RIDGE REGRESSION: STRUCTURE, CROSS-VALIDATION, AND SKETCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the following three fundamental problems about ridge regression: (1) what is the structure of the estimator? (2) how to correctly use cross-validation to choose the regularization parameter? and (3) how to accelerate computation without losing too much accuracy? We consider the three problems in a unified large-data linear model. We give a precise representation of ridge regression as a covariance matrix-dependent linear combination of the true parameter and the noise. We study the bias of K-fold cross-validation for choosing the regularization parameter, and propose a simple bias-correction. We analyze the accuracy of primal and dual sketching for ridge regression, showing they are surprisingly accurate. Our results are illustrated by simulations and by analyzing empirical data.

1 INTRODUCTION

Ridge or ℓ_2-regularized regression is a widely used method for prediction and estimation when the data dimension p is large compared to the number of datapoints n. This is especially so in problems with many good features, where sparsity assumptions may not be justified. A great deal is known about ridge regression. It is Bayes optimal for any quadratic loss in a Bayesian linear model where the parameters and noise are Gaussian. The asymptotic properties of ridge have been widely studied (e.g., Tulino & Verdú, 2004; Serdobolskii, 2007; Couillet & Debbah, 2011; Dicker, 2016; Dobriban & Wager, 2018, etc). For choosing the regularization parameter in practice, cross-validation (CV) is widely used. In addition, there is an exact shortcut (e.g., Hastie et al., 2009), which has good consistency properties (Hastie et al., 2019). There is also a lot of work on fast approximate algorithms for ridge, e.g., using sketching methods (e.g., el Alaoui & Mahoney, 2015; Chen et al., 2015; Wang et al., 2018; Chowdhury et al., 2018, among others).

Here we seek to develop a deeper understanding of ridge regression, going beyond existing work in multiple aspects. We work in linear models under a popular asymptotic regime where $n, p \to \infty$ at the same rate (Marchenko & Pastur, 1967; Serdobolskii, 2007; Couillet & Debbah, 2011; Yao et al., 2015). In this framework, we develop a fundamental representation for ridge regression, which shows that it is well approximated by a linear scaling of the true parameters perturbed by noise. The scaling matrices are functions of the population-level covariance of the features. As a consequence, we derive formulas for the training error and bias-variance tradeoff of ridge.

Second, we study commonly used methods for choosing the regularization parameter. Inspired by the observation that CV has a bias for estimating the error rate (e.g., Hastie et al., 2009, p. 243), we study the bias of CV for selecting the regularization parameter. We discover a surprisingly simple form for the bias, and propose a downward scaling bias correction procedure. Third, we study the accuracy loss of a class of randomized sketching algorithms for ridge regression. These algorithms approximate the sample covariance matrix by sketching or random projection. We show they can be surprisingly accurate, e.g., they can sometimes cut computational cost in half, only incurring 5% extra error. Even more, they can sometimes improve the MSE if a suboptimal regularization parameter is originally used.

Our work leverages recent results from asymptotic random matrix theory and free probability theory. One challenge in our analysis is to find the limit of the trace $\operatorname{tr}(\Sigma_1+\Sigma_2^{-1})^{-1}/p$, where Σ_1 and Σ_2 are $p\times p$ independent sample covariance matrices of Gaussian random vectors. The calculation requires nontrivial aspects of freely additive convolutions (e.g., Voiculescu et al., 1992; Nica & Speicher, 2006).

Our work is connected to prior works on ridge regression in high-dimensional statistics (Serdobolskii, 2007) and wireless communications (Tulino & Verdú, 2004; Couillet & Debbah, 2011). Among other related works, El Karoui & Kösters (2011) discuss the implications of the geometric sensitivity of random matrix theory for ridge regression, without considering our problems. El Karoui (2018) and Dicker (2016) study ridge regression estimators, but focus only on the risk for identity covariance. Hastie et al. (2019) study "ridgeless" regression, where the regularization parameter tends to zero.

Sketching is an increasingly popular research topic, see Vempala (2005); Halko et al. (2011); Mahoney (2011); Woodruff (2014); Drineas & Mahoney (2017) and references therein. For sketched ridge regression, Zhang et al. (2013a;b) study the dual problem in a complementary finite-sample setting, and their results are hard to compare. Chen et al. (2015) propose an algorithm combining sparse embedding and the subsampled randomized Hadamard transform (SRHT), proving relative approximation bounds. Wang et al. (2017) study iterative sketching algorithms from an optimization point of view, for both the primal and the dual problems. Dobriban & Liu (2018) study sketching using asymptotic random matrix theory, but only for unregularized linear regression. Chowdhury et al. (2018) propose a data-dependent algorithm in light of the ridge leverage scores. Other related works include Sarlos (2006); Ailon & Chazelle (2006); Drineas et al. (2006; 2011); Dhillon et al. (2013); Ma et al. (2015); Raskutti & Mahoney (2016); Gonen et al. (2016); Thanei et al. (2017); Ahfock et al. (2017); Lopes et al. (2018); Huang (2018).

The structure of the paper is as follows: We state our results on representation, risk, and bias-variance tradeoff in Section 2. We study the bias of cross-validation for choosing the regularization parameter in Section 3. We study the accuracy of randomized primal and dual sketching for both orthogonal and Gaussian sketches in Section 4. We provide proofs and additional simulations in the Appendix.

2 Ridge regression

We work in the usual linear regression model $Y = X\beta + \varepsilon$, where each row x_i of $X \in \mathbb{R}^{n\times p}$ is a datapoint in p dimensions, and so there are p features. The corresponding element y_i of $Y \in \mathbb{R}^n$ is its continous response (or outcome). We assume mean zero uncorrelated noise, so $\mathbb{E}\varepsilon = 0$, and $\operatorname{Cov}[\varepsilon] = \sigma^2 I_n$. We estimate the coefficient $\beta \in \mathbb{R}^p$ by ridge regression, solving the optimization problem

$$\hat\beta = \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2,$$

where $\lambda > 0$ is a regularization parameter. The solution has the closed form

$$\hat\beta = \left(X^\top X/n + \lambda I_p\right)^{-1} X^\top Y/n. \tag{1}$$

We work in a "big data" asymptotic limit, where both the dimension p and the sample size n tend to infinity, and their aspect ratio converges to a constant, $p/n \to \gamma \in (0,\infty)$. Our results can be interpreted for any n and p, using $\gamma = p/n$ as an approximation. We recall that the empirical spectral distribution (ESD) of a $p\times p$ symmetric matrix Σ is the distribution $\frac{1}{p}\sum_{i=1}^p \delta_{\lambda_i}$ where λ_i, $i = 1,\dots,p$ are the eigenvalues of Σ, and δ_x is the point mass at x. We say that the ESD of the $n\times p$ matrix X is the ESD of $X^\top X/n$. We will consider models for the data of the form $X = U\Sigma^{1/2}$, where U has iid entries of zero mean and unit variance. This means that the datapoints have the form $x_i = \Sigma^{1/2}u_i$, where u_i have iid entries. Then Σ is the true covariance matrix of the features, which is typically not observed.

We start by finding a precise representation of the ridge estimator. For vectors u_n, v_n of growing dimension, $u_n \asymp v_n$ means that for any sequence of fixed (or random and independent of u_n, v_n) vectors w_n such that

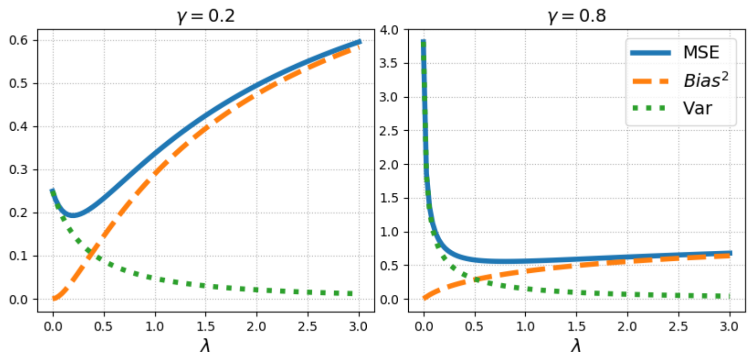

Figure 1: Ridge regression bias-variance tradeoff. Left: $\gamma = p/n = 0.2$; right: $\gamma = 0.8$. The data matrix X has iid Gaussian entries. The coefficient β has distribution $\beta \sim \mathcal{N}(0, I_p/p)$, while the noise $\varepsilon \sim \mathcal{N}(0, I_p)$.

$\|w_n\|_2 < \infty$ almost surely, we have $|w_n^\top(u_n - v_n)| \to 0$ almost surely. Thus linear combinations or u_n are well approximated by those of v_n. We extend scalar functions $f : \mathbb{R} \to \mathbb{R}$ to matrices by functional calculus, applying them to the eigenvalues and keeping the eigenvectors. We find the following result.

Theorem 2.1 (Representation of ridge estimator)**.** *Suppose the data matrix has the form $X = U\Sigma^{1/2}$, where $U \in \mathbb{R}^{n\times p}$ has iid entries of zero mean, unit variance and finite $8 + c$-th moment for some $c > 0$, and $\Sigma = \Sigma_{n,p} \in \mathbb{R}^{p\times p}$ is a deterministic positive definite matrix. Suppose that $n, p \to \infty$ with $p/n \to \gamma > 0$. Suppose the ESD of the sequence of Σs converges in distribution to a probability measure with compact support bounded away from the origin. Suppose that the noise is Gaussian, and that $\beta = \beta_{n,p}$ is an arbitrary sequence of deterministic vectors, such that* $\limsup \|\beta\|_2 < \infty$.

