Random Design Analysis of Ridge Regression

Daniel Hsu DAHSU@MICROSOFT.COM
Microsoft Research

Sham M. Kakade SKAKADE@MICROSOFT.COM
Microsoft Research

Tong Zhang TZHANG@STAT.RUTGERS.EDU
Rutgers University

Editor: Shie Mannor, Nathan Srebro, Bob Williamson

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 proof 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, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, 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 β (though this hypothesis need not be true). Here, the ϵ_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 β with accuracy measured under a particular norm.

The random design setting stands in contrast to the fixed design setting, where the covariates $x_1, x_2, \ldots, x_n$ are fixed (*i.e.*, deterministic), and only the responses $y_1, y_2, \ldots, y_n$ 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* estimator (Hoerl, 1962) 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, as well as the effect of approximating the true regression function by a linear function in the case the model is misspecified. Neither of these effects is present in the fixed design analysis of ridge regression. The random design analysis shows that the effect of errors in the estimated covariance structure is minimal—it is typically 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 with the noise variance when the approximation error vanishes.

One feature of the analysis in this work is that it applies to the ridge estimator with an arbitrary setting of λ. The estimation error is given in terms of the spectrum of the covariance $\mathbb{E}[x \otimes x]$ and the particular choice of λ. When $\lambda = 0$, we obtain an analysis of ordinary least squares, applicable when the spectrum is finite (*i.e.*, when the covariates live in a finite dimensional space). More generally, the convergence rate can be optimized by appropriately setting λ based on assumptions about the spectrum.

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. An application to smoothing splines is provided in Appendix A, and the proof of the main results are given in the Appendix B.

2. Preliminaries

2.1. Notation

Unless otherwise specified, all vectors in this work are assumed to live in a (possibly infinite dimensional) separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$. Let $\| \cdot \|_M$ for a self-adjoint positive semidefinite linear operator $M \succeq 0$ denote the vector norm given by $\|v\|_M := \sqrt{\langle v, Mv \rangle}$. When M is omitted, it is assumed to be the identity, 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 $\|M\|_{\mathrm{F}}$ denote its Frobenius norm, *i.e.*, $\|M\|_{\mathrm{F}} = \sqrt{\operatorname{tr}(M^* M)}$. If M is self-adjoint, $\|M\|_{\mathrm{F}} = \sqrt{\operatorname{tr}(M^2)}$. 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. Let $\{v_j\}$ be the eigenvectors of

$$\varSigma := \mathbb{E}[x \otimes x], \tag{1}$$

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

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

(assumed to be non-zero for convenience). Let β achieve the minimum *mean squared error* over all linear functions, *i.e.*,

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

so that:

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

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 21).

2.3. The ridge and ordinary least squares estimators

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

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

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

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

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

The following conditions on the covariate x ensure that the second-moment operator Σ can be estimated from a random sample with sufficient accuracy. The first requires that the spectrum of Σ decays sufficiently fast at regularization level λ.

Condition 1 (Spectral decay at λ) *For $p \in \{1, 2\}$,*

$$d_{p,\lambda} := \sum_j \left(\frac{\lambda_j}{\lambda_j + \lambda}\right)^p < \infty. \tag{5}$$

For technical reasons, we also use the quantity

$$\tilde{d}_{1,\lambda} := \max\{d_{1,\lambda}, 1\} \tag{6}$$

merely to simplify certain probability tail inequalities in the main result in the peculiar case that $\lambda \to \infty$ (upon which $d_{1,\lambda} \to 0$). We remark that $d_{2,\lambda}$ appears naturally arises in the standard fixed design analysis of ridge regression (see Proposition 5), and that $d_{1,\lambda}$ was also used by Zhang (2005) in his random design analysis of (kernel) ridge regression. It is easy to see that $d_{2,\lambda} \le d_{1,\lambda}$, and that in in covariate spaces of finite dimension $d < \infty$, we have $d_{p,\lambda} \le d$ with equality iff $\lambda = 0$.

