High-Dimensional Analysis of Double Descent for Linear Regression with Random Projections[*]

Francis Bach[†]

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 nonparametric 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. Overparameterized models estimated with some form of gradient descent come in various forms, such as linear regression with potentially nonlinear features, neural networks, or kernel methods. Originally characterized in simple models [41, 34] (see also [29] for a complete history), the double descent phenomenon can be seen empirically in these aforementioned models [6, 17]: 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 [31], random Fourier features [26], or linear regression [7, 19]. While the analysis of [31, 26] for random features corresponds to a single prediction problem with a sequence of increasingly larger prediction models, most of the analysis of [19] for linear regression does not consider a single problem, but varying problems, which does not actually lead to a double descent curve. Random variable subsampling on a single prediction problem was analyzed with a simpler model with isotropic covariance matrices in [7], but without a proper double descent as the model is too simple (e.g., isotropic covariance matrices) to

[†]INRIA - Département d'Informatique de l'Ecole Normale Supérieure, PSL Research University, Paris, France (francis.bach@inria.fr, https://www.di.ens.fr/~fbach/).
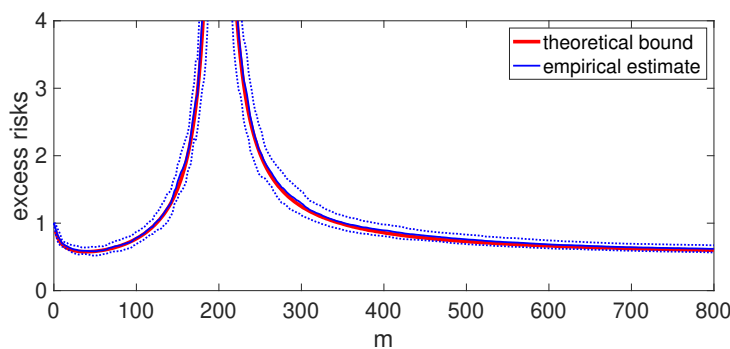
Figure 1. *Example of a double descent curve, for linear regression with random projections with $n = 200$ observations, in dimension $d = 400$ and a nonisotropic covariance matrix. The data are normalized so that predicting zero leads to an excess risk of 1 and the noise so that the optimal expected 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.

account for a U-shaped curve in the underparameterized regime. In work related to ours, principal component regression was analyzed by [43] with a double descent curve but with less general assumptions regarding the spectrum of the covariance matrix and the optimal predictor. 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. Earlier work on the same model ([31], [19, section 5.2], and [44]) considers only isotropic covariance matrices, where the U-shaped curve in the underparameterized regime is not observed. Our analysis follows the high-dimensional analysis of [19, 16, 36, 25, 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 [21], where a squared Euclidean penalty is added to the empirical risk.

The paper is organized as follows, first presenting earlier work on the high-dimensional analysis of regularized least-squares and then presenting in section 6 the novel analysis of random projections in the same context.

- 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, following [10], we reinterpret the results of [16, 36, 11, 42, 5] using classical notions from nonparametric statistics, namely, the degrees of freedom, a.k.a. effective dimensionality [45, 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 [32], 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 [19, 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 underparameterized regime ($m < n$), blow up at $m = n$, and then go 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 $x_1, \ldots, x_n \in \mathbb{R}^d$ 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 noncentered empirical covariance matrix, while $XX^\top \in \mathbb{R}^{n \times n}$ is the kernel matrix.

We will characterize the performance of an estimator $\hat{\theta}$ through its excess risk, that is, its expected square loss on unseen data minus the optimal such loss. It has a simple expression as a quadratic form $\mathcal{R}(\hat{\theta}) = (\hat{\theta} - \theta_*)\Sigma(\hat{\theta} - \theta_*)$. 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 $\mathbb{E}_\varepsilon[\mathcal{R}^{(\mathrm{var})}(\hat{\theta})]$ 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 [16, 36], that is, the following hold:

(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 $\sigma_1, \ldots, \sigma_d$ 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 [37]. 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 the following: for any bounded function $r : \mathbb{R}_+ \to \mathbb{R}$, $\theta_*^\top r(\Sigma)\theta_* \to \int_0^{+\infty} r(\sigma)d\nu(\sigma)$ and corresponds to ν having a density with respect to μ. Moreover, it is often replaced by θ_* being random with mean zero and covariance matrix proportional to identity [16] or a spectral variant of Σ [36].

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 [16, 19, 36, 25], but more refined ones will be needed along the lines of [12, 25] (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 reinterpretation 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 $\lambda > 0$,

$$(3.1) \qquad \mathrm{tr}\,[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-1}] \sim \mathrm{tr}\,[\Sigma(\Sigma + \kappa(\lambda)I)^{-1}],$$

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, 20] and denoted[2]

$$\widehat{df}_1(\lambda) = \mathrm{tr}[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-1}] \quad \text{and} \quad df_1(\kappa) = \mathrm{tr}[\Sigma(\Sigma + \kappa I)^{-1}].$$

In the limit when d tends to infinity, by the definition of μ in assumption (A3), then $\frac{1}{d}df_1(\kappa) \to \int_0^{+\infty} \frac{\sigma d\mu(\sigma)}{\sigma + \kappa}$, which is strictly decreasing in κ, with a value of 1 at $\kappa = 0$. Since $\mathrm{tr}\,[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-1}] \leqslant d$, this asymptotically defines uniquely $\kappa(\lambda)$.

The extra knowledge from random matrix theory will be the self-consistency equation $\kappa(\lambda) - \lambda = \kappa(\lambda) \cdot \gamma \int_0^{+\infty} \frac{\sigma d\mu(\sigma)}{\sigma + \kappa}$, which allows us to define $\kappa(\lambda)$, which we will write equivalently

$$\kappa(\lambda) - \lambda \sim \kappa(\lambda) \cdot \frac{1}{n} df_1(\kappa(\lambda)).$$

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 underparameterized models ($\gamma < 1$), while for overparameterized 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 almost surely when the dimensions n, d go to infinity. This allows us 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 (3.1), we will need asymptotic equivalents for the quantities $\text{tr}\,[A\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-1}]$ and $\text{tr}\,[A\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-1}B\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-1}]$ 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 Propositions 3.1 and 3.2 below).

These results recover existing work with $A, B = I$ or Σ [24, 16] and lead to the same formulas as [12, 25] 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 $z \in \mathbb{C} \backslash \mathbb{R}_+$:

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

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, 24]:

$$\frac{1}{\varphi(z)} + z = \gamma \int_0^{+\infty} \frac{\sigma d\mu(\sigma)}{1 + \sigma\varphi(z)}. \tag{3.2}$$

When $\Sigma = \sigma I$, this allows us 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 not convenient to describe for general Σ) and only access it through its Stieltjes transform.

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

$$\frac{1}{\varphi(-\lambda)} - \lambda = \gamma \int_0^{+\infty} \frac{\sigma d\mu(\sigma)}{1 + \sigma\varphi(-\lambda)}. \tag{3.3}$$

In the ridge regression context, as mentioned above, the quantity $\text{df}_1(\kappa) = \text{tr}\,[\Sigma(\Sigma + \kappa I)^{-1}] \in [0, d]$ is referred to as the "degrees of freedom." It is a strictly decreasing function of κ with $\text{df}_1(0) = \text{rank}(\Sigma)$. 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 (3.3) as

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

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

$$\kappa(\lambda) - \lambda \sim \kappa(\lambda) \cdot \frac{1}{n}\text{df}_1(\kappa(\lambda)) \quad \Leftrightarrow \quad \lambda \sim \kappa(\lambda)\Big(1 - \frac{1}{n}\text{df}_1(\kappa(\lambda))\Big). \tag{3.4}$$
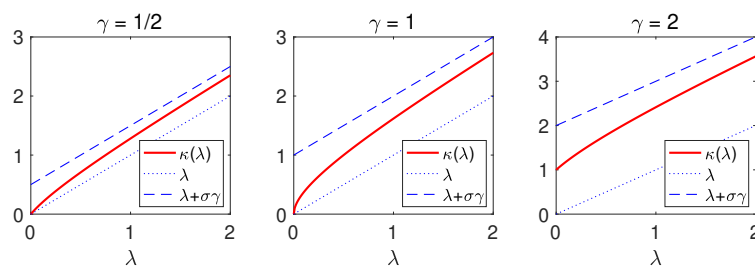
Figure 2. *Implicit regularization parameter $\kappa(\lambda)$ in the three regimes for isotropic covariance matrices, with $\sigma = 1$. See text for details.*

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 [25, 23, 10], and we refer to it as self-induced.