Then the ridge regression estimator is asymptotically equivalent to a random vector with the following representation:

$$\hat{\beta}(\lambda) \asymp A(\Sigma, \lambda) \cdot \beta + B(\Sigma, \lambda) \cdot \sigma \cdot \frac{Z}{p^{1/2}}.$$

Here $Z \sim \mathcal{N}(0, I_p)$ is a random vector whose distribution depends only on the noise and A, B are deterministic matrix functions of all problem parameters. They are defined for scalars as

$$A(x, \lambda) = (c_p x + \lambda)^{-2}(c_p + c_p')x, \qquad B(x, \lambda) = (c_p x + \lambda)^{-1} c_p x.$$

Here $c_p := c(n, p, \Sigma, \lambda)$ is the unique positive solution of the fixed point equation

$$1 - c_p = \frac{c_p}{n} \operatorname{tr}\left[\Sigma(c_p\Sigma + \lambda I)^{-1}\right].$$

It is known that c_p is well defined, and this follows by a simple monotonicity argument, see Hachem et al. (2007); Rubio & Mestre (2011). Also c_p' is the derivative of c_p with respect to $z := -\lambda$ and an explicit expression is provided in the proof in Section A.1. This result gives a precise representation of the ridge regression estimator. It is a sum of two terms: the true coefficient vector β scaled by the matrix $A(\Sigma, \lambda)$, and the noise vector Z scaled by the matrix $B(\Sigma, \lambda)$. However, the coefficients are not fully explicit, as they depend on the unknown population covariance matrix Σ, as well as on the fixed-point variable c_p.

Here we discuss some implications of this representation. For uncorrelated features, $\Sigma = I_p$, A, B reduce to multiplication by scalars. Hence, each coordinate of the ridge regression estimator is simply a scalar multiple of the corresponding coordinate of β. One can use this to find the bias in each individual coordinate.

One can use also use the representation to derive the training error of ridge (see Sec. A.2), in addition to its known estimation error (Serdobolskii, 2007; Tulino & Verdú, 2004). In that setting, we work in a *random-effects model*, where the p-dimensional regression parameter β is random, each coefficient has zero mean $\mathbb{E}\beta_i = 0$, and is normalized so that Var$\beta_i = \alpha^2/p$. This ensures that the signal strength $\mathbb{E}\|\beta\|^2 = \alpha^2$ is fixed for any p. The asymptotically optimal λ in this setting is always $\lambda^* = \gamma\sigma^2/\alpha^2$ see e.g., Tulino & Verdú (2004); Dicker (2016); Dobriban & Wager (2018). The ridge regression estimator with $\lambda = p\sigma^2/(n\alpha^2)$ is the posterior mean of β, when β and ε are normal random variables.

As a consequence, we can also find the bias and the variance of ridge. See Figure 1 for a plot and Sec. A.4 for the details. As far as we know, this is one of the few examples of high-dimensional asymptotic problems where the precise form of the bias and variance can be evaluated.

As a further consequence, we can find how the bias and variance change with the aspect ratio γ at the optimal $\lambda^* = \gamma\sigma^2/\alpha^2$ (see Figure 6). This can be viewed as the "pure" effect of dimensionality on the problem, keeping all other parameters fixed, and has intriguing properties. The variance first increases, then decreases with γ. In the "classical" low-dimensional case, most of the risk is due to variance, while in the "modern" high-dimensional case, most of it is due to bias. This is consistent with other phenomena in proportional-limit asymptotics, e.g., that the map between population and sample eigenvalue distributions is asymptotically deterministic (Marchenko & Pastur, 1967).

This fundamental representation may have applications to important statistical inference questions. For instance, inference on the regression coefficient β and the noise variance σ^2 are important and challenging problems. Can we use our representation to develop debiasing techniques for this task? This will be interesting to explore in future work.

3 CROSS-VALIDATION

How can we choose the regularization parameter? In practice, cross-validation (CV) is the most popular approach. However, it is well known that CV has a bias for estimating the error rate, because it uses a smaller number of samples than the full data size (e.g., Hastie et al., 2009, p. 243). Here we study related questions, proposing a bias-correction method for the optimal regularization parameter.

Suppose we split the n datapoints (samples) into K equal-sized subsets, each containing $n_0 = n/K$ samples. We use the k-th subset (X_k, Y_k) as the validation set and the other $K-1$ subsets (X_{-k}, Y_{-k}), with total sample size $n_1 = (K-1)n/K$ as the training set. We find the ridge regression estimator $\hat\beta_{-k}$, i.e.

$$\hat\beta_{-k}(\lambda) = \left(X_{-k}^\top X_{-k} + n_1\lambda I_p\right)^{-1} X_{-k}^\top Y_{-k}.$$

The expected cross-validation error is, for isotropic covariance, i.e., $\Sigma = I$,

$$CV(\lambda) = \mathbb{E}\widehat{CV}(\lambda) = \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|Y_k - X_k\hat\beta_{-k}(\lambda)\|_2^2/n_0\right] = \sigma^2 + \mathbb{E}\left[\|\hat\beta_{-k} - \beta\|_2^2\right].$$

When n, p tend to infinity so that $p/n \to \gamma > 0$, and in the random effects model with $\mathbb{E}\beta_i = 0$, Var$\beta_i = \alpha^2/p$ described above, the minimizer of $CV(\lambda)$ tends to $\lambda_k^* = \tilde\gamma\sigma^2/\alpha^2$, where $\tilde\gamma$ is the limiting aspect ratio of X_{-k}, i.e. $\tilde\gamma = \gamma K/(K-1)$. Since the aspect ratios of X_{-k} and X differ, the limiting minimizer of the cross-validation estimator of the test error is biased for the limiting minimizer of the actual test error, which is $\lambda^* = \gamma\sigma^2/\alpha^2$.

Suppose we have found $\hat\lambda_k^*$, the minimizer of $\widehat{CV}(\lambda)$. Afterwards, we usually refit ridge regression on the entire dataset, i.e., find

$$\hat\beta(\hat\lambda^*) = (X^\top X + \hat\lambda^* n I)^{-1} X^\top Y.$$

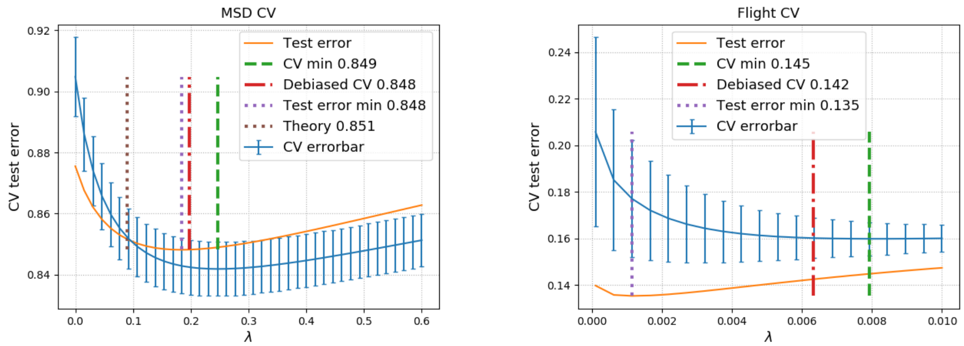

Figure 2: Left: Cross-validation on the Million Song Dataset (MSD, Bertin-Mahieux et al., 2011). For the error bar, we take $n = 1000$, $p = 90$, $K = 5$, and average over 90 different sub-datasets. For the test error, we train on 1000 training datapoints and fit on 9000 test datapoints. The debiased λ reduces the test error by 0.00024, and the minimal test error is 0.8480. Right: Cross-validation on the flights dataset Wickham (2018). For the error bar, we take $n = 300, p = 21, K = 5$, and average over 180 different sub-datasets. For the test error, we train on 300 datapoints and fit on 27000 test datapoints. The debiased λ reduces the test error by 0.0022, and the minimal test error is 0.1353.

Based on our bias calculation, we propose to use a *bias-corrected* parameter

$$\hat{\lambda}^* := \hat{\lambda}_k^* \frac{K-1}{K}.$$

So if we use 5 folds, we should multiply the CV-optimal λ by 0.8. We find it surprising that this theoretically justified bias-correction does not depend on any unknown parameters, such as $\beta, \alpha^2, \sigma^2$.While the bias of CV is widely known, we are not aware that this bias-correction for the regularization parameter has been proposed before.

Figure 2 shows on two empirical data examples that the debiased estimator gets closer to the optimal λ than the original minimizer of the CV. However, in this case it does not significantly improve the test error. See Section A.5 for similar phenomena in simulations.

The same bias-correction idea also applies to train-test validation. In addition, there is a special fast "short-cut" for leave-one-out cross-validation in ridge regression (e.g., Hastie et al., 2009), which has the same cost as one ridge regression. The minimizer converges to λ^* (Hastie et al., 2019). However, we think that the bias-correction idea is still valuable, as the idea applies beyond ridge regression: CV selects regularization parameters that are too large. See Section A.6 for more details and experiments comparing different ways of choosing the regularization parameter.

4 SKETCHING

The time complexity of computing ridge regression using the standard QR decomposition is $O(np \min(n, p))$. Sketching is a set of methods for reducing the time complexity by effectively reducing n or p (e.g., Mahoney, 2011; Woodruff, 2014; Drineas & Mahoney, 2016). Specifically, *primal sketching* approximates the sample covariance matrix $X^\top X/n$ by $X^\top L^\top LX/n$, where L is an $m \times n$ sketching matrix, and $m < n$. If L is chosen as a suitable random matrix, then this can still approximate the original sample

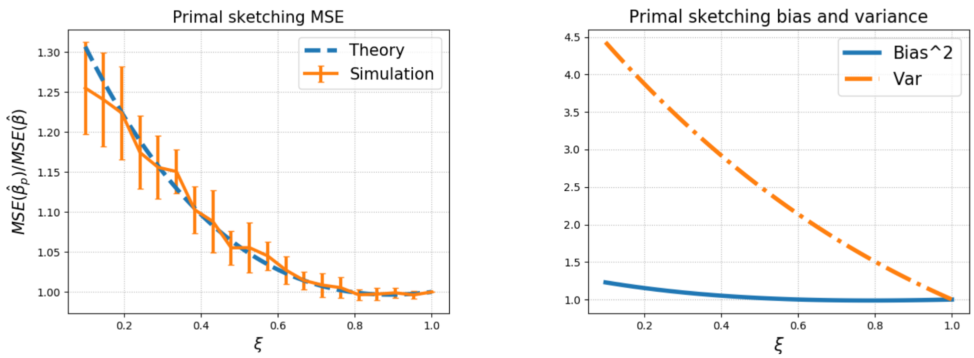

Figure 3: Primal sketching with $n = 500$, $\gamma = 1.5$, $\lambda = 1.5$, $\alpha = 3$, $\sigma = 1$. Left: MSE of primal sketching normalized by the MSE of ridge regression. The error bar is the standard deviation over 50 repetitions. Right: Bias and variance of primal sketching normalized by the bias and variance of ridge regression, respectively.

covariance matrix. Then the primal sketched ridge regression estimator is

$$\hat{\beta}_p = \left(X^\top L^\top LX/n + \lambda I_p\right)^{-1} X^\top Y/n. \tag{2}$$

Dual sketching reduces p instead. An equivalent expression for ridge regression is $\hat{\beta} = n^{-1}X^\top \left(XX^\top/n + \lambda I_n\right)^{-1} Y$. Dual sketched ridge regression reduces the computation cost of the Gram matrix $XX^\top$, approximating it by $XRR^\top X^\top$ for another sketching matrix $R \in \mathbb{R}^{p\times d}$ ($d < p$), so

$$\hat{\beta}_d = X^\top \left(XRR^\top X^\top/n + \lambda I_n\right)^{-1} Y/n. \tag{3}$$

The sketching matrices R and L are usually chosen as random matrices with iid entries (e.g., Gaussian ones) or as orthogonal matrices. In this section, we study the asymptotic MSE for both orthogonal and Gaussian sketching. We also mention *full sketching*, which performs ridge after projecting down both X and Y. In section A.12, we find its MSE. However, the other two methods have better tradeoffs, and we can empirically get better results for the same computational cost.