The second 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 λ-*whitening*, is that the expectation $\mathbb{E}[\|(\Sigma + \lambda I)^{-1/2} x\|^2]$ may only be small for sufficiently large λ (as in Condition 1), as

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

Condition 2 (Bounded statistical leverage at λ) *There exists finite $\rho_\lambda \ge 1$ such that, almost surely,*

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

The hard "almost sure" bound in Condition 2 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, in finite dimensional settings, it is easy to replace Condition 2 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 (Finite dimensional setting and $\lambda = 0$) *If $\lambda = 0$ and the dimension of the covariate space is d, then Condition 2 reduces to the requirement that there exists a finite $\rho_0 \ge 1$ such that, almost surely,*

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

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

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

in which case Condition 2 is satisfied with

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

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

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

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

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

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

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

Condition 3 (Subgaussian noise) *There exists finite $\sigma \geq 0$ such that, almost surely,*

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

Condition 3 is satisfied, for instance, if $\operatorname{noise}(x)$ is normally distributed with mean zero and variance σ^2.

For the next condition, define β_λ to be the minimizer of the regularized mean squared error, *i.e.*,

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

and also define

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

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

Condition 4 (Bounded approximation error at λ) *There exist finite $b_\lambda \geq 0$ such that, almost surely,*

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

The hard "almost sure" bound in Condition 4 can easily be relaxed to moment conditions, but we do not consider it here for sake of simplicity. We also remark that b_λ only appears in lower-order terms in the main bounds.

Remark 3 (Finite dimensional setting and $\lambda = 0$) *If $\lambda = 0$ and the dimension of the covariate space is d, then Condition 4 reduces to the requirement that there exists a finite $b_0 \geq 0$ such that, almost surely,*

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

Remark 4 (Bounded $|\operatorname{approx}(x)|$) *If $|\operatorname{approx}(x)| \leq a$ almost surely and Condition 2 (with parameter ρ_λ) holds, then*

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

where the first and last inequalities use Condition 2, the second inequality uses the definition of $\operatorname{approx}_\lambda(x)$ 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 23. In this case, Condition 4 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 2 and Condition 4 are satisfied with

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

as per Remark 2.

2.5. Related work

Many classical analyses of the ridge and ordinary least squares estimators in the random design setting (*e.g.*, in the context of non-parametric estimators) do not actually show non-asymptotic $O(d/n)$ 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 Bayes error, plus a $O(d/n)$ term (Györfi et al., 2004):

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

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 non-standard restrictions on applicability or additional dependencies on the spectrum of the second moment matrix (see the recent work of Audibert and Catoni (2010b) for a comprehensive survey of these results). For instance, a result of Catoni (2004, Proposition 5.9.1) 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 β (as a sequence) and in terms of the eigenspectrum of the of the second moment matrix Σ. Previous analyses of ridge regression make strong boundedness assumptions, or fail to give a bound in the case $\lambda = 0$ (*e.g.*, Zhang, 2005; Smale and Zhou, 2007; Caponnetto and Vito, 2007; Steinwart et al., 2009). For instance, Zhang assumes $\|x\| \le b_x$ and $|\langle \beta, x\rangle - y| \le b_{\text{approx}}$ almost surely, and gives the bound $\|\hat{\beta}_\lambda - \beta\|_\Sigma^2 \le \lambda \|\hat{\beta}_\lambda - \beta\|^2 + c \cdot \frac{d_{1,\lambda} \cdot (b_{\text{approx}} + b_x \|\hat{\beta}_\lambda - \beta\|)^2}{n}$ where $d_{1,\lambda}$ is a notion of effective dimension at scale λ (same as that in Condition 1). The quantity $\|\hat{\beta}_\lambda - \beta\|$ is then bounded by assuming $\|\beta\| < \infty$. Smale and Zhou assumes the more stringent conditions that $|y| \le b_y$ and $\|x\| \le b_x$ almost surely, and proves the bound $\|\hat{\beta}_\lambda - \beta_\lambda\|_\Sigma^2 \le c \cdot \frac{b_x^2 b_y^2}{\lambda^2 n}$ (note that the bound becomes trivial when $\lambda = 0$); this is then used to bound $\|\hat{\beta}_\lambda - \beta\|_\Sigma^2$ under explicit boundedness assumptions on β. Caponnetto and Vito crucially require boundedness of $|\beta\|$ and $\lambda > 0$ in their analysis (in particular, in their Theorem 4), and also have a worse tail behavior with a bound of the form $d_{1,\lambda} t^2 / n$ with probability $\ge 1 - e^{-t}$. Finally, Steinwart et al. explicitly require $|y| \le b_y$ and their bound depends on b_y in the dominant term; moreover, their bounds require explicit decay conditions on the eigenspectrum (Equation 6) and 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 β as a sequence, the eigenspectrum of Σ, and λ); this quantity vanishes when $\lambda = 0$ and can be bounded even when $\|\beta\|$ is unbounded. We note that $\|\beta_\lambda - \beta\|_\Sigma^2$ is precisely the bias term from the standard fixed design analysis of ridge regression, and therefore is natural to expect in a random design analysis.