Note that in order to compute $\kappa(\lambda)$, we can either solve (3.4) if we can compute $\mathrm{df}_1(\kappa(\lambda))$ or simply use that $\kappa(\lambda)^{-1}$ is the almost sure limit of $\mathrm{tr}\,[(XX^\top + n\lambda I)^{-1}]$, 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 (3.4) is equivalent to $\lambda = \kappa(\lambda)(1 - \frac{\gamma\sigma}{\sigma+\kappa})$. We can solve it in closed form as

$$(3.5) \qquad \kappa(\lambda) = \frac{1}{2}\Big(\lambda - \sigma(1 - \gamma) + \sqrt{(\sigma(1-\gamma) - \lambda)^2 + 4\lambda\sigma}\Big).$$

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 [25, 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 (3.5) is an upper-bound with σ replaced by $\frac{1}{d}\mathrm{tr}(\Sigma)$.

- *Underparameterized* ($\gamma < 1 \Leftrightarrow d < n$): 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 $\mathrm{rank}(\Sigma) = d$).
- *Overparameterized* ($\gamma > 1 \Leftrightarrow d > n$): 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 $\mathrm{df}_1(\kappa(\lambda)) \leqslant \frac{\mathrm{tr}\,\Sigma}{\kappa(\lambda) + \mathrm{tr}\,\Sigma/d} \leqslant \frac{\mathrm{tr}\,\Sigma}{\kappa(\lambda)}$. This in turn implies that $\kappa(\lambda) \in [\lambda, \lambda + \frac{\mathrm{tr}\,\Sigma}{n}]$ and also a finer bound based on (3.5) with σ replaced by $\frac{1}{d}\mathrm{tr}(\Sigma)$. Moreover, we have the bound $\kappa(0) \leqslant \frac{\mathrm{tr}\,\Sigma}{n}(1 - n/d) = \frac{\mathrm{tr}\,\Sigma}{d}(\gamma - 1)$.

"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 $k \in \{1, \ldots, d\}$ 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 $v = ud(\kappa/\tau)^{1/\alpha}$,

$$\mathrm{df}_1(\kappa d^\alpha) \sim d \int_0^1 \frac{\tau u^{-\alpha} du}{\tau u^{-\alpha} + d^\alpha \kappa} = d \int_0^1 \frac{du}{1 + (ud)^\alpha \kappa \tau^{-1}} \sim (\tau/\kappa)^{1/\alpha} \int_0^{+\infty} \frac{dv}{1 + v^\alpha}.$$

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 [24, 16], 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. See also [37, 15] for a proof of (3.6) and [14] for subcases of (3.7). The asymptotic equivalents will be expressed in terms of a different notion of degrees of freedom, that is, $\mathrm{df}_2(\kappa) = \mathrm{tr}\,[\Sigma^2(\Sigma + \kappa I)^{-2}]$ (see sections 4.2 and 5.2 for a discussion relating the two definitions of degrees of freedom).

Proposition 3.1. 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* (3.2),

$$(3.6) \qquad \mathrm{tr}\,[A\widehat{\Sigma}(\widehat{\Sigma} - zI)^{-1}] \sim \mathrm{tr}\,\Big[A\Sigma\Big(\Sigma + \frac{1}{\varphi(z)}I\Big)^{-1}\Big],$$

$$(3.7) \qquad \mathrm{tr}\,[A\widehat{\Sigma}(\widehat{\Sigma} - zI)^{-1}B\widehat{\Sigma}(\widehat{\Sigma} - zI)^{-1}] \sim \mathrm{tr}\,\Big[A\Sigma\Big(\Sigma + \frac{1}{\varphi(z)}I\Big)^{-1}B\Sigma\Big(\Sigma + \frac{1}{\varphi(z)}I\Big)^{-1}\Big]$$

$$+ \frac{1}{\varphi(z)^2} tr\,\Big[A\Big(\Sigma + \frac{1}{\varphi(z)}I\Big)^{-2}\Sigma\Big] \cdot \mathrm{tr}\,\Big[B\Big(\Sigma + \frac{1}{\varphi(z)}I\Big)^{-2}\Sigma\Big] \cdot \frac{1}{n - df_2(1/\varphi(z))}.$$

Equation (3.6) can formally be seen as the limit $\frac{1}{d}\mathrm{tr}\,[A\widehat{\Sigma}(\widehat{\Sigma} - zI)^{-1}] \to \int_0^{+\infty} \frac{\sigma d\nu_A(\sigma)}{\sigma + 1/\varphi(z)}$, and a similar result holds for (3.7). From (3.6) and (3.7), 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

$$(3.8) \qquad \mathrm{tr}\,[A(\widehat{\Sigma} - zI)^{-1}] \sim \frac{-1}{z\varphi(z)}\mathrm{tr}\,\Big[A\Big(\Sigma + \frac{1}{\varphi(z)}I\Big)^{-1}\Big],$$

$$(3.9) \qquad \mathrm{tr}\left[A(\widehat{\Sigma}-zI)^{-1}B(\widehat{\Sigma}-zI)^{-1}\right] \sim \frac{1}{z^2\varphi(z)^2}\mathrm{tr}\left[A\left(\Sigma+\frac{1}{\varphi(z)}I\right)^{-1}B\left(\Sigma+\frac{1}{\varphi(z)}I\right)^{-1}\right]$$

$$+ \frac{1}{z^2\varphi(z)^2}\mathrm{tr}\left[A\left(\Sigma+\frac{1}{\varphi(z)}I\right)^{-2}\Sigma\right]\cdot\mathrm{tr}\left[B\left(\Sigma+\frac{1}{\varphi(z)}I\right)^{-2}\Sigma\right]\cdot\frac{1}{n-\mathrm{df}_2(1/\varphi(z))}.$$

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 [25, Theorem 4.6].

Proposition 3.2. *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* (3.2),

$$(3.10)$$

$$\mathrm{tr}\left[AZ^\top(Z\Sigma Z^\top - nzI)^{-1}Z\right] \sim \mathrm{tr}\left[A\left(\Sigma+\frac{1}{\varphi(z)}I\right)^{-1}\right],$$

$$\mathrm{tr}\left[AZ^\top(Z\Sigma Z^\top - nzI)^{-1}ZBZ^\top(Z\Sigma Z^\top - nzI)^{-1}Z\right] \sim \mathrm{tr}\left[A\left(\Sigma+\frac{1}{\varphi(z)}I\right)^{-1}B\left(\Sigma+\frac{1}{\varphi(z)}I\right)^{-1}\right]$$

$$(3.11)$$

$$+ \frac{1}{\varphi(z)^2}\mathrm{tr}\left[A\left(\Sigma+\frac{1}{\varphi(z)}I\right)^{-2}\right]\cdot\mathrm{tr}\left[B\left(\Sigma+\frac{1}{\varphi(z)}I\right)^{-2}\right]\cdot\frac{1}{n-df_2(1/\varphi(z))}.$$

Like in [25, Theorem 4.6], (3.11) can be rewritten more intuitively as

$$Z^\top(Z\Sigma Z^\top - nzI)^{-1}ZBZ^\top(Z\Sigma Z^\top - nzI)^{-1}Z \sim \left(\Sigma+\frac{1}{\varphi(z)}I\right)^{-1}(B+\mu(z)I)\left(\Sigma+\frac{1}{\varphi(z)}I\right)^{-1}$$

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

Letting $\lambda \to 0$ for $\gamma > 1$. Following arguments from [16, 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 overparameterized 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 (3.8) and (3.9) instead, that is, $\mathrm{tr}\left[A(\widehat{\Sigma}+\lambda I)^{-1}\right] \sim \frac{\kappa(\lambda)}{\lambda}\mathrm{tr}\left[A(\Sigma+\kappa(\lambda)I)^{-1}\right]$ with $\frac{\kappa(\lambda)}{\lambda} \sim \frac{1}{1-\gamma}$ when λ goes to zero, and $\kappa(0)=0$, leading to

$$(3.12) \qquad\qquad \mathrm{tr}\left[A\widehat{\Sigma}^{-1}\right] \sim \frac{1}{1-d/n}\mathrm{tr}\left[A\Sigma^{-1}\right].$$

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 $x_1, \ldots, x_n$ are assumed deterministic, we obtain an expected excess risk, with Σ replaced with $\widehat{\Sigma}$, which considerably simplifies the analysis (see, e.g., [22]):

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

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 characterize this alignment [8].