4.1 Orthogonal sketching

First we consider primal sketching with orthogonal projections. These can be implemented by subsampling, Haar distributed matrices, or subsampled randomized Hadamard transforms. We recall that the standard *Marchenko-Pastur (MP) law* is the probability distribution which is the limit of the ESD of $X^\top X/n$, when the $n \times p$ matrix X has iid standard Gaussian entries, and $n, p \to \infty$ so that $p/n \to \gamma > 0$, which has an explicit density (Marchenko & Pastur, 1967; Bai & Silverstein, 2010).

Theorem 4.1 (Primal orthogonal sketching)**.** *Suppose β has iid entries with $\mathbb{E}\beta_i = 0$, $\mathrm{Var}[\beta_i] = \alpha^2/p$, $i = 1, \dots, p$ and β is independent of X and ε. Suppose X has iid standard normal entries.*

We compute primal sketched ridge regression (2) with an $m \times n$ orthogonal matrix L ($m < n$, $LL^\top = I_m$). Let n, p and m tend to infinity with $p/n \to \gamma \in (0, \infty)$ and $m/n \to \xi \in (0, 1)$. Then the MSE of $\hat{\beta}_p(\lambda)$ has the limit

$$M(\lambda) = \alpha^2 \frac{\left[(\lambda + \xi - 1)^2 + \gamma(1 - \xi)\right] \theta_2\left(\frac{\gamma}{\xi}, \frac{\lambda}{\xi}\right)}{\xi^2} + \gamma\sigma^2 \frac{\xi\theta_1\left(\frac{\gamma}{\xi}, \frac{\lambda}{\xi}\right) - (\lambda + \xi - 1)\theta_2\left(\frac{\gamma}{\xi}, \frac{\lambda}{\xi}\right)}{\xi^2},$$

where $\theta_i(\gamma, \lambda) = \int (x + \lambda)^{-i} dF_\gamma(x)$ and F_γ is the standard Marchenko-Pastur law with aspect ratio γ.

The proof is in Section A.7, with explicit formulas in Section A.7.1. The θ_i are related to the resolvent of the MP law and its derivatives. A simulation in Figure 3 shows a good match with our theory. It also shows that sketching does not increase the MSE too much. By reducing the sample size to half the original one, we only increase the MSE by a factor of 1.05. This shows sketching can be very effective. We also see that variance is compromised much more than bias.

The reader may wonder how strongly this depends on the choice of the regularization parameter λ. Perhaps ridge regression works poorly with this λ, so sketching cannot worsen it too much? What happens if we take the optimal λ instead of a fixed one? In experiments in Section A.13 we show that the behavior is quite robust to the choice of regularization parameter.

The next theorem states a result for dual sketching.

Theorem 4.2 (Dual orthogonal sketching)**.** *Under the conditions of Theorem 4.1, we compute the dual sketched ridge regression with an orthogonal $p \times d$ sketching matrix R ($d \leqslant p$, $R^\top R = I_d$). Let n, p and d go to infinity with $p/n \to \gamma \in (0, \infty)$ and $d/n \to \zeta \in (0, \gamma)$. Then the MSE of $\hat{\beta}_d(\lambda)$ has the limit*

$$\frac{\alpha^2}{\gamma}\left[\gamma - 1 + (\lambda - \gamma + \zeta)^2 \bar{\theta}_2(\zeta, \lambda) + (\gamma - \zeta)\bar{\theta}_1^2(\zeta, \lambda)\right] + \sigma^2\left[\bar{\theta}_1(\zeta, \lambda) - (\lambda + \zeta - \gamma)\bar{\theta}_2(\zeta, \lambda)\right],$$

where $\bar{\theta}_i(\zeta, \lambda) = (1 - \zeta)/\lambda^i + \zeta \int (x + \lambda)^{-i} dF_\zeta(x)$, and F_ζ is the standard Marchenko-Pastur law.

The proof is in Section A.8. Simulation results are shown in Figure 11 from Section A.13. They are similar to the ones before: sketching has favorable properties, and the bias increases less than the variance.

For both primal and dual sketching, the optimal regularization parameter minimizing the MSE seems analytically intractable. Instead, we use a numerical approach in our experiments, based on a binary search. Since this is one-dimensional problem, there are no numerical issues.

4.1.1 Extreme projection — marginal regression

It is of special interest to investigate *extreme projections*, where the sketching dimension is much reduced compared to the sample size, so $m \ll n$. This corresponds to $\xi = 0$, and the formulas simplify. This can also be viewed as a scaled *marginal regression* estimator, i.e., $\hat{\beta} \propto X^\top Y$. For dual sketching, the same case can be recovered with $\zeta = 0$.

Theorem 4.3 (Marginal regression)**.** *Under the same assumption as Theorem 4.1, let $\xi = 0$. Then the form of the MSE is $M(\lambda) = [\alpha^2\left[(\lambda - 1)^2 + \gamma\right] + \sigma^2\gamma]/\lambda^2$. Moreover, the optimal λ^* that minimizes this equals $\gamma\sigma^2/\alpha^2 + 1 + \gamma$ and the optimal MSE is $M(\lambda^*) = \alpha^2\left(1 - \alpha^2/[\alpha^2(1 + \gamma) + \gamma\sigma^2]\right)$.*

The proof is in Section A.9. When is the optimal MSE of marginal regression small? Compared to the MSE of the zero estimator α^2, it is small when $\gamma(\sigma^2/\alpha^2 + 1) + 1$ is large. In Figure 4, we compare marginal and ridge regression for different aspect ratios and SNR. When the signal to noise ratio (SNR) α^2/σ^2 is small or the aspect ratio γ is large, marginal regression does not increase the MSE much. As a concrete example, if we take $\alpha^2 = \sigma^2 = 1$ and $\gamma = 0.7$, the marginal MSE is $1 - 1/2.4 \approx 0.58$. The optimal ridge MSE is about 0.52, so their ratio is only ca. $0.58/0.52 \approx 1.1$. It seems quite surprising that a simple-minded method like marginal regression can work so well. However, the reason is that when the SNR is small, we cannot expect ridge regression to have good performance. Large γ can also be interpreted as small SNR, where ridge regression works poorly and sketching does not harm performance too much.

4.2 Gaussian sketching

In this section, we study Gaussian sketching. The following theorem states the bias of dual Gaussian sketching. The bias is enough to characterize the performance in the high SNR regime where $\alpha/\sigma \to \infty$, and we discuss the extension to low SNR after the proof.

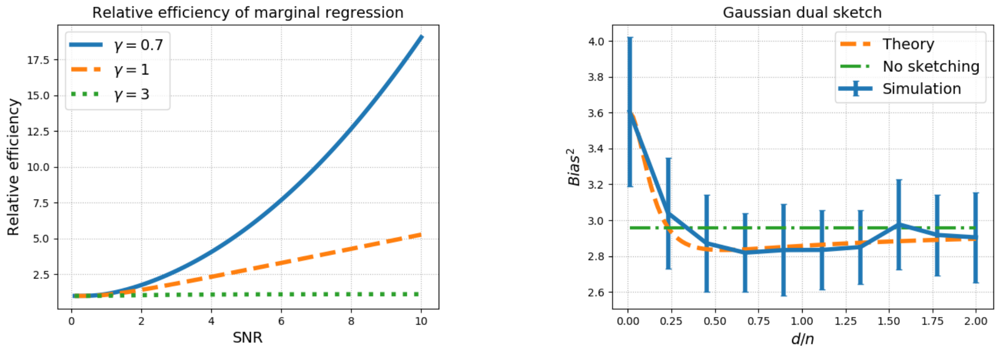

Figure 4: Left: Ratio of optimal MSE of marginal regression to that of optimally tuned ridge regression, for three values of $\gamma = p/n$, as a function of the SNR α^2/σ^2. Right: Gaussian dual sketch when there is no noise. $\gamma = 0.4$, $\alpha = 1$, $\lambda = 1$ (both for original and sketching). Standard error over 50 experiments.

Theorem 4.4 (Bias of dual Gaussian sketch). *Suppose X is an $n \times p$ standard Gaussian random matrix. Suppose also that R is a $p \times d$ matrix with i.i.d. $\mathcal{N}(0, 1/d)$ entries. Then the bias of dual sketch has the expression $MSE(\hat{\beta}_d) = \alpha^2 + \alpha^2/\gamma \cdot [m'(z) - 2m(z)]\,|_{z=0}$, where m is a function described below, and $m'(z)$ denotes the derivative of m w.r.t. z.*

The function m is characterized by its inverse function, which has the explicit formula $m^{-1}(z) = 1/[1 + z/\zeta] - [\gamma + 1 - \sqrt{(\gamma-1)^2 + 4\lambda z}]/(2z)$ for complex z with positive imaginary part.

The proof is in Section A.10. We use the branch of the square root with positive imaginary part. We mention that the same result holds when the matrices involved have iid non-Gaussian entries, but the proof is more technical. The current proof is already based on free probability theory (e.g., Voiculescu et al., 1992; Hiai & Petz, 2006; Couillet & Debbah, 2011). The function m is the Stieltjes transform of the free additive convolution of a standard MP law $F_{1/\xi}$ and a scaled inverse MP law $\lambda/\gamma \cdot F^{-1}_{1/\gamma}$ (see the proof).

