

# RANDOM DESIGN ANALYSIS OF RIDGE REGRESSION

DANIEL HSU, SHAM M. KAKADE, AND TONG ZHANG

ABSTRACT. This work gives a simultaneous analysis of both the ordinary least squares estimator and the ridge regression estimator in the random design setting under mild assumptions on the covariate/response distributions. In particular, the analysis provides sharp results on the “out-of-sample” prediction error, as opposed to the “in-sample” (fixed design) error. The analysis also reveals the effect of errors in the estimated covariance structure, as well as the effect of modeling errors, neither of which effects are present in the fixed design setting. The proofs of the main results are based on a simple decomposition lemma combined with concentration inequalities for random vectors and matrices.

## 1. INTRODUCTION

In the random design setting for linear regression, we are provided with samples of covariates and responses,  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , which are sampled independently from a population, where the  $x_i$  are random vectors and the  $y_i$  are random variables. Typically, these pairs are hypothesized to have the linear relationship

$$y_i = \langle \beta, x_i \rangle + \epsilon_i$$

for some linear function  $\beta$  (though this hypothesis need not be true). Here, the  $\epsilon_i$  are error terms, typically assumed to be normally distributed as  $\mathcal{N}(0, \sigma^2)$ . The goal of estimation in this setting is to find coefficients  $\hat{\beta}$  based on these  $(x_i, y_i)$  pairs such that the expected prediction error on a new draw  $(x, y)$  from the population, measured as  $\mathbb{E}[(\langle \hat{\beta}, x \rangle - y)^2]$ , is as small as possible. This goal can also be interpreted as estimating  $\beta$  with accuracy measured under a particular norm.

The random design setting stands in contrast to the fixed design setting, where the covariates  $x_1, x_2, \dots, x_n$  are fixed (*i.e.*, deterministic), and only the responses  $y_1, y_2, \dots, y_n$  treated as random. Thus, the covariance structure of the design points is completely known and need not be estimated, which simplifies the analysis of standard estimators. However, the fixed design setting does not directly address out-of-sample prediction, which is of primary concern in many applications; for instance, in prediction problems, the estimator  $\hat{\beta}$  is computed from an initial sample from the population, and the end-goal is to use  $\hat{\beta}$  as a predictor of  $y$  given  $x$  where  $(x, y)$  is a new draw from the population. A fixed design analysis only assesses the accuracy of  $\hat{\beta}$  on data already seen, while a random design analysis is concerned with the predictive performance on unseen data.

This work gives a detailed analysis of both the ordinary least squares and *ridge* estimators [9] in the random design setting that quantifies the essential differences between random and fixed design. In particular, the analysis reveals, through a simple decomposition:

- the effect of errors in the estimated covariance structure;
- the effect of errors in the estimated covariance structure, as well as the effect of approximating the true regression function by a linear function in the case the model is misspecified;
- the effect of errors due to noise in the response.

Neither of the first two effects is present in the fixed design analysis of ridge regression, and the random design analysis shows that the effect of errors in the estimated covariance structure is minimal—essentially a second-order effect as soon as the sample size is large enough. The analysis also isolates the effect of approximation error in the main terms of the estimation error bound so that the bound reduces to one that scales only with the noise variance when the approximation error vanishes.

---

2010 *Mathematics Subject Classification*. Primary 62J07; Secondary 62J05.

*Key words and phrases*. Linear regression, ordinary least squares, ridge regression, randomized approximation.

Another important feature of the analysis that distinguishes it from that of previous work is that it applies to the ridge estimator with an arbitrary setting of  $\lambda \geq 0$ . The estimation error is given in terms of the spectrum of the second moment of  $x$  and the particular choice of  $\lambda$ —the dimension of the covariate space does not enter explicitly except when  $\lambda = 0$ . When  $\lambda = 0$ , we immediately obtain an analysis of ordinary least squares; we are not aware of any other random design analysis of the ridge estimator with this characteristic. More generally, the convergence rate can be optimized by appropriately setting  $\lambda$  based on assumptions about the spectrum.

Finally, while our analysis is based on an operator-theoretical approach similar to that of [19] and [4], it relies on probabilistic tail inequalities in a modular way that gives explicit dependencies without additional boundedness assumptions other than those assumed by the probabilistic bounds.

**Outline.** Section 2 discusses the model, preliminaries, and related work. Section 3 presents the main results on the excess mean squared error of the ordinary least squares and ridge estimators under random design and discusses the relationship to the standard fixed design analysis. Section 4 discusses an application to accelerating least squares computations on large data sets. The proofs of the main results are given in Section 5.

## 2. PRELIMINARIES

**2.1. Notation.** Unless otherwise specified, all vectors in this work are assumed to live in a finite dimensional inner product space with inner product  $\langle \cdot, \cdot \rangle$ . The restriction to finite-dimensions is due to the probabilistic bounds used in the proofs; the main results of this work can be extended to (possibly infinite-dimensional) separable Hilbert spaces under mild assumptions by using suitable infinite-dimensional generalizations of these probabilistic bounds. We denote the dimensionality of this space by  $d$ , but stress that our results will not explicitly depend on  $d$  except when considering the special case of  $\lambda = 0$ . Let  $\|\cdot\|_M$  for a self-adjoint positive definite linear operator  $M \succ 0$  denote the vector norm given by  $\|v\|_M := \sqrt{\langle v, Mv \rangle}$ . When  $M$  is omitted, it is assumed to be the identity  $I$ , so  $\|v\| = \sqrt{\langle v, v \rangle}$ . Let  $u \otimes u$  denote the outer product of a vector  $u$ , which acts as the rank-one linear operator  $v \mapsto (u \otimes u)v = \langle v, u \rangle u$ . For a linear operator  $M$ , let  $\|M\|$  denote its spectral (operator) norm, *i.e.*,  $\|M\| = \sup_{v \neq 0} \|Mv\|/\|v\|$ , and let  $\|M\|_F$  denote its Frobenius norm, *i.e.*,  $\|M\|_F = \sqrt{\text{tr}(M^*M)}$ . If  $M$  is self-adjoint,  $\|M\|_F = \sqrt{\text{tr}(M^2)}$ . Let  $\lambda_{\max}[M]$  and  $\lambda_{\min}[M]$ , respectively, denote the largest and smallest eigenvalue of a self-adjoint linear operator  $M$ .

**2.2. Linear regression.** Let  $x$  be a random vector, and let  $y$  be a random variable. Throughout, it is assumed that  $x$  and  $y$  have finite second moments ( $\mathbb{E}[\|x\|^2] < \infty$  and  $\mathbb{E}[y^2] < \infty$ ). Let  $\{v_j\}$  be the eigenvectors of

$$(1) \quad \Sigma := \mathbb{E}[x \otimes x],$$

so that they form an orthonormal basis. The corresponding eigenvalues are

$$\lambda_j := \langle v_j, \Sigma v_j \rangle = \mathbb{E}[\langle v_j, x \rangle^2].$$

It is without loss of generality that we assume all eigenvalues  $\lambda_j$  are strictly positive, since otherwise we may restrict attention of all vectors to a subspace in which the assumption holds. Let  $\beta$  achieve the minimum *mean squared error* over all linear functions, *i.e.*,

$$\mathbb{E}[(\langle \beta, x \rangle - y)^2] = \min_w \{ \mathbb{E}[(\langle w, x \rangle - y)^2] \},$$

so that

$$(2) \quad \beta := \sum_j \beta_j v_j \quad \text{where} \quad \beta_j := \frac{\mathbb{E}[\langle v_j, x \rangle y]}{\mathbb{E}[\langle v_j, x \rangle^2]}.$$

We also have that the *excess* mean squared error of  $w$  over the minimum is

$$\mathbb{E}[(\langle w, x \rangle - y)^2] - \mathbb{E}[(\langle \beta, x \rangle - y)^2] = \|w - \beta\|_{\Sigma}^2$$

(see Proposition 5).

**2.3. The ridge and ordinary least squares estimators.** Let  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  be independent copies of  $(x, y)$ , and let  $\widehat{\mathbb{E}}$  denote the empirical expectation with respect to these  $n$  copies, *i.e.*,

$$(3) \quad \widehat{\mathbb{E}}[f] := \frac{1}{n} \sum_{i=1}^n f(x_i, y_i) \quad \widehat{\Sigma} := \widehat{\mathbb{E}}[x \otimes x] = \frac{1}{n} \sum_{i=1}^n x_i \otimes x_i.$$

Let  $\hat{\beta}_\lambda$  denote the *ridge estimator* with parameter  $\lambda \geq 0$ , defined as the minimizer of the  $\lambda$ -regularized empirical mean squared error, *i.e.*,

$$(4) \quad \hat{\beta}_\lambda := \arg \min_w \left\{ \widehat{\mathbb{E}}[(\langle w, x \rangle - y)^2] + \lambda \|w\|^2 \right\}.$$

The special case with  $\lambda = 0$  is the *ordinary least squares estimator*, which minimizes the empirical mean squared error. These estimators are uniquely defined if and only if  $\widehat{\Sigma} + \lambda I \succ 0$  (a sufficient condition is  $\lambda > 0$ ), in which case

$$\hat{\beta}_\lambda = (\widehat{\Sigma} + \lambda I)^{-1} \widehat{\mathbb{E}}[xy].$$

**2.4. Data model.** We now specify the conditions on the random pair  $(x, y)$  under which the analysis applies.

**2.4.1. Covariate model.** We first define the following effective dimensions of the covariate  $x$  based on the second moment operator  $\Sigma$  and the regularization level  $\lambda$ :

$$(5) \quad d_{p,\lambda} := \sum_j \left( \frac{\lambda_j}{\lambda_j + \lambda} \right)^p, \quad p \in \{1, 2\}.$$

It will become apparent in the analysis that these dimensions govern the sample size needed to ensure that  $\Sigma$  is estimated with sufficient accuracy. For technical reasons, we also use the quantity

$$(6) \quad \tilde{d}_{1,\lambda} := \max\{d_{1,\lambda}, 1\}$$

merely to simplify certain probability tail inequalities in the main result in the peculiar case that  $\lambda \rightarrow \infty$  (upon which  $d_{1,\lambda} \rightarrow 0$ ). We remark that  $d_{2,\lambda}$  appears naturally arises in the standard fixed design analysis of ridge regression (see Proposition 1), and that  $d_{1,\lambda}$  was also used by [23] and [4] in their random design analyses of (kernel) ridge regression. It is easy to see that  $d_{2,\lambda} \leq d_{1,\lambda}$ , and that  $d_{p,\lambda}$  is at most the dimension  $d$  of the inner product space (with equality iff  $\lambda = 0$ ).

Our main condition requires that the squared length of  $(\Sigma + \lambda I)^{-1/2}x$  is never more than a constant factor greater than its expectation (hence the name *bounded statistical leverage*). The linear mapping  $x \mapsto (\Sigma + \lambda I)^{-1/2}x$  is sometimes called *whitening* when  $\lambda = 0$ . The reason for considering  $\lambda > 0$ , in which case we call the mapping  $\lambda$ -*whitening*, is that the expectation  $\mathbb{E}[\|(\Sigma + \lambda I)^{-1/2}x\|^2]$  may only be small for sufficiently large  $\lambda$ , as

$$\mathbb{E}[\|(\Sigma + \lambda I)^{-1/2}x\|^2] = \text{tr}((\Sigma + \lambda I)^{-1/2} \Sigma (\Sigma + \lambda I)^{-1/2}) = \sum_j \frac{\lambda_j}{\lambda_j + \lambda} = d_{1,\lambda}.$$

**Condition 1** (Bounded statistical leverage at  $\lambda$ ). There exists finite  $\rho_\lambda \geq 1$  such that, almost surely,

$$\frac{\|(\Sigma + \lambda I)^{-1/2}x\|}{\sqrt{\mathbb{E}[\|(\Sigma + \lambda I)^{-1/2}x\|^2]}} = \frac{\|(\Sigma + \lambda I)^{-1/2}x\|}{\sqrt{d_{1,\lambda}}} \leq \rho_\lambda.$$

The hard “almost sure” bound in Condition 1 may be relaxed to moment conditions simply by using different probability tail inequalities in the analysis. We do not consider this relaxation for sake of simplicity. We also remark that it is possible to replace Condition 1 with a subgaussian condition (specifically, a requirement that every projection of  $(\Sigma + \lambda I)^{-1/2}x$  be subgaussian), which can lead to a sharper deviation bound in certain cases.

*Remark 1* (Ordinary least squares). If  $\lambda = 0$ , then Condition 1 reduces to the requirement that there exists a finite  $\rho_0 \geq 1$  such that, almost surely,

$$\frac{\|\Sigma^{-1/2}x\|}{\sqrt{\mathbb{E}[\|\Sigma^{-1/2}x\|^2]}} = \frac{\|\Sigma^{-1/2}x\|}{\sqrt{d}} \leq \rho_0.$$

*Remark 2* (Bounded covariates). If  $\|x\| \leq r$  almost surely, then

$$\frac{\|(\Sigma + \lambda I)^{-1/2} x\|}{\sqrt{d_{1,\lambda}}} \leq \frac{r}{\sqrt{(\inf\{\lambda_j\} + \lambda)d_{1,\lambda}}}$$

in which case Condition 1 holds with  $\rho_\lambda$  satisfying

$$\rho_\lambda \leq \frac{r}{\sqrt{\lambda d_{1,\lambda}}}.$$

**2.4.2. Response model.** The response model considered in this work is a relaxation of the typical Gaussian model; the model specifically allows for approximation error and general subgaussian noise. Define the random variables

$$(7) \quad \text{noise}(x) := y - \mathbb{E}[y|x] \quad \text{and} \quad \text{approx}(x) := \mathbb{E}[y|x] - \langle \beta, x \rangle$$

where  $\text{noise}(x)$  corresponds to the response noise, and  $\text{approx}(x)$  corresponds to the approximation error of  $\beta$ . This gives the following modeling equation:

$$y = \langle \beta, x \rangle + \text{approx}(x) + \text{noise}(x).$$

Conditioned on  $x$ ,  $\text{noise}(x)$  is random, while  $\text{approx}(x)$  is deterministic.