This leads us to introduce the two classical different notions of degrees of freedom, $\mathrm{df}_1(\lambda) = \mathrm{tr}\left[\Sigma(\Sigma + \lambda I)^{-1}\right]$ and $\mathrm{df}_2(\lambda) = \mathrm{tr}\left[\Sigma^2(\Sigma + \lambda I)^{-2}\right]$, as key quantities [22]. 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 [33] 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 [16, 33, 5] with an explicit interpretation in terms of degrees of freedom.

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

$$(4.1) \qquad \hat{\theta} = (X^\top X + n\lambda I)^{-1}X^\top y = (X^\top X + n\lambda I)^{-1}X^\top X\theta_* + (X^\top X + n\lambda I)^{-1}X^\top\varepsilon$$

$$= (\widehat{\Sigma} + \lambda I)^{-1}\widehat{\Sigma}\theta_* + (\widehat{\Sigma} + \lambda I)^{-1}\frac{X^\top\varepsilon}{n}.$$

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 4.1. *Assume* (A1), (A2), (A3), *and* (A4). *For the ridge regression estimator in* (4.1), *we have*

$$\mathbb{E}_\varepsilon[\mathcal{R}^{(\mathrm{var})}(\hat{\theta})] \sim \frac{\sigma^2}{n}df_2(\kappa(\lambda)) \cdot \frac{1}{1 - \frac{1}{n}df_2(\kappa(\lambda))},$$

$$\mathcal{R}^{(\mathrm{bias})}(\hat{\theta}) \sim \kappa(\lambda)^2\theta_*^\top\Sigma(\Sigma + \kappa(\lambda)I)^{-2}\theta_* \cdot \frac{1}{1 - \frac{1}{n}df_2(\kappa(\lambda))}$$

with $\kappa(\lambda)$ *related to* λ *by* $\kappa(\lambda)(1 - \frac{1}{n}df_1(\kappa(\lambda))) \sim \lambda$.

Proof. The variance term is exactly the same as the one from [16], and we simply provide here a reinterpretation with degrees of freedom. We obtain it by taking expectations starting from (4.1) to get $\mathbb{E}_\varepsilon[\mathcal{R}^{(\mathrm{var})}(\hat{\theta})] = \frac{\sigma^2}{n}\mathrm{tr}\,[\Sigma(\widehat{\Sigma} + \lambda I)^{-2}\widehat{\Sigma}]$. We can then use (3.8) and (3.9) with $A = I$, $B = \Sigma$, and $z = -\lambda$ to get, using $\kappa(\lambda)\mathrm{tr}\,[\Sigma(\Sigma + \kappa(\lambda)I)^{-2}] = \mathrm{df}_1(\kappa(\lambda)) - \mathrm{df}_2(\kappa(\lambda))$,

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

For the bias term, we have

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

We then apply (3.9) with $A = \Sigma$ and $B = \theta_*\theta_*^\top$, which applies because of assumption (A4), to get

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

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, and the fixed and random design analyses lead to the same expressions. For vanishing λ in section 5, the variance term will only be able to go to zero when the two degrees of freedom differ significantly from each other.

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 [18]. We consider first the underparameterized case ($\gamma < 1$) and then the overparameterized one ($\gamma > 1$).

5.1. Underparameterized 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

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

Using (3.12), we obtain the classical equivalent $\sigma^2 \frac{\gamma}{1-\gamma} \sim \sigma^2 \frac{d}{n-d}$, as derived, e.g., in [19]. Note that for Gaussian data, this is, in fact, almost an equality, that is, $\mathbb{E}_\varepsilon[\mathcal{R}^{(\text{var})}(\hat{\theta})] = \sigma^2 \frac{d}{n-d-1}$ for $n > d+1$.

5.2. Overparameterized 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 [19] (in a nonasymptotic framework), here with an interpretation in terms of degrees of freedom. We obtain, with $\kappa(0)$ such that $\text{df}_1(\kappa(0)) = n$,

$$\mathbb{E}_\varepsilon[\mathcal{R}^{(\text{var})}(\hat{\theta})] \sim \frac{\sigma^2}{n} \text{df}_2(\kappa(0)) \cdot \frac{1}{1 - \frac{1}{n}\text{df}_2(\kappa(0))},$$

$$\mathcal{R}^{(\text{bias})}(\hat{\theta}) \sim \kappa(0)^2 \theta_*^\top \Sigma (\Sigma + \kappa(0)I)^{-2} \theta_* \cdot \frac{1}{1 - \frac{1}{n}\text{df}_2(\kappa(0))}.$$

Following [4, 19], we can try to understand when the overparameterized limit with no regularization makes statistical sense, with two questions in mind: (1) does it lead to catastrophic overfitting? and (2) can it lead to a good performance? The answers to these questions will depend on how $\text{df}_1(\kappa(\lambda))$ and $\text{df}_2(\kappa(\lambda))$ are related. Since $\text{df}_1(\kappa(\lambda)) = n$, we have $\text{df}_2(\kappa(\lambda)) \leqslant \text{df}_1(\kappa(\lambda)) = n$, 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 underparameterized problems (see also section 6).

Unbalanced degrees of freedom. If $\text{df}_2(\kappa(\lambda)) \ll \text{df}_1(\kappa(\lambda)) = n$, then the variance term can indeed go to zero when n tends to infinity. This happens only in particular situations thoroughly described by [4, 19, 36].

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, 35].

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). See [28, Theorem 4.4] for a partial similar result for the bias term.

Proposition 6.1. *Assume* (A1), (A2), (A3), (A4), (A5), (A6). *For the minimum norm least-squares estimator $\hat{\theta}$ based on random projections, we have for the underparameterized regime ($m < n$)*

$$\mathbb{E}_\varepsilon[\mathcal{R}^{\mathrm{var}}(\hat{\theta})] \sim \frac{\sigma^2 m}{n - m} = \frac{1}{1 - \frac{m}{n}} \cdot \frac{\sigma^2 m}{n},$$

$$\mathcal{R}^{\mathrm{bias}}(\hat{\theta}) \sim \frac{1}{1 - \frac{m}{n}} \cdot \kappa_m \theta_*^\top \Sigma (\Sigma + \kappa_m I)^{-1} \theta_*$$

with κ_m defined by $df_1(\kappa_m) \sim m$. In the overparameterized regime, we get, for κ_n such that $df_1(\kappa_n) \sim n$,

$$\mathbb{E}_\varepsilon[\mathcal{R}^{\mathrm{var}}(\hat{\theta})] \sim \frac{\sigma^2}{n} \cdot \frac{df_2(\kappa_n)}{1 - \frac{1}{n} df_2(\kappa_n)} + \sigma^2 \frac{n}{m - n},$$

$$\mathcal{R}^{\mathrm{bias}}(\hat{\theta}) \sim \kappa_n^2 \theta_*^\top \Sigma (\Sigma + \kappa_n I)^{-2} \theta_* \cdot \frac{1}{1 - \frac{1}{n} df_2(\kappa_n)} + \kappa_n \theta_*^\top \Sigma (\Sigma + \kappa_n I)^{-1} \theta_* \cdot \frac{n}{m - n}.$$

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 [16, Lemma 6.2]. We thus consider

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

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

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

$$\begin{aligned}
\mathbb{E}_\varepsilon[\mathcal{R}^{(\mathrm{var})}(\hat{\theta})] &= \sigma^2 \mathrm{tr}\,[XS(S^\top X^\top X S + n\lambda I)^{-1} S^\top \Sigma S(S^\top X^\top X S + n\lambda I)^{-1} S^\top X^\top] \\
&= \sigma^2 \mathrm{tr}\,[S^\top \Sigma S(S^\top X^\top X S + n\lambda I)^{-2} S^\top X^\top X S],
\end{aligned}$$
(6.1)

while, for the bias, we have

$$\begin{aligned}
\mathcal{R}^{(\mathrm{bias})}(\hat{\theta}) &= (M\theta_* - \theta_*)^\top \Sigma (M\theta_* - \theta_*) = \theta_*^\top \Sigma \theta_* + \theta_*^\top M^\top \Sigma M \theta_* - 2\theta_*^\top M^\top \Sigma \theta_* \\
&= \theta_*^\top \Sigma \theta_* - 2\theta_*^\top X^\top X S(S^\top X^\top X S + n\lambda I)^{-1} S^\top \Sigma \theta_* \\
&\quad + \theta_*^\top X^\top X S(S^\top X^\top X S + n\lambda I)^{-1} S^\top \Sigma S(S^\top X^\top X S + n\lambda I)^{-1} S^\top X^\top X \theta_*.
\end{aligned}$$
(6.2)

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 underparameterized regime and vice versa for the overparameterized regime.

