance matrices We consider the case $\Sigma = \sigma I$ to first study the dependence between $\kappa(\lambda)$ and λ. By the use of Jensen's inequality, this will lead to bounds in the general case. In this isotropic situation, we have $\frac{1}{n}\mathrm{df}_1(\kappa) = \frac{\gamma\sigma}{\sigma+\kappa}$, and Eq. (4) is equivalent to $\lambda = \kappa(\lambda)\big(1 - \frac{\gamma\sigma}{\sigma+\kappa}\big)$. We can solve it in closed form as:

$$\kappa(\lambda) = \frac{1}{2}\Big(\lambda - \sigma(1-\gamma) + \sqrt{(\sigma(1-\gamma)-\lambda)^2 + 4\lambda\sigma}\Big). \tag{5}$$

We then have three cases, as illustrated in Figure 2. The function κ is always increasing with the same asymptote $\lambda + \sigma\gamma$ at infinity, but different behaviors at 0 (see a more thorough discussion in [23, Section 5.4.1]):

- $\gamma < 1$: $\kappa(0) = 0$ with $\kappa'(0) = 1/(1-\gamma)$.
- $\gamma > 1$: $\kappa(0) = (\gamma - 1)\sigma > 0$.
- $\gamma = 1$: $\kappa(0) = 0$ with $\kappa'(0) = +\infty$, and $\kappa \sim \sqrt{\lambda}$ around 0.

General case Beyond isotropic covariance matrices, we have a similar behavior in the general case, in particular, by Jensen's inequality, the expression in Eq. (5) is an upper-bound with σ replaced by $\frac{1}{d}\mathrm{tr}(\Sigma)$.

- *Under-parameterized* ($\gamma < 1 \Leftrightarrow d < n$): we then have $\mathrm{df}_1(\kappa(\lambda)) \leqslant d < n$, and the function $\lambda \mapsto \kappa(\lambda)$ is strictly increasing with $\kappa(0) = 0$ and $\kappa(\lambda) \in [\lambda, \lambda/(1-d/n)]$, with an equivalent $\kappa(\lambda) \sim \lambda + \frac{1}{n}\mathrm{tr}\,\Sigma$ when λ tends to infinity, and the equivalent $\kappa(\lambda) \sim \lambda/(1-d/n)$ when λ tends to zero (since we have assumed that $\mathrm{rank}(\Sigma) = d$).
- *Over-parameterized* ($\gamma > 1 \Leftrightarrow d > n$): we then have $\kappa(0) > 0$, which is defined by $\mathrm{df}_1(\kappa(0)) = n$. The function $\lambda \mapsto \kappa(\lambda)$ is still strictly increasing, with an equivalent $\kappa(\lambda) \sim \lambda + \frac{1}{n}\mathrm{tr}\,\Sigma$ when λ tends

to infinity. By Jensen's inequality, we have $\mathrm{df}_1(\kappa(\lambda)) \leqslant \frac{\operatorname{tr}\Sigma}{\kappa(\lambda)+\operatorname{tr}\Sigma/d} \leqslant \frac{\operatorname{tr}\Sigma}{\kappa(\lambda)}$. This in turn implies that $\kappa(\lambda) \in [\lambda, \lambda + \frac{\operatorname{tr}\Sigma}{n}]$, and also a finer bound based on Eq. (5) with σ replaced by $\frac{1}{d}\operatorname{tr}(\Sigma)$. Moreover, we have the bound $\kappa(0) \leqslant \frac{\operatorname{tr}\Sigma}{n}(1-n/d) = \frac{\operatorname{tr}\Sigma}{d}(\gamma-1)$.

"Classical" statistical asymptotic behaviors Within positive-definite kernel methods [8], it is common to have infinite-dimensional covariance operators, with a sequence of eigenvalues of the form $\lambda_k = \frac{\tau}{k^\alpha}$ with $\alpha > 1$ and $k \geqslant 1$. To make it correspond to the high-dimensional framework with $k \in \{1, \dots, d\}$ with d tending to infinity, we need to rescale the eigenvalues by d^α, so that the spectral measure is $\widehat{\mu} = \frac{1}{d}\sum_{k=1}^d \delta_{\tau(d/k)^\alpha}$, which converges to the distribution of τ/u^α for u uniform on $[0,1]$. The support of this distribution is bounded from below, but not from above, and thus does not satisfy our assumptions (but in our simulations, our asymptotic equivalents match the empirical behavior). See [10, Section 4.2] for an analysis that covers explicitly this spectral behavior.

In terms of degrees of freedom, we then have, using the same rescaling by d^α, and with the change of variable $v = ud(\kappa/\tau)^{1/\alpha}$:

$$\mathrm{df}_2(\kappa d^\alpha) \sim d\int_0^1 \frac{\tau u^{-\alpha} du}{\tau u^{-\alpha} + d^\alpha \kappa} = d\int_0^1 \frac{du}{1+(ud)^\alpha \kappa \tau^{-1}} \sim (\tau/\kappa)^{1/\alpha}\int_0^{+\infty} \frac{dv}{1+v^\alpha}.$$

We get the usual explosion of degrees of freedom in $\kappa^{-1/\alpha}$ [8]. It can then be shown, if our formulas apply, that $\kappa(0) \propto \frac{1}{n^\alpha}$. See [11] for a detailed analysis of the consequences of the ridge regression asymptotic equivalents when such assumptions are made.

3.3 Asymptotic equivalents for spectral functions

Following [22, 14], we can provide asymptotic equivalents for quantities depending on the spectrum of $\widehat{\Sigma}$. We prove in Appendix A.2 the following result, with two asymptotic equivalents matching the earlier work of [12, Lemma 10] that was obtained for the special case of Gaussian distributions.

Proposition 1 *Assume **(A1)**, **(A2)**, **(A3)**, that A and B are bounded in operator norm, and that the measures $\sum_{i=1}^d v_i^\top A v_i \cdot \delta_{\sigma_i}$ and $\sum_{i=1}^d v_i^\top B v_i \cdot \delta_{\sigma_i}$ converge to measures ν_A and ν_B with bounded total variation. Then, for $z \in \mathbb{C}\backslash\mathbb{R}_+$, with $\varphi(z)$ satisfying Eq. (2),*

$$\operatorname{tr}\left[A\widehat{\Sigma}(\widehat{\Sigma}-zI)^{-1}\right] \sim \operatorname{tr}\left[A\Sigma\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-1}\right] \tag{6}$$

$$\begin{aligned}\operatorname{tr}\left[A\widehat{\Sigma}(\widehat{\Sigma}-zI)^{-1}B\widehat{\Sigma}(\widehat{\Sigma}-zI)^{-1}\right] &\sim \operatorname{tr}\left[A\Sigma\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-1}B\Sigma\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-1}\right] \\ &\quad + \tfrac{1}{\varphi(z)^2}\operatorname{tr}\left[A\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-2}\Sigma\right]\cdot\operatorname{tr}\left[B\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-2}\Sigma\right]\cdot\tfrac{1}{n-\mathrm{df}_2(1/\varphi(z))}.\end{aligned} \tag{7}$$

Eq. (6) can formally be seen as the limit $\frac{1}{d}\operatorname{tr}\left[A\widehat{\Sigma}(\widehat{\Sigma}-zI)^{-1}\right] \to \int_0^{+\infty}\frac{\sigma d\nu_A(\sigma)}{\sigma+1/\varphi(z)}$, and a similar result holds for Eq. (7). From Eq. (6) and Eq. (7), as shown in Appendix A.2, we can also derive results for slightly modified traces, with $\widehat{\Sigma}(\widehat{\Sigma}-zI)^{-1}$ replaced by $(\widehat{\Sigma}-zI)^{-1}$, as:

$$\operatorname{tr}\left[A(\widehat{\Sigma}-zI)^{-1}\right] \sim \tfrac{-1}{z\varphi(z)}\operatorname{tr}\left[A\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-1}\right] \tag{8}$$

$$\begin{aligned}\operatorname{tr}\left[A(\widehat{\Sigma}-zI)^{-1}B\big(\widehat{\Sigma}-zI\big)^{-1}\right] &\sim \tfrac{1}{z^2\varphi(z)^2}\operatorname{tr}\left[A\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-1}B\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-1}\right] \\ &\quad + \tfrac{1}{z^2\varphi(z)^2}\operatorname{tr}\left[A\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-2}\Sigma\right]\cdot\operatorname{tr}\left[B\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-2}\Sigma\right]\cdot\tfrac{1}{n-\mathrm{df}_2(1/\varphi(z))}.\end{aligned} \tag{9}$$

Expectation of kernel matrices Through the matrix inversion lemma, we have $\widehat{\Sigma}(\widehat{\Sigma} - zI)^{-1} = X^\top X(X^\top X - nzI)^{-1} = X^\top(XX^\top - nzI)X$, and thus we obtain another set of asymptotic results, where we can replace $\Sigma^{1/2}A\Sigma^{1/2}$ by A, matching the earlier results of [23, Theorem 4.6].

Proposition 2 *Assume **(A1)**, **(A2)**, **(A3)**, that A and B are bounded in operator norm, and that the measures $\sum_{i=1}^d v_i^\top A v_i \cdot \delta_{\sigma_i}$ and $\sum_{i=1}^d v_i^\top B v_i \cdot \delta_{\sigma_i}$ converge to measures ν_A and ν_B with bounded total variation. Then, for $z \in \mathbb{C}\backslash\mathbb{R}_+$, with $\varphi(z)$ satisfying Eq. (2),*

$$\operatorname{tr}\big[AZ^\top(Z\Sigma Z^\top - nzI)^{-1}Z\big] \quad\sim\quad \operatorname{tr}\big[A\big(\Sigma + \tfrac{1}{\varphi(z)}I\big)^{-1}\big] \tag{10}$$

$$\begin{aligned}\operatorname{tr}\big[AZ^\top(Z\Sigma Z^\top - nzI)^{-1}ZBZ^\top(Z\Sigma Z^\top - nzI)^{-1}Z\big] \quad&\sim\quad \operatorname{tr}\big[A\big(\Sigma + \tfrac{1}{\varphi(z)}I\big)^{-1}B\big(\Sigma + \tfrac{1}{\varphi(z)}I\big)^{-1}\big] \\ &\quad + \tfrac{1}{\varphi(z)^2}\operatorname{tr}\big[A\big(\Sigma + \tfrac{1}{\varphi(z)}I\big)^{-2}\big] \cdot \operatorname{tr}\big[B\big(\Sigma + \tfrac{1}{\varphi(z)}I\big)^{-2}\big] \cdot \tfrac{1}{n - \mathrm{df}_2(1/\varphi(z))}.\end{aligned} \tag{11}$$

Like in [23, Theorem 4.6], Eq. (11) can be rewritten more intuitively as

$$Z^\top(Z\Sigma Z^\top - nzI)^{-1}ZBZ^\top(Z\Sigma Z^\top - nzI)^{-1}Z \sim \big(\Sigma + \tfrac{1}{\varphi(z)}I\big)^{-1}\big(B + \mu(z)I\big)\big(\Sigma + \tfrac{1}{\varphi(z)}I\big)^{-1},$$

with $\mu(z) = \frac{1}{\varphi(z)^2}\frac{\operatorname{tr}[B(\Sigma + \frac{1}{\varphi(z)}I)^{-2}]}{n - \mathrm{df}_2(1/\varphi(z))}$.

Letting $\lambda \to 0$ for $\gamma > 1$ Following arguments from [14, Lemma 6.2], in the high-dimensional situation where $\gamma > 1$, we can take the limit $\lambda = 0$, with the implicit regularization parameter $\kappa(0) > 0$ defined in Section 3.1, which is such that $\mathrm{df}_1(\kappa(0)) = n$. This works for the kernel version since we can write

$$\widehat{\Sigma}(\widehat{\Sigma} - zI)^{-1} = X^\top X(X^\top X - nzI)^{-1} = X^\top(XX^\top - nzI)^{-1}X = \Sigma^{1/2}Z^\top(Z\Sigma Z^\top - nzI)^{-1}\Sigma^{1/2},$$

which makes sense even with $z = 0$, as the kernel matrix $XX^\top$ is then asymptotically almost surely invertible (since Σ is invertible, and $ZZ^\top$ almost surely is [3]). This will be used in the over-parameterized regime in Section 5 and for random projections in Section 6.