To evaluate the formula, we note that $m^{-1}(m(0)) = 0$, so $m(0)$ is a root of m^{-1}. Also, $dm(0)/dz$ equals $1/(dm^{-1}(y)/dy|_{y=m(0)})$, the reciprocal of the derivative of m^{-1} evaluated at $m(0)$. We use a simple numerical binary search on the function m to find the solution. The theoretical result agrees with the simulation quite well, see Figure 4. This shows that increasing d will reduce the MSE.

Somewhat unexpectedly, the MSE of dual sketching can be below the MSE of ridge regression, see Figure 4. This can happen when the original regularization parameter is suboptimal. As d grows, the MSE of Gaussian dual sketching converges to that of ridge regression (see Figure 9).

We have also found the bias of primal Gaussian sketching. However, stating the result requires free probability theory, and so we present it in the Appendix, see Theorem A.2. To further validate our results, we present additional simulations in Sec. A.13, for both fixed and optimal regularization parameters after sketching. A detailed study of the computational cost for sketching in Sec. A.14 concludes, as expected, that primal sketching can reduce cost when $p < n$, while dual sketching can reduce it when $p > n$; and also provides a more detailed analysis.

REFERENCES

Daniel Ahfock, William J Astle, and Sylvia Richardson. Statistical properties of sketching algorithms. *arXiv preprint arXiv:1706.03665*, 2017.

Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pp. 557–563. ACM, 2006.

Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. *An Introduction to Random Matrices*. Number 118. Cambridge University Press, 2010.

Theodore W Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley New York, 2003.

Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, New York, 2nd edition, 2010.

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

Shouyuan Chen, Yang Liu, Michael R Lyu, Irwin King, and Shengyu Zhang. Fast relative-error approximation algorithm for ridge regression. In *UAI*, pp. 201–210, 2015.

Agniva Chowdhury, Jiasen Yang, and Petros Drineas. An iterative, sketching-based framework for ridge regression. In *International Conference on Machine Learning*, pp. 988–997, 2018.

Romain Couillet and Merouane Debbah. *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.

Paramveer Dhillon, Yichao Lu, Dean P Foster, and Lyle Ungar. New subsampling algorithms for fast least squares regression. In *Advances in neural information processing systems*, pp. 360–368, 2013.

Lee H Dicker. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22 (1):1–37, 2016.

Edgar Dobriban and Sifan Liu. A new theory for sketching in linear regression. *arXiv preprint arXiv:1810.06089, NeurIPS 2019*, 2018.

Edgar Dobriban and Yue Sheng. Distributed linear regression by averaging. *arXiv preprint arxiv:1810.00412*, 2018.

Edgar Dobriban and Yue Sheng. One-shot distributed ridge regression in high dimensions. *arXiv preprint arXiv:1903.09321*, 2019.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

Petros Drineas and Michael W Mahoney. RandNLA: randomized numerical linear algebra. *Communications of the ACM*, 59(6):80–90, 2016.

Petros Drineas and Michael W Mahoney. Lectures on randomized numerical linear algebra. *arXiv preprint arXiv:1712.08880*, 2017.

Petros Drineas, Michael W Mahoney, and S Muthukrishnan. Sampling algorithms for l 2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pp. 1127–1136. Society for Industrial and Applied Mathematics, 2006.

Petros Drineas, Michael W Mahoney, S Muthukrishnan, and Tamás Sarlós. Faster least squares approximation. *Numerische mathematik*, 117(2):219–249, 2011.

Ahmed el Alaoui and Michael W Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems*, pp. 775–783, 2015.

Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1-2):95–175, 2018.

Noureddine El Karoui and Holger Kösters. Geometric sensitivity of random matrix results: consequences for shrinkage estimators of covariance and related statistical methods. *arXiv preprint arXiv:1105.1404*, 2011.

Alon Gonen, Francesco Orabona, and Shai Shalev-Shwartz. Solving ridge regression using sketched preconditioned svrg. In *International Conference on Machine Learning*, pp. 1397–1405, 2016.

Walid Hachem, Philippe Loubaton, and Jamal Najim. Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3):875–930, 2007.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer series in statistics, 2009.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Fumio Hiai and Dénes Petz. *The semicircle law, free random variables and entropy*. Number 77. American Mathematical Soc., 2006.

Zengfeng Huang. Near optimal frequent directions for sketching dense and sparse matrices. In *International Conference on Machine Learning*, pp. 2053–2062, 2018.

Miles E Lopes, Shusen Wang, and Michael W Mahoney. Error estimation for randomized least-squares algorithms via the bootstrap. *arXiv preprint arXiv:1803.08021*, 2018.

Ping Ma, Michael W Mahoney, and Bin Yu. A statistical perspective on algorithmic leveraging. *The Journal of Machine Learning Research*, 16(1):861–911, 2015.

Michael W Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

Vladimir A Marchenko and Leonid A Pastur. Distribution of eigenvalues for some sets of random matrices. *Mat. Sb.*, 114(4):507–536, 1967.

Robb J Muirhead. *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons, 2009.

Alexandru Nica and Roland Speicher. *Lectures on the combinatorics of free probability*, volume 13. Cambridge University Press, 2006.

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

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

Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pp. 143–152. IEEE, 2006.

Vadim Ivanovich Serdobolskii. *Multiparametric Statistics*. Elsevier, 2007.

Gian-Andrea Thanei, Christina Heinze, and Nicolai Meinshausen. Random projections for large-scale regression. In *Big and complex data analysis*, pp. 51–68. Springer, 2017.

Antonio M Tulino and Sergio Verdú. Random matrix theory and wireless communications. *Communications and Information theory*, 1(1):1–182, 2004.

Santosh S Vempala. *The random projection method*, volume 65. American Mathematical Soc., 2005.

Dan V Voiculescu, Ken J Dykema, and Alexandru Nica. *Free random variables*. Number 1. American Mathematical Soc., 1992.

Jialei Wang, Jason D Lee, Mehrdad Mahdavi, Mladen Kolar, and Nathan Srebro. Sketching meets random projection in the dual: A provable recovery algorithm for big and high-dimensional data. *Electronic Journal of Statistics*, 11(2): 4896–4944, 2017.

Shusen Wang, Alex Gittens, and Michael W Mahoney. Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. *Journal of Machine Learning Research*, 18:1–50, 2018.

Hadley Wickham. *nycflights13: Flights that Departed NYC in 2013*, 2018. URL `https://CRAN.R-project.org/package=nycflights13`. R package version 1.0.0.

David P Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Jianfeng Yao, Zhidong Bai, and Shurong Zheng. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge University Press, New York, 2015.

Lijun Zhang, Mehrdad Mahdavi, and Rong Jin. Linear convergence with condition number independent access of full gradients. In *Advances in Neural Information Processing Systems*, pp. 980–988, 2013a.

Lijun Zhang, Mehrdad Mahdavi, Rong Jin, Tianbao Yang, and Shenghuo Zhu. Recovering the optimal solution by dual random projection. In *Conference on Learning Theory*, pp. 135–157, 2013b.

A Appendix

A.1 Proof of Theorem 2.1

If $p/n \to \gamma$ and the spectral distribution of Σ converges to H, we have by the general Marchenko-Pastur (MP) theorem of Rubio and Mestre (Rubio & Mestre, 2011), that

$$(\widehat{\Sigma} + \lambda I)^{-1} \asymp (c_p \Sigma + \lambda I)^{-1},$$

where $c_p := c(n, p, \Sigma, \lambda)$ is the unique positive solution of the fixed point equation

$$1 - c_p = \frac{c_p}{n} \operatorname{tr}\left[\Sigma (c_p \Sigma + \lambda I)^{-1}\right].$$

Here, using the terminology of the calculus of deterministic equivalents (Dobriban & Sheng, 2018), two sequences of (not necessarily symmetric) $n \times n$ matrices A_n, B_n of growing dimensions are *equivalent*, and we write

$$A_n \asymp B_n$$

if $\lim_{n\to\infty} \operatorname{tr}\left[C_n(A_n - B_n)\right] = 0$ almost surely, for any sequence C_n of (not necessarily symmetric) $n \times n$ deterministic matrices with bounded trace norm, i.e., such that $\limsup \|C_n\|_{tr} < \infty$ (Dobriban & Sheng, 2018). Informally, linear combinations of the entries of A_n can be approximated by the entries of B_n.

We start with

$$\begin{aligned}\hat\beta &= \left(X^\top X/n + \lambda I_p\right)^{-1} X^\top Y/n = \left(X^\top X/n + \lambda I_p\right)^{-1} \frac{X^\top(X\beta+\varepsilon)}{n} \\ &= (\widehat\Sigma + \lambda I_p)^{-1}\widehat\Sigma\beta + (\widehat\Sigma + \lambda I_p)^{-1}\frac{X^\top\varepsilon}{n}.\end{aligned}$$

Then, by the general MP law written in the language of the calculus of deterministic equivalents

$$(\widehat\Sigma + \lambda I_p)^{-1}\widehat\Sigma = I_p - \lambda(\widehat\Sigma + \lambda I_p)^{-1} \asymp I_p - \lambda(c_p\Sigma + \lambda I)^{-1} = c_p\Sigma(c_p\Sigma + \lambda I)^{-1}.$$

By the definition of equivalence for vectors,

$$(\widehat\Sigma + \lambda I_p)^{-1}\widehat\Sigma\beta \asymp c_p\Sigma(c_p\Sigma + \lambda I)^{-1}\beta.$$

We note a subtle point here. The rank of the matrix $M := (\widehat\Sigma + \lambda I_p)^{-1}\widehat\Sigma$ is at most n, and so it is not a full rank matrix when $n < p$. In contrast, $c_p\Sigma(c_p\Sigma + \lambda I)^{-1}$ can be a full rank matrix. Therefore, for the vectors β in the null space of $\widehat\Sigma$, which is also the null space of X, we certainly have that the two sides are not equal. However, here we assumed that the matrix X is random, and so its null space is a random $\max(p-n, 0)$ dimensional linear space. Therefore, for any fixed vector β, the random matrix M will not contain it in its null space with high probability, and so there is no contradiction.