The noise is assumed to satisfy the following subgaussian moment condition:

**Condition 2** (Subgaussian noise). There exists finite  $\sigma \geq 0$  such that, almost surely,

$$\mathbb{E}[\exp(\eta \text{noise}(x))|x] \leq \exp(\eta^2 \sigma^2 / 2) \quad \forall \eta \in \mathbb{R}.$$

Condition 2 is satisfied, for instance, if  $\text{noise}(x)$  is normally distributed with mean zero and variance  $\sigma^2$ .

For the next condition, define  $\beta_\lambda$  to be the minimizer of the regularized mean squared error, *i.e.*,

$$(8) \quad \beta_\lambda := \arg \min_w \{ \mathbb{E}[(\langle w, x \rangle - y)^2] + \lambda \|w\|^2 \} = (\Sigma + \lambda I)^{-1} \mathbb{E}[xy],$$

and also define

$$(9) \quad \text{approx}_\lambda(x) := \mathbb{E}[y|x] - \langle \beta_\lambda, x \rangle.$$

The final condition requires a bound on the size of  $\text{approx}_\lambda(x)$ .

**Condition 3** (Bounded approximation error at  $\lambda$ ). There exist finite  $b_\lambda \geq 0$  such that, almost surely,

$$\frac{\|(\Sigma + \lambda I)^{-1/2} x \text{approx}_\lambda(x)\|}{\sqrt{\mathbb{E}[\|(\Sigma + \lambda I)^{-1/2} x\|^2]}} = \frac{\|(\Sigma + \lambda I)^{-1/2} x \text{approx}_\lambda(x)\|}{\sqrt{d_{1,\lambda}}} \leq b_\lambda.$$

The hard ‘‘almost sure’’ bound in Condition 3 can easily be relaxed to moment conditions, but we do not consider it here for sake of simplicity. We also remark that  $b_\lambda$  only appears in lower-order terms in the main bounds.

*Remark 3* (Ordinary least squares). If  $\lambda = 0$  and the dimension of the covariate space is  $d$ , then Condition 3 reduces to the requirement that there exists a finite  $b_0 \geq 0$  such that, almost surely,

$$\frac{\|\Sigma^{-1/2} x \text{approx}(x)\|}{\sqrt{\mathbb{E}[\|\Sigma^{-1/2} x\|^2]}} = \frac{\|\Sigma^{-1/2} x \text{approx}(x)\|}{\sqrt{d}} \leq b_0.$$

*Remark 4* (Bounded approximation error). If  $|\text{approx}(x)| \leq a$  almost surely and Condition 1 (with parameter  $\rho_\lambda$ ) holds, then

$$\begin{aligned} \frac{\|(\Sigma + \lambda I)^{-1/2} x \text{approx}_\lambda(x)\|}{\sqrt{d_{1,\lambda}}} &\leq \rho_\lambda |\text{approx}_\lambda(x)| \\ &\leq \rho_\lambda (a + |\langle \beta - \beta_\lambda, x \rangle|) \\ &\leq \rho_\lambda (a + \|\beta - \beta_\lambda\|_{\Sigma + \lambda I} \|x\|_{(\Sigma + \lambda I)^{-1}}) \\ &\leq \rho_\lambda (a + \rho_\lambda \sqrt{d_{1,\lambda}} \|\beta - \beta_\lambda\|_{\Sigma + \lambda I}) \end{aligned}$$

where the first and last inequalities use Condition 1, the second inequality uses the definition of  $\text{approx}_\lambda(x)$  in (9) and the triangle inequality, and the third inequality follows from Cauchy-Schwarz. The quantity

$\|\beta - \beta_\lambda\|_{\Sigma + \lambda I}$  can be bounded by  $\sqrt{\lambda}\|\beta\|$  using the arguments in the proof of Proposition 7. In this case, Condition 3 is satisfied with

$$b_\lambda \leq \rho_\lambda(a + \rho_\lambda \sqrt{\lambda d_{1,\lambda}} \|\beta\|).$$

If in addition  $\|x\| \leq r$  almost surely, then Condition 1 and Condition 3 are satisfied with

$$\rho_\lambda \leq \frac{r}{\sqrt{\lambda d_{1,\lambda}}} \quad \text{and} \quad b_\lambda \leq \rho_\lambda(a + r\|\beta\|)$$

as per Remark 2.

**2.5. Related work.** The ridge and ordinary least squares estimators are classically studied in the fixed design setting: the covariates  $x_1, x_2, \dots, x_n$  are fixed vectors in  $\mathbb{R}^d$ , and the responses  $y_1, y_2, \dots, y_n$  are independent random variables, each with mean  $\mathbb{E}[y_i] = \langle \beta, x_i \rangle$  and variance  $\text{var}(y_i) \leq \sigma^2$  [16]. The analysis reviewed in Section 3.1 reveals the expected prediction error  $\mathbb{E}[\|\hat{\beta}_\lambda - \beta\|_\Sigma^2]$  is controlled by the sum of a bias term, which is zero when  $\lambda = 0$ , and a variance term, which is bounded by  $\sigma^2 d_{2,\lambda}/n$ . As discussed in the introduction, our random design analysis of the ridge estimator reveals the essential differences between fixed and random design by comparing with this classical analysis.

Many classical analyses of the ridge and ordinary least squares estimators in the random design setting (e.g., in the context of nonparametric estimators) do not actually show nonasymptotic  $O(d/n)$  convergence of the mean squared error to that of the best linear predictor, where  $d$  is the dimension of the covariate space. Rather, the error relative to the Bayes error is bounded by some multiple  $c > 1$  of the error of the optimal linear predictor relative to the Bayes error, plus a  $O(d/n)$  term [8]:

$$\mathbb{E}[(\langle \hat{\beta}, x \rangle - \mathbb{E}[y|x])^2] \leq c \cdot \mathbb{E}[(\langle \beta, x \rangle - \mathbb{E}[y|x])^2] + O(d/n).$$

Such bounds are appropriate in non-parametric settings where the error of the optimal linear predictor also approaches the Bayes error at an  $O(d/n)$  rate. Beyond these classical results, analyses of ordinary least squares often come with nonstandard restrictions on applicability or additional dependencies on the spectrum of the second moment operator (see the recent work of [2] for a comprehensive survey of these results); for instance, a result of [5] gives a bound on the excess mean squared error of the form

$$\|\hat{\beta} - \beta\|_\Sigma^2 \leq O\left(\frac{d + \log(\det(\hat{\Sigma})/\det(\Sigma))}{n}\right),$$

but the bound is only shown to hold when every linear predictor with low empirical mean squared error satisfies certain boundedness conditions.

This work provides ridge regression bounds explicitly in terms of the vector  $\beta$  (as a sequence) and in terms of the eigenspectrum of the second moment operator  $\Sigma$ . While the essential setting we study is not new, previous analyses make unnecessarily strong boundedness assumptions or fail to give a bound in the case  $\lambda = 0$ . Here we review the analyses of [23], [19], [4], and [20]. [23] assumes  $\|x\| \leq b_x$  and  $|\langle \beta, x \rangle - y| \leq b_{\text{approx}}$  almost surely, and gives the bound

$$\|\hat{\beta}_\lambda - \beta\|_\Sigma^2 \leq \lambda \|\hat{\beta}_\lambda - \beta\|^2 + c \cdot \frac{d_{1,\lambda} \cdot (b_{\text{approx}} + b_x \|\hat{\beta}_\lambda - \beta\|)^2}{n}$$

for some  $c > 0$ , where  $d_{1,\lambda}$  is the effective dimension at scale  $\lambda$  as defined in (5). The quantity  $\|\hat{\beta}_\lambda - \beta\|$  is then bounded by assuming  $\|\beta\| < \infty$ . Thus, the dominant terms of the final bound have explicit dependences on  $b_{\text{approx}}$  and  $b_x$ . [19] assume that  $|y| \leq b_y$  and  $\|x\| \leq b_x$  almost surely, and prove the bound

$$\|\hat{\beta}_\lambda - \beta\|_\Sigma^2 \leq c' \cdot \frac{b_x^2 b_y^2}{n \lambda^2}$$

for some  $c' > 0$  (and note that the bound becomes trivial when  $\lambda = 0$ ); this is then used to bound  $\|\hat{\beta}_\lambda - \beta\|_\Sigma^2$  under explicit assumptions on  $\beta$ . [4] assume  $\|x\| \leq b_x$  almost surely, and prove the bound (in their Theorem 4)

$$\|\hat{\beta}_\lambda - \beta\|_\Sigma^2 \leq c'' \cdot \left( \|\beta_\lambda - \beta\|_\Sigma^2 + \frac{b_x \|\beta_\lambda - \beta\|_\Sigma^2}{n \lambda} + \frac{\sigma^2 d_{1,\lambda}}{n} + o(1/n) \right).$$

Here, we also note that, if one desires the bound to hold with probability  $\geq 1 - e^{-t}$  for some  $t > 0$ , then the leading factor  $c'' > 1$  depends quadratically on  $t$ . Finally, [20] explicitly require  $|y| \leq b_y$  and their main

bound on  $\|\hat{\beta}_\lambda - \beta\|_\Sigma^2$  (specialized for the ridge estimator) depends on  $b_y$  in a dominant term. Moreover, this main bound contains  $c''' \cdot (\lambda\|\beta_\lambda\|^2 + \|\beta_\lambda - \beta\|_\Sigma^2)$  as a dominant term for some  $c''' > 1$ , and it is only given under explicit decay conditions on the eigenspectrum (their Equation 6). The bound is also trivial when  $\lambda = 0$ . Our result for ridge regression is given explicitly in terms of  $\|\beta_\lambda - \beta\|_\Sigma^2$  (and therefore explicitly in terms of  $\beta$  as a sequence, the eigenspectrum of  $\Sigma$ , and  $\lambda$ ); this quantity vanishes when  $\lambda = 0$  and can be small even when  $\|\beta\|$  itself is large. We note that  $\|\beta_\lambda - \beta\|_\Sigma^2$  is precisely the bias term from the classical fixed design analysis of ridge regression, and therefore is natural to expect in a random design analysis.

Recently, [3] derived sharp risk bounds for the ordinary least squares and ridge estimators (in addition to specially developed PAC-Bayesian estimators) in a random design setting under very mild moment assumptions using PAC-Bayesian techniques. Their nonasymptotic bound for ordinary least squares holds with probability at least  $1 - e^{-t}$  but only for  $t \leq \ln n$ ; this is essentially due to their weak moment assumptions. By relying on stronger moment assumptions, we allow the probability tail parameter  $t$  to be as large as  $\Omega(n/d)$ . Our analysis is also arguably more transparent and yields more reasonable quantitative bounds. The analysis of [3] for the ridge estimator is established only in an asymptotic sense and therefore are not directly comparable to those provided here.

Finally, although the focus of our present work is on understanding the ordinary least squares and ridge estimators, it should also be mentioned that a number of other estimators have been considered in the literature with nonasymptotic prediction error bounds [14, 3, 13]. Indeed, the works of [3] and [13] propose estimators that require considerably weaker moment conditions on  $x$  and  $y$  to obtain optimal rates.

### 3. RANDOM DESIGN REGRESSION

This section presents the main results of the paper on the excess mean squared error of the ridge estimator under random design (and its specialization to the ordinary least squares estimator). First, we review the standard fixed design analysis.

**3.1. Review of fixed design analysis.** It is informative to first review the fixed design analysis of the ridge estimator. Recall that, in this setting, the design points  $x_1, x_2, \dots, x_n$  are fixed (deterministic) vectors, and the responses  $y_1, y_2, \dots, y_n$  are independent random variables. Therefore, we define  $\Sigma := \widehat{\Sigma} = n^{-1} \sum_{i=1}^n x_i \otimes x_i$  (which is nonrandom), and assume it has eigenvectors  $\{v_j\}$  and corresponding eigenvalues  $\lambda_j := \langle v_j, \Sigma v_j \rangle$ . As in the random design setting, the linear function  $\beta := \sum_j \beta_j v_j$  where  $\beta_j := (n\lambda_j)^{-1} \sum_{i=1}^n \langle v_j, x_i \rangle \mathbb{E}[y_i]$  minimizes the expected mean squared error, *i.e.*,

$$\beta := \arg \min_w \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(\langle w, x_i \rangle - y_i)^2].$$

Similar to the random design setup, define noise( $x_i$ ) :=  $y_i - \mathbb{E}[y_i]$  and approx( $x_i$ ) :=  $\mathbb{E}[y_i] - \langle \beta, x_i \rangle$  for  $i = 1, 2, \dots, n$ , so the following modeling equation holds:

$$y_i = \langle \beta, x_i \rangle + \text{approx}(x_i) + \text{noise}(x_i)$$

for  $i = 1, 2, \dots, n$ . Because  $\Sigma = \widehat{\Sigma}$ , the ridge estimator  $\hat{\beta}_\lambda$  in the fixed design setting is an unbiased estimator of the minimizer of the regularized mean squared error, *i.e.*,

$$\mathbb{E}[\hat{\beta}_\lambda] = (\Sigma + \lambda I)^{-1} \left( \frac{1}{n} \sum_{i=1}^n x_i \mathbb{E}[y_i] \right) = \arg \min_w \left\{ \frac{1}{n} \sum_{i=1}^n \mathbb{E}[(\langle w, x_i \rangle - y_i)^2] + \lambda \|w\|^2 \right\}.$$

This unbiasedness implies that the expected mean squared error of  $\hat{\beta}_\lambda$  has the bias-variance decomposition

$$(10) \quad \mathbb{E}[\|\hat{\beta}_\lambda - \beta\|_\Sigma^2] = \|\mathbb{E}[\hat{\beta}_\lambda] - \beta\|_\Sigma^2 + \mathbb{E}[\|\hat{\beta}_\lambda - \mathbb{E}[\hat{\beta}_\lambda]\|_\Sigma^2].$$

The following bound on the expected excess mean squared error easily follows from this decomposition and the definition of  $\beta$  (see, *e.g.*, Proposition 7).