Letting $\lambda \to 0$ for $\gamma < 1$ In this situation, $\kappa(\lambda)$ tends to zero, and we can use Eq. (8) and Eq. (9) instead, that is, $\operatorname{tr}\big[A\big(\widehat{\Sigma} + \lambda I\big)^{-1}\big] \sim \frac{\kappa(\lambda)}{\lambda}\operatorname{tr}\big[A\big(\Sigma + \kappa(\lambda)I\big)^{-1}\big]$, with $\frac{\kappa(\lambda)}{\lambda} \sim \frac{1}{1-\gamma}$ when λ goes to zero, and $\kappa(0) = 0$, leading to

$$\operatorname{tr}\big[A\widehat{\Sigma}^{-1}\big] \quad\sim\quad \frac{1}{1 - d/n}\operatorname{tr}[A\Sigma^{-1}]. \tag{12}$$

Equipped with the proper random matrix theory tools, we can apply them to least-squares regression, starting with ridge regression in Section 4, its limit when $\lambda \to 0$ in Section 5, and then with random projections in Section 6.

4 Analysis of ridge regression

We consider the ridge regression estimator, obtained as the unique minimizer of $\frac{1}{n}\sum_{i=1}^n (y_i - x_i^\top\theta)^2 + \lambda\|\theta\|_2^2$, which is equal to:

$$\hat{\theta} = (X^\top X + n\lambda I)^{-1}X^\top y = X^\top(XX^\top + n\lambda I)^{-1}y.$$

In the fixed design framework, its analysis is explicit and leads to usual bias/variance trade-offs based on simple quantities.

4.1 Fixed design analysis of ridge regression

In the fixed design set-up where inputs $x_1, \dots, x_n$ are assumed deterministic, we obtain an expected excess risk, with Σ replaced with $\widehat{\Sigma}$, which considerably simplifies the analysis (see, e.g., [20]):

$$\mathbb{E}_\varepsilon\big[(\hat\theta - \theta_*)^\top \widehat{\Sigma}(\hat\theta - \theta_*)\big] = \lambda^2 \theta_*^\top (\widehat{\Sigma} + \lambda I)^{-2} \widehat{\Sigma} \theta_* + \frac{\sigma^2}{n} \operatorname{tr}\big[\widehat{\Sigma}^2 (\widehat{\Sigma} + \lambda I)^{-2}\big].$$

The (squared) bias term $\lambda^2 \theta_*^\top (\widehat{\Sigma} + \lambda I)^{-2} \widehat{\Sigma} \theta_*$ is increasing in λ, and depends on how the true θ_* aligns with eigenvectors of $\widehat{\Sigma}$, and "source conditions" are typically used to characterized this alignment [8].

This leads us to introduce the two classical different notions degrees of freedom $\mathrm{df}_1(\lambda) = \operatorname{tr}\big[\Sigma(\Sigma + \lambda I)^{-1}\big]$ and $\mathrm{df}_2(\lambda) = \operatorname{tr}\big[\Sigma^2(\Sigma + \lambda I)^{-2}\big]$ as key quantities [20]. Typically, they behave similarly when λ tends to zero (in particular, they are both equal to the rank of Σ for $\lambda = 0$). We will see in Section 5 that when they differ significantly, this has consequences regarding the relevance of the end of the double descent curve.

Our goal is to obtain similar results to those for fixed design, using degrees of freedom and (squared) bias of the form $\lambda^2 \theta_*^\top (\Sigma + \lambda I)^{-2} \Sigma \theta_*$. While bounds can be obtained in expectations [29] or high probability [8], we aim here at getting asymptotic equivalents.

4.2 Random design analysis of ridge regression

In this section, we recover the results from [14, 29, 5] with an explicit interpretation in terms of degrees of freedom.

We have, separating the noise from the part coming from θ_*:

$$\begin{aligned} \hat\theta &= (X^\top X + n\lambda I)^{-1} X^\top y = (X^\top X + n\lambda I)^{-1} X^\top X \theta_* + (X^\top X + n\lambda I)^{-1} X^\top \varepsilon. \\ &= (\widehat{\Sigma} + \lambda I)^{-1} \widehat{\Sigma} \theta_* + (\widehat{\Sigma} + \lambda I)^{-1} \frac{X^\top \varepsilon}{n}. \end{aligned} \tag{13}$$

This leads to the following proposition, with the same expressions as [5, Theorem 4.13] (see also [10] for the same expressions in a more general context):

Proposition 3 *Assume **(A1)**, **(A2)**, **(A3)**, and **(A4)**. For the ridge regression estimator in Eq. (13), we have:*

$$\begin{aligned} \mathbb{E}_\varepsilon\big[\mathcal{R}^{(\mathrm{var})}(\hat\theta)\big] &\sim \frac{\sigma^2}{n} \mathrm{df}_2(\kappa(\lambda)) \cdot \frac{1}{1 - \frac{1}{n}\mathrm{df}_2(\kappa(\lambda))} \\ \mathcal{R}^{(\mathrm{bias})}(\hat\theta) &\sim \kappa(\lambda)^2 \theta_*^\top \Sigma (\Sigma + \kappa(\lambda) I)^{-2} \theta_* \cdot \frac{1}{1 - \frac{1}{n}\mathrm{df}_2(\kappa(\lambda))}, \end{aligned}$$

with $\kappa(\lambda)$ related to λ by $\kappa(\lambda)\big(1 - \frac{1}{n}\mathrm{df}_1(\kappa(\lambda))\big) \sim \lambda$.

Proof The variance term is exactly the same as the one from [14], and we simply provide here a reinterpretation with degrees of freedom. We obtain it by taking expectations starting from Eq. (13) to get $\mathbb{E}_\varepsilon\big[\mathcal{R}^{(\mathrm{var})}(\hat\theta)\big] = \frac{\sigma^2}{n}\operatorname{tr}\big[\Sigma(\widehat\Sigma+\lambda I)^{-2}\widehat\Sigma\big]$. We can then use Eq. (8) and Eq. (9) with $A=I$, $B=\Sigma$, and $z=-\lambda$, to get, using $\kappa(\lambda)\operatorname{tr}\big[\Sigma(\Sigma+\kappa(\lambda)I)^{-2}\big] = \mathrm{df}_1(\kappa(\lambda)) - \mathrm{df}_2(\kappa(\lambda))$:

$$
\begin{aligned}
\mathbb{E}_\varepsilon\big[\mathcal{R}^{(\mathrm{var})}(\hat\theta)\big] &= \tfrac{\sigma^2}{n}\operatorname{tr}\big[\Sigma(\widehat\Sigma+\lambda I)^{-2}\widehat\Sigma\big] = \tfrac{\sigma^2}{n}\operatorname{tr}\big[\Sigma(\widehat\Sigma+\lambda I)^{-1}\big] - \lambda\tfrac{\sigma^2}{n}\operatorname{tr}\big[\Sigma(\widehat\Sigma+\lambda I)^{-2}\big] \\
&\sim \tfrac{\sigma^2}{n}\tfrac{\kappa(\lambda)}{\lambda}\operatorname{tr}\big[\Sigma(\Sigma+\kappa(\lambda)I)^{-1}\big] - \tfrac{\sigma^2}{n}\tfrac{\kappa(\lambda)^2}{\lambda}\operatorname{tr}\big[\Sigma\big(\Sigma+\kappa(\lambda)I\big)^{-2}\big] \\
&\quad -\tfrac{\sigma^2}{n}\tfrac{\kappa(\lambda)^2}{\lambda}\operatorname{tr}\big[\Sigma^2\big(\Sigma+\kappa(\lambda)I\big)^{-2}\big]\cdot\operatorname{tr}\big[\Sigma\big(\Sigma+\kappa(\lambda)I\big)^{-2}\big]\cdot\tfrac{1}{n-\mathrm{df}_2(\kappa(\lambda))} \\
&= \tfrac{\sigma^2}{n}\tfrac{\kappa(\lambda)}{\lambda}\mathrm{df}_2(\kappa(\lambda)) - \tfrac{\sigma^2}{n}\tfrac{\kappa(\lambda)^2}{\lambda}\operatorname{tr}\big[\Sigma\big(\Sigma+\kappa(\lambda)I\big)^{-2}\big]\cdot\tfrac{\mathrm{df}_2(\kappa(\lambda))}{n-\mathrm{df}_2(\kappa(\lambda))} \\
&= \tfrac{\sigma^2}{n}\tfrac{\kappa(\lambda)}{\lambda}\mathrm{df}_2(\kappa(\lambda)) - \tfrac{\sigma^2}{n}\tfrac{\kappa(\lambda)}{\lambda}\big(\mathrm{df}_1(\kappa(\lambda)) - \mathrm{df}_2(\kappa(\lambda))\big)\cdot\tfrac{\mathrm{df}_2(\kappa(\lambda))}{n-\mathrm{df}_2(\kappa(\lambda))} \\
&= \tfrac{\sigma^2}{n}\tfrac{\kappa(\lambda)}{\lambda}\tfrac{\mathrm{df}_2(\kappa(\lambda))\big(n-\mathrm{df}_1(\kappa(\lambda))\big)}{n-\mathrm{df}_2(\kappa(\lambda))} = \sigma^2\tfrac{\mathrm{df}_2(\kappa(\lambda))}{n-\mathrm{df}_2(\kappa(\lambda))}.
\end{aligned}
$$

For the bias term, we have:

$$
\mathcal{R}^{(\mathrm{bias})}(\hat\theta) = \big\|\Sigma^{1/2}\big((\widehat\Sigma+\lambda I)^{-1}\widehat\Sigma - I\big)\theta_*\big\|_2^2 = \lambda^2\theta_*^\top(\widehat\Sigma+\lambda I)^{-1}\Sigma(\widehat\Sigma+\lambda I)^{-1}\theta_*.
$$

We then apply Eq. (9) with $A=\Sigma$ and $B=\theta_*\theta_*^\top$, which applies because of Assumption **(A4)**, to get:

$$
\begin{aligned}
\mathcal{R}^{(\mathrm{bias})}(\hat\theta) &= \kappa(\lambda)^2\theta_*^\top\big(\Sigma+\kappa(\lambda)I\big)^{-2}\Sigma\theta_* \\
&\qquad +\kappa(\lambda)^2\operatorname{tr}\big[\Sigma^2\big(\Sigma+\kappa(\lambda)I\big)^{-2}\big]\cdot\theta_*^\top\big(\Sigma+\kappa(\lambda)I\big)^{-2}\Sigma\theta_*\cdot\tfrac{1}{n-\mathrm{df}_2(\kappa(\lambda))} \\
&= \kappa(\lambda)^2\theta_*^\top\big(\Sigma+\kappa(\lambda)I\big)^{-2}\Sigma\theta_*\cdot\big(1+\tfrac{\mathrm{df}_2(\kappa(\lambda))}{n-\mathrm{df}_2(\kappa(\lambda))}\big),
\end{aligned}
$$

which concludes the proof. ■

Up to the term $\frac{1}{1-\mathrm{df}_2(\kappa(\lambda))/n}$, we exactly recover the fixed design analysis for the new larger regularization parameter $\kappa(\lambda)$. Note that in most situations, for the optimal regularization parameter, we usually have $\mathrm{df}_1(\kappa(\lambda)) \ll n$ and $\mathrm{df}_2(\kappa(\lambda)) \ll n$ so that the exploding term disappears.

We thus see two effects when we go from fixed design to random design: (1) an additional self-induced regularization due to moving from λ to $\kappa(\lambda) \geqslant \lambda$, and (2) an explosion of the excess risk if the degrees of freedom get too large.