Underparameterized 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 underparameterized results apply, and we get the following for the variance term (first term above), for S fixed, where we can directly consider $\lambda = 0$ (because of cancellations):

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

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

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

Using the matrix inversion lemma, we get

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

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

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

To find expansions of the red terms above, we can directly use the results from section 3.3, using (3.10) with $A = C\Sigma^{1/2}\theta_*\theta_*^\top \Sigma^{1/2}$, and (3.11) 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 Proposition 3.2 since $C = T T^\top = \Sigma^{1/2} S S^\top \Sigma^{1/2}$ has almost surely a limiting spectral measure and the resulting needed traces involving the matrices A and B have well-defined limits. We get, for any λ,

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

---

[3] We use the notation $\tilde\kappa(\lambda)$ to avoid confusion with $\kappa\lambda$, which is defined from the covariance matrix Σ.

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

$$
\mathcal{R}^{(\mathrm{bias})}(\hat{\theta}) \sim \theta_*^\top\Sigma^{1/2}(I - \Sigma^{1/2}S(S^\top\Sigma S)^{-1}S^\top\Sigma^{1/2})^2\Sigma^{1/2}\theta_*
$$
$$
+ \frac{\theta_*^\top\Sigma^{1/2}(I - T(T^\top T)^{-1}T^\top)^2\Sigma^{1/2}\theta_* \cdot m}{n - m}
$$
$$
(6.3) \qquad = \theta_*^\top(\Sigma - \Sigma S(S^\top\Sigma S)^{-1}S^\top\Sigma)\theta_* \cdot \left(1 + \frac{m}{n-m}\right).
$$

Underparameterized 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 Proposition 3.2 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 (6.3), we get the desired result: $\mathcal{R}^{\mathrm{bias}}(\hat{\theta}) \sim \frac{1}{1-m/n}\kappa_m\theta_*^\top\Sigma(\Sigma + \kappa_m I)^{-1}\theta_*$.

Overparameterized regime: Expansion with respect to S. We have, from (6.1),

$$
\mathbb{E}_\varepsilon[\mathcal{R}^{(\mathrm{var})}(\hat{\theta})] = \frac{\sigma^2}{n}\mathrm{tr}\left(\Sigma S(S^\top\widehat{\Sigma}S + \lambda I)^{-1}S^\top\widehat{\Sigma}S(S^\top\widehat{\Sigma}XS + \lambda I)^{-1}S^\top\right).
$$

To obtain an expansion of the red term, we can use Proposition 3.2 with covariance matrix $\widehat{\Sigma}$ and thus degrees of freedom and $\tilde{\kappa}$ associated to $\widehat{\Sigma}$:

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

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 $\tilde{\kappa}(\lambda)^2(\widehat{\Sigma} + \tilde{\kappa}(\lambda)I)^{-2} = n^2\tilde{\kappa}(\lambda)^2(X^\top X + n\tilde{\kappa}(\lambda)I)^{-2} = (I - X^\top(XX^\top + n\tilde{\kappa}(\lambda)I)^{-1}X)^2 \to (I - X^\top(XX^\top)^{-1}X)^2 = (I - X^\top(XX^\top)^{-1}X)$, and we thus get

$$
(6.4) \qquad \mathbb{E}_\varepsilon[\mathcal{R}^{(\mathrm{var})}(\hat{\theta})] \sim \sigma^2\mathrm{tr}\left[\Sigma X^\top(XX^\top)^{-2}X\right] + \frac{\mathrm{tr}\left[\Sigma(I - X^\top(XX^\top)^{-1}X)\right] \cdot \mathrm{tr}\left[(XX^\top)^{-1}\right]}{m - n}.
$$

We can now take care of the (squared) bias term with the same technique, with $\lambda \to 0$, starting from (6.2):

$$
\mathcal{R}^{(\mathrm{bias})}(\hat{\theta}) = \theta_*^\top\Sigma\theta_* - 2\theta_*^\top\Sigma^{1/2}S(S^\top\widehat{\Sigma}S + \lambda I)^{-1}S^\top\widehat{\Sigma}\theta_*
$$
$$
+ \theta_*^\top\widehat{\Sigma}S(S^\top\widehat{\Sigma}S + \lambda I)^{-1}S^\top\Sigma S(S^\top X^\top XS + n\lambda I)^{-1}S^\top\widehat{\Sigma}\theta_*
$$
$$
\sim \theta_*^\top\Sigma\theta_* - 2\theta_*^\top\Sigma^{1/2}(\widehat{\Sigma} + \tilde{\kappa}(\lambda)I)^{-1}\widehat{\Sigma}\theta_* + \theta_*^\top\widehat{\Sigma}(\widehat{\Sigma} + \tilde{\kappa}(\lambda)I)^{-1}\Sigma(\widehat{\Sigma} + \tilde{\kappa}(\lambda)I)^{-1}\widehat{\Sigma}\theta_*
$$
$$
+ \tilde{\kappa}(\lambda)^2\frac{\mathrm{tr}\left[\Sigma(\widehat{\Sigma} + \tilde{\kappa}(\lambda)I)^{-2}\right] \cdot \theta_*^\top\widehat{\Sigma}(\widehat{\Sigma} + \tilde{\kappa}(\lambda)I)^{-2}\widehat{\Sigma}\theta_*}{m - \widetilde{\mathrm{df}}_2(\tilde{\kappa}(\lambda))}
$$
$$
\sim \|\Sigma^{1/2}(I - X^\top(XX^\top + \tilde{\kappa}(\lambda)I)^{-1}X)\theta_*\|_2^2
$$
$$
+ \tilde{\kappa}(\lambda)^2\frac{\mathrm{tr}\left[\Sigma(X^\top X + \tilde{\kappa}(\lambda)I)^{-2}\right] \cdot \theta_*^\top X^\top X(X^\top X + \tilde{\kappa}(\lambda)I)^{-2}X^\top X\theta_*}{m - \widetilde{\mathrm{df}}_2(\tilde{\kappa}(\lambda))}
$$
$$
\sim \theta_*^\top(I - X^\top(XX^\top)^{-1}X)\Sigma(I - X^\top(XX^\top)^{-1}X)\theta_*
$$
$$
(6.5) \qquad + \frac{1}{m-n}\theta_*^\top X^\top(XX^\top)^{-1}X\theta_* \cdot \mathrm{tr}\left[\Sigma(I - X^\top(XX^\top)^{-1}X)\right].
$$

Overparameterized 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 Proposition 3.2, with (6.4) and (6.5),

$$\mathbb{E}[\mathcal{R}^{\mathrm{var}}(\hat{\theta})] \sim \sigma^2 \frac{\mathrm{df}_2(\kappa_n)}{\mathrm{df}_1(\kappa_n) - \mathrm{df}_2(\kappa_n)} + \sigma^2 \frac{n}{m - n},$$

$$\mathcal{R}^{\mathrm{bias}}(\hat{\theta}) \sim \kappa_n^2 \theta_*^\top \Sigma (\Sigma + \kappa_n I)^{-2} \theta_* \cdot \frac{\mathrm{df}_1(\kappa_n)}{\mathrm{df}_1(\kappa_n) - \mathrm{df}_2(\kappa_n)} + \kappa_n \theta_*^\top \Sigma (\Sigma + \kappa_n I)^{-1} \theta_* \cdot \frac{n}{m - n},$$

which is the desired result. ∎

We can make the following observations:

- In the underparameterized regime, we recover the traditional bias and variance terms[4] divided by $1 - \frac{m}{n}$, which leads to the expected catastrophic overfitting 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 $(1 - \frac{m}{n})^{-1}$. 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 underparameterized regime.
- In the overparameterized 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 $\hat{\theta} = X^\top (XX^\top)^{-1} y$, 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.[5]

Testing the asymptotic limit. We consider a fixed spectral measure $\mu = \pi_1 \delta_{\sigma_1} + \pi_2 \delta_{\sigma_2}$ already considered by [19, 36] 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 $\delta \in [0, 1]$,

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

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

---

[4] For the variance term, they match the results for ridge regression, up to df_2 being replaced by df_1, with a similar relationship for the bias term, where a slightly different notion of bias is used.