**Proposition 1** (Ridge regression: fixed design). *Fix  $\lambda \geq 0$ , and assume  $\Sigma + \lambda I$  is invertible. If there exists  $\sigma \geq 0$  such that  $\text{var}(y_i^2) \leq \sigma^2$  for all  $i = 1, 2, \dots, n$ , then*

$$\mathbb{E}[\|\hat{\beta}_\lambda - \beta\|_\Sigma^2] \leq \sum_j \frac{\lambda_j}{(\frac{\lambda_j}{\lambda} + 1)^2} \beta_j^2 + \frac{\sigma^2}{n} \sum_j \left( \frac{\lambda_j}{\lambda_j + \lambda} \right)^2$$

with equality iff  $\text{var}(y_i) = \sigma^2$  for all  $i = 1, 2, \dots, n$ .

*Remark 5* (Effect of approximation error in fixed design). Observe that  $\text{approx}(x_i)$  has no effect on the expected excess mean squared error.

*Remark 6* (Effective dimension). The second sum in the bound is equal to  $d_{2,\lambda}$ , a notion of effective dimension at regularization level  $\lambda$ .

*Remark 7* (Ordinary least squares in fixed design). Setting  $\lambda = 0$  gives the following bound for the ordinary least squares estimator  $\hat{\beta}_0$ :

$$\mathbb{E}[\|\hat{\beta}_0 - \beta\|_{\Sigma}^2] \leq \frac{\sigma^2 d}{n}$$

where, as before, equality holds iff  $\text{var}(y_i) = \sigma^2$  for all  $i = 1, 2, \dots, n$ .

**3.2. Ordinary least squares.** Our analysis of the ordinary least squares estimator (under random design) is based on a simple decomposition of the excess mean squared error, similar to the one from the fixed design analysis. To state the decomposition, first let  $\bar{\beta}_0$  denote the conditional expectation of the least squares estimator  $\hat{\beta}_0$  conditioned on  $x_1, x_2, \dots, x_n$ , i.e.,

$$\bar{\beta}_0 := \mathbb{E}[\hat{\beta}_0 | x_1, x_2, \dots, x_n] = \hat{\Sigma}^{-1} \widehat{\mathbb{E}}[x \mathbb{E}[y | x]].$$

Also, define the bias and variance as:

$$\varepsilon_{\text{bs}} := \|\bar{\beta}_0 - \beta\|_{\Sigma}^2, \quad \varepsilon_{\text{vr}} := \|\hat{\beta}_0 - \bar{\beta}_0\|_{\Sigma}^2$$

**Proposition 2** (Random design decomposition). *We have*

$$\begin{aligned} \|\hat{\beta}_0 - \beta\|_{\Sigma}^2 &\leq \varepsilon_{\text{bs}} + 2\sqrt{\varepsilon_{\text{bs}}\varepsilon_{\text{vr}}} + \varepsilon_{\text{vr}} \\ &\leq 2(\varepsilon_{\text{bs}} + \varepsilon_{\text{vr}}) \end{aligned}$$

*Proof.* The claim follows from the triangle inequality and the fact  $(a + b)^2 \leq 2(a^2 + b^2)$ .  $\square$

*Remark 8.* Note that, in general,  $\mathbb{E}[\hat{\beta}_0] \neq \beta$  (unlike in the fixed design setting where  $\mathbb{E}[\hat{\beta}_0] = \beta$ ). Hence, our decomposition differs from that in the fixed design analysis (see (10)).

Our first main result characterizes the excess loss of the ordinary least squares estimator.

**Theorem 1** (Ordinary least squares regression). *Pick any  $t > \max\{0, 2.6 - \log d\}$ . Assume Condition 1 (with parameter  $\rho_0$ ), Condition 2 (with  $\sigma$ ), and Condition 3 (with  $b_0$ ) hold and that*

$$n \geq 6\rho_0^2 d(\log d + t).$$

*With probability at least  $1 - 3e^{-t}$ , the following holds:*

- (1) Relative spectral norm error in  $\hat{\Sigma}$ :  $\hat{\Sigma}$  is invertible, and

$$\|\Sigma^{1/2} \hat{\Sigma}^{-1} \Sigma^{1/2}\| \leq (1 - \delta_s)^{-1},$$

where  $\Sigma$  is defined in (1),  $\hat{\Sigma}$  is defined in (3), and

$$\delta_s := \sqrt{\frac{4\rho_0^2 d(\log d + t)}{n}} + \frac{2\rho_0^2 d(\log d + t)}{3n}$$

(note that the lower bound on  $n$  ensures  $\delta_s \leq 0.93 < 1$ ).

- (2) Effect of bias due to random design:

$$\begin{aligned} \varepsilon_{\text{bs}} &\leq \frac{2}{(1 - \delta_s)^2} \left( \frac{\mathbb{E}[\|\Sigma^{-1/2} x \text{approx}(x)\|^2]}{n} (1 + \sqrt{8t})^2 + \frac{16b_0^2 dt^2}{9n^2} \right) \\ &\leq \frac{2}{(1 - \delta_s)^2} \left( \frac{\rho_0^2 d \mathbb{E}[\text{approx}(x)^2]}{n} (1 + \sqrt{8t})^2 + \frac{16b_0^2 dt^2}{9n^2} \right), \end{aligned}$$

and  $\text{approx}(x)$  is defined in (9).

(3) Effect of noise:

$$\varepsilon_{\text{vr}} \leq \frac{1}{1 - \delta_s} \cdot \frac{\sigma^2(d + 2\sqrt{dt} + 2t)}{n}.$$

*Remark 9* (Simplified form). Suppressing the terms that are  $o(1/n)$ , the overall bound from Theorem 1 is

$$\|\hat{\beta}_0 - \beta\|_{\Sigma}^2 \leq \frac{2\mathbb{E}[\|\Sigma^{-1/2}x \text{ approx}(x)\|^2]}{n} (1 + \sqrt{8t})^2 + \frac{\sigma^2(d + 2\sqrt{dt} + 2t)}{n} + o(1/n)$$

(so  $b_0$  appears only in the  $o(1/n)$  terms). If the linear model is correct (*i.e.*,  $\mathbb{E}[y|x] = \langle \beta, x \rangle$  almost surely), then

$$(11) \quad \|\hat{\beta}_0 - \beta\|_{\Sigma}^2 \leq \frac{\sigma^2(d + 2\sqrt{dt} + 2t)}{n} + o(1/n).$$

One can show that the constants in the first-order term in (11) are the same as those that one would obtain for a fixed design tail bound.

*Remark 10* (Tightness of the bound). Since

$$\|\bar{\beta}_0 - \beta\|_{\Sigma}^2 = \|(\Sigma^{1/2} \hat{\Sigma}^{-1} \Sigma^{1/2}) \widehat{\mathbb{E}}[\Sigma^{-1/2}x \text{ approx}(x)]\|^2$$

and

$$\|\Sigma^{1/2} \hat{\Sigma}^{-1} \Sigma^{1/2} - I\| \rightarrow 0$$

as  $n \rightarrow \infty$  (Lemma 2),  $\|\bar{\beta}_0 - \beta\|_{\Sigma}^2$  is within constant factors of  $\|\widehat{\mathbb{E}}[\Sigma^{-1/2}x \text{ approx}(x)]\|^2$  for sufficiently large  $n$ . Moreover,

$$\mathbb{E}[\|\widehat{\mathbb{E}}[\Sigma^{-1/2}x \text{ approx}(x)]\|^2] = \frac{\mathbb{E}[\|\Sigma^{-1/2}x \text{ approx}(x)\|^2]}{n},$$

which is the main term that appears in the bound for  $\varepsilon_{\text{bs}}$ . Similarly,  $\|\hat{\beta}_0 - \bar{\beta}_0\|_{\Sigma}^2$  is within constant factors of  $\|\hat{\beta}_0 - \bar{\beta}_0\|_{\Sigma}^2$  for sufficiently large  $n$ , and

$$\mathbb{E}[\|\hat{\beta}_0 - \bar{\beta}_0\|_{\Sigma}^2] \leq \frac{\sigma^2 d}{n}$$

with equality iff  $\text{var}(y) = \sigma^2$  (this comes from the fixed design risk bound in Remark 7). Therefore, in this case where  $\text{var}(y) = \sigma^2$ , we conclude that the bound Theorem 1 is tight up to constant factors and lower-order terms.

**3.3. Random design ridge regression.** The analysis of the ridge estimator under random design is again based on a simple decomposition of the excess mean squared error. Here, let  $\bar{\beta}_{\lambda}$  denote the conditional expectation of  $\hat{\beta}_{\lambda}$  given  $x_1, x_2, \dots, x_n$ , *i.e.*,

$$(12) \quad \bar{\beta}_{\lambda} := \mathbb{E}[\hat{\beta}_{\lambda} | x_1, x_2, \dots, x_n] = (\hat{\Sigma} + \lambda I)^{-1} \widehat{\mathbb{E}}[x \mathbb{E}[y|x]].$$

Define the bias from regularization, the bias from the random design, and the variance as:

$$\varepsilon_{\text{rg}} := \|\beta_{\lambda} - \beta\|_{\Sigma}^2, \quad \varepsilon_{\text{bs}} := \|\bar{\beta}_{\lambda} - \beta_{\lambda}\|_{\Sigma}^2, \quad \varepsilon_{\text{vr}} := \|\hat{\beta}_{\lambda} - \bar{\beta}_{\lambda}\|_{\Sigma}^2,$$

where  $\beta_{\lambda}$  is the minimizer of the regularized mean squared error (see (8)).

**Proposition 3** (General random design decomposition).

$$\begin{aligned} \|\hat{\beta}_{\lambda} - \beta\|_{\Sigma}^2 &\leq \varepsilon_{\text{rg}} + \varepsilon_{\text{bs}} + \varepsilon_{\text{vr}} + 2(\sqrt{\varepsilon_{\text{rg}}\varepsilon_{\text{bs}}} + \sqrt{\varepsilon_{\text{rg}}\varepsilon_{\text{vr}}} + \sqrt{\varepsilon_{\text{bs}}\varepsilon_{\text{vr}}}) \\ &\leq 3(\varepsilon_{\text{rg}} + \varepsilon_{\text{bs}} + \varepsilon_{\text{vr}}) \end{aligned}$$

*Proof.* The claim follows from the triangle inequality and the fact  $(a + b)^2 \leq 2(a^2 + b^2)$ .  $\square$

*Remark 11.* Again, note that  $\mathbb{E}[\bar{\beta}_{\lambda}] \neq \beta_{\lambda}$  in general, so the bias-variance decomposition in (10) from the fixed design analysis is not directly applicable in the random design setting.

The following theorem is the main result of the paper:

**Theorem 2** (Ridge regression). Fix some  $\lambda \geq 0$ , and pick any  $t > \max\{0, 2.6 - \log \tilde{d}_{1,\lambda}\}$ . Assume Condition 1 (with parameter  $\rho_\lambda$ ), Condition 2 (with parameter  $\sigma$ ), and Condition 3 (with parameter  $b_\lambda$ ) hold; and that

$$n \geq 6\rho_\lambda^2 d_{1,\lambda}(\log \tilde{d}_{1,\lambda} + t),$$

where  $d_{p,\lambda}$  for  $p \in \{1, 2\}$  is defined in (5), and  $\tilde{d}_{1,\lambda}$  is defined in (6).

With probability at least  $1 - 4e^{-t}$ , the following holds:

- (1) Relative spectral norm error in  $\widehat{\Sigma} + \lambda I$ :  $\widehat{\Sigma} + \lambda I$  is invertible, and

$$\|(\Sigma + \lambda I)^{1/2}(\widehat{\Sigma} + \lambda I)^{-1}(\Sigma + \lambda I)^{1/2}\| \leq (1 - \delta_s)^{-1},$$

where  $\Sigma$  is defined in (1),  $\widehat{\Sigma}$  is defined in (3), and

$$\delta_s := \sqrt{\frac{4\rho_\lambda^2 d_{1,\lambda}(\log \tilde{d}_{1,\lambda} + t)}{n} + \frac{2\rho_\lambda^2 d_{1,\lambda}(\log \tilde{d}_{1,\lambda} + t)}{3n}}$$

(note that the lower bound on  $n$  ensures  $\delta_s \leq 0.93 < 1$ ).