We should also derive an asymptotic equivalent for

$$(\widehat\Sigma + \lambda I_p)^{-1}\frac{X^\top\varepsilon}{n}.$$

Suppose we have Gaussian noise, and let $Z \sim \mathcal{N}(0, I_p)$. Then we can write

$$(\widehat\Sigma + \lambda I_p)^{-1}\frac{X^\top\varepsilon}{n} =_d (\widehat\Sigma + \lambda I_p)^{-1}\widehat\Sigma^{1/2}\frac{\sigma Z}{n^{1/2}}.$$

So the question reduces to finding a deterministic equivalent for $h(\widehat\Sigma)$, where $h(x) = (x+\lambda)^{-2}x$. Note that

$$h(x) = (x+\lambda)^{-2}x = (x+\lambda)^{-2}(x+\lambda-\lambda) = (x+\lambda)^{-1} - \lambda(x+\lambda)^{-2}.$$

By the calculus of determinstic equivalents: $(\widehat\Sigma + \lambda)^{-1} \asymp (c_p\Sigma + \lambda I)^{-1}$. Moreover, fortunately the limit of the second part was recently calculated in (Dobriban & Sheng, 2019). This used the so-called "differentiation rule" of the calculus of deterministic equivalents to find

$$(\widehat\Sigma + \lambda)^{-2} \asymp (c_p\Sigma + \lambda I)^{-2}(I - c_p'\Sigma).$$

The derivative $c_p' = dc_p/dz$ has been found in Dobriban & Sheng (2019), in the proof of Theorem 3.1, part 2b. The result is (with $\gamma_p = p/n$, H_p the spectral distribution of Σ, and T a random variable distributed according to H_p)

$$c_p' = \frac{\gamma_p \mathbb{E}_{H_p}\frac{c_pT}{(c_pT-z)^2}}{-1 + \gamma_p z\mathbb{E}_{H_p}\frac{T}{(c_pT-z)^2}}. \tag{4}$$

So, we find the final answer

$$(\widehat\Sigma + \lambda I_p)^{-1}\widehat\Sigma^{1/2} \asymp A(\Sigma,\lambda) := (c_p\Sigma + \lambda I)^{-1} - \lambda(c_p\Sigma + \lambda I)^{-2}(I - c_p'\Sigma).$$

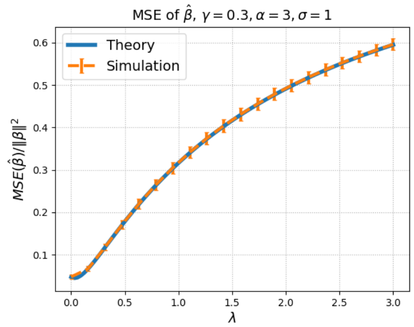

Figure 5: Simulation for ridge regression. We take $n = 1000$, $\lambda = 0.3$. Also, X has iid $\mathcal{N}(0, 1)$ entries, $\beta_i \sim_{iid} \mathcal{N}(0, \alpha^2/p)$, $\varepsilon_i \sim_{iid} \mathcal{N}(0, \sigma^2)$, with $\alpha = 3, \sigma = 1$. The standard deviations are over 50 repetitions. The theoretical lines are plotted according to Theorem A.1. The MSE is normalized by the norm of β.

A.2 Risk analysis

For a distribution F, we define the quantities

$$\theta_i(\lambda) = \int \frac{1}{(x+\lambda)^i} dF_\gamma(x),$$

($i = 1, 2, \ldots$). These are the moments of the resolvent and its derivatives (up to constants). We work in a random-effects model, where the p-dimensional regression parameter β is random, each coefficient has zero mean $\mathbb{E}\beta_i = 0$, and is normalized so that Var$\beta_i = \alpha^2/p$. This ensures that $\mathbb{E}\|\beta\|^2 = \alpha^2$. We use the following loss functions: mean squared estimation error: $MSE(\hat{\beta}) = \mathbb{E}\|\hat{\beta} - \beta\|_2^2$, and residual or training error: $Res(\hat{\beta}) = \mathbb{E}\left[\|\right] Y - X\hat{\beta}\|_2^2$.

Theorem A.1 (MSE, training error and test error of ridge regression)**.** *Suppose β has iid entries with $\mathbb{E}\beta_i = 0$, Var $[\beta_i] = \alpha^2/p$, $i = 1, \ldots, p$ and β is independent of X and ε. Suppose X is an arbitrary $n \times p$ matrix depending on n and p, and the ESD of X converges weakly to a deterministic distribution F as $n, p \to \infty$ and $p/n \to \gamma$. Then the asymptotic MSE, residual and test error of the ridge regression estimator $\hat{\beta}(\lambda)$ has the form*

$$\lim_{n\to\infty} M(\hat{\beta}(\lambda)) = \alpha^2\lambda^2\theta_2 + \gamma\sigma^2[\theta_1 - \lambda\theta_2], \tag{5}$$

$$\lim_{n\to\infty} R(\hat{\beta}(\lambda)) = \alpha^2\lambda^2[\theta_1 - \lambda\theta_2] + \sigma^2\left[1 - \gamma(1 + \lambda\theta_1 - \lambda^2\theta_2)\right], \tag{6}$$

The proof is in Section A.3. Figure 5 shows the simulation result. We see a good match between theory and simulation.

The asymptotically optimal λ in this setting is always $\lambda = \gamma\sigma^2/\alpha^2$. This follows from a Bayesian argument. The ridge regression estimator with $\lambda = p\sigma^2/(n\alpha^2)$ can be viewed as the Bayes estimator in a Gaussian model where β and ε are normal random variables. See Dobriban & Wager (2018).

A.3 Proof of Theorem A.1

Proof. The MSE of $\hat{\beta}$ has the form

$$\mathbb{E}\|\hat{\beta} - \beta\|^2 = \text{bias}^2 + \delta^2,$$

where

$$\text{bias}^2 = \mathbb{E}\left[\left\|\left(X^\top X/n + \lambda I_p\right)^{-1} X^\top X/n\beta - \beta\right\|_2^2\right],$$
$$\delta^2 = \sigma^2 \mathbb{E}\left[\left\|\left(X^\top X/n + \lambda I_p\right)^{-1} n^{-1} X^\top\right\|_F^2\right].$$

We assume that X has iid entries of zero mean and unit variance, and that $\mathbb{E}\beta = 0$, $\text{Var}\left[\beta\right] = \alpha^2/pI_p$. As $p/n \to \gamma$ as n goes to infinity, the ESD of $\frac{1}{n}X^\top X$ converges to the MP law F_γ. So we have

$$\begin{aligned}
\text{bias}^2 &= \mathbb{E}\left[\left\|\lambda\left(X^\top X/n + \lambda I_p\right)^{-1}\beta\right\|_2^2\right] \\
&= \alpha^2\lambda^2\mathbb{E}\left[\frac{1}{p}\operatorname{tr}[\left(X^\top X/n + \lambda I_p\right)^{-2}]\right] \to \alpha^2\lambda^2\int\frac{1}{(x+\lambda)^2}dF_\gamma(x),
\end{aligned}$$

and

$$\begin{aligned}
\delta^2 &= \frac{\sigma^2}{n^2}\mathbb{E}\left[\operatorname{tr}[\left(X^\top X/n + \lambda I_p\right)^{-2} X^\top X]\right] \\
&= \frac{\sigma^2}{n}\mathbb{E}\left[\operatorname{tr}[\left(X^\top X/n + \lambda I_p\right)^{-1} - \lambda\left(X^\top X/n + \lambda I_p\right)^{-2}]\right] \\
&\to \sigma^2\gamma\left[\int\frac{1}{x+\lambda}dF_\gamma(x) - \lambda\int\frac{1}{(x+\lambda)^2}dF_\gamma(x)\right].
\end{aligned}$$

Denoting $\theta_i(\gamma,\lambda) = \int\frac{1}{(x+\lambda)^i}dF_\gamma(x)$, then

$$AMSE(\hat{\beta}) = \alpha^2\lambda^2\theta_2 + \gamma\sigma^2[\theta_1 - \lambda\theta_2]. \tag{7}$$

For the standard Marchenko-Pastur law (i.e., when $\Sigma = I_p$), we have the explicit forms of θ_1 and θ_2. Specifically,

$$\theta_1 = \int\frac{1}{x+\lambda}dF_\gamma(x) = -\frac{1}{2}\left[\frac{2(1+\lambda)}{\lambda\gamma} + \frac{2}{\sqrt{\gamma}\lambda}z_2\right]$$

where

$$z_2 = -\frac{1}{2}\left[(\sqrt{\gamma} + \frac{1+\lambda}{\sqrt{\gamma}}) + \sqrt{(\sqrt{\gamma} + \frac{1+\lambda}{\sqrt{\gamma}})^2 - 4}\right].$$

It is known that the limiting Stieltjes transform $m_{F_\gamma} := m_\gamma$ of $\widehat{\Sigma}$ has the explicit form (Marchenko & Pastur, 1967):

$$m_\gamma(z) = \frac{(z+\gamma-1) + \sqrt{(z+\gamma-1)^2 - 4z\gamma}}{-2z\gamma}.$$

As usual in the area, we use the principal branch of the square root of complex numbers. Hence $\theta_1 = \frac{(-\lambda+\gamma-1)+\sqrt{(-\lambda+\gamma-1)^2+4\lambda\gamma}}{2\lambda\gamma}$. Also

$$\theta_2(\gamma,\lambda) = \int \frac{1}{(x+\lambda)^2} dF_\gamma(x) = -\int \frac{d}{d\lambda}\frac{1}{x+\lambda} dF_\gamma(x)$$
$$= -\frac{d}{d\lambda}\theta_1 = -\frac{1}{\gamma\lambda^2} + \frac{1}{\sqrt{\gamma}}\frac{d}{d\lambda}\frac{z_2}{\lambda}$$
$$= -\frac{1}{\gamma\lambda^2} + \frac{\gamma+1}{2\gamma\lambda^2} - \frac{1}{2\sqrt{\gamma}}[\frac{\lambda+\gamma+1}{\gamma\lambda\sqrt{(\sqrt{\gamma}+\frac{1+\lambda}{\sqrt{\gamma}})^2-4}} - \frac{\sqrt{(\sqrt{\gamma}+\frac{1+\lambda}{\sqrt{\gamma}})^2-4}}{\lambda^2}]$$