Recently, Audibert and Catoni (2010a,b) 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 assumptions. Their bounds are proved using PAC-Bayesian techniques, which allows them to achieve exponential tail inequalities under remarkably minimal moment conditions. Their non-asymptotic bound for ordinary least squares holds with probability at least $1 - e^{-t}$ but only for $t \le \ln n$. Our result requires stronger assumptions in some respects, but it avoids this restriction on the probability tail parameter t, and the analysis is arguably more transparent and yields more reasonable quantitative bounds. The analysis of Audibert and Catoni (2010a) for the ridge estimator is established only in an asymptotic sense and bounds the excess regularized mean squared error rather than the excess mean squared error itself. Therefore, the results are not directly comparable to those provided here. It should also be mentioned that a number of other linear estimators have been considered in the literature with non-asymptotic prediction error bounds (*e.g.*, Koltchinskii, 2006; Audibert and Catoni, 2010a,b), but the focus of our work is on the ordinary least squares and ridge estimators.

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 $x_1, x_2, \ldots, x_n$ are fixed (deterministic) vectors, and the responses $y_1, y_2, \ldots, y_n$ are independent random variables. Therefore, we define $\Sigma := \widehat{\Sigma} = n^{-1} \sum_{i=1}^n x_i \otimes x_i$ (which is non-random), 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 $\operatorname{noise}(x_i) := y_i - \mathbb{E}[y_i]$ and $\operatorname{approx}(x_i) := \mathbb{E}[y_i] - \langle \beta, x_i \rangle$ for $i = 1, 2, \ldots, n$, so the following modeling equations holds:

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

for $i = 1, 2, \ldots, n$. 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

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

The following bound on the expected excess mean squared error easily follows from this decomposition and the definition of β (see, *e.g.*, Proposition 23).

Proposition 5 (Ridge regression: fixed design) *Fix $\lambda \geq 0$, and assume $\Sigma + \lambda I$ is invertible. If there exists $\sigma \geq 0$ such that $\operatorname{var}(y_i^2) \leq \sigma^2$ for all $i = 1, 2, \ldots, n$, 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 $\operatorname{var}(y_i) = \sigma^2$ for all $i = 1, 2, \ldots, n$.

Remark 6 (Effect of approximation error in fixed design) *Observe that $\operatorname{approx}(x_i)$ has no effect on the expected excess mean squared error.*

Remark 7 (Effective dimension) *The second sum in the bound is equal to $d_{2,\lambda}$ from Condition 1, which implies a notion of effective dimension at regularization level λ.*

Remark 8 (Ordinary least squares in fixed design) *In finite dimensional spaces of dimension d, Σ has only d non-zero eigenvalues λ_j, and therefore 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 $\operatorname{var}(y_i) = \sigma^2$ for all $i = 1, 2, \ldots, n$.

3.2. Ordinary least squares in finite dimensions

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 $x_1, x_2, \ldots, x_n$, *i.e.*,

$$\bar{\beta}_0 := \mathbb{E}[\hat{\beta}_0 | x_1, x_2, \ldots, x_n] = \widehat{\Sigma}^{-1} \widehat{\mathbb{E}}[x \mathbb{E}[y|x]]. \tag{11}$$

Also, define the bias and variance as:

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

Proposition 9 (Random design decomposition) *We have:*

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

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

Remark 10 *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 11 (Ordinary least squares regression) *Let d be the dimension of the covariate space. Pick any $t > \max\{0, 2.6 - \log d\}$. Assume Condition 1, Condition 2 (with parameter ρ_0), Condition 3 (with σ), and Condition 4 (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 $\widehat{\Sigma}$: $\widehat{\Sigma}$ is invertible, and*

$$\|\Sigma^{1/2}\widehat{\Sigma}^{-1}\Sigma^{1/2}\| \leq (1 - \delta_{\mathrm{s}})^{-1}$$

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

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

(note that the lower-bound on n ensures $\delta_{\mathrm{s}} \leq 0.93 < 1$).