- (2) Frobenius norm error in  $\widehat{\Sigma}$ :

$$\|(\Sigma + \lambda I)^{-1/2}(\widehat{\Sigma} - \Sigma)(\Sigma + \lambda I)^{-1/2}\|_F \leq \sqrt{d_{1,\lambda}}\delta_f,$$

where

$$\delta_f := \sqrt{\frac{\rho_\lambda^2 d_{1,\lambda} - d_{2,\lambda}/d_{1,\lambda}}{n}(1 + \sqrt{8t}) + \frac{4\sqrt{\rho_\lambda^4 d_{1,\lambda} + d_{2,\lambda}/d_{1,\lambda}}t}{3n}}.$$

- (3) Effect of regularization:

$$\varepsilon_{\text{rg}} \leq \sum_j \frac{\lambda_j}{(\frac{\lambda_j}{\lambda} + 1)^2} \beta_j^2.$$

If  $\lambda = 0$ , then  $\varepsilon_{\text{rg}} = 0$ .

- (4) Effect of bias due to random design:

$$\begin{aligned} \varepsilon_{\text{bs}} &\leq \frac{2}{(1 - \delta_s)^2} \left( \frac{\mathbb{E}[\|(\Sigma + \lambda I)^{-1/2}(x \text{ approx}_\lambda(x) - \lambda\beta_\lambda)\|^2]}{n} (1 + \sqrt{8t})^2 + \frac{16(b_\lambda \sqrt{d_{1,\lambda}} + \sqrt{\varepsilon_{\text{rg}}})^2 t^2}{9n^2} \right) \\ &\leq \frac{4}{(1 - \delta_s)^2} \left( \frac{\rho_\lambda^2 d_{1,\lambda} \mathbb{E}[\text{approx}_\lambda(x)^2] + \varepsilon_{\text{rg}} (1 + \sqrt{8t})^2 + \frac{(b_\lambda \sqrt{d_{1,\lambda}} + \sqrt{\varepsilon_{\text{rg}}})^2 t^2}{n^2}}{n} \right), \end{aligned}$$

and  $\text{approx}_\lambda(x)$  is defined in (9). If  $\lambda = 0$ , then  $\text{approx}_\lambda(x) = \text{approx}(x)$  as defined in (7).

- (5) Effect of noise:

$$\varepsilon_{\text{vr}} \leq \frac{\sigma^2 (d_{2,\lambda} + \sqrt{d_{1,\lambda} d_{2,\lambda}} \delta_f)}{n(1 - \delta_s)^2} + \frac{2\sigma^2 \sqrt{(d_{2,\lambda} + \sqrt{d_{1,\lambda} d_{2,\lambda}} \delta_f) t}}{n(1 - \delta_s)^{3/2}} + \frac{2\sigma^2 t}{n(1 - \delta_s)}.$$

We now discuss various aspects of Theorem 2.

*Remark 12* (Simplified form). Ignoring the terms that are  $o(1/n)$  and treating  $t$  as a constant, the overall bound from Theorem 2 is

$$\begin{aligned} \|\hat{\beta}_\lambda - \beta\|_\Sigma^2 &\leq \|\beta_\lambda - \beta\|_\Sigma^2 + O\left(\frac{\mathbb{E}[\|(\Sigma + \lambda I)^{-1/2}(x \text{ approx}_\lambda(x) - \lambda\beta_\lambda)\|^2] + \sigma^2 d_{2,\lambda}}{n}\right) \\ &\leq \|\beta_\lambda - \beta\|_\Sigma^2 + O\left(\frac{\rho_\lambda^2 d_{1,\lambda} \mathbb{E}[\text{approx}_\lambda(x)^2] + \|\beta_\lambda - \beta\|_\Sigma^2 + \sigma^2 d_{2,\lambda}}{n}\right) \\ &\leq \|\beta_\lambda - \beta\|_\Sigma^2 + O\left(\frac{\rho_\lambda^2 d_{1,\lambda} \mathbb{E}[\text{approx}(x)^2] + (\rho_\lambda^2 d_{1,\lambda} + 1)\|\beta_\lambda - \beta\|_\Sigma^2 + \sigma^2 d_{2,\lambda}}{n}\right) \end{aligned}$$

where the last inequality follows from the fact  $\sqrt{\mathbb{E}[\text{approx}_\lambda(x)^2]} \leq \sqrt{\mathbb{E}[\text{approx}(x)^2]} + \|\beta_\lambda - \beta\|_\Sigma$ .

*Remark 13* (Effect of errors in  $\widehat{\Sigma}$ ). The accuracy of  $\widehat{\Sigma}$  has a relatively mild effect on the bound—it appears essentially through multiplicative factors  $(1 - \delta_s)^{-1} = 1 + O(\delta_s)$  and  $1 + \delta_f$ , where both  $\delta_s$  and  $\delta_f$  are decreasing with  $n$  (as  $n^{-1/2}$ ), and therefore only contribute to lower-order terms overall.

*Remark 14* (Effect of approximation error). The effect of approximation error is isolated in the term  $\|\bar{\beta}_\lambda - \beta_\lambda\|_\Sigma^2$ . The bound  $\varepsilon_{\text{bs}}$  scales with a fourth-moment quantity  $\mathbb{E}[\|(\Sigma + \lambda I)^{-1/2}(x \text{ approx}_\lambda(x) - \lambda \beta_\lambda)\|^2]$ ; when using the looser bound  $O(\rho_\lambda^2 d_{1,\lambda} \mathbb{E}[\text{approx}(x)^2] + (\rho^2 d_{1,\lambda} + 1) \|\beta_\lambda - \beta\|_\Sigma^2)$ , the overall simplified bound from Remark 12 can be viewed as

$$\begin{aligned} & \mathbb{E}[(\langle \hat{\beta}_\lambda, x \rangle - \mathbb{E}[y|x])^2 | \hat{\beta}_\lambda] \\ & \leq \mathbb{E}[(\langle \beta, x \rangle - \mathbb{E}[y|x])^2] \left(1 + \frac{c_1 \rho_\lambda^2 d_{1,\lambda}}{n}\right) + \mathbb{E}[\langle \beta_\lambda - \beta, x \rangle^2] \left(1 + \frac{c_2 (\rho_\lambda^2 d_{1,\lambda} + 1)}{n}\right) \\ & \quad + \text{terms due to stochastic noise} \end{aligned}$$

for some positive constants  $c_1$  and  $c_2$ . Therefore, the (bound on the) mean squared error of  $\hat{\beta}_\lambda$  is the sum of two contributions (up to lower-order terms): the first is a scaling of the approximation errors  $\mathbb{E}[(\langle \beta, x \rangle - \mathbb{E}[y|x])^2] + \mathbb{E}[\langle \beta_\lambda - \beta, x \rangle^2]$ , where the scaling  $1 + O((\rho_\lambda^2 d_{1,\lambda} + 1)/n)$  tends to one as  $n \rightarrow \infty$ ; and the second is the stochastic noise contribution. The approximation error contribution is unique to random design, while the stochastic noise appears in both random and fixed design.

*Remark 15* (Bounded covariates). Suppose  $\text{approx}(x) = 0$  and that there exists  $r > 0$  such that  $\|x\| \leq r$  almost surely. This is the setting of a well-specified model with bounded covariates; the minimax risk over the class of models  $\beta$  with  $\|\beta\| \leq B$  for some  $B > 0$  is at least  $\Omega(\sqrt{\sigma^2 r^2 B^2/n})$  [17]. In this case, using the inequalities  $\|\beta_\lambda - \beta\|_\Sigma^2 \leq \lambda \|\beta\|^2/2$  and  $d_{2,\lambda} \leq \text{tr}(\Sigma)/(2\lambda)$ , the simplified bound from Remark 12 reduces to

$$\|\hat{\beta}_\lambda - \beta\|_\Sigma^2 \leq \left(1 + O\left(\frac{1 + r^2/\lambda}{n}\right)\right) \cdot \frac{\lambda \|\beta\|^2}{2} + \frac{\sigma^2}{n} \cdot \frac{\text{tr}(\Sigma)}{2\lambda}.$$

Choosing  $\lambda > 0$  to minimize the bound and using the fact  $\text{tr}(\Sigma) \leq r^2$  gives

$$\|\hat{\beta}_\lambda - \beta\|_\Sigma^2 \leq \sqrt{\frac{\sigma^2 r^2 B^2}{n} \cdot \left(1 + O(1/n)\right)} + O\left(\frac{r^2 B^2}{n}\right),$$

which matches the lower bound up to constant factors and lower-order terms.

*Remark 16* (Application to smoothing splines). The applications of ridge regression considered by [23] can also be analyzed using Theorem 2 (although technically our result is only proved in the finite-dimensional setting). We specifically consider the problem of approximating a periodic function with smoothing splines, which are functions  $f: \mathbb{R} \rightarrow \mathbb{R}$  whose  $s$ -th derivatives  $f^{(s)}$ , for some  $s > 1/2$ , satisfy

$$\int \left(f^{(s)}(t)\right)^2 dt < \infty.$$

The one-dimensional covariate  $t \in \mathbb{R}$  can be mapped to the infinite-dimensional representation  $x := \phi(t) \in \mathbb{R}^\infty$  where

$$x_{2k} := \frac{\sin(kt)}{(k+1)^s} \quad \text{and} \quad x_{2k+1} := \frac{\cos(kt)}{(k+1)^s}, \quad k \in \{0, 1, 2, \dots\}.$$

Assume that the regression function is

$$\mathbb{E}[y|x] = \langle \beta, x \rangle$$

so  $\text{approx}(x) = 0$  almost surely. Observe that  $\|x\|^2 \leq \frac{2s}{2s-1}$ , so Condition 1 is satisfied with

$$\rho_\lambda := \left(\frac{2s}{2s-1}\right)^{1/2} \frac{1}{\sqrt{\lambda d_{1,\lambda}}}$$

as per Remark 2. Therefore, the simplified bound from Remark 12 becomes in this case

$$\begin{aligned} \|\hat{\beta}_\lambda - \beta\|_\Sigma^2 &\leq \|\beta_\lambda - \beta\|_\Sigma^2 + C \cdot \left( \frac{2s}{2s-1} \cdot \frac{\|\beta_\lambda - \beta\|_\Sigma^2}{\lambda n} + \frac{\|\beta_\lambda - \beta\|_\Sigma^2 + \sigma^2 d_{2,\lambda}}{n} \right) \\ &\leq \frac{\lambda \|\beta\|^2}{2} + C \cdot \frac{\sigma^2 d_{2,\lambda}}{n} + C \cdot \left( \frac{2s}{2s-1} + \frac{\lambda}{2} \right) \cdot \frac{\|\beta\|^2}{n} \end{aligned}$$

for some constant  $C > 0$ , where we have used the inequality  $\|\beta_\lambda - \beta\|_\Sigma^2 \leq \lambda \|\beta\|^2 / 2$ . [23] shows that

$$d_{1,\lambda} \leq \inf_{k \geq 1} \left\{ 2k + \frac{2/\lambda}{(2s-1)k^{2s-1}} \right\}.$$

Since  $d_{2,\lambda} \leq d_{1,\lambda}$ , it follows that setting  $\lambda := k^{-2s}$  where  $k = \lfloor ((2s-1)n/(2s))^{1/(2s+1)} \rfloor$  gives the bound

$$\|\hat{\beta}_\lambda - \beta\|_\Sigma^2 \leq \left( \frac{\|\beta\|^2}{2} + 2C\sigma^2 \right) \cdot \left( \frac{2s-1}{2s} \cdot n \right)^{-\frac{2s}{2s+1}} + \text{lower-order terms}$$

which has the optimal data-dependent rate of  $n^{-\frac{2s}{2s+1}}$  [22].

*Remark 17* (Comparison with fixed design). As already discussed, the ridge estimator behaves similarly under fixed and random designs, with the main differences being the lack of errors in  $\hat{\Sigma}$  under fixed design, and the influence of approximation error under random design. These are revealed through the quantities  $\rho_\lambda$  and  $d_{1,\lambda}$  (and  $b_\lambda$  in lower-order terms), which are needed to apply the probability tail inequalities. Therefore, the scaling of  $\rho_\lambda^2 d_{1,\lambda}$  with  $\lambda$  crucially controls the effect of random design compared with fixed design.

#### 4. APPLICATION TO ACCELERATING LEAST SQUARES COMPUTATIONS

Our results for the ordinary least squares estimator can be used to analyze a randomized approximation scheme for overcomplete least squares problems [7, 18]. The goal of these randomized methods is to approximately solve the least squares problem

$$\min_{w \in \mathbb{R}^d} \frac{1}{m} \|Aw - b\|^2$$

for some large, full-rank design matrix  $A \in \mathbb{R}^{m \times d}$  ( $m \gg d$ ) and vector  $b \in \mathbb{R}^m$ . Note that using a standard method to exactly solve the least squares problem requires  $\Omega(md^2)$  operations, which can be prohibitive for large-scale problems. However, when an approximate solution is satisfactory, significant computational savings can be achieved through the use of randomization.

**4.1. A randomized approximation scheme for least squares.** The approximation scheme is as follows:

- (1) The columns of  $A$  and the vector  $b$  are first subjected to a randomly chosen rotation matrix (*i.e.*, an orthogonal transformation)  $\Theta \in \mathbb{R}^{m \times m}$ . The distribution over rotation matrices that may be used is discussed below.
- (2) A sample of  $n$  rows of  $[\Theta A, \Theta b] \in \mathbb{R}^{m \times (d+1)}$  are then selected uniformly at random with replacement; let  $\{(x_i^\top, y_i) : i = 1, 2, \dots, n\}$  (where  $x_i \in \mathbb{R}^d$  and  $y_i \in \mathbb{R}$ ) be the  $n$  selected rows of  $[\Theta A, \Theta b]$ .
- (3) Finally, the least squares problem

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n (\langle w, x_i \rangle - y_i)^2$$

is solved by computing the ordinary least squares estimator  $\hat{\beta}_0$  on the sample  $\{(x_i, y_i) : i = 1, 2, \dots, n\}$ .