In the next section, we consider the limit when λ tends to zero.

5 Minimum norm least-square estimation

The ridge regression estimator converges to the minimum ℓ_2-norm estimator when λ tends to zero. It turns out that this is precisely the estimator found by gradient descent started from zero [16]. We consider first the under-parameterized case ($\gamma < 1$) and then the over-parameterized one ($\gamma > 1$).

5.1 Under-parameterized regime (ordinary least-squares)

When $\gamma < 1$ (that is, $n > d$), then the OLS estimator is $\hat{\theta} = (X^\top X)^{-1} X^\top y = (X^\top X)^{-1} X^\top (X\theta_* + \varepsilon) = \theta_* + (X^\top X)^{-1} X^\top \varepsilon$, and thus we have $\mathcal{R}^{(\text{bias})}(\hat{\theta}) = 0$, and:

$$\mathbb{E}_\varepsilon\big[\mathcal{R}^{(\text{var})}(\hat{\theta})\big] = \Sigma^2 \operatorname{tr}\big[X(X^\top X)^{-1}\Sigma(X^\top X)^{-1}X^\top\big] = \tfrac{\sigma^2}{n}\operatorname{tr}\big[\Sigma\widehat{\Sigma}^{-1}\big].$$

Using Eq. (12), we obtain the classical equivalent $\sigma^2 \frac{\gamma}{1-\gamma} \sim \sigma^2 \frac{d}{n-d}$, as derived, e.g., in [17]. Note that for Gaussian data, this is, in fact, almost an equality, that is, $\mathbb{E}_\varepsilon\big[\mathcal{R}^{(\text{var})}(\hat{\theta})\big] = \sigma^2 \frac{d}{n-d-1}$ for $n > d+1$.

5.2 Over-parameterized regime

We now consider the case $\gamma > 1$ (that is, $d > n$). We can see it as the limit when λ tends to zero within ridge regression. This is exactly what was obtained in [17] (in a non-asymptotic framework), here with an interpretation in terms of degrees of freedom. We obtain, with $\kappa(0)$ such that $\mathrm{df}_1(\kappa(0)) = n$:

$$\begin{aligned}
\mathbb{E}_\varepsilon\big[\mathcal{R}^{(\text{var})}(\hat{\theta})\big] &\sim \frac{\sigma^2}{n}\mathrm{df}_2(\kappa(0)) \cdot \frac{1}{1 - \frac{1}{n}\mathrm{df}_2(\kappa(0))} \\
\mathcal{R}^{(\text{bias})}(\hat{\theta}) &\sim \kappa(0)^2 \theta_*^\top \Sigma(\Sigma + \kappa(0) I)^{-2}\theta_* \cdot \frac{1}{1 - \frac{1}{n}\mathrm{df}_2(\kappa(0))}.
\end{aligned}$$

Following [4, 17], we can try to understand when the over-parameterized limit with no regularization makes statistical sense, with two questions in mind: (1) does it lead to catastrophic over-fitting? (2) can it lead to a good performance? The answers to these questions will depend on how $\mathrm{df}_1(\kappa(\lambda))$ and $\mathrm{df}_2(\kappa(\lambda))$ are related. Since $\mathrm{df}_1(\kappa(\lambda)) = n$, we have $\mathrm{df}_2(\kappa(\lambda)) \leqslant \mathrm{df}_1(\kappa(\lambda)) = n$, but how much smaller?

Equivalent degrees of freedom In many standard situations, the two degrees of freedom are constants away from each other, in particular in the infinite-dimensional cases described at the end of Section 3.2. Thus the variance term is proportional to σ^2, while the bias term is proportional to $\kappa(\lambda)^2\theta_*^\top\Sigma(\Sigma + \kappa(\lambda)I)^{-2}\theta_*$. There is no catastrophic overfitting, but the variance term cannot go to zero as n tends to infinity, and we cannot expect a good performance when σ is far from zero. However, in noiseless problems where $\sigma = 0$, the bias term can lead to a better performance than what can be obtained with under-parameterized problems (see also Section 6).

Unbalanced degrees of freedom If $\mathrm{df}_2(\kappa(\lambda)) \ll \mathrm{df}_1(\kappa(\lambda)) = n$, then the variance term can indeed go to zero when n tends to infinity. This happens only in particular situations thoroughly described by [4, 17, 31].

6 Random projections

We consider a random projection matrix $S \in \mathbb{R}^{d\times m}$, sampled independently from X with the following assumptions:

(A5) $S \in \mathbb{R}^{d\times m}$ has sub-Gaussian i.i.d. components with mean zero and unit variance.

(A6) The number of projections m tends to infinity with $\frac{m}{n}$ tending to $\delta > 0$.

As for the linear regression assumptions, we do not assume Gaussian random projections, and in all of our experiments, we used Rademacher random variables in $\{-1, 1\}$. Given the matrix S, we consider projecting each covariate $x \in \mathbb{R}^d$ to $S^\top x \in \mathbb{R}^m$. Thus, if $\hat{\eta} \in \mathbb{R}^m$ is the minimum-norm minimizer of $\|y - XS\eta\|_2^2$, we consider $\hat{\theta} = S\hat{\eta} \in \mathbb{R}^d$. Note that this is different from applying the random projection on the left of y and X, which is often referred to as "sketching" [13, 30].

The asymptotic performance can be characterized as follows (again, apart from the expectation with respect to the noise variable ε, all results are meant almost surely).

Proposition 4 *Assume **(A1)**, **(A2)**, **(A3)**, **(A4)**, **(A5)**, **(A6)**. For the minimum norm least-squares estimator $\hat{\theta}$ based on random projections, we have for the under-parameterized regime ($m < n$):*

$$\begin{aligned}\mathbb{E}_\varepsilon\big[\mathcal{R}^{\rm var}(\hat{\theta})\big] &\sim \frac{\sigma^2 m}{n-m} = \frac{1}{1-\frac{m}{n}}\cdot\frac{\sigma^2 m}{n}\\ \mathcal{R}^{\rm bias}(\hat{\theta}) &\sim \frac{1}{1-\frac{m}{n}}\cdot \kappa_m \theta_*^\top \Sigma(\Sigma+\kappa_m I)^{-1}\theta_*,\end{aligned}$$

with κ_m defined by $\mathrm{df}_1(\kappa_m) \sim m$. In the over-parameterized regime, we get, for κ_n such that $\mathrm{df}_1(\kappa_n) \sim n$:

$$\begin{aligned}\mathbb{E}_\varepsilon\big[\mathcal{R}^{\rm var}(\hat{\theta})\big] &\sim \frac{\sigma^2}{n}\cdot\frac{\mathrm{df}_2(\kappa_n)}{1-\frac{1}{n}\mathrm{df}_2(\kappa_n)} + \sigma^2\frac{n}{m-n}\\ \mathcal{R}^{\rm bias}(\hat{\theta}) &\sim \kappa_n^2\theta_*^\top\Sigma(\Sigma+\kappa_n I)^{-2}\theta_*\cdot\frac{1}{1-\frac{1}{n}\mathrm{df}_2(\kappa_n)} + \kappa_n\theta_*^\top\Sigma(\Sigma+\kappa_n I)^{-1}\theta_*\cdot\frac{n}{m-n}.\end{aligned}$$

Proof We will consider the ℓ_2-regularized estimator, with a regularization parameter λ that we will let go to zero. The validity of such limits follows from the same arguments as [14, Lemma 6.2]. We thus consider:

$$\begin{aligned}\hat{\theta} &= S(S^\top X^\top XS + n\lambda I)^{-1}S^\top X^\top y\\ &= S(S^\top X^\top XS + n\lambda I)^{-1}S^\top X^\top X\theta_* + S(S^\top X^\top XS + n\lambda I)^{-1}S^\top X^\top\varepsilon\\ &= M\theta_* + S(S^\top X^\top XS + n\lambda I)^{-1}S^\top X^\top\varepsilon,\end{aligned}$$

with $M = S(S^\top X^\top XS + n\lambda I)^{-1}S^\top X^\top X$.

Conditioned on S and X, the expected risk is equal to, for the variance part:

$$\begin{aligned}\mathbb{E}_\varepsilon\big[\mathcal{R}^{(\rm var)}(\hat{\theta})\big] &= \sigma^2\,\mathrm{tr}\,\big[XS(S^\top X^\top XS + n\lambda I)^{-1}S^\top\Sigma S(S^\top X^\top XS + n\lambda I)^{-1}S^\top X^\top\big]\\ &= \sigma^2\,\mathrm{tr}\,\big[S^\top\Sigma S(S^\top X^\top XS + n\lambda I)^{-2}S^\top X^\top XS\big],\end{aligned}\tag{14}$$

while, for the bias, we have:

$$\begin{aligned}\mathcal{R}^{(\rm bias)}(\hat{\theta}) &= \big(M\theta_* - \theta_*\big)^\top\Sigma\big(M\theta_* - \theta_*\big) = \theta_*^\top\Sigma\theta_* + \theta_*^\top M^\top\Sigma M\theta_* - 2\theta_*^\top M^\top\Sigma\theta_*\\ &= \theta_*^\top\Sigma\theta_* - 2\theta_*^\top X^\top XS(S^\top X^\top XS + n\lambda I)^{-1}S^\top\Sigma\theta_*\\ &\qquad + \theta_*^\top X^\top XS(S^\top X^\top XS + n\lambda I)^{-1}S^\top\Sigma S(S^\top X^\top XS + n\lambda I)^{-1}S^\top X^\top X\theta_*.\end{aligned}\tag{15}$$

For the proof, we separate the two regimes $m < n$ and $m < n$. For both of them, we provide asymptotic expansions in two steps, first with respect to X and then S in the under-parameterized regime and vice-versa for the over-parameterized regime.

Under-parameterized regime: expansion with respect to X We consider S fixed and use the random matrix theory arguments from Section 3 for X. We have a covariance matrix $S^\top \Sigma S \in \mathbb{R}^{m\times m}$ of rank m, so under-parameterized results apply, and we get for the variance term (first term above), for S fixed, where we can directly consider $\lambda = 0$ (because of cancellations):

$$\mathbb{E}_\varepsilon\big[\mathcal{R}^{(\mathrm{var})}(\hat\theta)\big] = \sigma^2 \operatorname{tr}\big(S^\top\Sigma S(S^\top X^\top X S)^{-1}\big) \sim \tfrac{\sigma^2}{n-m}\operatorname{tr}\big(S^\top\Sigma S(S^\top\Sigma S)^{-1}\big) = \tfrac{\sigma^2 m}{m-n},$$

independently of the sketching matrix S. Note here that $S^\top\Sigma S$ is a random kernel matrix satisfying assumptions of Section 3; thus, its spectral measure has a limit.