2. *Effect of bias due to random design:*

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

and approx(x) *is defined in (9).*

3. *Effect of noise:*

$$\varepsilon_{\mathrm{vr}} \le \frac{1}{1-\delta_{\mathrm{s}}} \cdot \frac{\sigma^2(d+2\sqrt{dt}+2t)}{n}.$$

Remark 12 (Simplified form) *Suppressing the terms that are $o(1/n)$, the overall bound from Theorem 11 is*

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

*(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*

$$\|\hat{\beta}_0-\beta\|_{\Sigma}^2 \le \frac{\sigma^2(d+2\sqrt{dt}+2t)}{n} + o(1/n). \tag{12}$$

One can show that the constants in the first-order term in (12) are the same as those that one would obtain for a fixed design tail bound.

Remark 13 (Tightness of the bound) *Since*

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

and

$$\|\Sigma^{1/2}\widehat{\Sigma}^{-1}\Sigma^{1/2} - I\| \to 0$$

as $n\to\infty$ (Lemma 24), $\|\bar{\beta}_0-\beta\|_{\Sigma}^2$ is within constant factors of $\|\widehat{\mathbb{E}}[\Sigma^{-1/2}x\,\mathrm{approx}(x)]\|^2$ for sufficiently large n. Moreover,

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

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

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

with equality iff $\mathrm{var}(y)=\sigma^2$ (this comes from the fixed design risk bound in Remark 8). Therefore, in this case where $\mathrm{var}(y)=\sigma^2$, we conclude that the bound Theorem 11 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 $x_1, x_2, \ldots, x_n$, *i.e.*,

$$\bar{\beta}_\lambda := \mathbb{E}[\hat{\beta}_\lambda | x_1, x_2, \ldots, x_n] = (\widehat{\Sigma}+\lambda I)^{-1}\widehat{\mathbb{E}}[x\mathbb{E}[y|x]]. \tag{13}$$

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

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

where β_λ is the minimizer of the regularized mean squared error (see (8)).

Proposition 14 (General random design decomposition)

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

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

Remark 15 *Again, note that $\mathbb{E}[\hat{\beta}_\lambda] \ne \beta_\lambda$ 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 16 (Ridge regression) *Fix some $\lambda \ge 0$, and pick any $t > \max\{0, 2.6 - \log \tilde{d}_{1,\lambda}\}$. Assume Condition 1, Condition 2 (with parameter ρ_λ), Condition 3 (with parameter σ), and Condition 4 (with parameter b_λ) hold; and that*

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

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*

$$\|(\Sigma + \lambda I)^{1/2}(\widehat{\Sigma} + \lambda I)^{-1}(\Sigma + \lambda I)^{1/2}\| \le (1 - \delta_{\mathrm{s}})^{-1}$$

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

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

(note that the lower-bound on n ensures $\delta_{\mathrm{s}} \le 0.93 < 1$).

2. *Frobenius norm error in $\widehat{\Sigma}$:*

$$\|(\Sigma + \lambda I)^{-1/2}(\widehat{\Sigma} - \Sigma)(\Sigma + \lambda I)^{-1/2}\|_{\mathrm{F}} \le \sqrt{d_{1,\lambda}}\delta_{\mathrm{f}}$$

where

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

3. *Effect of regularization:*

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

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

4. *Effect of bias due to random design:*

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

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

5. *Effect of noise:*

$$\varepsilon_{\mathrm{vr}} \leq \frac{\sigma^2\Big(d_{2,\lambda}+\sqrt{d_{1,\lambda}d_{2,\lambda}}\delta_{\mathrm{f}}\Big)}{n(1-\delta_{\mathrm{s}})^2}+\frac{2\sigma^2\sqrt{\Big(d_{2,\lambda}+\sqrt{d_{1,\lambda}d_{2,\lambda}}\delta_{\mathrm{f}}\Big)t}}{n(1-\delta_{\mathrm{s}})^{3/2}}+\frac{2\sigma^2 t}{n(1-\delta_{\mathrm{s}})}.$$

We now discuss various aspects of Theorem 16.