For the residual,

$$\mathbb{E}\left[\frac{1}{n}\|Y - X\hat{\beta}\|_2^2 | X\right] = \alpha^2\lambda^2\frac{1}{p}\operatorname{tr}[\left(X^\top X/n + \lambda I_p\right)^{-1} - \lambda\left(X^\top X/n + \lambda I_p\right)^{-2}]$$
$$+\sigma^2\frac{1}{n}[\operatorname{tr}(I_n) - 2\operatorname{tr}\left(X^\top X/n + \lambda I_p\right)^{-1} X^\top X/n + \operatorname{tr}\left(\left(X^\top X/n + \lambda I_p\right)^{-1} X^\top X/n\right)^2].$$

Next,

$$\mathbb{E}\left[\frac{1}{p}\operatorname{tr}[\left(\left(X^\top X/n + \lambda I_p\right)^{-1} X^\top X/n\right)^2]\right] = \mathbb{E}\left[\frac{1}{p}\operatorname{tr}[\left(I_p - \lambda\left(X^\top X/n + \lambda I_p\right)^{-1}\right)^2]\right]$$
$$\to 1 - 2\lambda\theta_1 + \lambda^2\theta_2.$$

Therefore

$$\mathbb{E}\left[\frac{1}{n}\|Y - X\hat{\beta}\|_2^2\right] \to \alpha^2\lambda^2[\theta_1 - \lambda\theta_2] + \sigma^2\left[1 - 2\gamma(1-\lambda\theta_1) + \gamma(1 - 2\lambda\theta_1 + \lambda^2\theta_2)\right]$$
$$= \alpha^2\lambda^2[\theta_1 - \lambda\theta_2] + \sigma^2\left[1 - \gamma(1 + \lambda\theta_1 - \lambda^2\theta_2)\right].$$

□

A.4 BIAS-VARIANCE TRADEOFF

The limiting MSE decomposes into a limiting squared bias and variance. The specific forms of these are

$$\text{bias}^2 = \alpha^2 \int \frac{\lambda^2}{(x+\lambda)^2} dF_\gamma(x), \qquad \text{var} = \gamma\sigma^2 \int \frac{x}{(x+\lambda)^2} dF_\gamma(x).$$

See Figure 1 for a plot. We can make several observations.

1. The bias increases with λ, starting out at zero for $\lambda = 0$ (linear regression), and increasing to α^2 as $\lambda \to \infty$ (zero estimator).
2. The variance decreases with λ, from $\gamma\sigma^2 \int x^{-1} dF_\gamma(x)$ to zero.
3. In the setting plotted in the figure, when α^2 and σ^2 are roughly comparable, there are additional qualitative properties we can investigate. When γ is small, the regularization parameter λ influences the bias more strongly than the variance (i.e., the derivative of the normalized quantities in the range plotted is generally larger for the normalized squared bias). In contrast when γ is large, the variance is influenced more.

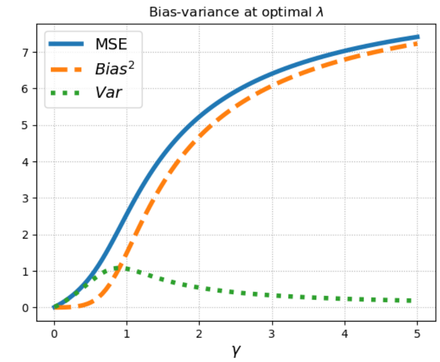

Figure 6: Bias-variance tradeoff at optimal $\lambda^* = \gamma\sigma^2/\alpha^2$, when $\alpha = 3, \sigma = 1$.

Next we consider how bias and variance change with γ at the optimal $\lambda^* = \gamma\sigma^2/\alpha^2$. This can be viewed as the "pure" effects of dimensionality on the problem, keeping all other parameters fixed. Ineed, α^2/σ^2 can be viewed as the signal-to-noise ratio (SNR), and is fixed. This analysis allows us to study for the best possible estimator (ridge regression, a Bayes estimator), behaves with the dimension. We refer to Figure 6, where we make some specific choices of α and σ.

1. Clearly the overall risk increases, as the problem becomes harder with increasing dimension. This is in line with our intuition.
2. The classical bias-variance tradeoff can be summarized by the equation
$$\text{bias}^2(\lambda) + \text{var}(\lambda) \geqslant M^*(\alpha, \gamma),$$
where we made explicit the dependence of the bias and variance on λ, and where $M^*(\alpha, \gamma)$ is the minimum MSE achievable, also known as the Bayes error, for which there are explicit formulas available (Tulino & Verdú, 2004; Dobriban & Wager, 2018).
3. The variance first increases, then decreases with γ. This shows that in the "classical" low-dimensional case, most of the risk is due to variance, while in the "modern" high-dimensional case, most of it is due to bias. This observation is consistent with other phenomena in proportional-limit asymptotics, for instance that the map between population and sample eigenvalue distributions is asymptotically deterministic (Marchenko & Pastur, 1967; Bai & Silverstein, 2010).

A.5 Simulations with cross-validation

See Figure 7. It is also shown that the one-standard-error rule (e.g., Hastie et al., 2009) does not perform well here.

A.6 Choosing the regularization parameter- additional details

Another possible prediction method is to use the average of the ridge estimators computed during cross-validation. Here it is also natural to use the CV-optimal regularization parameters, averaging $\hat\beta_{-k}(\hat\lambda_k^*)$, i.e.

$$\hat\beta_{avg}(\hat\lambda_k^*) = \frac{1}{K}\sum_{k=1}^{K}\hat\beta_{-k}(\hat\lambda_k^*).$$

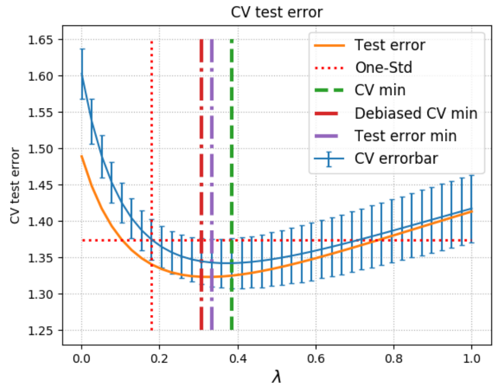

Figure 7: We generate a train set ($n = 3000$) and a test set ($n_{test} = 5000$) from the same distribution. We split the train set into $K = 5$ equally sized folds and do cross-validation. The blue error bars plot the mean and standard error of the K test errors. The red dotted lines indicates the "one-standard-error" location. The blue dashed line indicates the minimal λ^*_{CV} obtained by cross-validation, while the red dashed-dotted line indicates the debiased version $\frac{K-1}{K}\lambda^*_{CV}$. The orange line plots the test error when training on the whole train set and fit on the whole test set, and the purple dashed-dotted line indicates the minimal λ^*_{test}. The debiased λ reduces the test error by 0.00166. The minimal test error is 1.32332.

This has the advantage that it does not require refitting the ridge regression estimator, and also that we use the optimal regularization parameter.

A.6.1 Train-test validation

The same bias in the regularization parameter also applies to train-test validation. Since the number of samples is changed when restricting to the training set, the optimal λ chosen by train-test validation is also biased for the true regularization parameter minimizing the test error. We will later see in simulations (Figure 8) that retraining the ridge regression estimator on the whole data will still significantly improve the performance (this is expected based on our results on CV). For prediction, here we can also use ridge regression on the training set. This effectively reduces sample size $n \to n_{train}$, where n_{train} is the sample size of the train set. However, if the training set grows such that $n/n_{train} \to 1$ while $n_{train} \to \infty$, the train-test split has asymptotically optimal performance.

A.6.2 Leave-one-out

There is a special "short-cut" for leave-one-out in ridge regression, which saves us from burdensome computation. Write $loo(\lambda)$ for the leave-one-out estimator of prediction error with parameter λ. Instead of doing ridge regression n times, we can calculate the error explicitly as

$$loo(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i - X_i^\top \hat{\beta}(\lambda)}{1 - S_{ii}(\lambda)}\right]^2.$$

where $S(\lambda) = X(X^\top X + n\lambda I)^{-1}X^\top$. The minimizer of $loo(\lambda)$ is asymptotically optimal, i.e., it converges to λ^* (Hastie et al., 2019). However, the computational cost of this shortcut is the same as that of a train-test split. Therefore, the method described above has the same asymptotic performance.

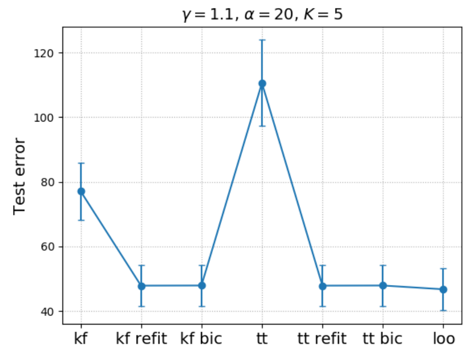

Figure 8: Comparing different ways of doing cross-validation. We take $n = 500$, $p = 550$, $\alpha = 20$, $\sigma = 1$, $K = 5$. As for train-test validation, we take 80% of samples to be training set and the rest 20% be test set. The error bars are the mean and standard deviation over 20 repetitions.

Simulations: Figure 8 shows simulation results comparing different cross-validation methods:

1. kf — k-fold cross-validation by taking the average of the ridge estimators at the CV-optimal regularization parameter.
2. kf refit — k-fold cross-validation by refitting ridge regression on the whole dataset using the CV-optimal regularization parameter.
3. kf bic — k-fold cross-validation by refitting ridge regression on the whole dataset using the CV-optimal regularization parameter, with bias correction.
4. tt — train-test validation, by using the ridge estimator computed on the train data, at the validation-optimal regularization parameter. Note: we expect this to be similar, but worse than the "kf" estimator.
5. tt refit — train-test validation by refitting ridge regression on the whole dataset, using the validation-optimal regularization parameter. Note: we expect this to be similar, but slightly worse than the "kf refit" estimator.
6. tt bic — train-test validation by refitting ridge regression on the whole dataset using the CV-optimal regularization parameter, with bias correction.
7. loo — leave-one-out

Figure 8 shows that the naive estimators (kf and tt) can be quite inaccurate without refitting or bias correction. However, if we either refit or bias-correct, the accuracy improves. In this case, there seems to be no significant difference between the various methods.