For the bias term, the computation is more involved. With $T = \Sigma^{1/2}S$, and $X = \Sigma^{1/2}Z$, it is equal to:

$$\begin{aligned}\mathcal{R}^{(\mathrm{bias})}(\hat\theta) &= \theta_*^\top\Sigma\theta_* - 2\theta_*^\top\Sigma^{1/2}Z^\top ZT(T^\top Z^\top ZT + n\lambda I)^{-1}T^\top\Sigma^{1/2}\theta_* \\ &\quad + \theta_*^\top\Sigma^{1/2}Z^\top ZT(T^\top Z^\top ZT + n\lambda I)^{-1}T^\top T(T^\top Z^\top ZT + n\lambda I)^{-1}T^\top Z^\top Z\Sigma^{1/2}\theta_*.\end{aligned}$$

Using the matrix inversion lemma, we get:

$$\begin{aligned}\mathcal{R}^{(\mathrm{bias})}(\hat\theta) &= \theta_*^\top\Sigma\theta_* - 2\theta_*^\top\Sigma^{1/2}Z^\top(ZTT^\top Z^\top + n\lambda I)^{-1}ZTT^\top\Sigma^{1/2}\theta_* \\ &\quad + \theta_*^\top\Sigma^{1/2}Z^\top(ZTT^\top Z^\top + n\lambda I)^{-1}ZTT^\top TT^\top Z^\top(ZTT^\top Z^\top + n\lambda I)^{-1}Z\Sigma^{1/2}\theta_*.\end{aligned}$$

Denoting $C = TT^\top$, we then have

$$\begin{aligned}\mathcal{R}^{(\mathrm{bias})}(\hat\theta) &= \theta_*^\top\Sigma\theta_* - 2\theta_*^\top\Sigma^{1/2}Z^\top(ZCZ^\top + n\lambda I)^{-1}ZC\Sigma^{1/2}\theta_* \\ &\quad + \theta_*^\top\Sigma^{1/2}Z^\top(ZCZ^\top + n\lambda I)^{-1}ZC^2Z^\top(ZCZ^\top + n\lambda I)^{-1}Z\Sigma^{1/2}\theta_*.\end{aligned}$$

To find expansions of the red terms above, we can directly use the results from Section 3.3, using Eq. (10) with $A = C\Sigma^{1/2}\theta_*\theta_*^\top\Sigma^{1/2}$, and Eq. (11) with $A = \Sigma^{1/2}\theta_*\theta_*^\top\Sigma^{1/2}$ and $B = C^2$, with the covariance matrix C, and thus with degrees of freedom and the implicit regularization parameter $\tilde\kappa(\lambda)$ associated to C.[3] We can apply Prop. 2 since $C = TT^\top = \Sigma^{1/2}SS^\top\Sigma^{1/2}$ has almost surely a limiting spectral measure and the resulting needed traces involving the matrices A and B have well-defined limits. We get:

$$\begin{aligned}\mathcal{R}^{(\mathrm{bias})}(\hat\theta) &\sim \theta_*^\top\Sigma\theta_* - 2\theta_*^\top\Sigma^{1/2}(C+\tilde\kappa(\lambda)I)^{-1}C\Sigma^{1/2}\theta_* \\ &\quad + \theta_*^\top\Sigma^{1/2}(C+\tilde\kappa(\lambda)I)^{-1}CC(C+\tilde\kappa(\lambda)I)^{-1}\Sigma^{1/2}\theta_* \\ &\quad + \tilde\kappa(\lambda)^2\frac{\theta_*^\top\Sigma^{1/2}(C+\tilde\kappa(\lambda)I)^{-2}\Sigma^{1/2}\theta_*\cdot\operatorname{tr}\big[C^2(C+\tilde\kappa(\lambda)I)^{-2}\big]}{n-\widetilde{\mathrm{df}}_2(\tilde\kappa(\lambda))} \\ &\sim \theta_*^\top\Sigma^{1/2}\big(I - C(C+\tilde\kappa(\lambda)I)^{-1}\big)^2\Sigma^{1/2}\theta_* \\ &\quad + \tilde\kappa(\lambda)^2\frac{\theta_*^\top\Sigma^{1/2}(C+\tilde\kappa(\lambda)I)^{-2}\Sigma^{1/2}\theta_*\cdot\operatorname{tr}\big[C^2(C+\tilde\kappa(\lambda)I)^{-2}\big]}{n-\widetilde{\mathrm{df}}_2(\tilde\kappa(\lambda))}.\end{aligned}$$

When λ goes to zero, we have $\tilde\kappa(\lambda)\to 0$, $\widetilde{\mathrm{df}}_2(\tilde\kappa(\lambda))\to m$, as well as $C(C+\tilde\kappa(\lambda)I)^{-1} = TT^\top(TT^\top + \tilde\kappa(\lambda)I)^{-1} = T(T^\top T + \tilde\kappa(\lambda)I)^{-1}T^\top \to \Sigma^{1/2}S(S^\top\Sigma S)^{-1}S^\top\Sigma^{1/2}$, and $\tilde\kappa(\lambda)(C+\tilde\kappa(\lambda)I)^{-1} = I - C(C+\tilde\kappa(\lambda)I)^{-1} \to I - \Sigma^{1/2}S(S^\top\Sigma S)^{-1}S^\top\Sigma^{1/2}$. This leads to:

$$\begin{aligned}\mathcal{R}^{(\mathrm{bias})}(\hat\theta) &\sim \theta_*^\top\Sigma^{1/2}\big(I - \Sigma^{1/2}S(S^\top\Sigma S)^{-1}S^\top\Sigma^{1/2}\big)^2\Sigma^{1/2}\theta_* \\ &\quad + \frac{\theta_*^\top\Sigma^{1/2}(I - T(T^\top T)^{-1}T^\top)^2\Sigma^{1/2}\theta_*\cdot m}{n-m} \\ &= \theta_*^\top\big(\Sigma - \Sigma S(S^\top\Sigma S)^{-1}S^\top\Sigma\big)\theta_*\cdot\big(1+\frac{m}{n-m}\big). \qquad (16)\end{aligned}$$

[3] We use a different notation with $\tilde{}$, to avoid confusion with the same quantities with Σ.

Under-parameterized regime: full expansion Using results from Section 3.3, this time with $Z = S^\top$ and the covariance matrix Σ, with κ_m defined by $\mathrm{df}_1(\kappa_m) = m$, we get from Prop. 2 the equivalent $\Sigma - \Sigma S(S^\top \Sigma S)^{-1} S^\top \Sigma \sim \Sigma - \Sigma^{1/2}(\Sigma + \kappa_m I)^{-1}\Sigma^{1/2}$, and thus, from Eq. (16), we get the desired result: $\mathcal{R}^{\text{bias}}(\hat\theta) \sim \frac{1}{1-m/n}\kappa_m \theta_*^\top \Sigma(\Sigma + \kappa_m I)^{-1}\theta_*$.

Over-parameterized regime: expansion with respect to S We have, from Eq. (14):

$$\mathbb{E}_\varepsilon\big[\mathcal{R}^{(\text{var})}(\hat\theta)\big] \;=\; \frac{\sigma^2}{n}\operatorname{tr}\big(\Sigma S(S^\top \widehat\Sigma S + \lambda I)^{-1} S^\top \widehat\Sigma S(S^\top \widehat\Sigma X S + \lambda I)^{-1} S^\top\big).$$

To obtain an expansion of the red term, we can use Prop. 2 with covariance matrix $\widehat\Sigma$ and thus degrees of freedom and $\tilde\kappa$ associated to $\widehat\Sigma$:

$$\mathbb{E}_\varepsilon\big[\mathcal{R}^{(\text{var})}(\hat\theta)\big] \;\sim\; \frac{\sigma^2}{n}\operatorname{tr}\big[\Sigma\widehat\Sigma(\widehat\Sigma + \tilde\kappa(\lambda)I)^{-2}\big] + \frac{\sigma^2}{n}\tilde\kappa(\lambda)^2 \frac{\operatorname{tr}\big[\Sigma(\widehat\Sigma + \tilde\kappa(\lambda)I)^{-2}\big]\cdot \operatorname{tr}\big[\widehat\Sigma(\widehat\Sigma + \tilde\kappa(\lambda)I)^{-2}\big]}{m - \widetilde{\mathrm{df}}_2(\tilde\kappa(\lambda))}.$$

Using that $\tilde\kappa(\lambda) \to 0$ when $\lambda \to 0$, $\widehat\Sigma(\widehat\Sigma + \tilde\kappa(\lambda)I)^{-2} = nX^\top X(X^\top X + n\tilde\kappa(\lambda)I)^{-2}$ can be rewritten as $nX^\top(XX^\top + n\tilde\kappa(\lambda)I)^{-2}X \to nX^\top(XX^\top)^{-2}X$, and $\tilde\kappa(\lambda)^2(\widehat\Sigma + \tilde\kappa(\lambda)I)^{-2} = n^2\tilde\kappa(\lambda)^2(X^\top X + n\tilde\kappa(\lambda)I)^{-2} = (I - X^\top(XX^\top + n\tilde\kappa(\lambda)I)^{-1}X)^2 \to (I - X^\top(XX^\top)^{-1}X)^2 = (I - X^\top(XX^\top)^{-1}X)$, we thus get:

$$\mathbb{E}_\varepsilon\big[\mathcal{R}^{(\text{var})}(\hat\theta)\big] \sim \sigma^2 \operatorname{tr}\big[\Sigma X^\top(XX^\top)^{-2}X\big] + \frac{\operatorname{tr}\big[\Sigma(I - X^\top(XX^\top)^{-1}X)\big]\cdot \operatorname{tr}\big[(XX^\top)^{-1}\big]}{m-n}. \tag{17}$$

We can now take care of the (squared) bias term with the same technique, with $\lambda \to 0$, starting from Eq. (15):

$$\begin{aligned}
\mathcal{R}^{(\text{bias})}(\hat\theta) \;&=\; \theta_*^\top\Sigma\theta_* - 2\theta_*^\top\Sigma^{1/2}S(S^\top\widehat\Sigma S + \lambda I)^{-1}S^\top\widehat\Sigma\theta_* \\
&\qquad + \theta_*^\top\widehat\Sigma S(S^\top\widehat\Sigma S + \lambda I)^{-1}S^\top\Sigma S(S^\top X^\top X S + n\lambda I)^{-1}S^\top\widehat\Sigma\theta_* \\
&\sim\; \theta_*^\top\Sigma\theta_* - 2\theta_*^\top\Sigma^{1/2}(\widehat\Sigma + \tilde\kappa(\lambda)I)^{-1}\widehat\Sigma\theta_* + \theta_*^\top\widehat\Sigma(\widehat\Sigma + \tilde\kappa(\lambda)I)^{-1}\Sigma(\widehat\Sigma + \tilde\kappa(\lambda)I)^{-1}\widehat\Sigma\theta_* \\
&\qquad + \tilde\kappa(\lambda)^2\frac{\operatorname{tr}\big[\Sigma(\widehat\Sigma + \tilde\kappa(\lambda)I)^{-2}\big]\cdot\theta_*^\top\widehat\Sigma(\widehat\Sigma + \tilde\kappa(\lambda)I)^{-2}\widehat\Sigma\theta_*}{m - \widetilde{\mathrm{df}}_2(\tilde\kappa(\lambda))} \\
&\sim\; \big\|\Sigma^{1/2}\big(I - X^\top(XX^\top + \tilde\kappa(\lambda)I)^{-1}X\big)\theta_*\big\|_2^2 \\
&\qquad + \tilde\kappa(\lambda)^2\frac{\operatorname{tr}\big[\Sigma(X^\top X + \tilde\kappa(\lambda)I)^{-2}\big]\cdot\theta_*^\top X^\top X(X^\top X + \tilde\kappa(\lambda)I)^{-2}X^\top X\theta_*}{m - \widetilde{\mathrm{df}}_2(\tilde\kappa(\lambda))} \\
&\sim\; \theta_*^\top(I - X^\top(XX^\top)^{-1}X)\Sigma(I - X^\top(XX^\top)^{-1}X)\theta_* \\
&\qquad + \frac{1}{m-n}\theta_*^\top X^\top(XX^\top)^{-1}X\theta_*\cdot\operatorname{tr}\big[\Sigma(I - X^\top(XX^\top)^{-1}X)\big].
\end{aligned} \tag{18}$$

Over-parameterized regime: full expansion For κ_n defined as $\mathrm{df}_1(\kappa_n) = n$ for the full covariance matrix Σ (which is exactly the value of κ_m above for $m = n$), we get, using Prop. 2, with Eq. (17) and Eq. (18):

$$\begin{aligned}
\mathbb{E}\big[\mathcal{R}^{\text{var}}(\hat\theta)\big] \;&\sim\; \sigma^2\frac{\mathrm{df}_2(\kappa_n)}{\mathrm{df}_1(\kappa_n) - \mathrm{df}_2(\kappa_n)} + \sigma^2\frac{n}{m-n} \\
\mathcal{R}^{\text{bias}}(\hat\theta) \;&\sim\; \kappa_n^2\theta_*^\top\Sigma(\Sigma + \kappa_n I)^{-2}\theta_*\cdot\frac{\mathrm{df}_1(\kappa_n)}{\mathrm{df}_1(\kappa_n) - \mathrm{df}_2(\kappa_n)} + \kappa_n\theta_*^\top\Sigma(\Sigma + \kappa_n I)^{-1}\theta_*\cdot\frac{n}{m-n},
\end{aligned}$$

which is the desired result. ■

We can make the following observations:

- In the under-parameterized regime, we recover the traditional bias and variance terms divided by $1 - \frac{m}{n}$, which leads to the expected catastrophic over-fitting when m is close to n. Moreover, while the variance term goes up from $m = 0$ to $m = n$, the bias term has one decreasing term $\kappa_m \theta_*^\top \Sigma (\Sigma + \kappa_m I)^{-1} \theta_*$ and one increasing term $\big(1 - \frac{m}{n}\big)^{-1}$. In some cases (e.g., for θ_* and Σ isotropic), the overall performance always goes up, but in many situations, we obtain the traditional U-shaped curve in the under-parameterized regime.
- In the over-parameterized regime, the limit when m tends to infinity is exactly the same as the limit λ tending to zero for ridge regression in Section 5.2, since κ_n is exactly what was referred to as $\kappa(0)$. Moreover, we have, for both variance and bias, a decreasing function of m. Thus, once in this regime, it is always best to take m as large as possible. Note that to achieve the performance for $m = \infty$, we can simply take $\hat{\theta} = X^\top (X X^\top)^{-1} y$, and there is no need to solve a problem in dimension m with m large.
- Combining the two regimes, we indeed see an actual double descent in many scenarios. See illustrative experiments in Section 7.