Remark 17 (Simplified form) *Ignoring the terms that are* $o(1/n)$ *and treating* t *as a constant, the overall bound from Theorem 16 is*

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

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

Remark 18 (Effect of errors in $\widehat{\varSigma}$) *The accuracy of* $\widehat{\varSigma}$ *has a relatively mild effect on the bound—it appears essentially through multiplicative factors* $(1-\delta_{\mathrm{s}})^{-1} = 1+O(\delta_{\mathrm{s}})$ *and* $1+\delta_{\mathrm{f}}$*, where both* δ_{s} *and* δ_{f} *are decreasing with* n *(as* $n^{-1/2}$*), and therefore only contribute to lower-order terms overall.*

Remark 19 (Comparison to 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* $\widehat{\varSigma}$ *under fixed design, and the influence of approximation error under random design. These are revealed through the quantities* ρ_λ *and* $d_{1,\lambda}$ *(and* b_λ *in lower-order terms), which are needed to apply the probability tail inequalities. Therefore, the scaling of* $\rho_\lambda^2 d_{1,\lambda}$ *with* λ *crucially controls the effect of random design compared to fixed design.*

Acknowledgments

We thank Dean Foster, David McAllester, and Robert Stine for many insightful discussions.

References

J.-Y. Audibert and O. Catoni. Robust linear least squares regression, 2010a. arXiv:1010.0074.

J.-Y. Audibert and O. Catoni. Robust linear regression through PAC-Bayesian truncation, 2010b. arXiv:1010.0072.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

O. Catoni. *Statistical Learning Theory and Stochastic Optimization, Lectures on Probability and Statistics, Ecole d'Eté de Probabilitiés de Saint-Flour XXXI – 2001*, volume 1851 of *Lecture Notes in Mathematics*. Springer, 2004.

L. Györfi, M. Kohler, A. Kryżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer, 2004.

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

R. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, 1985.

D. Hsu, S. M. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors, 2011. arXiv:1110.2842.

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.

V. Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000.

S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximations*, 26:153–172, 2007.

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.

G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory.* Academic Press, 1990.

C. J. Stone. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10:1040–1053, 1982.

T. Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17:2077–2098, 2005.

Appendix A. Application to smoothing splines

The applications of ridge regression considered by Zhang (2005) can also be analyzed using Theorem 16. We specifically consider the problem of approximating a periodic function with smoothing splines, which are functions $f\colon \mathbb{R} \to \mathbb{R}$ 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 $\operatorname{approx}(x) = 0$ almost surely. Observe that $\|x\|^2 \leq \frac{2s}{2s-1}$, so Condition 2 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 17 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 $\|\beta_\lambda - \beta\|_\Sigma^2 \leq \lambda \|\beta\|^2/2$. Zhang (2005, Section 5.3) 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 $n^{-\frac{2s}{2s+1}}$ (Stone, 1982).

Appendix B. Proofs of Theorem 11 and Theorem 16

The proof of Theorem 16 uses the decomposition of $\|\hat{\beta}_\lambda - \beta\|_\Sigma^2$ in Proposition 14, and then bounds each term using the lemmas proved in this section.

The proof of Theorem 11 omits one term from the decomposition in Proposition 14 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 28 rather than Lemma 29), which reduces the number of lower-order terms. Other than these differences, the proof is the same as that for Theorem 16 in the special case of $\lambda = 0$.

Define

$$\begin{aligned} \Sigma_\lambda &:= \Sigma + \lambda I, && (14) \\ \widehat{\Sigma}_\lambda &:= \widehat{\Sigma} + \lambda I, \quad \text{and} && (15) \\ \Delta_\lambda &:= \Sigma_\lambda^{-1/2}(\widehat{\Sigma} - \Sigma)\Sigma_\lambda^{-1/2} && (16) \\ &= \Sigma_\lambda^{-1/2}(\widehat{\Sigma}_\lambda - \Sigma_\lambda)\Sigma_\lambda^{-1/2}. \end{aligned}$$

Recall the basic decomposition from Proposition 14:

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

Section B.1 first establishes basic properties of β and β_λ, 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 B.2 employs probability tail inequalities for the spectral and Frobenius norms of random matrices to bound the matrix errors in estimating Σ with $\widehat{\Sigma}$. Finally, Section B.3 and Section B.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 $\widehat{\Sigma}$.