The motivation for the random rotation  $\Theta$  is captured in Lemma 1, which shows that, if  $\Theta$  is chosen randomly from certain distributions over rotation matrices, then applying  $\Theta$  to  $A$  and  $b$  creates an equivalent least squares problem for which the statistical leverage parameter (the quantity  $\rho_0$  in Condition 1) is small. Consequently, the new least squares problem can be approximately solved with a small random sample, as per Theorems 2 and 1. Without the random rotation, the statistical leverage parameter could be so large that small random sample of the rows will likely miss a row crucial for obtaining an accurate approximation. The role of statistical leverage in this setting was also pointed out by [6], although Lemma 1 makes the

connection more direct. We note that Lemma 1 and the analysis below can be generalized to the case where  $\Theta$  is only approximately orthogonal; for most standard distributions over rotation matrices, the additional error terms that arise do not affect the overall analysis.

The running time of the approximation scheme is given by (i) the time required to apply the  $m \times m$  random rotation operator  $\Theta$  to the original  $m \times (d + 1)$  matrix  $[A, b]$  and randomly sample  $n$  rows, plus (ii) the time to solve the least squares problem on the smaller design matrix of size  $n \times d$ . For (i), naïvely applying an arbitrary  $m \times m$  rotation matrix requires  $\Omega(m^2 d)$  operations; however, there are (distributions over) rotation matrices for which this running time can be reduced to  $O(md \log m)$  (see Example 2 in Section 4.3 below), which is a considerable speed-up when  $m$  is large. In fact, because only  $n$  out of  $m$  rows are to be retained anyway, this computation can be reduced to  $O(md \log n)$  [1]. For (ii), standard methods can produce the ordinary least squares estimator or the ridge regression estimator with  $O(nd^2)$  operations. Therefore, we are interested in the sample size  $n$  that suffices to yield an accurate approximation.

**4.2. Analysis of the approximation scheme.** Our approach to analyzing the above approximation scheme is to treat it as a random design regression problem. We apply Theorem 1 in this setting to give error bounds for the solution produced by the approximation scheme.

Let  $(x, y) \in \mathbb{R}^d \times \mathbb{R}$  be a random pair distributed uniformly over the rows of  $[\Theta A, \Theta b]$ , where we assume that  $\Theta$  is randomly chosen from a suitable distribution over rotation matrices such as those described in Example 1 and Example 2. Lemma 1 (below) implies that there exists a constant  $c_0 > 0$  such that Condition 1 is satisfied with

$$\rho_0^2 \leq c_0 \cdot \left(1 + \frac{\log m + \tau}{d}\right)$$

with probability at least  $1 - e^{-\tau}$  over the choice of the random rotation matrix  $\Theta$ . Henceforth, we condition on the event that this holds.

Let  $\beta \in \mathbb{R}^d$  be the solution to the original least squares problem (*i.e.*,  $\beta := \arg \min_w \|Aw - b\|^2/m$ ), and let  $\hat{\beta}_0 \in \mathbb{R}^d$  be the ordinary least squares estimator computed on the random sample of the rows of  $[\Theta A, \Theta b]$ . Note that, for any  $w \in \mathbb{R}^d$ ,

$$\mathbb{E}[(\langle w, x \rangle - y)^2] = \frac{1}{m} \|\Theta Aw - \Theta b\|^2 = \frac{1}{m} \|Aw - b\|^2.$$

Moreover, we may assume for simplicity that  $y - \langle \beta, x \rangle = \text{approx}(x)$  (*i.e.*, there is no stochastic noise), so  $\mathbb{E}[\text{approx}(x)^2] = \mathbb{E}[(\langle \beta, x \rangle - y)^2] = \|A\beta - b\|^2/m$ .

By Theorem 1, if at least

$$n \geq 6(d + c_0(\log m + \tau))(\log d + t)$$

rows of  $[\Theta A, \Theta b]$  are sampled, then the ordinary least squares estimator  $\hat{\beta}_0$  satisfies the following approximation error guarantee (with probability at least  $1 - 3e^{-t}$  over the random sample of rows):

$$\frac{1}{m} \|A\hat{\beta}_0 - b\|^2 \leq \frac{1}{m} \|A\beta - b\|^2 \cdot \left(1 + c_1 \frac{(d + \log m + \tau)t}{n}\right) + o(1/n)$$

for some constant  $c_1 > 0$ . We note that the  $o(1/n)$  terms can be removed if one only requires constant probability of success (*i.e.*,  $\tau$  and  $t$  treated as constants), as is considered by [7]. In this case, we achieve an error bound of

$$\frac{1}{m} \|A\hat{\beta}_0 - b\|^2 \leq \frac{1}{m} \|A\beta - b\|^2 \cdot (1 + \epsilon)$$

for  $\epsilon > 0$  provided that the number of rows sampled is

$$n \geq c_2(d + \log m) \left(\frac{1}{\epsilon} + \log d\right)$$

for some constant  $c_2 > 0$ .

**4.3. Random rotations and bounding statistical leverage.** The following lemma gives a simple condition on the distribution of the random orthogonal matrix  $\Theta \in \mathbb{R}^{n \times n}$  used to preprocess a data matrix  $A$  so that Condition 1 is applicable to a random vector  $x$  drawn uniformly from the rows of  $\Theta A$ . Its proof is a straightforward application of Lemma 8.

**Lemma 1.** *Fix any  $\tau > 0$  and  $\lambda \geq 0$ . Suppose  $\Theta \in \mathbb{R}^{m \times m}$  is a random orthogonal matrix and  $\kappa > 0$  is a constant such that*

$$(13) \quad \mathbb{E} \left[ \exp \left( \alpha^\top (\sqrt{m} \Theta^\top e_i) \right) \right] \leq \exp \left( \kappa \|\alpha\|^2 / 2 \right), \quad \forall \alpha \in \mathbb{R}^m, \forall i = 1, 2, \dots, m,$$

where  $e_i$  is the  $i$ -th coordinate vector in  $\mathbb{R}^m$ . Let  $A \in \mathbb{R}^{m \times d}$  be any matrix of rank  $d$ , and let  $\Sigma := (1/m)(\Theta A)^\top (\Theta A) = (1/m)A^\top A$ . There exists

$$\rho_\lambda^2 \leq \kappa \left( 1 + 2 \sqrt{\frac{\log m + \tau}{d_{1,\lambda}}} + \frac{2(\log m + \tau)}{d_{1,\lambda}} \right)$$

such that

$$\Pr \left[ \max_{i=1,2,\dots,m} \|(\Sigma + \lambda I)^{-1/2} (\Theta A)^\top e_i\|^2 > \rho_\lambda^2 d_{1,\lambda} \right] \leq e^{-\tau}$$

where  $d_{1,\lambda} := \sum_{j=1}^d \frac{\lambda_j}{\lambda_j + \lambda}$  and  $\{\lambda_1, \lambda_2, \dots, \lambda_d\}$  are the eigenvalues of  $\Sigma$ .

*Proof.* Let  $z_i := \sqrt{m} \Theta^\top e_i$  for each  $i = 1, 2, \dots, m$ . Let  $U \in \mathbb{R}^{m \times d}$  be a matrix of left orthonormal singular vectors of  $(1/\sqrt{m})A$ , and let  $D_\lambda := \text{diag}(\frac{\lambda_1}{\lambda_1 + \lambda}, \frac{\lambda_2}{\lambda_2 + \lambda}, \dots, \frac{\lambda_d}{\lambda_d + \lambda})$ . Note that  $D_\lambda = I$  if  $\lambda = 0$ . We have

$$\|(\Sigma + \lambda I)^{-1/2} (\Theta A)^\top e_i\| = \|\sqrt{m} D_\lambda^{1/2} U^\top \Theta^\top e_i\| = \|D_\lambda^{1/2} U^\top z_i\|.$$

Since  $\text{tr}(UD_\lambda U^\top) = d_{1,\lambda}$ ,  $\text{tr}(UD_\lambda^2 U^\top) \leq d_{1,\lambda}$ , and  $\lambda_{\max}[UD_\lambda U^\top] \leq 1$ , Lemma 8 implies

$$\Pr \left[ \|D_\lambda^{1/2} U^\top z_i\|^2 > \kappa \left( d_{1,\lambda} + 2 \sqrt{d_{1,\lambda}(\log m + \tau)} + 2(\log m + \tau) \right) \right] \leq e^{-\tau} / m.$$

Therefore, by a union bound,

$$\Pr \left[ \max_{i=1,2,\dots,m} \|(\Sigma + \lambda I)^{-1/2} (\Theta A)^\top e_i\|^2 > \kappa \left( d_{1,\lambda} + 2 \sqrt{d_{1,\lambda}(\log m + \tau)} + 2(\log m + \tau) \right) \right] \leq e^{-\tau}. \quad \square$$

Below, we give two simple examples under which the condition (13) in Lemma 1 holds.

**Example 1.** Let  $\Theta$  be distributed uniformly over all  $m \times m$  orthogonal matrices. Fix any  $i = 1, 2, \dots, m$ . The random vector  $v := \Theta^\top e_i$  is distributed uniformly on the unit sphere  $\mathbb{S}^{m-1}$ . Let  $l$  be a  $\chi$  random variable with  $m$  degrees of freedom, so  $z := lv$  follows an isotropic multivariate Gaussian distribution. By Jensen's inequality and the fact that  $\mathbb{E}[\exp(q^\top z)] \leq \exp(\|q\|^2/2)$  for any vector  $q \in \mathbb{R}^m$ ,

$$\begin{aligned} \mathbb{E} \left[ \exp \left( \alpha^\top (\sqrt{m} \Theta^\top e_i) \right) \right] &= \mathbb{E} \left[ \exp \left( \alpha^\top (\sqrt{m} v) \right) \right] \\ &= \mathbb{E} \left[ \mathbb{E} \left[ \exp \left( \frac{\sqrt{m}}{\mathbb{E}[l]} \alpha^\top (\mathbb{E}[l] v) \right) \mid v \right] \right] \\ &\leq \mathbb{E} \left[ \exp \left( \frac{\sqrt{m}}{\mathbb{E}[l]} \alpha^\top (lv) \right) \right] \\ &= \mathbb{E} \left[ \exp \left( \frac{\sqrt{m}}{\mathbb{E}[l]} \alpha^\top z \right) \right] \\ &\leq \exp \left( \frac{\|\alpha\|^2 m}{2 \mathbb{E}[l]^2} \right) \\ &\leq \exp \left( \frac{\|\alpha\|^2}{2} \left( 1 - \frac{1}{4m} - \frac{1}{360m^3} \right)^{-2} \right) \end{aligned}$$

where the last inequality is due to the following lower estimate for  $\chi$  random variables:

$$\mathbb{E}[l] \geq \sqrt{m} \left( 1 - \frac{1}{4m} - \frac{1}{360m^3} \right).$$

Therefore, the condition (13) is satisfied with  $\kappa = 1 + O(1/m)$ .

**Example 2.** Let  $m$  be a power of two, and let  $\Theta := H \text{diag}(s)/\sqrt{m}$ , where  $H \in \{\pm 1\}^{m \times m}$  is the  $m \times m$  Hadamard matrix, and  $s := (s_1, s_2, \dots, s_m) \in \{\pm 1\}^m$  is a vector of  $m$  Rademacher variables (i.e.,  $s_1, s_2, \dots, s_m$  are i.i.d. with  $\Pr[s_1 = 1] = \Pr[s_1 = -1] = 1/2$ ). It is easy to check that  $\Theta$  is an orthogonal matrix. The random rotation  $\Theta$  is a key component of the fast Johnson-Lindenstrauss transform of [1], also used by [7]. It is especially important for the present application because it can be applied to vectors with  $O(m \log m)$  operations, which is significantly faster than the  $\Omega(m^2)$  running time of naïve matrix-vector multiplication.

For each  $i = 1, 2, \dots, m$ , the distribution of  $\sqrt{m}\Theta^\top e_i$  is the same as that of  $s$ , and therefore

$$\mathbb{E} \left[ \exp \left( \alpha^\top \left( \sqrt{m}\Theta^\top e_i \right) \right) \right] = \mathbb{E} \left[ \exp \left( \alpha^\top s \right) \right] \leq \exp(\|\alpha\|^2/2)$$

where the last step follows by Hoeffding's inequality. Therefore, the condition (13) is satisfied with  $\kappa = 1$ .

## 5. PROOFS OF THEOREM 1 AND THEOREM 2

The proof of Theorem 2 uses the decomposition of  $\|\hat{\beta}_\lambda - \beta\|_\Sigma^2$  in Proposition 3, and then bounds each term using the lemmas proved in this section.

The proof of Theorem 1 omits one term from the decomposition in Proposition 3 due to the fact that  $\beta = \beta_\lambda$  when  $\lambda = 0$ ; and it uses a slightly simpler argument to handle the effect of noise (Lemma 6 rather than Lemma 7), which reduces the number of lower-order terms. Other than these differences, the proof is the same as that for Theorem 2 in the special case of  $\lambda = 0$ .