A.7 Proof of Theorem 4.1

Proof. Suppose $m/n \to \xi$ as n goes to infinity. For $\hat{\beta}_p$, we have

$$\mathrm{bias}^2 = \mathbb{E}\left[\left\| \left(X^\top L^\top L X/n + \lambda I_p\right)^{-1} X^\top X/n\beta - \beta \right\|_2^2\right],$$
$$\delta^2 = \sigma^2 \mathbb{E}\left[\left\| \left(X^\top L^\top L X/n + \lambda I_p\right)^{-1} n^{-1} X^\top \right\|_F^2\right].$$

Denote $M = \left(X^\top L^\top LX/n + \lambda I_p\right)^{-1}$, the resolvent of the sketched matrix. We further assume that X has iid $\mathcal{N}(0,1)$ entries and $LL^\top = I_m$. Let L_1 be an orthogonal complementary matrix of L, such that $L^\top L + L_1^\top L_1 = I_n$. We also denote $N = \frac{X^\top L_1^\top L_1 X}{n}$. Then

$$MX^\top X/n = M\frac{X^\top L^\top LX + X^\top L_1^\top L_1 X}{n} = I_p - \lambda M + MN.$$

Therefore, using that Cov $[\beta] = \alpha^2/p \cdot I_p$, we find the bias as

$$\begin{aligned}
\text{bias}^2 &= \frac{\alpha^2}{p}\mathbb{E}\left[\operatorname{tr}(M - I_p)(M^\top - I_p)\right] \\
&= \frac{\alpha^2}{p}\left\{\lambda^2\mathbb{E}\left[\operatorname{tr}[M^2]\right] + \mathbb{E}\left[\operatorname{tr} M^2\frac{(X^\top L_1^\top L_1 X)^2}{n^2}\right] - 2\lambda\mathbb{E}\left[\operatorname{tr} M^2 N\right]\right\}.
\end{aligned}$$

By the properties of Wishart matrices (e.g., Anderson, 2003; Muirhead, 2009), we have

$$\begin{aligned}
\mathbb{E}\left[N\right] &= \frac{n-m}{n}I_p, \\
\mathbb{E}\left[(N)^2\right] &= \frac{1}{n^2}\mathbb{E}\left[Wishart(I_p, n-m)^2\right] = \frac{1}{n^2}[n - m + p(n-m) + (n-m)^2]I_p.
\end{aligned}$$

Recalling that $m, n \to \infty$ such that $m/n \to \xi$, and that $\theta_i(\gamma,\lambda) = \int (x+\lambda)^{-i} dF_\gamma(x)$,

$$\begin{aligned}
\text{bias}^2 &= \frac{\alpha^2}{p}\left[\lambda^2 + \frac{n - m + p(n-m) + (n-m)^2}{n^2} - 2\lambda\frac{n-m}{n}\right]\mathbb{E}\left[\operatorname{tr}[M^2]\right] \\
&\to \alpha^2[(\lambda + \xi - 1)^2 + \gamma(1-\xi)]\theta_2(\gamma,\xi,\lambda).
\end{aligned}$$

Moreover,

$$\begin{aligned}
\delta^2 &= \frac{\sigma^2}{n^2}\mathbb{E}\left[\operatorname{tr}[M^2 X^\top X]\right] \\
&= \frac{\sigma^2}{n}\cdot\left\{\mathbb{E}\left[\operatorname{tr}[M]\right] - \lambda\mathbb{E}\left[\operatorname{tr}[M^2]\right] + \mathbb{E}\left[\operatorname{tr}[M^2 N]\right]\right\} \\
&\to \gamma\sigma^2[\theta_1(\gamma,\xi,\lambda) - \lambda\theta_2(\gamma,\xi,\lambda) + (1-\xi)\theta_2(\gamma,\xi,\lambda)].
\end{aligned}$$

Here we used the additional definitions

$$\begin{aligned}
\theta_i(\gamma,\xi,\lambda) &= \int\frac{1}{(\xi x + \lambda)^i}dF_{\gamma/\xi}(x) \\
\theta_i(\gamma,\lambda) &= \theta_i(\gamma,\xi = 1,\lambda).
\end{aligned}$$

Note that these can be connected to the previous definitions by

$$\begin{aligned}
\theta_1(\gamma,\xi,\lambda) &= \frac{1}{\xi}\int\frac{1}{x + \lambda/\xi}dF_{\gamma/\xi}(x) = \frac{1}{\xi}\theta_1\left(\frac{\gamma}{\xi},\frac{\lambda}{\xi}\right) \\
\theta_2(\gamma,\xi,\lambda) &= \frac{1}{\xi^2}\theta_2\left(\frac{\gamma}{\xi},\frac{\lambda}{\xi}\right).
\end{aligned}$$

Therefore the AMSE of $\hat\beta_p$ is

$$\begin{aligned}
AMSE(\hat\beta_p) &= \alpha^2[(\lambda+\xi-1)^2 + \gamma(1-\xi)]\theta_2(\gamma,\xi,\lambda) + \gamma\sigma^2[\theta_1(\gamma,\xi,\lambda) - (\lambda+\xi-1)\theta_2(\gamma,\xi,\lambda)] \\
&= \alpha^2[(\lambda+\xi-1)^2 + \gamma(1-\xi)]\frac{1}{\xi^2}\theta_2\left(\frac{\gamma}{\xi},\frac{\lambda}{\xi}\right) \\
&+ \gamma\sigma^2\left[\frac{1}{\xi}\theta_1\left(\frac{\gamma}{\xi},\frac{\lambda}{\xi}\right) - (\lambda+\xi-1)\frac{1}{\xi^2}\theta_2\left(\frac{\gamma}{\xi},\frac{\lambda}{\xi}\right)\right].
\end{aligned} \tag{8}$$

$\square$

A.7.1 ISOTROPIC CASE

Consider the special case where $\Gamma = I$, that is, X has iid $\mathcal{N}(0,1)$ entries. Then F_γ is the standard MP law, and we have the explicit forms for $\theta_i = \theta_i(\gamma,\lambda) = \int \frac{1}{(x+\lambda)^i} dF_\gamma$:

$$\theta_1(\gamma,\lambda) = -\frac{1+\lambda}{\gamma\lambda} + \frac{1}{2\sqrt{\gamma}\lambda}[\sqrt{\gamma} + \frac{1+\lambda}{\sqrt{\gamma}} + \sqrt{(\sqrt{\gamma}+\frac{1+\lambda}{\sqrt{\gamma}})^2 - 4}],$$

$$\theta_2(\gamma,\lambda) = -\frac{1}{\gamma\lambda^2} + \frac{\gamma+1}{2\gamma\lambda^2} - \frac{1}{2\sqrt{\gamma}}(\frac{\lambda+1}{\gamma}+1)\frac{1}{\lambda\sqrt{(\sqrt{\gamma}+\frac{1+\lambda}{\sqrt{\gamma}})^2-4}} + \frac{1}{2\sqrt{\gamma}}\sqrt{(\sqrt{\gamma}+\frac{1+\lambda}{\sqrt{\gamma}})^2-4}\frac{1}{\lambda^2},$$

$$\bar\theta_1(\zeta,\lambda) = \zeta\theta_1(\zeta,\lambda) + \frac{1-\zeta}{\lambda},$$

$$\bar\theta_2(\zeta,\lambda) = \zeta\theta_2(\zeta,\lambda) + \frac{1-\zeta}{\lambda^2},$$

The results are obtained by the contour integral formula

$$\int f(x)dF_\gamma(x) = -\frac{1}{4\pi i}\oint_{|z|=1} \frac{f(|1+\gamma z|^2)(1-z^2)^2}{z^2(1+\sqrt{\gamma}z)(z+\sqrt{\gamma})}dz.$$

See Proposition 2.10 of Yao et al. (2015).

A.8 PROOF OF THEOREM 4.2

Proof. Suppose $d/p \to \zeta$ as n goes to infinity. For $\hat\beta_d$, we have

$$\mathrm{bias}^2 = \mathbb{E}\left[\left\|n^{-1}X^\top\left(XRR^\top X^\top/n + \lambda I_n\right)^{-1}X\beta - \beta\right\|_2^2\right],$$

$$\delta^2 = \sigma^2\,\mathrm{tr}[\left(XRR^\top X^\top/n + \lambda I_n\right)^{-2}\frac{XX^\top}{n^2}].$$

Denote $M = \left(XRR^\top X^\top/n + \lambda I_n\right)^{-1}$. Note that, using that $\mathrm{Cov}\,[\beta] = \alpha^2/p\cdot I_p$

$$\mathrm{bias}^2 = \frac{\alpha^2}{p}\mathbb{E}\left[\mathrm{tr}[MXX^\top/n]^2\right] - 2\frac{\alpha^2}{p}\mathbb{E}\left[\mathrm{tr}[MXX^\top/n]\right] + \frac{\alpha^2}{p}\,\mathrm{tr}(I_p).$$

Moreover, letting R_1 to be an orthogonal complementary matrix of R, such that $RR^\top + R_1R_1^\top = I_n$, and $N = \frac{XR_1R_1^\top X^\top}{n}$,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{p}\,\mathrm{tr}[MXX^\top/n]\right] &= \frac{1}{p}\,\mathrm{tr}[I_n - \lambda\mathbb{E}\left[\mathrm{tr}[M]\right] + \mathbb{E}\left[MN\right]]\\ &\to \frac{1}{\gamma} - \frac{\lambda}{\gamma}\int\frac{1}{x+\lambda}d\bar F_\zeta(x) + \frac{\gamma-\zeta}{\gamma}\int\frac{1}{x+\lambda}d\bar F_\zeta(x),\end{aligned}$$

where $\bar F_\zeta$ is the companion MP law, that is, $\bar F_\zeta = (1-\gamma)\delta_0 + \gamma F_\zeta$. The third term calculated by using that XR and XR_1 are independent for a Gaussian random matrix X, so that M, N are independent, and that