B.1. Basic properties of β and β_λ, and the effect of regularization

Proposition 20 (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 β in (2). ■

Proposition 21 (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 20. ■

Proposition 22 (Shrinkage) *For any j,*

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

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

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

■

Proposition 23 (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 22,

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

Therefore,

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

■

B.2. Effect of errors in $\widehat{\Sigma}$

Lemma 24 (Spectral norm error in $\widehat{\Sigma}$) *Assume Condition 1 and Condition 2 (with parameter ρ_λ) hold. 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 Δ_λ is defined in (16).

Proof The claim is a consequence of the tail inequality from Lemma 32. First, define

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

(where Σ_λ is defined in (14)), and let

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

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

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

where the second inequality follows from Condition 2. Moreover,

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

so

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

The claim now follows from Lemma 32 (recall that $\tilde{d}_{1,\lambda} = \max\{1, d_{1,\lambda}\}$). ■

Lemma 25 (Relative spectral norm error in $\widehat{\Sigma}_\lambda$) *If $\|\Delta_\lambda\| < 1$ where Δ_λ is defined in (16), then*

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

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

Proof Observe that

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

and that

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

by the assumption $\|\Delta_\lambda\| < 1$ and Weyl's theorem (Horn and Johnson, 1985, Theorem 4.3.1). Therefore

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

■

Lemma 26 (Frobenius norm error in $\widehat{\Sigma}$) *Assume Condition 1 and Condition 2 (with parameter ρ_λ) hold. Pick any $t > 0$. With probability at least $1 - e^{-t}$,*

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

where Δ_λ is defined in (16).

Proof The claim is a consequence of the tail inequality in Lemma 31. As in the proof of Lemma 24, define $\tilde{x} := \Sigma_\lambda^{-1/2}x$ and $\widetilde{\Sigma} := \Sigma_\lambda^{-1/2}\Sigma\Sigma_\lambda^{-1/2}$, and let $Z := \tilde{x} \otimes \tilde{x} - \widetilde{\Sigma}$ so $\Delta_\lambda = \widehat{\mathbb{E}}[Z]$. Now endow the space of self-adjoint linear operators with the inner product given by $\langle A, B\rangle_{\mathrm{F}} := \mathrm{tr}(AB)$, and note that this inner product induces the Frobenius norm $\|M\|_{\mathrm{F}} = \langle M, M\rangle_{\mathrm{F}}$. Observe that $\mathbb{E}[Z] = 0$ and

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

where the inequality follows from Condition 2. Moreover,

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

where the inequality again uses Condition 2. The claim now follows from Lemma 31. ■

B.3. Effect of approximation error

Lemma 27 (Effect of approximation error) *Assume Condition 1, Condition 2 (with parameter ρ_λ), and Condition 4 (with parameter b_λ) hold. Pick any $t > 0$. If $\|\Delta_\lambda\| < 1$ where Δ_λ is defined in (16), then*

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

where $\bar\beta_\lambda$ is defined in (13), β_λ is defined in (8), $\operatorname{approx}_\lambda(x)$ is defined in (9), and Σ_λ is defined in (14). Moreover, with probability at least $1 - e^{-t}$,

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

Proof By the definitions of $\bar\beta_\lambda$ and β_λ,

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

Therefore, using the sub-multiplicative property of the spectral norm,

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

where the second inequality follows from Lemma 25 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 31. Observe that $\mathbb{E}[x \operatorname{approx}(x)] = \mathbb{E}[x(\mathbb{E}[y|x] - \langle\beta, x\rangle)] = 0$ by Proposition 20, and that $\mathbb{E}[x\langle\beta - \beta_\lambda, x\rangle] - \lambda\beta_\lambda = \Sigma\beta - (\Sigma + \lambda I)\beta_\lambda = 0$. Therefore,

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

Moreover, by Proposition 22 and Proposition 23,

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

Combining the inequality from (17) with Condition 4 and the triangle inequality, it follows that

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

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

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

The claim now follows from Lemma 31. ■

B.4. Effect of noise

Lemma 28 (Effect of noise, $\lambda = 0$) *Assume the dimension of the covariate space is $d < \infty$ and that $\lambda = 0$. Assume Condition 3 (with parameter σ) holds. Pick any $t > 0$. With probability at least $1 - e^{-t}$, either $\|\Delta_0\| \geq 1$, or*

$$\|\Delta_0\| < 1 \quad \textit{and} \quad \|\bar\beta_0 - \hat\beta_0\|_\Sigma^2 \leq \frac{1}{1-\|\Delta_0\|}\cdot\frac{\sigma^2(d + 2\sqrt{dt} + 2t)}{n},$$

where Δ_0 is defined in (16).