Define

$$(14) \quad \Sigma_\lambda := \Sigma + \lambda I,$$

$$(15) \quad \hat{\Sigma}_\lambda := \hat{\Sigma} + \lambda I, \quad \text{and}$$

$$(16) \quad \begin{aligned} \Delta_\lambda &:= \Sigma_\lambda^{-1/2} (\hat{\Sigma} - \Sigma) \Sigma_\lambda^{-1/2} \\ &= \Sigma_\lambda^{-1/2} (\hat{\Sigma}_\lambda - \Sigma_\lambda) \Sigma_\lambda^{-1/2}. \end{aligned}$$

Recall the basic decomposition from Proposition 3:

$$\|\hat{\beta}_\lambda - \beta\|_\Sigma^2 \leq \left( \|\beta_\lambda - \beta\|_\Sigma + \|\bar{\beta}_\lambda - \beta_\lambda\|_\Sigma + \|\hat{\beta}_\lambda - \bar{\beta}_\lambda\|_\Sigma \right)^2.$$

Section 5.1 first establishes basic properties of  $\beta$  and  $\beta_\lambda$ , which are then used to bound  $\|\beta_\lambda - \beta\|_\Sigma^2$ ; this part is exactly the same as the standard fixed design analysis of ridge regression. Section 5.2 employs probability tail inequalities for the spectral and Frobenius norms of random matrices to bound the matrix errors in estimating  $\Sigma$  with  $\hat{\Sigma}$ . Finally, Section 5.3 and Section 5.4 bound the contributions of approximation error (in  $\|\bar{\beta}_\lambda - \beta_\lambda\|_\Sigma^2$ ) and noise (in  $\|\hat{\beta}_\lambda - \bar{\beta}_\lambda\|_\Sigma^2$ ), respectively, using probability tail inequalities for random vectors as well as the matrix error bounds for  $\hat{\Sigma}$ .

**5.1. Basic properties of  $\beta$  and  $\beta_\lambda$ , and the effect of regularization.** The following propositions are well known in the study of inverse problems:

**Proposition 4** (Normal equations).  $\mathbb{E}[\langle w, x \rangle y] = \mathbb{E}[\langle w, x \rangle \langle \beta, x \rangle]$  for any  $w$ .

*Proof.* It suffices to prove the claim for  $w = v_j$ . Since  $\mathbb{E}[\langle v_j, x \rangle \langle v_{j'}, x \rangle] = 0$  for  $j' \neq j$ , it follows that  $\mathbb{E}[\langle v_j, x \rangle \langle \beta, x \rangle] = \sum_{j'} \beta_{j'} \mathbb{E}[\langle v_j, x \rangle \langle v_{j'}, x \rangle] = \beta_j \mathbb{E}[\langle v_j, x \rangle^2] = \mathbb{E}[\langle v_j, x \rangle y]$ , where the last equality follows from the definition of  $\beta$  in (2).  $\square$

**Proposition 5** (Excess mean squared error).  $\mathbb{E}[(\langle w, x \rangle - y)^2] - \mathbb{E}[(\langle \beta, x \rangle - y)^2] = \mathbb{E}[(\langle w - \beta, x \rangle)^2]$  for any  $w$ .

*Proof.* Directly expanding the squares in the expectations reveals that

$$\begin{aligned}
& \mathbb{E}[(\langle w, x \rangle - y)^2] - \mathbb{E}[(\langle \beta, x \rangle - y)^2] \\
&= \mathbb{E}[\langle w, x \rangle^2] - 2\mathbb{E}[\langle w, x \rangle y] + 2\mathbb{E}[\langle \beta, x \rangle y] - \mathbb{E}[\langle \beta, x \rangle^2] \\
&= \mathbb{E}[\langle w, x \rangle^2] - 2\mathbb{E}[\langle w, x \rangle \langle \beta, x \rangle] + 2\mathbb{E}[\langle \beta, x \rangle \langle \beta, x \rangle] - \mathbb{E}[\langle \beta, x \rangle^2] \\
&= \mathbb{E}[\langle w, x \rangle^2 - 2\langle w, x \rangle \langle \beta, x \rangle + \langle \beta, x \rangle^2] \\
&= \mathbb{E}[\langle w - \beta, x \rangle^2]
\end{aligned}$$

where the third equality follows from Proposition 4.  $\square$

**Proposition 6** (Shrinkage). *For any  $j$ ,*

$$\langle v_j, \beta_\lambda \rangle = \frac{\lambda_j}{\lambda_j + \lambda} \beta_j.$$

*Proof.* Since  $(\Sigma + \lambda I)^{-1} = \sum_j (\lambda_j + \lambda)^{-1} v_j \otimes v_j$ ,

$$\langle v_j, \beta_\lambda \rangle = \langle v_j, (\Sigma + \lambda I)^{-1} \mathbb{E}[xy] \rangle = \frac{1}{\lambda_j + \lambda} \mathbb{E}[\langle v_j, x \rangle y] = \frac{\lambda_j}{\lambda_j + \lambda} \frac{\mathbb{E}[\langle v_j, x \rangle y]}{\langle v_j, x \rangle^2} = \frac{\lambda_j}{\lambda_j + \lambda} \beta_j.$$

$\square$

**Proposition 7** (Effect of regularization).

$$\|\beta - \beta_\lambda\|_\Sigma^2 = \sum_j \frac{\lambda_j}{(\frac{\lambda_j}{\lambda} + 1)^2} \beta_j^2.$$

*Proof.* By Proposition 6,

$$\langle v_j, \beta - \beta_\lambda \rangle = \beta_j - \frac{\lambda_j}{\lambda_j + \lambda} \beta_j = \frac{\lambda}{\lambda_j + \lambda} \beta_j.$$

Therefore,

$$\|\beta - \beta_\lambda\|_\Sigma^2 = \sum_j \lambda_j \left( \frac{\lambda}{\lambda_j + \lambda} \beta_j \right)^2 = \sum_j \frac{\lambda_j}{(\frac{\lambda_j}{\lambda} + 1)^2} \beta_j^2.$$

$\square$

## 5.2. Effect of errors in $\widehat{\Sigma}$ .

**Lemma 2** (Spectral norm error in  $\widehat{\Sigma}$ ). *Assume Condition 1 (with parameter  $\rho_\lambda$ ) holds. Pick  $t > \max\{0, 2.6 - \log \tilde{d}_{1,\lambda}\}$ . With probability at least  $1 - e^{-t}$ ,*

$$\|\Delta_\lambda\| \leq \sqrt{\frac{4\rho_\lambda^2 d_{1,\lambda} (\log \tilde{d}_{1,\lambda} + t)}{n} + \frac{2\rho_\lambda^2 d_{1,\lambda} (\log \tilde{d}_{1,\lambda} + t)}{3n}}$$

where  $\Delta_\lambda$  is defined in (16).

*Proof.* The claim is a consequence of the tail inequality from Lemma 10. First, define

$$\tilde{x} := \Sigma_\lambda^{-1/2} x \quad \text{and} \quad \tilde{\Sigma} := \Sigma_\lambda^{-1/2} \Sigma \Sigma_\lambda^{-1/2}$$

(where  $\Sigma_\lambda$  is defined in (14)), and let

$$\begin{aligned}
Z &:= \tilde{x} \otimes \tilde{x} - \tilde{\Sigma} \\
&= \Sigma_\lambda^{-1/2} (x \otimes x - \Sigma) \Sigma_\lambda^{-1/2}
\end{aligned}$$

so  $\Delta_\lambda = \widehat{\mathbb{E}}[Z]$ . Observe that  $\mathbb{E}[Z] = 0$  and

$$\|Z\| = \max\{\lambda_{\max}[Z], \lambda_{\max}[-Z]\} \leq \max\{\|\tilde{x}\|^2, 1\} \leq \rho_\lambda^2 d_{1,\lambda}$$

where the second inequality follows from Condition 1. Moreover,

$$\mathbb{E}[Z^2] = \mathbb{E}[(\tilde{x} \otimes \tilde{x})^2] - \tilde{\Sigma}^2 = \mathbb{E}[\|\tilde{x}\|^2 (\tilde{x} \otimes \tilde{x})] - \tilde{\Sigma}^2$$

so

$$\begin{aligned}\lambda_{\max}[\mathbb{E}[Z^2]] &\leq \lambda_{\max}[\mathbb{E}[(\tilde{x} \otimes \tilde{x})^2]] \leq \rho_\lambda^2 d_{1,\lambda} \lambda_{\max}[\tilde{\Sigma}] \leq \rho_\lambda^2 d_{1,\lambda} \\ \text{tr}(\mathbb{E}[Z^2]) &\leq \text{tr}(\mathbb{E}[\|\tilde{x}\|^2(\tilde{x} \otimes \tilde{x})]) \leq \rho_\lambda^2 d_{1,\lambda} \text{tr}(\tilde{\Sigma}) = \rho_\lambda^2 d_{1,\lambda}^2.\end{aligned}$$

The claim now follows from Lemma 10 (recall that  $\tilde{d}_{1,\lambda} = \max\{1, d_{1,\lambda}\}$ ).  $\square$

**Lemma 3** (Relative spectral norm error in  $\hat{\Sigma}_\lambda$ ). *If  $\|\Delta_\lambda\| < 1$  where  $\Delta_\lambda$  is defined in (16), then*

$$\|\Sigma_\lambda^{1/2} \hat{\Sigma}_\lambda^{-1} \Sigma_\lambda^{1/2}\| \leq \frac{1}{1 - \|\Delta_\lambda\|}$$

where  $\Sigma_\lambda$  is defined in (14) and  $\hat{\Sigma}_\lambda$  is defined in (15).

*Proof.* Observe that

$$\begin{aligned}\Sigma_\lambda^{-1/2} \hat{\Sigma}_\lambda \Sigma_\lambda^{-1/2} &= \Sigma_\lambda^{-1/2} (\Sigma_\lambda + \hat{\Sigma}_\lambda - \Sigma_\lambda) \Sigma_\lambda^{-1/2} \\ &= I + \Sigma_\lambda^{-1/2} (\hat{\Sigma}_\lambda - \Sigma_\lambda) \Sigma_\lambda^{-1/2} \\ &= I + \Delta_\lambda,\end{aligned}$$

and that

$$\lambda_{\min}[I + \Delta_\lambda] \geq 1 - \|\Delta_\lambda\| > 0$$

by the assumption  $\|\Delta_\lambda\| < 1$  and Weyl's theorem [10]. Therefore,

$$\|\Sigma_\lambda^{1/2} \hat{\Sigma}_\lambda^{-1} \Sigma_\lambda^{1/2}\| = \lambda_{\max}[(\Sigma_\lambda^{-1/2} \hat{\Sigma}_\lambda \Sigma_\lambda^{-1/2})^{-1}] = \lambda_{\max}[(I + \Delta_\lambda)^{-1}] = \frac{1}{\lambda_{\min}[I + \Delta_\lambda]} \leq \frac{1}{1 - \|\Delta_\lambda\|}.$$

$\square$

**Lemma 4** (Frobenius norm error in  $\hat{\Sigma}$ ). *Assume Condition 1 (with parameter  $\rho_\lambda$ ) holds. Pick any  $t > 0$ . With probability at least  $1 - e^{-t}$ ,*

$$\begin{aligned}\|\Delta_\lambda\|_{\text{F}} &\leq \sqrt{\frac{\mathbb{E}[\|\Sigma_\lambda^{-1/2} x\|^4] - d_{2,\lambda}}{n}} (1 + \sqrt{8t}) + \frac{4\sqrt{\rho_\lambda^4 d_{1,\lambda}^2 + d_{2,\lambda} t}}{3n} \\ &\leq \sqrt{\frac{\rho_\lambda^2 d_{1,\lambda}^2 - d_{2,\lambda}}{n}} (1 + \sqrt{8t}) + \frac{4\sqrt{\rho_\lambda^4 d_{1,\lambda}^2 + d_{2,\lambda} t}}{3n}\end{aligned}$$

where  $\Delta_\lambda$  is defined in (16).