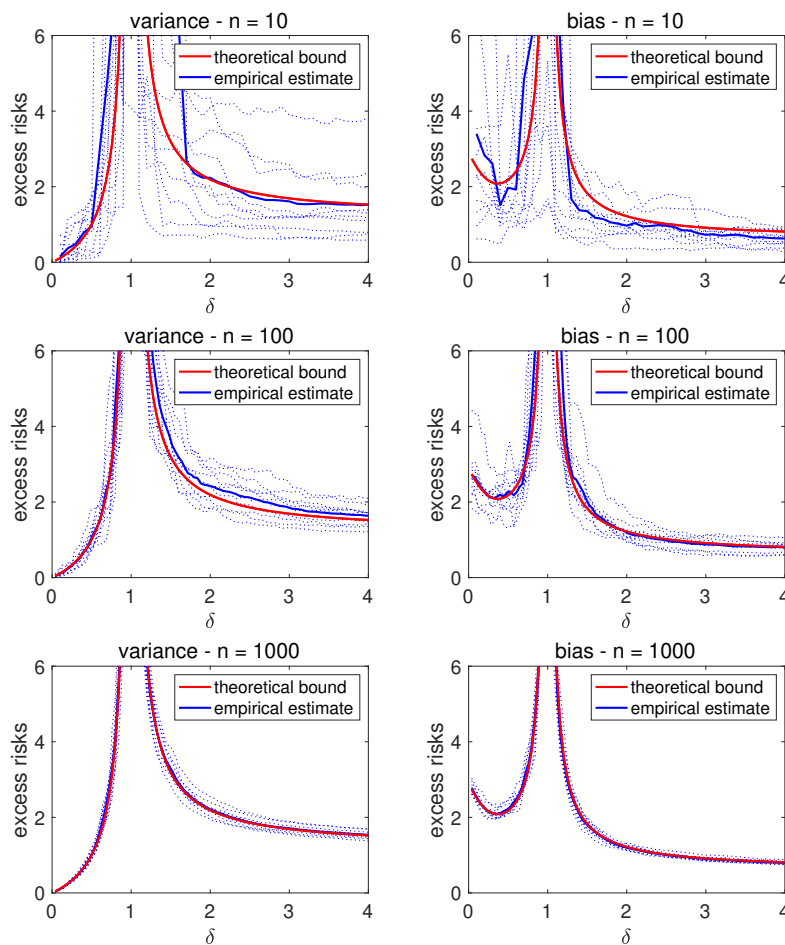[5] MATLAB code to reproduce figures can be downloaded from https://www.di.ens.fr/~fbach/dd_rp.zip.

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.*

$O(1/\sqrt{n})$), 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 $k \in \{1, \dots, d\}$ (nonisotropic) or constant (isotropic). We normalize the matrix so that $\operatorname{tr}(\Sigma) = 1$. We generate a vector $\theta_* \in \mathbb{R}^d$ from 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 $\kappa_m^{-1} = \mathbb{E}[\operatorname{tr}[(S^\top \Sigma S)^{-1}]]$, using an average over 40 replications.

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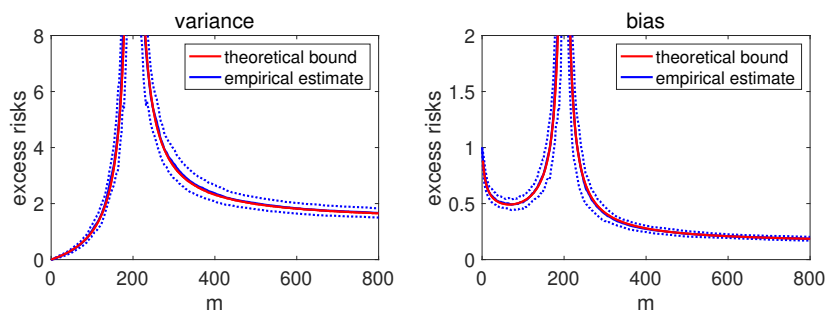
Figure 4. *(Left) Variance with $\sigma = 1$ and $\operatorname{tr}(\Sigma) = 1$. (Right) Bias with $\theta_*^\top \Sigma \theta_* = 1$. We consider $n = 200$, $d = 400$, with Z and S sampled from Rademacher random variables, and eigenvalues of Σ proportional to $1/k$. For the empirical curve, we plot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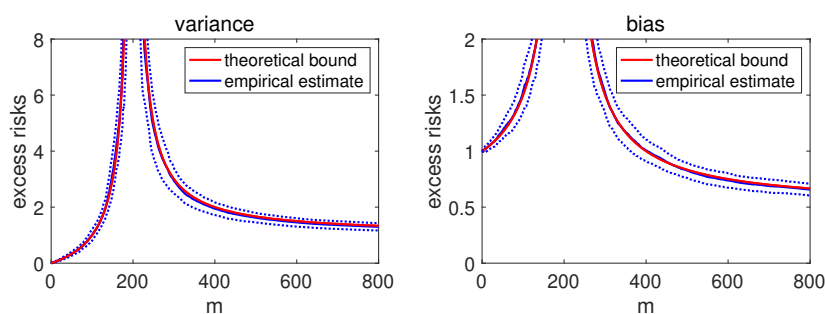

Figure 5. *(Left) Variance with $\sigma = 1$ and $\operatorname{tr}(\Sigma) = 1$. (Right) Bias with $\theta_*^\top \Sigma \theta_* = 1$. We consider $n = 200$, $d = 400$, with Z and S sampled from Rademacher random variables and uniform eigenvalues for Σ. For the empirical curve, we plot the average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

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 [27, 30], characterizing how quickly our asymptotic analysis kicks in using tools from [1], looking at more general random projection matrices [25], 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, 35]; see [9] for a recent work in this direction.

Appendix A. Random matrix theory results. In this appendix, we provide a sketch of proof for classical random matrix theory results presented in sections 3.1 and 3.2, with a proof for the new results from section 3.3. For more details, see [40, 2].

A.1. Self-consistency equation. We follow the proof of [39] 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,

$$\operatorname{tr}[X^\top X(X^\top X - nzI)^{-1}] = \sum_{i=1}^n \operatorname{tr}\left[x_i x_i^\top \left(\sum_{j\neq i} x_j x_j^\top - nzI + x_i x_i^\top\right)^{-1}\right]$$

$$= \sum_{i=1}^n \frac{x_i^\top (\sum_{j\neq i} x_j x_j^\top - nzI)^{-1} x_i}{1 + x_i^\top (\sum_{j\neq i} x_j x_j^\top - nzI)^{-1} x_i}$$

$$= n - \sum_{i=1}^n \frac{1}{1 + x_i^\top (\sum_{j\neq i} x_j x_j^\top - nzI)^{-1} x_i}.$$

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

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

We also have more generally

$$(A.2) \quad \widehat{\Sigma}(\widehat{\Sigma} - zI)^{-1} = \sum_{i=1}^n x_i x_i^\top \left(\sum_{j\neq i} x_j x_j^\top - nzI + x_i x_i^\top\right)^{-1} = \sum_{i=1}^n \frac{x_i x_i^\top (\sum_{j\neq i} x_j x_j^\top - nzI)^{-1}}{1 + x_i^\top (\sum_{j\neq i} x_j x_j^\top - nzI)^{-1} x_i}.$$

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

$$(\widehat{\Sigma} - zI)^{-1} - (-z\widehat{\varphi}(z)\Sigma - zI)^{-1} = (z\widehat{\varphi}(z)\Sigma + zI)^{-1}\left(\widehat{\Sigma} - (-z\widehat{\varphi}(z)\Sigma)\right)(\widehat{\Sigma} - zI)^{-1}$$

$$= (z\widehat{\varphi}(z)\Sigma + zI)^{-1}\left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\top - (-z\widehat{\varphi}(z)\Sigma)\right)(\widehat{\Sigma} - zI)^{-1}$$

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

$$= (z\widehat{\varphi}(z)\Sigma + zI)^{-1} \frac{1}{n} \sum_{i=1}^n \frac{x_i x_i^\top (\widehat{\Sigma}_{-i} - zI)^{-1} - \Sigma(\widehat{\Sigma} - zI)^{-1}}{1 + x_i^\top (n\widehat{\Sigma}_{-i} - nzI)^{-1} x_i}.$$

We thus get

$$(A.3) \quad (\widehat{\Sigma} - zI)^{-1} = (-z\widehat{\varphi}(z)\Sigma - zI)^{-1}\left(I - \frac{1}{n}\sum_{i=1}^n \frac{x_i x_i^\top (\widehat{\Sigma}_{-i} - zI)^{-1} - \Sigma(\widehat{\Sigma} - zI)^{-1}}{1 + x_i^\top (n\widehat{\Sigma}_{-i} - nzI)^{-1} x_i}\right)$$

$$= (-z\widehat{\varphi}(z)\Sigma - zI)^{-1}(I - \Delta).$$

The main property we will leverage is that Δ will almost certainly be "negligible." For this, we need that $\operatorname{tr}[(-z\widehat{\varphi}(z)\Sigma - zI)^{-1}\Delta] = o(d)$, and we simply need to study each of

the n terms and show that they are $o(d)$. The key is that $(\widehat{\Sigma}_{-i} - zI)^{-1}$ is independent of x_i and that for any deterministic (or independent random bounded) matrix, $\text{tr}\,[(z_i z_i^\top - I)N]$ is small enough with a strong probabilistic control [40, 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 on the Hanson–Wright inequality [38], which leads to, for a constant $c > 0$,

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

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 [39] for a detailed proof.