7 Experiments

In this section, we present illustrative experiments to showcase our asymptotic equivalents from Section 6.[4]

Testing the asymptotic limit We consider a fixed spectral measure $\mu = \pi_1 \delta_{\sigma_1} + \pi_2 \delta_{\sigma_2}$ already considered by [17, 31] and the fixed measure $\nu = \mu$ for the optimal predictor, for which we can compute all of the asymptotic equivalents in Section 6 in closed form. We take $\gamma = d/n = 2$ and plot bias and variance as functions of $\delta = m/n$. We then compare them to experiments with finite n (and the corresponding $d = \gamma n$ and $m = \delta n$), where we sample θ_* and Σ from their distributions (with a matrix of eigenvectors uniformly at random in the set of orthogonal matrices). We have here, for $\delta \in [0, 1]$,

$$\kappa(\delta) = \frac{1}{2}\Big(\frac{\gamma}{\delta}(\pi_1\sigma_1 + \pi_2\sigma_2) - \sigma_1 - \sigma_2 + \Big[\big(\frac{\gamma}{\delta}(\pi_1\sigma_1 + \pi_2\sigma_2) - \sigma_1 - \sigma_2\big)^2 + 4\sigma_1\sigma_2\big(\frac{\gamma}{\delta} - 1\big) \Big]^{1/2} \Big).$$

In Figure 3, we can see that as n gets larger, each realization of the experiment tends to the asymptotic limit, illustrating almost sure convergence (which we conjecture to be of order $O(1/\sqrt{n})$), while, when we consider expectations with respect to several realizations, we get a faster convergence (which we conjecture to be of order $O(1/n)$).

Illustration of the double descent phenomenon We consider a fixed covariance matrix Σ of size d, with uniformly random eigenvectors and eigenvalues proportional to $1/k$, for $k \in \{1, \dots, d\}$ (non-isotropic), or constant (isotropic). We normalize the matrix so that $\mathrm{tr}(\Sigma) = 1$. We generate a vector $\theta_* \in \mathbb{R}^d$ from

[4] Matlab code to reproduce figures can be downloaded from https://www.di.ens.fr/~fbach/dd_rp.zip.

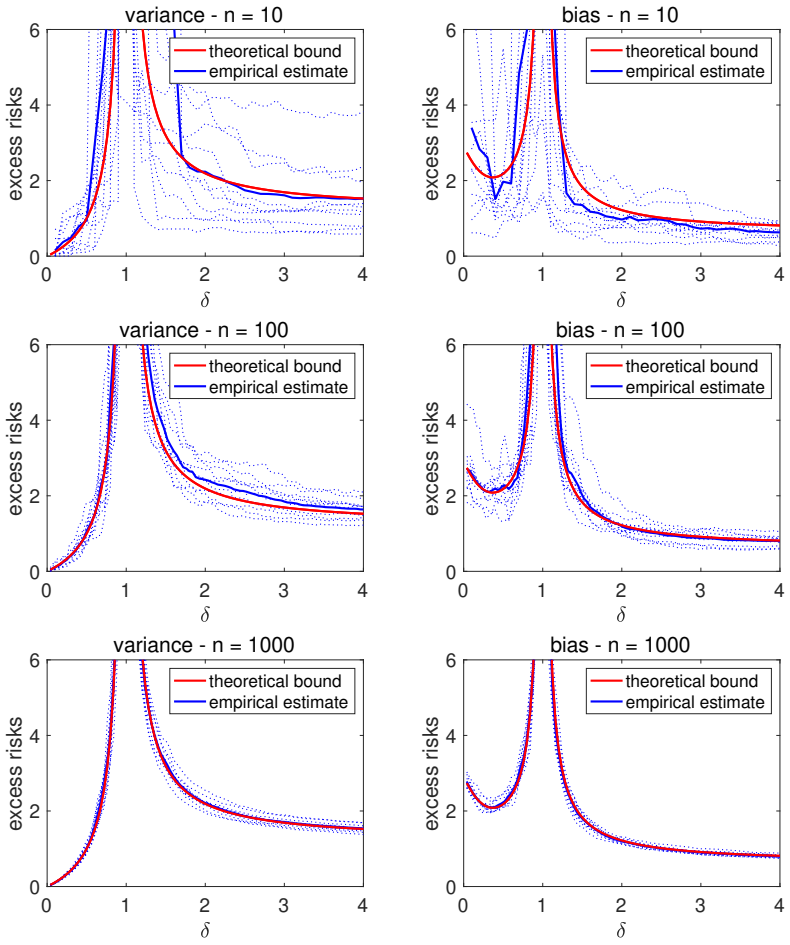

Figure 3: Comparison of theoretical bounds and empirical estimates for a spectral measure with two Diracs (see text for details): (left) variance, (right) bias, with three different numbers of observations, with $n = 10$ (top), $n = 100$ (middle), and $n = 1000$ (bottom). We plot ten realizations with the same spectral properties, as well as the average excess risk.

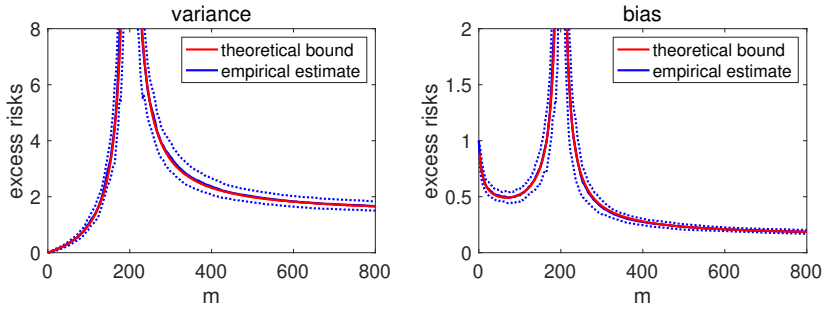

Figure 4: (Left) Variance with $\sigma = 1$ and $\mathrm{tr}(\Sigma) = 1$. (Right) Bias with $\theta_*^\top \Sigma \theta_* = 1$. We consider $n = 200$, $d = 400$, with Z and S sampled from Rademacher random variables, and eigenvalues of Σ proportional to $1/k$. For the empirical curve, we plot the average performance over 40 replications as well as the standard deviation in dotted.

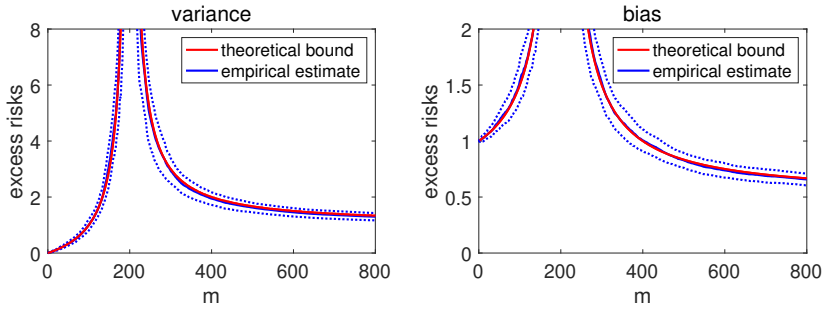

Figure 5: (Left) Variance with $\sigma = 1$ and $\mathrm{tr}(\Sigma) = 1$. (Right) Bias with $\theta_*^\top \Sigma \theta_* = 1$. We consider $n = 200$, $d = 400$, with Z and S sampled from Rademacher random variables and uniform eigenvalues for Σ. For the empirical curve, we plot the average performance over 40 replications as well as the standard deviation in dotted.

a standard Gaussian distribution and then normalize it so that $\theta_*^\top \Sigma \theta_* = 1$. Given this unique prediction problem, we generate 40 replications of Z and S from Rademacher random variables and plot the empirical performance for the bias and the variance. For the bounds, we compute κ_m from $\kappa_m^{-1} = \mathbb{E}\big[\,\mathrm{tr}[(S^\top \Sigma S)^{-1}]\big]$, using an average over 40 replications.

In Figure 4, we show the results for the non-isotropic covariance matrix, where we see a U-shaped curve for the bias term. In contrast, in Figure 5, we show the results for the isotropic covariance matrix, where we do not see a U-shaped curve for the bias term (and thus, there cannot be a U-shaped curve when summing bias and variance). The asymptotic limits from Section 6 closely match the empirical behavior in both cases.

8 Conclusion

In this paper, we have provided a high-dimensional asymptotic analysis of the double descent phenomenon for random projections. This was done using an interpretation of random matrix theory results for empirical covariance matrices based on degrees of freedom. Several avenues are worth exploring, such as going beyond

least-squares using tools from [25, 26], characterizing how quickly our asymptotic analysis kicks in using tools from [1], looking at more general random projection matrices [23], or relating it to the related sketching procedures that perform linear regression on $Ty \in \mathbb{R}^m$ and $TX \in \mathbb{R}^{m \times d}$, where the random matrix $T \in \mathbb{R}^{m \times n}$ now acts on the left of the design matrix rather than on the right, leading to a form of downsampling often referred to as sketching [13, 30]; see [9] for a recent work in this direction.

Acknowledgements

The author thanks Daniel LeJeune, Andrea Montanari, Bruno Loureiro, Ryan Tibshirani, and Florent Krzakala for feedback on the first version of the manuscript. He also acknowledges support from the French government under the management of the Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), as well as from the European Research Council (grant SEQUOIA 724063).

A Random matrix theory results

In this appendix, we provide a sketch of proof for classical random matrix theory results presented in Section 3.1 and Section 3.2, with a proof for the new results from Section 3.3. For more details, see [35, 2].

A.1 Self-consistency equation

We follow the proof of [34] and derive it in three steps.