*Proof.* The claim is a consequence of the tail inequality in Lemma 9. As in the proof of Lemma 2, define  $\tilde{x} := \Sigma_\lambda^{-1/2} x$  and  $\tilde{\Sigma} := \Sigma_\lambda^{-1/2} \Sigma \Sigma_\lambda^{-1/2}$ , and let  $Z := \tilde{x} \otimes \tilde{x} - \tilde{\Sigma}$  so  $\Delta_\lambda = \widehat{\mathbb{E}}[Z]$ . Now endow the space of self-adjoint linear operators with the inner product given by  $\langle A, B \rangle_{\text{F}} := \text{tr}(AB)$ , and note that this inner product induces the Frobenius norm  $\|M\|_{\text{F}} = \langle M, M \rangle_{\text{F}}$ . Observe that  $\mathbb{E}[Z] = 0$  and

$$\begin{aligned}\|Z\|_{\text{F}}^2 &= \langle \tilde{x} \otimes \tilde{x} - \tilde{\Sigma}, \tilde{x} \otimes \tilde{x} - \tilde{\Sigma} \rangle_{\text{F}} \\ &= \langle \tilde{x} \otimes \tilde{x}, \tilde{x} \otimes \tilde{x} \rangle_{\text{F}} - 2\langle \tilde{x} \otimes \tilde{x}, \tilde{\Sigma} \rangle_{\text{F}} + \langle \tilde{\Sigma}, \tilde{\Sigma} \rangle_{\text{F}} \\ &= \|\tilde{x}\|^4 - 2\|\tilde{x}\|_{\tilde{\Sigma}}^2 + \text{tr}(\tilde{\Sigma}^2) \\ &= \|\tilde{x}\|^4 - 2\|\tilde{x}\|_{\tilde{\Sigma}}^2 + d_{2,\lambda} \\ &\leq \rho_\lambda^4 d_{1,\lambda}^2 + d_{2,\lambda},\end{aligned}$$

where the inequality follows from Condition 1. Moreover,

$$\begin{aligned}\mathbb{E}[\|Z\|_{\text{F}}^2] &= \mathbb{E}[\langle \tilde{x} \otimes \tilde{x}, \tilde{x} \otimes \tilde{x} \rangle_{\text{F}}] - \langle \tilde{\Sigma}, \tilde{\Sigma} \rangle_{\text{F}} \\ &= \mathbb{E}[\|\tilde{x}\|^4] - d_{2,\lambda} \\ &\leq \rho_\lambda^2 d_{1,\lambda} \mathbb{E}[\|\tilde{x}\|^2] - d_{2,\lambda} \\ &= \rho_\lambda^2 d_{1,\lambda}^2 - d_{2,\lambda},\end{aligned}$$

where the inequality again uses Condition 1. The claim now follows from Lemma 9.  $\square$

### 5.3. Effect of approximation error.

**Lemma 5** (Effect of approximation error). *Assume Condition 1 (with parameter  $\rho_\lambda$ ) and Condition 3 (with parameter  $b_\lambda$ ) hold. Pick any  $t > 0$ . If  $\|\Delta_\lambda\| < 1$  where  $\Delta_\lambda$  is defined in (16), then*

$$\|\bar{\beta}_\lambda - \beta_\lambda\|_\Sigma \leq \frac{1}{1 - \|\Delta_\lambda\|} \|\widehat{\mathbb{E}}[x \text{ approx}_\lambda(x) - \lambda\beta_\lambda]\|_{\Sigma_\lambda^{-1}},$$

where  $\bar{\beta}_\lambda$  is defined in (12),  $\beta_\lambda$  is defined in (8),  $\text{approx}_\lambda(x)$  is defined in (9), and  $\Sigma_\lambda$  is defined in (14). Moreover, with probability at least  $1 - e^{-t}$ ,

$$\begin{aligned} & \|\widehat{\mathbb{E}}[x \text{ approx}_\lambda(x) - \lambda\beta_\lambda]\|_{\Sigma_\lambda^{-1}} \\ & \leq \sqrt{\frac{\mathbb{E}[\|\Sigma_\lambda^{-1/2}(x \text{ approx}_\lambda(x) - \lambda\beta_\lambda)\|^2]}{n}}(1 + \sqrt{8t}) + \frac{4(b_\lambda\sqrt{d_{1,\lambda}} + \|\beta - \beta_\lambda\|_\Sigma)t}{3n} \\ & \leq \sqrt{\frac{2(\rho_\lambda^2 d_{1,\lambda} \mathbb{E}[\text{approx}_\lambda(x)^2] + \|\beta - \beta_\lambda\|_\Sigma^2)}{n}}(1 + \sqrt{8t}) + \frac{4(b_\lambda\sqrt{d_{1,\lambda}} + \|\beta - \beta_\lambda\|_\Sigma)t}{3n}. \end{aligned}$$

*Proof.* By the definitions of  $\bar{\beta}_\lambda$  and  $\beta_\lambda$ ,

$$\begin{aligned} \bar{\beta}_\lambda - \beta_\lambda &= \widehat{\Sigma}_\lambda^{-1} \left( \widehat{\mathbb{E}}[x\mathbb{E}[y|x]] - \widehat{\Sigma}_\lambda\beta_\lambda \right) \\ &= \Sigma_\lambda^{-1/2} (\Sigma_\lambda^{1/2} \widehat{\Sigma}_\lambda^{-1} \Sigma_\lambda^{1/2}) \Sigma_\lambda^{-1/2} \left( \widehat{\mathbb{E}}[x(\text{approx}(x) + \langle \beta, x \rangle)] - \widehat{\Sigma}_\lambda\beta_\lambda - \lambda\beta_\lambda \right) \\ &= \Sigma_\lambda^{-1/2} (\Sigma_\lambda^{1/2} \widehat{\Sigma}_\lambda^{-1} \Sigma_\lambda^{1/2}) \Sigma_\lambda^{-1/2} \left( \widehat{\mathbb{E}}[x(\text{approx}(x) + \langle \beta, x \rangle - \langle \beta_\lambda, x \rangle)] - \lambda\beta_\lambda \right) \\ &= \Sigma_\lambda^{-1/2} (\Sigma_\lambda^{1/2} \widehat{\Sigma}_\lambda^{-1} \Sigma_\lambda^{1/2}) \Sigma_\lambda^{-1/2} \left( \widehat{\mathbb{E}}[x \text{ approx}_\lambda(x) - \lambda\beta_\lambda] \right). \end{aligned}$$

Therefore, using the submultiplicative property of the spectral norm,

$$\begin{aligned} \|\bar{\beta}_\lambda - \beta_\lambda\|_\Sigma &\leq \|\Sigma^{1/2} \Sigma_\lambda^{-1/2}\| \|\Sigma_\lambda^{1/2} \widehat{\Sigma}_\lambda^{-1} \Sigma_\lambda^{1/2}\| \|\widehat{\mathbb{E}}[x \text{ approx}_\lambda(x) - \lambda\beta_\lambda]\|_{\Sigma_\lambda^{-1}} \\ &\leq \frac{1}{1 - \|\Delta_\lambda\|} \|\widehat{\mathbb{E}}[x \text{ approx}_\lambda(x) - \lambda\beta_\lambda]\|_{\Sigma_\lambda^{-1}} \end{aligned}$$

where the second inequality follows from Lemma 3 and because

$$\|\Sigma^{1/2} \Sigma_\lambda^{-1/2}\|^2 = \lambda_{\max}[\Sigma_\lambda^{-1/2} \Sigma \Sigma_\lambda^{-1/2}] = \max_i \frac{\lambda_i}{\lambda_i + \lambda} \leq 1.$$

The second part of the claim is a consequence of the tail inequality in Lemma 9. Observe that  $\mathbb{E}[x \text{ approx}(x)] = \mathbb{E}[x(\mathbb{E}[y|x] - \langle \beta, x \rangle)] = 0$  by Proposition 4, and that  $\mathbb{E}[x(\beta - \beta_\lambda, x)] - \lambda\beta_\lambda = \Sigma\beta - (\Sigma + \lambda I)\beta_\lambda = 0$ . Therefore,

$$\mathbb{E}[\Sigma_\lambda^{-1/2}(x \text{ approx}_\lambda(x) - \lambda\beta_\lambda)] = \Sigma_\lambda^{-1/2} \mathbb{E}[x(\text{approx}(x) + \langle \beta - \beta_\lambda, x \rangle) - \lambda\beta_\lambda] = 0.$$

Moreover, by Proposition 6 and Proposition 7,

$$\begin{aligned} \|\lambda \Sigma_\lambda^{-1/2} \beta_\lambda\|^2 &= \sum_j \frac{\lambda^2}{\lambda_j + \lambda} \langle v_j, \beta_\lambda \rangle^2 \\ &= \sum_j \frac{\lambda^2}{\lambda_j + \lambda} \left( \frac{\lambda_j}{\lambda_j + \lambda} \beta_j \right)^2 \\ &\leq \sum_j \frac{\lambda^2}{\lambda_j + \lambda} \left( \frac{\lambda_j}{\lambda_j + \lambda} \right) \beta_j^2 \\ &= \sum_j \frac{\lambda_j}{(\frac{\lambda_j}{\lambda} + 1)^2} \beta_j^2 \\ &= \|\beta - \beta_\lambda\|_\Sigma^2. \end{aligned} \tag{17}$$

Combining the inequality from (17) with Condition 3 and the triangle inequality, it follows that

$$\begin{aligned}\|\Sigma_\lambda^{-1/2}(x \text{ approx}_\lambda(x) - \lambda\beta_\lambda)\| &\leq \|\Sigma_\lambda^{-1/2}x \text{ approx}_\lambda(x)\| + \|\lambda\Sigma_\lambda^{-1/2}\beta_\lambda\| \\ &\leq b_\lambda\sqrt{d_{1,\lambda}} + \|\beta - \beta_\lambda\|_\Sigma.\end{aligned}$$

Finally, by the triangle inequality, the fact  $(a+b)^2 \leq 2(a^2+b^2)$ , the inequality from (17), and Condition 1,

$$\begin{aligned}\mathbb{E}[\|\Sigma_\lambda^{-1/2}(x \text{ approx}_\lambda(x) - \lambda\beta_\lambda)\|^2] &\leq 2(\mathbb{E}[\|\Sigma_\lambda^{-1/2}x \text{ approx}_\lambda(x)\|^2] + \|\beta_\lambda - \beta\|_\Sigma^2) \\ &\leq 2(\rho_\lambda^2 d_{1,\lambda} \mathbb{E}[\text{approx}_\lambda(x)^2] + \|\beta_\lambda - \beta\|_\Sigma^2).\end{aligned}$$

The claim now follows from Lemma 9.  $\square$

#### 5.4. Effect of noise.

**Lemma 6** (Effect of noise,  $\lambda = 0$ ). *Assume  $\lambda = 0$ . Assume Condition 2 (with parameter  $\sigma$ ) holds. Pick any  $t > 0$ . With probability at least  $1 - e^{-t}$ , either  $\|\Delta_0\| \geq 1$ , or*

$$\|\Delta_0\| < 1 \quad \text{and} \quad \|\bar{\beta}_0 - \hat{\beta}_0\|_\Sigma^2 \leq \frac{1}{1 - \|\Delta_0\|} \cdot \frac{\sigma^2(d + 2\sqrt{dt} + 2t)}{n},$$

where  $\Delta_0$  is defined in (16).

*Proof.* Observe that

$$\|\bar{\beta}_0 - \hat{\beta}_0\|_\Sigma^2 \leq \|\Sigma^{1/2}\hat{\Sigma}^{-1/2}\|^2 \|\bar{\beta}_0 - \hat{\beta}_0\|_\Sigma^2 = \|\Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2}\| \|\bar{\beta}_0 - \hat{\beta}_0\|_\Sigma^2;$$

and if  $\|\Delta_0\| < 1$ , then  $\|\Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2}\| \leq 1/(1 - \|\Delta_0\|)$  by Lemma 3.

Let  $\xi := (\text{noise}(x_1), \text{noise}(x_2), \dots, \text{noise}(x_n))$  be the random vector whose  $i$ -th component is  $\text{noise}(x_i) = y_i - \mathbb{E}[y_i|x_i]$ . By the definition of  $\hat{\beta}_0$  and  $\bar{\beta}_0$

$$\|\hat{\beta}_0 - \bar{\beta}_0\|_\Sigma^2 = \|\hat{\Sigma}^{-1/2}\hat{\mathbb{E}}[x(y - \mathbb{E}[y|x])]\|^2 = \xi^\top \hat{K} \xi,$$

where  $\hat{K} \in \mathbb{R}^{n \times n}$  is the symmetric matrix whose  $(i, j)$ -th entry is  $\hat{K}_{i,j} := n^{-2} \langle \hat{\Sigma}^{-1/2}x_i, \hat{\Sigma}^{-1/2}x_j \rangle$ . Note that the nonzero eigenvalues of  $\hat{K}$  are the same as those of

$$\frac{1}{n} \hat{\mathbb{E}} \left[ (\hat{\Sigma}^{-1/2}x) \otimes (\hat{\Sigma}^{-1/2}x) \right] = \frac{1}{n} \hat{\Sigma}^{-1/2} \hat{\Sigma} \hat{\Sigma}^{-1/2} = \frac{1}{n} I.$$

By Lemma 8, with probability at least  $1 - e^{-t}$  (conditioned on  $x_1, x_2, \dots, x_n$ ),

$$\xi^\top \hat{K} \xi \leq \sigma^2(\text{tr}(\hat{K}) + 2\sqrt{\text{tr}(\hat{K}^2)t} + 2\lambda_{\max}(\hat{K})t) = \frac{\sigma^2(d + 2\sqrt{dt} + 2t)}{n}.$$

The claim follows.  $\square$

**Lemma 7** (Effect of noise,  $\lambda \geq 0$ ). *Assume Condition 2 (with parameter  $\sigma$ ) holds. Pick any  $t > 0$ . Let  $K$  be the  $n \times n$  symmetric matrix whose  $(i, j)$ -th entry is*

$$K_{i,j} := \frac{1}{n^2} \langle \Sigma^{1/2}\hat{\Sigma}_\lambda^{-1}x_i, \Sigma^{1/2}\hat{\Sigma}_\lambda^{-1}x_j \rangle,$$

where  $\hat{\Sigma}_\lambda$  is defined in (15). With probability at least  $1 - e^{-t}$ ,

$$\|\bar{\beta}_\lambda - \hat{\beta}_\lambda\|_\Sigma^2 \leq \sigma^2(\text{tr}(K) + 2\sqrt{\text{tr}(K)\lambda_{\max}(K)t} + 2\lambda_{\max}(K)t).$$

Moreover, if  $\|\Delta_\lambda\| < 1$  where  $\Delta_\lambda$  is defined in (16), then

$$\lambda_{\max}(K) \leq \frac{1}{n(1 - \|\Delta_\lambda\|)} \quad \text{and} \quad \text{tr}(K) \leq \frac{d_{2,\lambda} + \sqrt{d_{2,\lambda}\|\Delta_\lambda\|_{\mathbb{F}}^2}}{n(1 - \|\Delta_\lambda\|)^2}.$$

*Proof.* Let  $\xi := (\text{noise}(x_1), \text{noise}(x_2), \dots, \text{noise}(x_n))$  be the random vector whose  $i$ -th component is  $\text{noise}(x_i) = y_i - \mathbb{E}[y_i|x_i]$ . By the definition of  $\hat{\beta}_\lambda$ ,  $\bar{\beta}_\lambda$ , and  $K$ ,

$$\|\hat{\beta}_\lambda - \bar{\beta}_\lambda\|_\Sigma^2 = \|\hat{\Sigma}_\lambda^{-1} \widehat{\mathbb{E}}[x(y - \mathbb{E}[y|x])]\|_\Sigma^2 = \xi^\top K \xi.$$

By Lemma 8, with probability at least  $1 - e^{-t}$  (conditioned on  $x_1, x_2, \dots, x_n$ ),

$$\begin{aligned} \xi^\top K \xi &\leq \sigma^2 (\text{tr}(K) + 2\sqrt{\text{tr}(K^2)t} + 2\lambda_{\max}(K)t) \\ &\leq \sigma^2 (\text{tr}(K) + 2\sqrt{\text{tr}(K)\lambda_{\max}(K)t} + 2\lambda_{\max}(K)t), \end{aligned}$$

where the second inequality follows from von Neumann's theorem [10].