Proof Observe that

$$\|\bar\beta_0 - \hat\beta_0\|_\Sigma^2 \leq \|\Sigma^{1/2}\widehat\Sigma^{-1/2}\|^2\|\bar\beta_0 - \hat\beta_0\|_{\widehat\Sigma}^2 = \|\Sigma^{1/2}\widehat\Sigma^{-1}\Sigma^{1/2}\|\|\bar\beta_0 - \hat\beta_0\|_{\widehat\Sigma}^2;$$

and if $\|\Delta_0\| < 1$, then $\|\Sigma^{1/2}\widehat\Sigma^{-1}\Sigma^{1/2}\| \leq 1/(1-\|\Delta_0\|)$ by Lemma 25.

Let $\xi := (\operatorname{noise}(x_1), \operatorname{noise}(x_2), \ldots, \operatorname{noise}(x_n))$ be the random vector whose i-th component is $\operatorname{noise}(x_i) = y_i - \mathbb{E}[y_i|x_i]$. By the definition of $\hat\beta_0$ and $\bar\beta_0$

$$\|\hat\beta_0 - \bar\beta_0\|_{\widehat\Sigma}^2 = \|\widehat\Sigma^{-1/2}\widehat{\mathbb{E}}[x(y - \mathbb{E}[y|x])]\|^2 = \xi^\top\widehat K\xi$$

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

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

By Lemma 30, with probability at least $1 - e^{-t}$ (conditioned on $x_1, x_2, \ldots, x_n$),

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

The claim follows. ■

Lemma 29 (Effect of noise, $\lambda \geq 0$) *Assume Condition 1 and Condition 3 (with parameter σ) hold. Pick any $t > 0$. Let K be the $n \times n$ symmetric matrix whose (i, j)-th entry is*

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

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

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

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

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

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

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

By Lemma 30, with probability at least $1 - e^{-t}$ (conditioned on $x_1, x_2, \ldots, x_n$),

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

where the second inequality follows from von Neumann's theorem (Horn and Johnson, 1985, page 423).

Note that the non-zero eigenvalues of K are the same as that of

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

To bound $\lambda_{\max}(K)$, observe that by the sub-multiplicative property of the spectral norm and Lemma 25,

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

To bound $\operatorname{tr}(K)$, first define the λ-whitened versions of Σ, $\widehat{\Sigma}$, and $\widehat{\Sigma}_\lambda$:

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

Using these definitions with the cycle property of the trace,

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

Let $\{\lambda_j[M]\}$ denote the eigenvalues of a linear operator M. By von Neumann's theorem (Horn and Johnson, 1985, page 423),

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

and by Ostrowski's theorem (Horn and Johnson, 1985, Theorem 4.5.9),

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

Therefore

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

where the second inequality follows from Lemma 25, the third inequality follows from Cauchy-Schwarz, and the fourth inequality follows from Mirsky's theorem (Stewart and Sun, 1990, Corollary 4.13). ■

Appendix C. 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 setup 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 χ^2 random variables (Laurent and Massart, 2000). We give the proof for completeness.

Lemma 30 (Quadratic forms of a subgaussian random vector; Hsu et al., 2011) *Let ξ 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 $t > 0$,

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

Proof Let $z \in \mathbb{R}^n$ be a vector of n i.i.d. standard normal random variables (independent of ξ). For any $\tau \geq 0$ and $\lambda \geq 0$, let $\eta := c\lambda^2/2$, so