First step We consider $n\widehat{\Sigma} = X^\top X = \sum_{i=1}^n x_i x_i^\top$, for $x_i \in \mathbb{R}^d$ sampled with covariance matrix Σ (but not necessarily Gaussian) and write, using the matrix inversion lemma:

$$\begin{aligned}
\operatorname{tr}\left[X^\top X (X^\top X - nzI)^{-1}\right] &= \sum_{i=1}^n \operatorname{tr}\left[x_i x_i^\top \Big(\sum_{j \neq i} x_j x_j^\top - nzI + x_i x_i^\top \Big)^{-1}\right] \\
&= \sum_{i=1}^n \frac{x_i^\top \big(\sum_{j \neq i} x_j x_j^\top - nzI \big)^{-1} x_i}{1 + x_i^\top \big(\sum_{j \neq i} x_j x_j^\top - nzI \big)^{-1} x_i} \\
&= n - \sum_{i=1}^n \frac{1}{1 + x_i^\top \big(\sum_{j \neq i} x_j x_j^\top - nzI \big)^{-1} x_i}.
\end{aligned}$$

Together with $\operatorname{tr}\left[X^\top X (X^\top X - nzI)^{-1}\right] = \operatorname{tr}\left[(XX^\top - nzI + nzI)(XX^\top - nzI)^{-1}\right] = n + nz\widehat{\varphi}(z)$, this leads to the identity

$$-z\widehat{\varphi}(z) = \frac{1}{n} \sum_{i=1}^n \frac{1}{1 + x_i^\top \big(\sum_{j \neq i} x_j x_j^\top - nzI \big)^{-1} x_i}. \tag{19}$$

We also have more generally:

$$\widehat{\Sigma}(\widehat{\Sigma} - zI)^{-1} = \sum_{i=1}^n x_i x_i^\top \Big(\sum_{j \neq i} x_j x_j^\top - nzI + x_i x_i^\top \Big)^{-1} = \sum_{i=1}^n \frac{x_i x_i^\top \big(\sum_{j \neq i} x_j x_j^\top - nzI \big)^{-1}}{1 + x_i^\top \big(\sum_{j \neq i} x_j x_j^\top - nzI \big)^{-1} x_i}. \tag{20}$$

Second step We have, owing to Eq. (19), with the notation $\widehat{\Sigma}_{-i} = \frac{1}{n}\sum_{j\neq i} x_j x_j^\top$ for $i \in \{1,\dots,n\}$, and using $A^{-1} - B^{-1} = -B^{-1}(A-B)B^{-1}$:

$$\begin{aligned}
\big(\widehat{\Sigma}-zI\big)^{-1} - \big(-z\widehat{\varphi}(z)\Sigma - zI\big)^{-1} &= \big(z\widehat{\varphi}(z)\Sigma + zI\big)^{-1}\Big(\widehat{\Sigma} - \big(-z\widehat{\varphi}(z)\Sigma\big)\Big)\big(\widehat{\Sigma}-zI\big)^{-1} \\
&= \big(z\widehat{\varphi}(z)\Sigma + zI\big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top - \big(-z\widehat{\varphi}(z)\Sigma\big)\Big)\big(\widehat{\Sigma}-zI\big)^{-1} \\
&= \big(z\widehat{\varphi}(z)\Sigma + zI\big)^{-1}\sum_{i=1}^n \frac{x_i x_i^\top \big(\sum_{j\neq i} x_j x_j^\top - nzI\big)^{-1} - \Sigma(n\widehat{\Sigma} - nzI)^{-1}}{1 + x_i^\top \big(\sum_{j\neq i} x_j x_j^\top - nzI\big)^{-1} x_i} \\
&= \big(z\widehat{\varphi}(z)\Sigma + zI\big)^{-1}\frac{1}{n}\sum_{i=1}^n \frac{x_i x_i^\top \big(\widehat{\Sigma}_{-i} - zI\big)^{-1} - \Sigma(\widehat{\Sigma}-zI)^{-1}}{1 + x_i^\top \big(n\widehat{\Sigma}_{-i} - nzI\big)^{-1} x_i}.
\end{aligned}$$

We thus get

$$\begin{aligned}
\big(\widehat{\Sigma} - zI\big)^{-1} &= \big(-z\widehat{\varphi}(z)\Sigma - zI\big)^{-1}\Big(I - \frac{1}{n}\sum_{i=1}^n \frac{x_i x_i^\top \big(\widehat{\Sigma}_{-i} - zI\big)^{-1} - \Sigma(\widehat{\Sigma}-zI)^{-1}}{1 + x_i^\top \big(n\widehat{\Sigma}_{-i} - nzI\big)^{-1} x_i}\Big) \\
&= \big(-z\widehat{\varphi}(z)\Sigma - zI\big)^{-1}(I - \Delta).
\end{aligned} \tag{21}$$

The main property we will leverage is that Δ will almost certainly be "negligible". For this, we need that $\operatorname{tr}\big[\big(-z\widehat{\varphi}(z)\Sigma - zI\big)^{-1}\Delta\big] = o(d)$, and we simply need to study each of the n terms, and show that they are $o(d)$. The key is that $\big(\widehat{\Sigma}_{-i} - zI\big)^{-1}$ is independent of x_i and that for any deterministic (or independent random bounded) matrix, $\operatorname{tr}[(z_i z_i^\top - I)N]$ is small enough with a strong probabilistic control [35, Lemma 3.1]. This is where we need i.i.d. components for z_i with sufficient moments (we assumed sub-Gaussian for simplicity, but weaker assumptions could be used to obtain the same almost-sure result). We can for example rely the on Hanson-Wright inequality [33], which leads to, for a constant $c > 0$:

$$\mathbb{P}\Big[\big|z_i^\top N z_i - \operatorname{tr}(N)\big| \leqslant c\big(t\|N\|_{\mathrm{op}} + \sqrt{t}\|N\|_F\big)\Big] \geqslant 1 - 2e^{-t}.$$

This is then applied to N dominated by Σ, and thus $\|N\|_F = O(\|\Sigma\|_F) = O(\sqrt{d}) = o(d)$, which is sufficient for the asymptotic result and hints at a rate in $O(1/\sqrt{d})$ [1]. See [34] for a detailed proof.

Overall, once we can neglect the term in Δ, we get: $\operatorname{tr}\big[(-z\widehat{\varphi}(z)\Sigma - zI)^{-1}\big] \sim \operatorname{tr}\big[(\widehat{\Sigma} - zI)^{-1}\big]$, and thus

$$\operatorname{tr}\big[\big(\widehat{\Sigma} - zI\big)^{-1}\big] \sim \frac{-1}{z\widehat{\varphi}(z)}\operatorname{tr}\Big[\Big(\Sigma + \frac{1}{\widehat{\varphi}(z)}I\Big)^{-1}\Big] = \frac{-d}{z} + \frac{1}{z}\operatorname{tr}\Big[\Sigma\Big(\Sigma + \frac{1}{\widehat{\varphi}(z)}I\Big)^{-1}\Big]. \tag{22}$$

Third step We can rewrite

$$\begin{aligned}
\operatorname{tr}\big[\big(\widehat{\Sigma} - zI\big)^{-1}\big] &= \frac{1}{z}\operatorname{tr}\big[\big(zI - \widehat{\Sigma} + \widehat{\Sigma}\big)\big(\widehat{\Sigma} - zI\big)^{-1}\big] \\
&= -\frac{d}{z} + \frac{1}{z}\operatorname{tr}\big[\widehat{\Sigma}\big(\widehat{\Sigma} - zI\big)^{-1}\big] = -\frac{d}{z} + \frac{1}{z}\operatorname{tr}\big[XX^\top\big(XX^\top - nzI\big)^{-1}\big] \\
&= -\frac{d}{z} + \frac{1}{z}\operatorname{tr}\big[\big(XX^\top - nzI + nzI\big)\big(XX^\top - nzI\big)^{-1}\big] = \frac{n-d}{z} + n\widehat{\varphi}(z).
\end{aligned} \tag{23}$$

Following [34] and combining Eq. (22) and Eq. (23), this leads to $\widehat{\varphi}(z) \to \varphi(z)$, with

$$\varphi(z) + \frac{1}{z} = \frac{1}{nz} \operatorname{tr}\Big[\Sigma\Big(\Sigma + \frac{1}{\varphi(z)} I\Big)^{-1}\Big], \tag{24}$$

which is the desired self-consistent equation in Eq. (3) in Section 3.2.

And even more intuitively, since $\operatorname{tr}\big[\widehat{\Sigma}(\widehat{\Sigma} - zI)^{-1}\big] = d + z \operatorname{tr}\big[(\widehat{\Sigma} - zI)^{-1}\big] = nz(\widehat{\varphi}(z) + \frac{1}{z})$, we get Eq. (1) from Section 3.1:

$$\operatorname{tr}\big[\widehat{\Sigma}\big(\widehat{\Sigma} - zI\big)^{-1}\big] \sim \operatorname{tr}\Big[\Sigma\Big(\Sigma + \frac{1}{\varphi(z)} I\Big)^{-1}\Big]. \tag{25}$$

We have for $z = -\lambda$, with $\lambda > 0$: $\operatorname{tr}\big[\widehat{\Sigma}\big(\widehat{\Sigma} + \lambda I\big)^{-1}\big] \sim \operatorname{tr}\Big[\Sigma\Big(\Sigma + \frac{1}{\varphi(-\lambda)} I\Big)^{-1}\Big]$, and thus

$$\varphi(-\lambda) - \frac{1}{\lambda} = -\frac{1}{n\lambda} \operatorname{tr}\Big[\Sigma\Big(\Sigma + \frac{1}{\varphi(-\lambda)} I\Big)^{-1}\Big] = -\frac{1}{n\lambda} \mathrm{df}_1\Big(\frac{1}{\varphi(-\lambda)}\Big),$$

leading to $\lambda\varphi(-\lambda) = 1 - \frac{1}{n}\mathrm{df}_1\Big(\frac{1}{\varphi(-\lambda)}\Big)$, and thus the desired inequality with $\kappa(\lambda) = \frac{1}{\varphi(-\lambda)}$, presented in Section 3.2.

A.2 Equivalents of spectral functions

In this section, we prove Prop. 1 and Prop. 2. Following [14], we start with an asymptotic equivalent based on differentiation (see formal justification in [14]). See [12, 23] for similar results based more strongly on differentiation (which is only used here to derive an equivalent for $\operatorname{tr}\big[\big(\widehat{\Sigma} - zI\big)^{-2}\big]$).

Using differentiation We have, by differentiating Eq. (24) with respect to z:

$$\varphi(z) + z\varphi'(z) \;=\; \frac{1}{n} \operatorname{tr}\Big[\Sigma\Big(\Sigma + \frac{1}{\varphi(z)} I\Big)^{-2}\Big] \frac{\varphi'(z)}{\varphi(z)^2},$$

which leads to $\dfrac{\varphi(z)}{\varphi'(z)} = \dfrac{1}{n} \operatorname{tr}\Big[\Sigma\Big(\Sigma + \dfrac{1}{\varphi(z)} I\Big)^{-2}\Big] \dfrac{1}{\varphi(z)^2} - z$. Thus, differentiating Eq. (25) with respect to z and using the bound on $\frac{\varphi(z)}{\varphi'(z)}$ above, we get:

$$\begin{aligned}
\operatorname{tr}\big[\widehat{\Sigma}\big(\widehat{\Sigma} - zI\big)^{-2}\big] \;&\sim\; \operatorname{tr}\Big[\Sigma\Big(\Sigma + \frac{1}{\varphi(z)} I\Big)^{-2}\Big] \frac{\varphi'(z)}{\varphi(z)^2} = \frac{n \operatorname{tr}\big[\Sigma\big(\Sigma + \frac{1}{\varphi(z)} I\big)^{-2}\big]}{\operatorname{tr}\big[\Sigma\big(\Sigma + \frac{1}{\varphi(z)} I\big)^{-2}\big] \frac{1}{\varphi(z)} - nz\varphi(z)} \\
&= \frac{n \operatorname{tr}\big[\Sigma\big(\Sigma + \frac{1}{\varphi(z)} I\big)^{-2}\big]}{\operatorname{tr}\big[\Sigma\big(\Sigma + \frac{1}{\varphi(z)} I\big)^{-2}\big] \frac{1}{\varphi(z)} + n - \operatorname{tr}\big[\Sigma\big(\Sigma + \frac{1}{\varphi(z)} I\big)^{-1}\big]} = \frac{n \operatorname{tr}\big[\Sigma\big(\Sigma + \frac{1}{\varphi(z)} I\big)^{-2}\big]}{n - \operatorname{tr}\big[\Sigma^2\big(\Sigma + \frac{1}{\varphi(z)} I\big)^{-2}\big]}.
\end{aligned}$$

This leads to the asymptotic equivalent

$$\begin{aligned}
\operatorname{tr}\big[\big(\widehat{\Sigma} - zI\big)^{-2}\big] \;&=\; \frac{1}{z} \operatorname{tr}\big[(zI - \widehat{\Sigma} + \widehat{\Sigma})\big(\widehat{\Sigma} - zI\big)^{-2}\big] = \frac{1}{z} \operatorname{tr}\big[\widehat{\Sigma}\big(\widehat{\Sigma} - zI\big)^{-2}\big] - \frac{1}{z} \operatorname{tr}\big[\big(\widehat{\Sigma} - zI\big)^{-1}\big] \\
&\sim\; \frac{1}{z} \frac{n \operatorname{tr}\big[\Sigma\big(\Sigma + \frac{1}{\varphi(z)} I\big)^{-2}\big]}{n - \operatorname{tr}\big[\Sigma^2\big(\Sigma + \frac{1}{\varphi(z)} I\big)^{-2}\big]} - \frac{1}{z} \operatorname{tr}\big[\big(- z\varphi(z)\Sigma - zI\big)^{-1}\big],
\end{aligned} \tag{26}$$

which we will need later.