Note that the nonzero eigenvalues of  $K$  are the same as that of

$$\frac{1}{n} \widehat{\mathbb{E}} \left[ (\Sigma^{1/2} \hat{\Sigma}_\lambda^{-1} x) \otimes (\Sigma^{1/2} \hat{\Sigma}_\lambda^{-1} x) \right] = \frac{1}{n} \Sigma^{1/2} \hat{\Sigma}_\lambda^{-1} \widehat{\Sigma} \hat{\Sigma}_\lambda^{-1} \Sigma^{1/2}.$$

To bound  $\lambda_{\max}(K)$ , observe that by the submultiplicative property of the spectral norm and Lemma 3,

$$\begin{aligned} n\lambda_{\max}(K) &= \|\Sigma^{1/2} \hat{\Sigma}_\lambda^{-1} \widehat{\Sigma}^{1/2}\|^2 \\ &\leq \|\Sigma^{1/2} \Sigma_\lambda^{-1/2}\|^2 \|\Sigma_\lambda^{1/2} \hat{\Sigma}_\lambda^{-1/2}\|^2 \|\hat{\Sigma}_\lambda^{-1/2} \widehat{\Sigma}^{1/2}\|^2 \\ &\leq \|\Sigma_\lambda^{1/2} \hat{\Sigma}_\lambda^{-1/2}\|^2 \\ &= \|\Sigma_\lambda^{1/2} \hat{\Sigma}_\lambda^{-1} \Sigma_\lambda^{1/2}\| \\ &\leq \frac{1}{1 - \|\Delta_\lambda\|}. \end{aligned}$$

To bound  $\text{tr}(K)$ , first define the  $\lambda$ -whitened versions of  $\Sigma$ ,  $\widehat{\Sigma}$ , and  $\hat{\Sigma}_\lambda$  as

$$\begin{aligned} \Sigma_w &:= \Sigma_\lambda^{-1/2} \Sigma \Sigma_\lambda^{-1/2}, \\ \widehat{\Sigma}_w &:= \Sigma_\lambda^{-1/2} \widehat{\Sigma} \Sigma_\lambda^{-1/2}, \\ \widehat{\Sigma}_{\lambda,w} &:= \Sigma_\lambda^{-1/2} \widehat{\Sigma}_\lambda \Sigma_\lambda^{-1/2}. \end{aligned}$$

Using these definitions with the cycle property of the trace,

$$\begin{aligned} n \text{tr}(K) &= \text{tr}(\Sigma^{1/2} \hat{\Sigma}_\lambda^{-1} \widehat{\Sigma} \hat{\Sigma}_\lambda^{-1} \Sigma^{1/2}) \\ &= \text{tr}(\widehat{\Sigma}_\lambda^{-1} \widehat{\Sigma} \hat{\Sigma}_\lambda^{-1} \Sigma) \\ &= \text{tr}(\widehat{\Sigma}_{\lambda,w}^{-1} \widehat{\Sigma}_w \widehat{\Sigma}_{\lambda,w}^{-1} \Sigma_w). \end{aligned}$$

Let  $\{\lambda_j[M]\}$  denote the eigenvalues of a linear operator  $M$ . By von Neumann's theorem [10],

$$\text{tr}(\widehat{\Sigma}_{\lambda,w}^{-1} \widehat{\Sigma}_w \widehat{\Sigma}_{\lambda,w}^{-1} \Sigma_w) \leq \sum_j \lambda_j[\widehat{\Sigma}_{\lambda,w}^{-1} \widehat{\Sigma}_w \widehat{\Sigma}_{\lambda,w}^{-1}] \lambda_j[\Sigma_w]$$

and by Ostrowski's theorem [10],

$$\lambda_j[\widehat{\Sigma}_{\lambda,w}^{-1} \widehat{\Sigma}_w \widehat{\Sigma}_{\lambda,w}^{-1}] \leq \lambda_{\max}[\widehat{\Sigma}_{\lambda,w}^{-2}] \lambda_j[\widehat{\Sigma}_w].$$

Therefore

$$\begin{aligned}
\mathrm{tr}(\widehat{\Sigma}_{\lambda,w}^{-1} \widehat{\Sigma}_w \widehat{\Sigma}_{\lambda,w}^{-1} \Sigma_w) &\leq \lambda_{\max}[\widehat{\Sigma}_{\lambda,w}^{-2}] \sum_j \lambda_j[\widehat{\Sigma}_w] \lambda_j[\Sigma_w] \\
&\leq \frac{1}{(1 - \|\Delta_\lambda\|)^2} \sum_j \lambda_j[\widehat{\Sigma}_w] \lambda_j[\Sigma_w] \\
&= \frac{1}{(1 - \|\Delta_\lambda\|)^2} \sum_j \left( \lambda_j[\Sigma_w]^2 + (\lambda_j[\widehat{\Sigma}_w] - \lambda_j[\Sigma_w]) \lambda_j[\Sigma_w] \right) \\
&\leq \frac{1}{(1 - \|\Delta_\lambda\|)^2} \left( \sum_j \lambda_j[\Sigma_w]^2 + \sqrt{\sum_j (\lambda_j[\widehat{\Sigma}_w] - \lambda_j[\Sigma_w])^2} \sqrt{\sum_j \lambda_j[\Sigma_w]^2} \right) \\
&= \frac{1}{(1 - \|\Delta_\lambda\|)^2} \left( d_{2,\lambda} + \sqrt{\sum_j (\lambda_j[\widehat{\Sigma}_w] - \lambda_j[\Sigma_w])^2} \sqrt{d_{2,\lambda}} \right) \\
&\leq \frac{1}{(1 - \|\Delta_\lambda\|)^2} \left( d_{2,\lambda} + \|\widehat{\Sigma}_w - \Sigma_w\|_{\mathrm{F}} \sqrt{d_{2,\lambda}} \right) \\
&= \frac{1}{(1 - \|\Delta_\lambda\|)^2} \left( d_{2,\lambda} + \|\Delta_\lambda\|_{\mathrm{F}} \sqrt{d_{2,\lambda}} \right),
\end{aligned}$$

where the second inequality follows from Lemma 3, the third inequality follows from Cauchy-Schwarz, and the fourth inequality follows from Mirsky's theorem [21].  $\square$

*Acknowledgements.* The authors thank Dean Foster, David McAllester, and Robert Stine for many insightful discussions.

#### REFERENCES

- [1] N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. *SIAM J. Comput.*, 39(1):302–322, 2009.
- [2] J.-Y. Audibert and O. Catoni. Linear regression through PAC-Bayesian truncation, 2010. arXiv:1010.0072.
- [3] J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *The Annals of Statistics*, 30(5):2766–2794, 2011.
- [4] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.
- [5] O. Catoni. *Statistical Learning Theory and Stochastic Optimization, Lectures on Probability and Statistics, Ecole d'Été de Probabilités de Saint-Flour XXXI – 2001*, volume 1851 of *Lecture Notes in Mathematics*. Springer, 2004.
- [6] P. Drineas and M. W. Mahoney. Effective resistances, statistical leverage, and applications to linear equation solving, 2010. arXiv:1005.3097.
- [7] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2010.
- [8] L. Györfi, M. Kohler, A. Kryžak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2004.
- [9] A. E. Hoerl. Application of ridge analysis to regression problems. *Chemical Engineering Progress*, 58:54–59, 1962.
- [10] R. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.
- [11] D. Hsu, S. M. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors, 2011. arXiv:1110.2842.
- [12] D. Hsu, S. M. Kakade, and T. Zhang. Tail inequalities for sums of random matrices that depend on the intrinsic dimension. *Electronic Communications in Probability*, 17(14):1–13, 2012.
- [13] D. Hsu and S. Sabato. Loss minimization and parameter estimation with heavy tails, 2013. arXiv:1307.1827.
- [14] V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.
- [15] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.
- [16] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, second edition, 1998.
- [17] M. Nussbaum. Minimax risk: Pinsker bound. In S. Kotz, editor, *Encyclopedia of Statistical Sciences, Update Volume 3*, pages 451–460. Wiley, New York, 1999.
- [18] V. Rokhlin and M. Tygert. A fast randomized algorithm for overdetermined linear least-squares regression. *Proc. Natl. Acad. Sci. USA*, 105(36):13212–13217, 2008.
- [19] S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximations*, 26:153–172, 2007.

- [20] I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *Proceedings of the 22nd Annual Conference on Learning Theory*, pages 79–93, 2009.
- [21] G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. Academic Press, 1990.
- [22] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10:1040–1053, 1982.
- [23] T. Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17:2077–2098, 2005.

## APPENDIX A. PROBABILITY TAIL INEQUALITIES

The following probability tail inequalities are used in our analysis. These specific inequalities were chosen in order to satisfy the general conditions set up in Section 2.4; however, our analysis can specialize or generalize with the availability of other tail inequalities of these sorts.

The first tail inequality is for positive semidefinite quadratic forms of a subgaussian random vector. It generalizes a standard tail inequality for Gaussian random vectors based on linear combinations of  $\chi^2$  random variables [15].

**Lemma 8** (Quadratic forms of a subgaussian random vector; [11]). *Let  $\xi$  be a random vector taking values in  $\mathbb{R}^n$  such that for some  $c \geq 0$ ,*

$$\mathbb{E}[\exp(\langle u, \xi \rangle)] \leq \exp(c\|u\|^2/2), \quad \forall u \in \mathbb{R}^n.$$

*For all symmetric positive semidefinite matrices  $K \succeq 0$ , and all  $t > 0$ ,*

$$\Pr \left[ \xi^\top K \xi > c \left( \text{tr}(K) + 2\sqrt{\text{tr}(K^2)t} + 2\|K\|t \right) \right] \leq e^{-t}.$$

The next lemma is a tail inequality for sums of bounded random vectors; it is a standard application of Bernstein’s inequality.

**Lemma 9** (Vector Bernstein bound; see, e.g., [11]). *Let  $x_1, x_2, \dots, x_n$  be independent random vectors such that*

$$\sum_{i=1}^n \mathbb{E}[\|x_i\|^2] \leq v \quad \text{and} \quad \|x_i\| \leq r$$

*for all  $i = 1, 2, \dots, n$ , almost surely. Let  $s := x_1 + x_2 + \dots + x_n$ . For all  $t > 0$ ,*

$$\Pr \left[ \|s\| > \sqrt{v}(1 + \sqrt{8t}) + (4/3)rt \right] \leq e^{-t}$$

The last tail inequality concerns the spectral accuracy of an empirical second moment matrix.

**Lemma 10** (Matrix Bernstein bound; [12]). *Let  $X$  be a random matrix, and  $r > 0$ ,  $v > 0$ , and  $k > 0$  be such that, almost surely,*

$$\mathbb{E}[X] = 0, \quad \lambda_{\max}[X] \leq r, \quad \lambda_{\max}[\mathbb{E}[X^2]] \leq v, \quad \text{tr}(\mathbb{E}[X^2]) \leq vk.$$

*If  $X_1, X_2, \dots, X_n$  are independent copies of  $X$ , then for any  $t > 0$ ,*

$$\Pr \left[ \lambda_{\max} \left[ \frac{1}{n} \sum_{i=1}^n X_i \right] > \sqrt{\frac{2vt}{n}} + \frac{rt}{3n} \right] \leq kt(e^t - t - 1)^{-1}.$$

*If  $t \geq 2.6$ , then  $t(e^t - t - 1)^{-1} \leq e^{-t/2}$ .*

(D. Hsu) DEPARTMENT OF COMPUTER SCIENCE, COLUMBIA UNIVERSITY, 450 COMPUTER SCIENCE BUILDING, 1214 AMSTERDAM AVENUE, MAILCODE: 0401, NEW YORK, NY 10027-7003

*E-mail address*, D. Hsu: [djhsu@cs.columbia.edu](mailto:djhsu@cs.columbia.edu)

(S. M. Kakade) MICROSOFT RESEARCH, ONE MEMORIAL DRIVE, CAMBRIDGE, MA, 02142

*E-mail address*, S.M. Kakade: [skakade@microsoft.com](mailto:skakade@microsoft.com)

(T. Zhang) DEPARTMENT OF STATISTICS, RUTGERS UNIVERSITY, 501 HILL CENTER, 110 FRELINGHUYSEN ROAD, PISCATAWAY, NJ 08854

*E-mail address*, T. Zhang: [tzhang@stat.rutgers.edu](mailto:tzhang@stat.rutgers.edu)