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

$$(A.4) \qquad \text{tr}\,[(\widehat{\Sigma} - zI)^{-1}] \sim \frac{-1}{z\widehat{\varphi}(z)}\text{tr}\,\Big[\Big(\Sigma + \frac{1}{\widehat{\varphi}(z)}I\Big)^{-1}\Big] = \frac{-d}{z} + \frac{1}{z}\text{tr}\,\Big[\Sigma\Big(\Sigma + \frac{1}{\widehat{\varphi}(z)}I\Big)^{-1}\Big].$$

Third step. We can rewrite

$$\text{tr}\,[(\widehat{\Sigma} - zI)^{-1}] = \frac{1}{z}\text{tr}\,[(zI - \widehat{\Sigma} + \widehat{\Sigma})(\widehat{\Sigma} - zI)^{-1}]$$

$$= -\frac{d}{z} + \frac{1}{z}\text{tr}\,[\widehat{\Sigma}(\widehat{\Sigma} - zI)^{-1}] = -\frac{d}{z} + \frac{1}{z}\text{tr}\,[XX^\top(XX^\top - nzI)^{-1}]$$

$$(A.5) \qquad = -\frac{d}{z} + \frac{1}{z}\text{tr}\,[(XX^\top - nzI + nzI)(XX^\top - nzI)^{-1}] = \frac{n-d}{z} + n\widehat{\varphi}(z).$$

Following [39] and combining (A.4) and (A.5), this leads to $\widehat{\varphi}(z) \sim \varphi(z)$, with

$$(A.6) \qquad\qquad \varphi(z) + \frac{1}{z} = \frac{1}{nz}\text{tr}\,\Big[\Sigma\Big(\Sigma + \frac{1}{\varphi(z)}I\Big)^{-1}\Big],$$

which is the desired self-consistent equation in (3.3) in section 3.2. Note that the quantity $\varphi(z)$ still depends on n.

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

$$(A.7) \qquad\qquad \text{tr}\,[\widehat{\Sigma}(\widehat{\Sigma} - zI)^{-1}] \sim \text{tr}\,\Big[\Sigma\Big(\Sigma + \frac{1}{\varphi(z)}I\Big)^{-1}\Big].$$

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

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

leading to $\lambda\varphi(-\lambda) = 1 - \frac{1}{n}\text{df}_1\Big(\frac{1}{\varphi(-\lambda)}\Big)$, 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 Propositions 3.1 and 3.2. Following [16], we start with an asymptotic equivalent based on differentiation (see formal justification in [16]). See [12, 25] for similar results based more strongly on differentiation (which is only used here to derive an equivalent for $\mathrm{tr}\,[(\widehat{\Sigma} - zI)^{-2}]$).

Using differentiation. We have, by differentiating (A.6) with respect to z,

$$\varphi(z) + z\varphi'(z) = \frac{1}{n}\mathrm{tr}\left[\Sigma\left(\Sigma + \frac{1}{\varphi(z)}I\right)^{-2}\right]\frac{\varphi'(z)}{\varphi(z)^2},$$

which leads to $\frac{\varphi(z)}{\varphi'(z)} = \frac{1}{n}\mathrm{tr}\left[\Sigma\left(\Sigma + \frac{1}{\varphi(z)}I\right)^{-2}\right]\frac{1}{\varphi(z)^2} - z$. Thus, differentiating (A.7) with respect to z and using the bound on $\frac{\varphi(z)}{\varphi'(z)}$ above, we get

$$\mathrm{tr}\,[\widehat{\Sigma}(\widehat{\Sigma} - zI)^{-2}] \sim \mathrm{tr}\left[\Sigma\left(\Sigma + \frac{1}{\varphi(z)}I\right)^{-2}\right]\frac{\varphi'(z)}{\varphi(z)^2} = \frac{n\,\mathrm{tr}\,[\Sigma(\Sigma + \frac{1}{\varphi(z)}I)^{-2}]}{\mathrm{tr}\,[\Sigma(\Sigma + \frac{1}{\varphi(z)}I)^{-2}]\frac{1}{\varphi(z)} - nz\varphi(z)}$$

$$= \frac{n\,\mathrm{tr}\,[\Sigma(\Sigma + \frac{1}{\varphi(z)}I)^{-2}]}{\mathrm{tr}\,[\Sigma(\Sigma + \frac{1}{\varphi(z)}I)^{-2}]\frac{1}{\varphi(z)} + n - \mathrm{tr}\,[\Sigma(\Sigma + \frac{1}{\varphi(z)}I)^{-1}]} = \frac{n\,\mathrm{tr}\,[\Sigma(\Sigma + \frac{1}{\varphi(z)}I)^{-2}]}{n - \mathrm{tr}\,[\Sigma^2(\Sigma + \frac{1}{\varphi(z)}I)^{-2}]}.$$

This leads to the asymptotic equivalent

$$\mathrm{tr}\,[(\widehat{\Sigma} - zI)^{-2}] = \frac{1}{z}\mathrm{tr}\,[(zI - \widehat{\Sigma} + \widehat{\Sigma})(\widehat{\Sigma} - zI)^{-2}] = \frac{1}{z}\mathrm{tr}\,[\widehat{\Sigma}(\widehat{\Sigma} - zI)^{-2}] - \frac{1}{z}\mathrm{tr}\,[(\widehat{\Sigma} - zI)^{-1}]$$

$$\text{(A.8)} \qquad \sim \frac{1}{z}\frac{n\,\mathrm{tr}\,[\Sigma(\Sigma + \frac{1}{\varphi(z)}I)^{-2}]}{n - \mathrm{tr}\,[\Sigma^2(\Sigma + \frac{1}{\varphi(z)}I)^{-2}]} - \frac{1}{z}\mathrm{tr}\,[(- z\varphi(z)\Sigma - zI)^{-1}],$$

which we will need later.

Proof of (3.6) and (3.8). We now first show

$$\mathrm{tr}\,[A(\widehat{\Sigma} - zI)^{-1}] \sim \mathrm{tr}\,[A(- z\widehat{\varphi}(z)\Sigma - zI)^{-1}] \sim \mathrm{tr}\,[A(- z\varphi(z)\Sigma - zI)^{-1}]$$

$$= \frac{-1}{z\varphi(z)}\mathrm{tr}\left[A\left(\Sigma + \frac{1}{\varphi(z)}I\right)^{-1}\right],$$

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 (A.3),

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

which is negligible as soon as $\|A\|_{\mathrm{op}}$ is bounded (using the same arguments as in Appendix A.1). We can then express $\mathrm{tr}\,[A(\Sigma + \frac{1}{\varphi(z)}I)^{-1}]$ as $\int_0^{+\infty}\frac{d\nu_A(\sigma)}{\sigma + \frac{1}{\varphi(z)}}$. This leads to the desired result in Proposition 3.1.