Proof of Eq. (6) and Eq. (8) We now first show

$$\begin{aligned}\operatorname{tr}\big[A\big(\widehat{\Sigma}-zI\big)^{-1}\big] &\sim \operatorname{tr}\big[A\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}\big]\sim \operatorname{tr}\big[A\big(-z\varphi(z)\Sigma-zI\big)^{-1}\big],\\ &= \frac{-1}{z\varphi(z)}\operatorname{tr}\Big[A\Big(\Sigma+\frac{1}{\varphi(z)}I\Big)^{-1}\Big],\end{aligned}$$

where the last quantity is equivalent to $-\frac{d}{z}\int_0^{+\infty}\frac{d\nu_A(\sigma)}{1+\sigma\varphi(z)}$. We have, using Eq. (21):

$$\operatorname{tr}\big[A\big(\widehat{\Sigma}-zI\big)^{-1}\big]-\operatorname{tr}\big[A\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}\big]=-\operatorname{tr}\big[A\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}\Delta\big],$$

which is negligible as soon as $\|A\|_{\rm op}$ is bounded (using the same arguments as in Appendix A.1). We can then express $\operatorname{tr}\big[A\big(\Sigma+\frac{1}{\varphi(z)}I\big)^{-1}\big]$ as $\int_0^{+\infty}\frac{d\nu_A(\sigma)}{\sigma+\frac{1}{\varphi(z)}}$. This leads to the desired result in Prop. 1.

Proof of Eq. (7) and Eq. (9) For the quadratic form, we have for any matrices A and B, still using Eq. (21):

$$\begin{aligned}&\operatorname{tr}\big[A\big(\widehat{\Sigma}-zI\big)^{-1}B\big(\widehat{\Sigma}-zI\big)^{-1}\big]\\ =\;&\operatorname{tr}\big[A\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}(I-\Delta)B\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}(I-\Delta)\big]\\ =\;&\operatorname{tr}\big[A\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}B\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}\big]+\operatorname{tr}\big[A\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}\Delta B\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}\Delta\big]\\ &-\operatorname{tr}\big[A\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}\Delta B\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}-\operatorname{tr}\big[A\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}B\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}\Delta\big].\end{aligned}$$

The last two terms are negligible with the same arguments as in Appendix A.1 as soon as $\|A\|_{\rm op}$ and $\|B\|_{\rm op}$ are bounded. We have, for the second term:

$$\begin{aligned}&\operatorname{tr}\big[A\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}\Delta B\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}\Delta\big]\\ =\;&\frac{1}{n^2}\sum_{i,j=1}^n\operatorname{tr}\Big[A\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}\frac{(x_ix_i^\top-\Sigma)\big(\widehat{\Sigma}_{-i}-zI\big)^{-1}}{1+x_i^\top\big(n\widehat{\Sigma}_{-i}-nzI\big)^{-1}x_i}B\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}\frac{(x_jx_j^\top-\Sigma)\big(\widehat{\Sigma}_{-j}-zI\big)^{-1}}{1+x_j^\top\big(n\widehat{\Sigma}_{-j}-nzI\big)^{-1}x_j}\Big]\\ =\;&\frac{1}{n^2}\sum_{i,j=1}^n\frac{\operatorname{tr}\big[A\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}(x_ix_i^\top-\Sigma)\big(\widehat{\Sigma}_{-i}-zI\big)^{-1}B\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}(x_jx_j^\top-\Sigma)\big(\widehat{\Sigma}_{-j}-zI\big)^{-1}\big]}{\big(1+x_i^\top\big(n\widehat{\Sigma}_{-i}-nzI\big)^{-1}x_i\big)\big(1+x_j^\top\big(n\widehat{\Sigma}_{-j}-nzI\big)^{-1}x_j\big)}.\end{aligned}$$

When $i\neq j$, then we can separate terms with $x_ix_i^\top-\Sigma$ and $x_jx_j^\top-\Sigma$, which end up being negligible, thus leading to an equivalent

$$\frac{1}{n^2}\sum_{i=1}^n\frac{\operatorname{tr}\big[A\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}(x_ix_i^\top-\Sigma)\big(\widehat{\Sigma}_{-i}-zI\big)^{-1}B\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}(x_ix_i^\top-\Sigma)\big(\widehat{\Sigma}_{-i}-zI\big)^{-1}\big]}{\big(1+x_i^\top\big(n\widehat{\Sigma}_{-i}-nzI\big)^{-1}x_i\big)^2}.$$

To study its asymptotic limit, we need to characterize the asymptotic equivalent of the quantity $\operatorname{tr}\big[C(x_ix_i^\top-\Sigma)D(x_ix_i^\top-\Sigma)\big]=\operatorname{tr}\big[\Sigma^{1/2}C\Sigma^{1/2}(z_iz_i^\top-I)\Sigma^{1/2}D\Sigma^{1/2}(z_iz_i^\top-I)\big]$, with C and D bounded in operator norm. For $M=\Sigma^{1/2}C\Sigma^{1/2}$, and $N=\Sigma^{1/2}D\Sigma^{1/2}$, we can write:

$$\begin{aligned}\operatorname{tr}\big[M(z_iz_i^\top-I)N(z_iz_i^\top-I)\big]-\operatorname{tr}(M)\operatorname{tr}(N) &= (z_i^\top Mz_i-\operatorname{tr}(M))(z_i^\top Nz_i-\operatorname{tr}(N))\\ &\quad+\operatorname{tr}(M)(z_i^\top Nz_i-\operatorname{tr}(N))+\operatorname{tr}(N)(z_i^\top Mz_i-\operatorname{tr}(M))\\ &\quad-\operatorname{tr}[(MN+NM)(z_iz_i^\top-I)]\\ &= O_p(\|M\|_F\cdot\|N\|_F+\operatorname{tr}(M)\|N\|_F+\operatorname{tr}(N)\|M\|_F+\|NM\|_F).\end{aligned}$$

using the property from Appendix A.1 obtain from the i.i.d. assumption on the components of z_i, which is negligible compared to the term $\operatorname{tr}(M)\operatorname{tr}(N)$. Thus, using in addition that $\widehat{\Sigma}_{-j}$ is asymptotically equivalent to $\widehat{\Sigma}$, we get the equivalent

$$\frac{1}{n^2}\sum_{i=1}^n \frac{\operatorname{tr}\left[(\widehat{\Sigma}-zI)^{-1}A\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}\Sigma\right]\cdot\operatorname{tr}\left[(\widehat{\Sigma}-zI)^{-1}B\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}\Sigma\right]}{\big(1+x_i^\top\big(n\widehat{\Sigma}_{-i}-nzI\big)^{-1}x_i\big)^2}.$$

We thus overall have

$$\begin{aligned}
&\operatorname{tr}\left[A\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}\Delta B\Delta^\top\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}\right]\\
\sim\ &\operatorname{tr}\left[(\widehat{\Sigma}-zI)^{-1}A\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}\Sigma\right]\cdot\operatorname{tr}\left[(\widehat{\Sigma}-zI)^{-1}B\big(-z\widehat{\varphi}(z)\Sigma-zI\big)^{-1}\Sigma\right]\cdot\square\\
\sim\ &\operatorname{tr}\left[A\big(-z\varphi(z)\Sigma-zI\big)^{-2}\Sigma\right]\cdot\operatorname{tr}\left[B\big(-z\varphi(z)\Sigma-zI\big)^{-2}\Sigma\right]\cdot\square
\end{aligned}$$

with $\square=\dfrac{1}{n^2}\displaystyle\sum_{i=1}^n\frac{1}{\big(1+x_i^\top\big(n\widehat{\Sigma}_{-i}-nzI\big)^{-1}x_i\big)^2}$. This leads to:

$$\begin{aligned}
\operatorname{tr}\left[A\big(\widehat{\Sigma}-zI\big)^{-1}B\big(\widehat{\Sigma}-zI\big)^{-1}\right]\ \sim\ &\operatorname{tr}\left[A\big(-z\varphi(z)\Sigma-zI\big)^{-1}B\big(-z\varphi(z)\Sigma-zI\big)^{-1}\right]\\
&+\operatorname{tr}\left[A\big(-z\varphi(z)\Sigma-zI\big)^{-2}\Sigma\right]\cdot\operatorname{tr}\left[B\big(-z\varphi(z)\Sigma-zI\big)^{-2}\Sigma\right]\cdot\square.
\end{aligned}$$

To obtain an equivalent of $\square$, we consider the case $A=B=I$, to get:

$$\operatorname{tr}\left[\big(\widehat{\Sigma}-zI\big)^{-2}\right]\ \sim\ \operatorname{tr}\left[\big(-z\varphi(z)\Sigma-zI\big)^{-2}\right]+\big(\operatorname{tr}\left[\big(-z\varphi(z)\Sigma-zI\big)^{-2}\Sigma\right]\big)^2\cdot\square,$$

which allows to compute an equivalent of $\square$, as, using Eq. (26), with $z\varphi(z)\sim\operatorname{df}_1(1/\varphi(z))-\frac{1}{n}$.