$$\begin{aligned}
&\mathbb{E}\big[\exp(\lambda\langle z, K^{1/2}\xi\rangle)\big] \\
&\geq \mathbb{E}\big[\exp(\lambda\langle z, K^{1/2}\xi\rangle) | \|K^{1/2}\xi\|^2 > c(\operatorname{tr}(K)+\tau)\big] \cdot \Pr\big[\|K^{1/2}\xi\|^2 > c(\operatorname{tr}(K)+\tau)\big] \\
&\geq \exp(\lambda^2 c(\operatorname{tr}(K)+\tau)/2) \cdot \Pr\big[\|K^{1/2}\xi\|^2 > c(\operatorname{tr}(K)+\tau)\big] \\
&= \exp(\eta(\operatorname{tr}(K)+\tau)) \cdot \Pr\big[\|K^{1/2}\xi\|^2 > c(\operatorname{tr}(K)+\tau)\big] \qquad (18)
\end{aligned}$$

since $\mathbb{E}[\exp(\langle u, z\rangle)] = \exp(\|u\|^2/2)$ for any $u \in \mathbb{R}^n$. Moreover, by independence of ξ and z,

$$\begin{aligned}
\mathbb{E}\big[\exp(\lambda\langle z, K^{1/2}\xi\rangle)\big] &= \mathbb{E}\Big[\mathbb{E}\big[\exp(\lambda\langle K^{1/2}z, \xi\rangle) | z\big]\Big] \\
&\leq \mathbb{E}\big[\exp(c\lambda^2\|K^{1/2}z\|^2/2)\big] \\
&= \mathbb{E}\big[\exp(\eta\|K^{1/2}z\|^2)\big].
\end{aligned}$$

Since K is symmetric and positive semidefinite, $K = VDV^\top$ for some orthogonal matrix $V = [u_1|u_2|\cdots|u_r]$ and diagonal matrix $D = \operatorname{diag}(\rho_1, \rho_2, \ldots, \rho_r)$, where r is the rank of K. By rotational symmetry, the vector $V^\top z$ is equal in distribution to a vector of r i.i.d. standard normal random variables $q_1, q_2, \ldots, q_r$, and $\|K^{1/2}z\|^2 = \|D^{1/2}V^\top z\|^2 = \rho_1 q_1^2 + \rho_2 q_2^2 + \cdots + \rho_r q_r^2$. Therefore,

$$\mathbb{E}\big[\exp(\lambda\langle z, K^{1/2}\xi\rangle)\big] \leq \mathbb{E}\big[\exp(\eta\|K^{1/2}z\|^2)\big] = \mathbb{E}\big[\exp(\eta(\rho_1 q_1^2 + \rho_2 q_2^2 + \cdots + \rho_r q_r^2))\big]. \qquad (19)$$

Combining (18) and (19) gives

$$\Pr\big[\|K^{1/2}\xi\|^2 > c(\operatorname{tr}(K)+\tau)\big] \leq \exp(-\eta(\operatorname{tr}(K)+\tau)) \cdot \mathbb{E}\big[\exp(\eta(\rho_1 q_1^2 + \rho_2 q_2^2 + \cdots + \rho_r q_r^2))\big].$$

The expectation on the right-hand side is the moment generating function for a linear combination of r independent χ^2 random variables, each with one degree of freedom. Since $\operatorname{tr}(K) = \rho_1 + \rho_2 + \cdots + \rho_r$, $\operatorname{tr}(K^2) = \rho_1^2 + \rho_2^2 + \cdots + \rho_r^2$, and $\|K\| = \max\{\rho_1, \rho_2, \ldots, \rho_r\}$, the conclusion follows from standard facts about χ^2 random variables (Laurent and Massart, 2000):

$$\Pr\big[\|K^{1/2}\xi\|^2 > c(\operatorname{tr}(K) + \tau)\big] \leq \exp\left(-\frac{\operatorname{tr}(K^2)}{2\|K\|} \cdot h_1\left(\frac{\|K\|\tau}{\operatorname{tr}(K^2)}\right)\right)$$

where $h_1(a) := 1 + a - \sqrt{1 + 2a}$. ■

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

Lemma 31 (Vector Bernstein bound; see, *e.g.*, Hsu et al., 2011) *Let $x_1, x_2, \ldots, x_n$ 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 $i = 1, 2, \ldots, n$, almost surely. Let $s := x_1 + x_2 + \cdots + x_n$. 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 32 (Matrix Bernstein bound; Hsu et al., 2012) *Let X be a random matrix, and $r > 0$, $v > 0$, and $k > 0$ be such that, almost surely,*

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

If $X_1, X_2, \ldots, X_n$ 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 $t(e^t - t - 1)^{-1} \leq e^{-t/2}$.