Proof of (3.7) and (3.9). For the quadratic form, we have for any matrices A and B, still using (A.3),

$$\operatorname{tr}\left[A(\widehat{\Sigma}-zI)^{-1}B(\widehat{\Sigma}-zI)^{-1}\right]$$

$$= \operatorname{tr}\left[A(-z\widehat{\varphi}(z)\Sigma-zI)^{-1}(I-\Delta)B(-z\widehat{\varphi}(z)\Sigma-zI)^{-1}(I-\Delta)\right]$$

$$= \operatorname{tr}\left[A(-z\widehat{\varphi}(z)\Sigma-zI)^{-1}B(-z\widehat{\varphi}(z)\Sigma-zI)^{-1}\right]$$

$$\quad + \operatorname{tr}\left[A(-z\widehat{\varphi}(z)\Sigma-zI)^{-1}\Delta B(-z\widehat{\varphi}(z)\Sigma-zI)^{-1}\Delta\right]$$

$$\quad - \operatorname{tr}\left[A(-z\widehat{\varphi}(z)\Sigma-zI)^{-1}\Delta B(-z\widehat{\varphi}(z)\Sigma-zI)^{-1}\right]$$

$$\quad - \operatorname{tr}\left[A(-z\widehat{\varphi}(z)\Sigma-zI)^{-1}B(-z\widehat{\varphi}(z)\Sigma-zI)^{-1}\Delta\right].$$

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

$$\operatorname{tr}\left[A(-z\widehat{\varphi}(z)\Sigma-zI)^{-1}\Delta B(-z\widehat{\varphi}(z)\Sigma-zI)^{-1}\Delta\right]$$

$$= \tfrac{1}{n^2}\sum_{i,j=1}^{n}\operatorname{tr}\left[A(-z\widehat{\varphi}(z)\Sigma-zI)^{-1}\tfrac{(x_ix_i^\top-\Sigma)(\widehat{\Sigma}_{-i}-zI)^{-1}}{1+x_i^\top(n\widehat{\Sigma}_{-i}-nzI)^{-1}x_i}B(-z\widehat{\varphi}(z)\Sigma-zI)^{-1}\tfrac{(x_jx_j^\top-\Sigma)(\widehat{\Sigma}_{-j}-zI)^{-1}}{1+x_j^\top(n\widehat{\Sigma}_{-j}-nzI)^{-1}x_j}\right]$$

$$= \tfrac{1}{n^2}\sum_{i,j=1}^{n}\tfrac{\operatorname{tr}\left[A(-z\widehat{\varphi}(z)\Sigma-zI)^{-1}(x_ix_i^\top-\Sigma)(\widehat{\Sigma}_{-i}-zI)^{-1}B(-z\widehat{\varphi}(z)\Sigma-zI)^{-1}(x_jx_j^\top-\Sigma)(\widehat{\Sigma}_{-j}-zI)^{-1}\right]}{\left(1+x_i^\top(n\widehat{\Sigma}_{-i}-nzI)^{-1}x_i\right)\left(1+x_j^\top(n\widehat{\Sigma}_{-j}-nzI)^{-1}x_j\right)}.$$

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

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

To study its asymptotic limit, we need to characterize the asymptotic equivalent of the quantity $\operatorname{tr}\left[C(x_ix_i^\top-\Sigma)D(x_ix_i^\top-\Sigma)\right] = \operatorname{tr}\left[\Sigma^{1/2}C\Sigma^{1/2}(z_iz_i^\top-I)\Sigma^{1/2}D\Sigma^{1/2}(z_iz_i^\top-I)\right]$ 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

$$\operatorname{tr}\left[M(z_iz_i^\top-I)N(z_iz_i^\top-I)\right]-\operatorname{tr}(M)\operatorname{tr}(N)$$

$$= (z_i^\top Mz_i-\operatorname{tr}(M))(z_i^\top Nz_i-\operatorname{tr}(N))$$

$$\quad + \operatorname{tr}(M)(z_i^\top Nz_i-\operatorname{tr}(N)) + \operatorname{tr}(N)(z_i^\top Mz_i-\operatorname{tr}(M))$$

$$\quad - \operatorname{tr}\left[(MN+NM)(z_iz_i^\top-I)\right]$$

$$= O_p(\|M\|_F\cdot\|N\|_F+\operatorname{tr}(M)\|N\|_F+\operatorname{tr}(N)\|M\|_F+\|NM\|_F),$$

using the property from Appendix A.1, obtained 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

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

We thus overall have

$$\operatorname{tr}\left[A(-z\widehat{\varphi}(z)\Sigma-zI)^{-1}\Delta B\Delta^\top(-z\widehat{\varphi}(z)\Sigma-zI)^{-1}\right]$$

$$\sim \operatorname{tr}\left[(\widehat{\Sigma}-zI)^{-1}A(-z\widehat{\varphi}(z)\Sigma-zI)^{-1}\Sigma\right]\cdot\operatorname{tr}\left[(\widehat{\Sigma}-zI)^{-1}B(-z\widehat{\varphi}(z)\Sigma-zI)^{-1}\Sigma\right]\cdot\square$$

$$\sim \operatorname{tr}\left[A(-z\varphi(z)\Sigma-zI)^{-2}\Sigma\right]\cdot\operatorname{tr}\left[B(-z\varphi(z)\Sigma-zI)^{-2}\Sigma\right]\cdot\square$$

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

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

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

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

which allows us to compute an equivalent of $\square$, as, using (A.8), with $z\varphi(z) \sim \mathrm{df}_1(1/\varphi(z)) - \frac{1}{n}$:

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

$$\sim \frac{\frac{1}{z}\frac{n\operatorname{tr}\left[\Sigma\left(\Sigma + \frac{1}{\varphi(z)}I\right)^{-2}\right]}{n - \mathrm{df}_2(1/\varphi(z))} - \frac{1}{z}\operatorname{tr}\left[(-z\varphi(z)\Sigma - zI)^{-1}\right] - \operatorname{tr}\left[(-z\varphi(z)\Sigma - zI)^{-2}\right]}{\left(\operatorname{tr}\left[(-z\varphi(z)\Sigma - zI)^{-2}\Sigma\right]\right)^2}$$

$$\sim \frac{\frac{1}{z}\frac{n\operatorname{tr}\left[\Sigma\left(\Sigma + \frac{1}{\varphi(z)}I\right)^{-2}\right]}{n - \mathrm{df}_2(1/\varphi(z))} + \frac{1}{z^2\varphi(z)}\operatorname{tr}\left[\left(\Sigma + \frac{1}{\varphi(z)}I\right)^{-1}\right] - \frac{1}{z^2\varphi(z)}\operatorname{tr}\left[\frac{1}{\varphi(z)}\left(\Sigma + \frac{1}{\varphi(z)}I\right)^{-2}\right]}{\left(\frac{1}{z^2\varphi(z)}\operatorname{tr}\left[\frac{1}{\varphi(z)}\left(\Sigma + \frac{1}{\varphi(z)}I\right)^{-2}\Sigma\right]\right)^2}$$

$$\sim \frac{\frac{1}{z}\frac{n\operatorname{tr}\left[\Sigma\left(\Sigma + \frac{1}{\varphi(z)}I\right)^{-2}\right]}{n - \mathrm{df}_2(1/\varphi(z))} + \frac{1}{z^2\varphi(z)}\operatorname{tr}\left[\Sigma\left(\Sigma + \frac{1}{\varphi(z)}I\right)^{-2}\right]}{\left(\frac{1}{z^2\varphi(z)}\operatorname{tr}\left[\frac{1}{\varphi(z)}\left(\Sigma + \frac{1}{\varphi(z)}I\right)^{-2}\Sigma\right]\right)^2} \sim \frac{\frac{1}{z}\frac{n}{n - \mathrm{df}_2(1/\varphi(z))} + \frac{1}{z^2\varphi(z)}}{\left(\frac{1}{z^2\varphi(z)}\right)^2\frac{1}{\varphi(z)}\operatorname{tr}\left[\frac{1}{\varphi(z)}\left(\Sigma + \frac{1}{\varphi(z)}I\right)^{-2}\Sigma\right]}$$

$$\sim \frac{\frac{nz\varphi(z)}{n - \mathrm{df}_2(1/\varphi(z))} + 1}{\frac{1}{z^2\varphi(z)^2}\operatorname{tr}\left[\frac{1}{\varphi(z)}\left(\Sigma + \frac{1}{\varphi(z)}I\right)^{-2}\Sigma\right]} = \frac{\frac{\mathrm{df}_1(1/\varphi(z)) - n}{n - \mathrm{df}_2(1/\varphi(z))} + 1}{\frac{1}{z^2\varphi(z)^2}\operatorname{tr}\left[\frac{1}{\varphi(z)}\left(\Sigma + \frac{1}{\varphi(z)}I\right)^{-2}\Sigma\right]}$$

$$\sim \frac{\frac{\mathrm{df}_1(1/\varphi(z)) - n}{n - \mathrm{df}_2(1/\varphi(z))} + 1}{\frac{1}{z^2\varphi(z)^2}(\mathrm{df}_1(1/\varphi(z)) - \mathrm{df}_2(1/\varphi(z)))} = \frac{z^2\varphi(z)^2}{n - \mathrm{df}_2(1/\varphi(z))}.$$

Overall, we get

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

which is (3.9).

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

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

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

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

which is (3.7).

REFERENCES

[1] Z. BAI AND J. W. SILVERSTEIN, *CLT for linear spectral statistics of large-dimensional sample covariance matrices*, in Advances in Statistics, World Scientific, River Edge, NJ, 2008, pp. 281–333.