$$\begin{aligned}
\square\ &\sim\ \frac{\operatorname{tr}\left[\big(\widehat{\Sigma}-zI\big)^{-2}\right]-\operatorname{tr}\left[\big(-z\varphi(z)\Sigma-zI\big)^{-2}\right]}{\big(\operatorname{tr}\left[\big(-z\varphi(z)\Sigma-zI\big)^{-2}\Sigma\right]\big)^2}\\
&\sim\ \frac{\frac{1}{z}\frac{n\operatorname{tr}\left[\Sigma\left(\Sigma+\frac{1}{\varphi(z)}I\right)^{-2}\right]}{n-\operatorname{df}_2(1/\varphi(z))}-\frac{1}{z}\operatorname{tr}\left[\big(-z\varphi(z)\Sigma-zI\big)^{-1}\right]-\operatorname{tr}\left[\big(-z\varphi(z)\Sigma-zI\big)^{-2}\right]}{\big(\operatorname{tr}\left[\big(-z\varphi(z)\Sigma-zI\big)^{-2}\Sigma\right]\big)^2}\\
&\sim\ \frac{\frac{1}{z}\frac{n\operatorname{tr}\left[\Sigma\left(\Sigma+\frac{1}{\varphi(z)}I\right)^{-2}\right]}{n-\operatorname{df}_2(1/\varphi(z))}+\frac{1}{z^2\varphi(z)}\operatorname{tr}\left[\big(\Sigma+\frac{1}{\varphi(z)}I\big)^{-1}\right]-\frac{1}{z^2\varphi(z)}\operatorname{tr}\left[\frac{1}{\varphi(z)}\big(\Sigma+\frac{1}{\varphi(z)}I\big)^{-2}\right]}{\big(\frac{1}{z^2\varphi(z)}\operatorname{tr}\left[\frac{1}{\varphi(z)}\big(\Sigma+\frac{1}{\varphi(z)}I\big)^{-2}\Sigma\right]\big)^2}\\
&\sim\ \frac{\frac{1}{z}\frac{n\operatorname{tr}\left[\Sigma\left(\Sigma+\frac{1}{\varphi(z)}I\right)^{-2}\right]}{n-\operatorname{df}_2(1/\varphi(z))}+\frac{1}{z^2\varphi(z)}\operatorname{tr}\left[\Sigma\big(\Sigma+\frac{1}{\varphi(z)}I\big)^{-2}\right]}{\big(\frac{1}{z^2\varphi(z)}\operatorname{tr}\left[\frac{1}{\varphi(z)}\big(\Sigma+\frac{1}{\varphi(z)}I\big)^{-2}\Sigma\right]\big)^2}\sim\frac{\frac{1}{z}\frac{n}{n-\operatorname{df}_2(1/\varphi(z))}+\frac{1}{z^2\varphi(z)}}{\big(\frac{1}{z^2\varphi(z)}\big)^2\frac{1}{\varphi(z)}\operatorname{tr}\left[\frac{1}{\varphi(z)}\big(\Sigma+\frac{1}{\varphi(z)}I\big)^{-2}\Sigma\right]}\\
&\sim\ \frac{\frac{nz\varphi(z)}{n-\operatorname{df}_2(1/\varphi(z))}+1}{\frac{1}{z^2\varphi(z)^2}\operatorname{tr}\left[\frac{1}{\varphi(z)}\big(\Sigma+\frac{1}{\varphi(z)}I\big)^{-2}\Sigma\right]}=\frac{\frac{\operatorname{df}_1(1/\varphi(z))-n}{n-\operatorname{df}_2(1/\varphi(z))}+1}{\frac{1}{z^2\varphi(z)^2}\operatorname{tr}\left[\frac{1}{\varphi(z)}\big(\Sigma+\frac{1}{\varphi(z)}I\big)^{-2}\Sigma\right]}\\
&\sim\ \frac{\frac{\operatorname{df}_1(1/\varphi(z))-n}{n-\operatorname{df}_2(1/\varphi(z))}+1}{\frac{1}{z^2\varphi(z)^2}\big(\operatorname{df}_1(1/\varphi(z))-\operatorname{df}_2(1/\varphi(z))\big)}=\frac{z^2\varphi(z)^2}{n-\operatorname{df}_2(1/\varphi(z))}.
\end{aligned}$$

Overall, we get

$$\begin{aligned}
& \operatorname{tr}\left[A\big(\widehat{\Sigma}-zI\big)^{-1}B\big(\widehat{\Sigma}-zI\big)^{-1}\right] \\
\sim\ & \tfrac{1}{z^2\varphi(z)^2}\operatorname{tr}\left[A\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-1}B\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-1}\right] \\
& +\tfrac{1}{z^4\varphi(z)^4}\operatorname{tr}\left[A\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-2}\Sigma\right]\cdot\operatorname{tr}\left[B\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-2}\Sigma\right]\cdot\tfrac{z^2\varphi(z)^2}{n-\mathrm{df}_2(1/\varphi(z))} \\
\sim\ & \tfrac{1}{z^2\varphi(z)^2}\operatorname{tr}\left[A\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-1}B\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-1}\right] \\
& +\tfrac{1}{z^2\varphi(z)^2}\operatorname{tr}\left[A\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-2}\Sigma\right]\cdot\operatorname{tr}\left[B\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-2}\Sigma\right]\cdot\tfrac{1}{n-\mathrm{df}_2(1/\varphi(z))},
\end{aligned}$$

which is Eq. (9).

We also have, by writing $\widehat{\Sigma}\big(\widehat{\Sigma}-zI\big)^{-1}=I+z\big(\widehat{\Sigma}-zI\big)^{-1}$:

$$\begin{aligned}
& \operatorname{tr}\left[A\widehat{\Sigma}\big(\widehat{\Sigma}-zI\big)^{-1}B\big(\widehat{\Sigma}-zI\big)^{-1}\right] \\
=\ & z\operatorname{tr}\left[A\big(\widehat{\Sigma}-zI\big)^{-1}B\big(\widehat{\Sigma}-zI\big)^{-1}\right]+\operatorname{tr}\left[AB\big(\widehat{\Sigma}-zI\big)^{-1}\right] \\
\sim\ & -\frac{1}{z\varphi(z)}\operatorname{tr}\left[AB\big(\Sigma+\frac{1}{\varphi(z)}I\big)^{-1}\right]+\frac{1}{z\varphi(z)^2}\operatorname{tr}\left[A\big(\Sigma+\frac{1}{\varphi(z)}I\big)^{-1}B\big(\Sigma+\frac{1}{\varphi(z)}I\big)^{-1}\right] \\
& +\tfrac{1}{z\varphi(z)^2}\operatorname{tr}\left[A\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-2}\Sigma\right]\cdot\operatorname{tr}\left[B\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-2}\Sigma\right]\cdot\tfrac{1}{n-\mathrm{df}_2(1/\varphi(z))} \\
\sim\ & -\tfrac{1}{z\varphi(z)}\operatorname{tr}\left[AB\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-1}\right]+\tfrac{1}{z\varphi(z)}\operatorname{tr}\left[A\tfrac{1}{\varphi(z)}\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-1}B\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-1}\right] \\
& +\tfrac{1}{z\varphi(z)^2}\operatorname{tr}\left[A\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-2}\Sigma\right]\cdot\operatorname{tr}\left[B\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-2}\Sigma\right]\cdot\tfrac{1}{n-\mathrm{df}_2(1/\varphi(z))} \\
\sim\ & -\tfrac{1}{z\varphi(z)}\operatorname{tr}\left[A\Sigma\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-1}B\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-1}\right] \\
& +\tfrac{1}{z\varphi(z)^2}\operatorname{tr}\left[A\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-2}\Sigma\right]\cdot\operatorname{tr}\left[B\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-2}\Sigma\right]\cdot\tfrac{1}{n-\mathrm{df}_2(1/\varphi(z))}.
\end{aligned}$$

We also finally have by using again $\widehat{\Sigma}\big(\widehat{\Sigma}-zI\big)^{-1}=I+z\big(\widehat{\Sigma}-zI\big)^{-1}$:

$$\begin{aligned}
& \operatorname{tr}\left[A\widehat{\Sigma}\big(\widehat{\Sigma}-zI\big)^{-1}B\widehat{\Sigma}\big(\widehat{\Sigma}-zI\big)^{-1}\right] \\
\sim\ & \operatorname{tr}\left[A\Sigma\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-1}B\Sigma\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-1}\right] \\
& +\tfrac{1}{\varphi(z)^2}\operatorname{tr}\left[A\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-2}\Sigma\right]\cdot\operatorname{tr}\left[B\big(\Sigma+\tfrac{1}{\varphi(z)}I\big)^{-2}\Sigma\right]\cdot\tfrac{1}{n-\mathrm{df}_2(1/\varphi(z))},
\end{aligned}$$

which is Eq. (7).

References

[1] Zhidong Bai and Jack W. Silverstein. CLT for linear spectral statistics of large-dimensional sample covariance matrices. In *Advances In Statistics*, pages 281–333. World Scientific, 2008. (cited on pages 18 and 19)

[2] Zhidong Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*, volume 20. Springer, 2010. (cited on pages 2, 5, and 18)

[3] Zhidong Bai and Yong-Qua Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. In *Advances In Statistics*, pages 108–127. World Scientific, 2008. (cited on page 8)

[4] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020. (cited on pages 2 and 11)

[5] Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta Numerica*, 30:87–201, 2021. (cited on pages 2 and 9)

[6] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. (cited on page 1)

[7] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020. (cited on page 1)

[8] Andrea Caponnetto and Ernesto de Vito. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007. (cited on pages 2, 4, 7, and 9)

[9] Xin Chen, Yicheng Zeng, Siyue Yang, and Qiang Sun. Sketched ridgeless linear regression: The role of downsampling. Technical Report 2302.01088, arXiv, 2023. (cited on page 18)

[10] Chen Cheng and Andrea Montanari. Dimension free ridge regression. Technical Report 2210.08571, arXiv, 2022. (cited on pages 3, 6, 7, and 9)

[11] Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural Information Processing Systems*, 34:10131–10143, 2021. (cited on pages 2 and 7)

[12] Yehuda Dar, Daniel LeJeune, and Richard G. Baraniuk. The common intuition to transfer learning can win or lose: Case studies for linear regression. Technical Report 2103.05621, arXiv, 2021. (cited on pages 4, 5, 7, and 20)

[13] Edgar Dobriban and Sifan Liu. Asymptotics for sketching in least squares regression. *Advances in Neural Information Processing Systems*, 32, 2019. (cited on pages 12 and 18)

[14] Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018. (cited on pages 2, 3, 4, 5, 7, 8, 9, 10, 12, and 20)

[15] Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, (2):023401, 2020. (cited on page 1)

[16] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, 2018. (cited on page 10)

[17] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022. (cited on pages 1, 2, 4, 11, and 15)

[18] Trevor J. Hastie and Robert J. Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1990. (cited on page 4)

[19] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. (cited on page 2)

[20] Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on Learning Theory*, 2012. (cited on page 9)

[21] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Implicit regularization of random feature models. In *International Conference on Machine Learning*, 2020. (cited on page 6)

[22] Olivier Ledoit and Sandrine Péché. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1-2):233–264, 2011. (cited on pages 5 and 7)

[23] Daniel LeJeune, Pratik Patil, Hamid Javadi, Richard G. Baraniuk, and Ryan J. Tibshirani. Asymptotics of the sketched pseudoinverse. Technical Report 2211.03751, arXiv, 2022. (cited on pages 2, 4, 5, 6, 8, 18, and 20)

[24] Zhenyu Liao, Romain Couillet, and Michael W. Mahoney. A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. *Advances in Neural Information Processing Systems*, 33, 2020. (cited on page 1)

[25] Zhenyu Liao and Michael W. Mahoney. Hessian eigenspectra of more realistic nonlinear models. *Advances in Neural Information Processing Systems*, 34, 2021. (cited on page 18)

[26] Bruno Loureiro, Cédric Gerbelot, Maria Refinetti, Gabriele Sicuro, and Florent Krzakala. Fluctuations, bias, variance & ensemble of learners: Exact asymptotics for convex losses in high-dimension. In *International Conference on Machine Learning*, 2022. (cited on page 18)

[27] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022. (cited on page 1)

[28] Andrea Montanari and Yiqiao Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *The Annals of Statistics*, 50(5):2816–2847, 2022. (cited on page 2)

[29] Jaouad Mourtada and Lorenzo Rosasco. An elementary analysis of ridge regression with random design. *Comptes Rendus. Mathématique*, 360(G9):1055–1063, 2022. (cited on page 9)

[30] Garvesh Raskutti and Michael W. Mahoney. A statistical perspective on randomized sketching for ordinary least-squares. *Journal of Machine Learning Research*, 17(1):7508–7538, 2016. (cited on pages 12 and 18)

[31] Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge (less) regression under general source condition. In *International Conference on Artificial Intelligence and Statistics*, 2021. (cited on pages 2, 3, 4, 11, and 15)

[32] Francisco Rubio and Xavier Mestre. Spectral convergence for a general class of random matrices. *Statistics & Probability Letters*, 81(5):592–602, 2011. (cited on page 3)

[33] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013. (cited on page 19)

[34] Jack W. Silverstein. Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2):331–339, 1995. (cited on pages 18, 19, and 20)

[35] Jack W. Silverstein and Zhidong Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate analysis*, 54(2):175–192, 1995. (cited on pages 18 and 19)

[36] Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33, 2020. (cited on page 2)

[37] Ji Xu and Daniel J. Hsu. On the number of variables to use in principal component regression. *Advances in Neural Information Processing Systems*, 32, 2019. (cited on page 1)

[38] Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098, 2005. (cited on page 2)