[2] Z. BAI AND J. W. SILVERSTEIN, *Spectral Analysis of Large Dimensional Random Matrices*, Springer Ser. Statist. 20, Springer, New York, 2010.

[3] Z. BAI AND Y.-Q. YIN, *Limit of the smallest eigenvalue of a large dimensional sample covariance matrix*, in Advances in Statistics, World Scientific, River Edge, NJ, 2008, pp. 108–127.

[4] P. L. BARTLETT, P. M. LONG, G. LUGOSI, AND A. TSIGLER, *Benign overfitting in linear regression*, Proc. Natl. Acad. Sci. USA, 117 (2020), pp. 30063–30070.

[5] P. L. BARTLETT, A. MONTANARI, AND A. RAKHLIN, *Deep learning: A statistical viewpoint*, Acta Numer., 30 (2021), pp. 87–201.

[6] M. BELKIN, D. HSU, S. MA, AND S. MANDAL, *Reconciling modern machine-learning practice and the classical bias–variance trade-off*, Proc. Natl. Acad. Sci. USA, 116 (2019), pp. 15849–15854.

[7] M. BELKIN, D. HSU, AND J. XU, *Two models of double descent for weak features*, SIAM J. Math. Data Sci., 2 (2020), pp. 1167–1180.

[8] A. CAPONNETTO AND E. DE VITO, *Optimal rates for regularized least-squares algorithm*, Found. Comput. Math., 7 (2007), pp. 331–368.

[9] X. CHEN, Y. ZENG, S. YANG, AND Q. SUN, *Sketched Ridgeless Linear Regression: The Role of Downsampling*, https://arxiv.org/abs/2302.01088, 2023.

[10] C. CHENG AND A. MONTANARI, *Dimension Free Ridge Regression*, https://arxiv.org/abs/2210.08571, 2022.

[11] H. CUI, B. LOUREIRO, F. KRZAKALA, AND L. ZDEBOROVÁ, *Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime*, Adv. Neural Inf. Process. Syst., 34 (2021), pp. 10131–10143.

[12] Y. DAR, D. LEJEUNE, AND R. G. BARANIUK, *The Common Intuition to Transfer Learning Can Win or Lose: Case Studies for Linear Regression*, https://arxiv.org/abs/2103.05621, 2021.

[13] E. DOBRIBAN AND S. LIU, *Asymptotics for sketching in least squares regression*, Adv. Neural Inf. Process. Syst., 32 (2019).

[14] E. DOBRIBAN AND Y. SHENG, *WONDER: Weighted one-shot distributed ridge regression in high dimensions*, J. Mach. Learn. Res., 21 (2020), pp. 2483–2534.

[15] E. DOBRIBAN AND Y. SHENG, *Distributed linear regression by averaging*, Ann. Statist., 49 (2021), pp. 918–943.

[16] E. DOBRIBAN AND S. WAGER, *High-dimensional asymptotics of prediction: Ridge regression and classification*, Ann. Statist., 46 (2018), pp. 247–279.

[17] M. GEIGER, A. JACOT, S. SPIGLER, F. GABRIEL, L. SAGUN, S. D'ASCOLI, G. BIROLI, C. HONGLER, AND M. WYART, *Scaling description of generalization with number of parameters in deep learning*, J. Stat. Mech. Theory Exp., 2020, 023401.

[18] S. GUNASEKAR, J. LEE, D. SOUDRY, AND N. SREBRO, *Characterizing implicit bias in terms of optimization geometry*, in Proceedings of the International Conference on Machine Learning, 2018.

[19] T. HASTIE, A. MONTANARI, S. ROSSET, AND R. J. TIBSHIRANI, *Surprises in high-dimensional ridgeless least squares interpolation*, Ann. Statist., 50 (2022), pp. 949–986.

[20] T. J. HASTIE AND R. J. TIBSHIRANI, *Generalized Additive Models*, Chapman and Hall, London, 1990.

[21] A. E. HOERL AND R. W. KENNARD, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics, 12 (1970), pp. 55–67.

[22] D. HSU, S. M. KAKADE, AND T. ZHANG, *Random design analysis of ridge regression*, in Proceedings of the Conference on Learning Theory, 2012.

[23] A. JACOT, B. SIMSEK, F. SPADARO, C. HONGLER, AND F. GABRIEL, *Implicit regularization of random feature models*, in Proceedings of the International Conference on Machine Learning, 2020.

[24] O. LEDOIT AND S. PÉCHÉ, *Eigenvectors of some large sample covariance matrix ensembles*, Probab. Theory Related Fields, 151 (2011), pp. 233–264.

[25] D. LEJEUNE, P. PATIL, H. JAVADI, R. G. BARANIUK, AND R. J. TIBSHIRANI, *Asymptotics of the Sketched Pseudoinverse*, https://arxiv.org/abs/2211.03751, 2022.

[26] Z. LIAO, R. COUILLET, AND M. W. MAHONEY, *A random matrix analysis of random Fourier features: Beyond the Gaussian kernel, a precise phase transition, and the corresponding double descent*, Adv. Neural Inf. Process. Syst., 33 (2020).

[27] Z. LIAO AND M. W. MAHONEY, *Hessian eigenspectra of more realistic nonlinear models*, Adv. Neural Inf. Process. Syst., 34 (2021).

[28] S. LIU AND E. DOBRIBAN, *Ridge regression: Structure, cross-validation, and sketching*, in Proceedings of the International Conference on Learning Representations, 2019.

[29] M. LOOG, T. VIERING, A. MEY, J. H. KRIJTHE, AND D. M. J. TAX, *A brief prehistory of double descent*, Proc. Natl. Acad. Sci. USA, 117 (2020), pp. 10625–10626.

[30] B. LOUREIRO, C. GERBELOT, M. REFINETTI, G. SICURO, AND F. KRZAKALA, *Fluctuations, bias, variance & ensemble of learners: Exact asymptotics for convex losses in high-dimension*, in Proceedings of the International Conference on Machine Learning, 2022.

[31] S. MEI AND A. MONTANARI, *The generalization error of random features regression: Precise asymptotics and the double descent curve*, Comm. Pure Appl. Math., 75 (2022), pp. 667–766.

[32] A. MONTANARI AND Y. ZHONG, *The interpolation phase transition in neural networks: Memorization and generalization under lazy training*, Ann. Statist., 50 (2022), pp. 2816–2847.

[33] J. MOURTADA AND L. ROSASCO, *An elementary analysis of ridge regression with random design*, C. R. Math., 360 (2022), pp. 1055–1063.

[34] M. Opper, W. Kinzel, J. Kleinz, and R. Nehl, *On the ability of the optimal perceptron to generalise*, J. Phys. A, 23 (1990), L581.

[35] G. Raskutti and M. W. Mahoney, *A statistical perspective on randomized sketching for ordinary least-squares*, J. Mach. Learn. Res., 17 (2016), pp. 7508–7538.

[36] D. Richards, J. Mourtada, and L. Rosasco, *Asymptotics of ridge (less) regression under general source condition*, in Proceedings of the International Conference on Artificial Intelligence and Statistics, 2021.

[37] F. Rubio and X. Mestre, *Spectral convergence for a general class of random matrices*, Statist. Probab. Lett., 81 (2011), pp. 592–602.

[38] M. Rudelson and R. Vershynin, *Hanson-Wright inequality and sub-Gaussian concentration*, Electron. Commun. Probab., 18 (2013), pp. 1–9.

[39] J. W. Silverstein, *Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices*, J. Multivariate Anal., 55 (1995), pp. 331–339.

[40] J. W. Silverstein and Z. Bai, *On the empirical distribution of eigenvalues of a class of large dimensional random matrices*, J. Multivariate Anal., 54 (1995), pp. 175–192.

[41] F. Vallet, J.-G. Cailton, and P. Refregier, *Linear and nonlinear extension of the pseudo-inverse solution for learning Boolean functions*, Europhys. Lett., 9 (1989), 315.

[42] D. Wu and J. Xu, *On the optimal weighted ℓ_2 regularization in overparameterized linear regression*, Adv. Neural Inf. Process. Syst., 33 (2020).

[43] J. Xu and D. J. Hsu, *On the number of variables to use in principal component regression*, Adv. Neural Inf. Process. Syst., 32 (2019).

[44] Z. Yang, Y. Yu, C. You, J. Steinhardt, and Y. Ma, *Rethinking bias-variance trade-off for generalization of neural networks*, in Proceedings of the International Conference on Machine Learning, 2020, pp. 10767–10777.

[45] T. Zhang, *Learning bounds for kernel regression using effective data dimensionality*, Neural Comput., 17 (2005), pp. 2077–2098.