$\mathbb{E}[N] = \frac{p-d}{n} I_n$. Thus

$$\mathbb{E}\left[\frac{1}{p}\operatorname{tr}[MXX^\top/n]\right] \to \frac{1}{\gamma} - \frac{\lambda+\zeta-\gamma}{\gamma}\bar{\theta}_1(\zeta,\lambda)$$
$$= \frac{1}{\gamma} - \frac{\lambda+\zeta-\gamma}{\gamma}\left[\frac{1-\zeta}{\lambda} + \zeta\theta_1(\zeta,\lambda)\right].$$

Then

$$\mathbb{E}\left[\frac{1}{p}\operatorname{tr}[MXX^\top/n]^2\right] = \frac{1}{p}\mathbb{E}\left[\operatorname{tr}[I_n + \lambda^2 M^2 + MNMN - 2\lambda M + 2MN - \lambda M^2 N - \lambda MNM\right].$$

Note that

$$\mathbb{E}[MNMN|M] = M[(p-d)(M^\top + \operatorname{tr}(M)I_n) + (p-d)^2 M]/n^2$$
$$= \frac{p-d+(p-d)^2}{n^2}M^2 + \frac{p-d}{n^2}\operatorname{tr}(M)M,$$

so

$$\mathbb{E}\left[\frac{1}{p}\operatorname{tr}[MXX^\top/n]^2\right] \to \frac{1}{\gamma}[1 + (\lambda^2 - 2\lambda(\gamma-\zeta) + (\gamma-\zeta)^2)\bar{\theta}_2(\zeta,\lambda)$$
$$+ 2(\gamma-\zeta-\lambda)\bar{\theta}_1(\zeta,\lambda) + (\gamma-\zeta)\bar{\theta}_1^2(\zeta,\lambda)].$$

Thus we find the following exprssion for the limiting squared bias:

$$\text{bias}^2 \to \frac{\alpha^2}{\gamma}[\gamma - 1 + (\lambda-\gamma+\zeta)^2\bar{\theta}_2 + (\gamma-\zeta)\bar{\theta}_1^2].$$

With similar calculations (that we omit for brevity), we can find

$$\delta^2 \to \sigma^2(\bar{\theta}_1(\zeta,\lambda) - (\lambda+\zeta-\gamma)\bar{\theta}_2(\zeta,\lambda)).$$

Therefore the AMSE of $\hat{\beta}_d$ is

$$AMSE = \frac{\alpha^2}{\gamma}[\gamma - 1 + (\lambda-\gamma+\zeta)^2\bar{\theta}_2 + (\gamma-\zeta)\bar{\theta}_1^2] + \sigma^2[\bar{\theta}_1(\zeta,\lambda) - (\lambda+\zeta-\gamma)\bar{\theta}_2(\zeta,\lambda)]. \tag{9}$$

□

A.9 Proof of Theorem 4.3

Proof. Recall that we have $m, n \to \infty$, such that $m/n \to \xi$. Then we need to take $\xi \to 0$. However, we find it more convenient to do the calculation directly from the finite sample results as $m, n, p \to \infty$ with $m/n \to 0$, $p/n \to \gamma$, It is not hard to check that computing the results in the other way (i.e., interchanging the limits), leads to the same results. Starting from our bias formula for primal sketching, we first get

$$\text{bias}^2 = \frac{\alpha^2}{p}\left[\lambda^2 + \frac{n-m+p(n-m)+(n-m)^2}{n^2} - 2\lambda\frac{n-m}{n}\right]\mathbb{E}\left[\operatorname{tr}[\left(X^\top L^\top LX/n + \lambda I_p\right)^{-2}]\right]$$
$$\to \alpha^2[(\lambda-1)^2 + \gamma]/\lambda^2.$$

The limit of the trace term is not entirely trivial, but it can be calculated by (1) observing that the $m \times p$ sketched data matrix $P = LX$ has iid normal entries (2) thus the operator norm of $P^\top P/n$ vanishes, (3) and so by a simple matrix perturbation argument the trace concentrates around p/λ^2. This gives the rough steps of finding the above limit. Moreover,

$$\delta^2 = \frac{\sigma^2}{n^2}\mathbb{E}\left[\operatorname{tr}[\left(X^\top L^\top LX/n + \lambda I_p\right)^{-2} X^\top X]\right] \to \gamma\sigma^2/\lambda^2 \cdot \mathbb{E}_{F_\gamma} X^2 = \gamma\sigma^2/\lambda^2$$

So the MSE is $M(\lambda) = \alpha^2[(\lambda-1)^2 + \gamma]/\lambda^2 + \sigma^2 \cdot \gamma/\lambda^2$. From this it is elementary to find the optimal λ and its objective value. □

A.10 PROOF OF THEOREM 4.4

Proof. Note that the bias can be written as

$$\begin{aligned}\text{bias}^2 = &\frac{\alpha^2}{p}\mathbb{E}\left[\text{tr}[\left(\frac{XRR^\top X^\top}{nd} + \lambda I_n\right)^{-1}\frac{XX^\top}{nd}]^2\right] \\ &- 2\frac{\alpha^2}{p}\mathbb{E}\left[\text{tr}[\left(XRR^\top X^\top/n + \lambda I_n\right)^{-1}XX^\top/n]\right] + \alpha^2.\end{aligned}$$

Write $G = XX^\top$. Since $RR^\top \sim \mathcal{W}_p(I_p, d)$, we have $XRR^\top X^\top \sim \mathcal{W}_n(G, d)$. So $XRR^\top X^\top \stackrel{d}{=} G^{1/2}WG^{1/2}$, where $W \sim \mathcal{W}_n(I_n, d)$.

$$\begin{aligned}\mathbb{E}\left[\text{tr}[\left(\frac{XRR^\top X^\top}{nd} + \lambda I_n\right)^{-1}XX^\top/n]\right] &= \mathbb{E}\left[\text{tr}[(G^{1/2}WG^{1/2}/d + n\lambda I_n)^{-1}G]\right] \\ &= \mathbb{E}\left[\text{tr}[(\frac{W}{d} + \lambda(\frac{G}{n})^{-1})^{-1}]\right].\end{aligned}$$

So we need to find the law of $\frac{W}{d} + \frac{\lambda}{\gamma}(\frac{G}{p})^{-1}$. Suppose first that $G = XX^\top \sim \mathcal{W}_n(I_n, p)$. Then W and G^{-1} are asymptotically freely independent. The l.s.d. of W/d is the MP law $F_{1/\xi}$ while the l.s.d. of G/p is the MP law $F_{1/\gamma}$. We need to find the additive free convolution $W \boxplus \bar{G}$, where $\bar{G} = \frac{\lambda}{\gamma}G^{-1}$.

Recall that the R-transform of a distribution F is defined by

$$R_F(z) = m_F^{-1}(-z) - \frac{1}{z},$$

where $m_F^{-1}(z)$ is the inverse function of the Stieltjes transform of F (e.g., Voiculescu et al., 1992; Hiai & Petz, 2006; Couillet & Debbah, 2011). We can find the R-transform by solving

$$m_F(R_F(z) + \frac{1}{z}) = -z.$$

Note that the R-transform of W/d is

$$R_W(z) = \frac{1}{1 - z/\xi}.$$

The Stieltjes transform of G^{-1} is

$$\begin{aligned}m_{G^{-1}}(z) &= \int \frac{1}{1/x - z}dF_{1/\gamma}(x) = -\frac{1}{z} - \frac{1}{z^2}m_{1/\gamma}(\frac{1}{z}) \\ &= -\frac{1}{z} - \frac{1 - \frac{1}{\gamma} - \frac{1}{z} + \sqrt{(1 + \frac{1}{\gamma} + \frac{1}{z})^2 - \frac{4}{\gamma}}}{2\frac{z}{\gamma}} \\ &= -\frac{1 + \frac{1}{\gamma} - \frac{1}{z} + \sqrt{(1 + \frac{1}{\gamma} - \frac{1}{z})^2 - \frac{4}{\gamma}}}{2\frac{z}{\gamma}}.\end{aligned}$$

Then the R-transform of G^{-1} is

$$\begin{aligned}R_{G^{-1}}(z) &= -\frac{1}{z} + \frac{\gamma + 1 - \sqrt{(\gamma+1)^2 - 4\gamma(z+1)}}{2z} \\ &= \frac{\gamma - 1 - \sqrt{(\gamma-1)^2 - 4\gamma z}}{2z}.\end{aligned}$$

Since we have the property that $R_{a\mu}(z) = aR_\mu(az)$,

$$R_{\bar{G}} = R_{\frac{\lambda}{\gamma}G^{-1}}(z) = \frac{\gamma - 1 - \sqrt{(\gamma-1)^2 - 4\lambda z}}{2z}.$$

Hence we have

$$R_{W \boxplus \bar{G}} = R_W + R_{\bar{G}} = \frac{1}{1 - z/\xi} + \frac{\gamma - 1 - \sqrt{(\gamma-1)^2 - 4\lambda z}}{2z}.$$

Moreover, the Stieltjes transform of $\mu = W \boxplus \bar{G}$ satisfies

$$m_\mu^{-1}(z) = m_{W\boxplus\bar{G}}^{-1}(z) = R_F(-z) - \frac{1}{z} = \frac{1}{1+z/\xi} + \frac{\gamma - 1 - \sqrt{(\gamma-1)^2 + 4\lambda z}}{-2z} - \frac{1}{z}.$$

Note that

$$2\frac{\alpha^2}{p}\mathbb{E}\left[\mathrm{tr}[\left(\frac{XRR^\top X^\top}{nd} + \lambda I_n\right)^{-1} XX^\top/n]\right] \to 2\frac{\alpha^2}{\gamma}\mathbb{E}_\mu\left[\frac{1}{x}\right] = 2\frac{\alpha^2}{\gamma}\lim_{z\to 0} m(z),$$

$$\frac{\alpha^2}{p}\mathbb{E}\left[\mathrm{tr}[\left(\frac{XRR^\top X^\top}{nd} + \lambda I_n\right)^{-1} \frac{XX^\top}{nd}]^2\right] \to \frac{\alpha^2}{\gamma}\mathbb{E}_\mu\left[\frac{1}{x^2}\right] = \frac{\alpha^2}{\gamma}\lim_{z\to 0}\frac{d}{dz}m(z).$$

So it suffices to find $m(z)$ and $\frac{d}{dz}m(z)$ evaluated at zero. $\square$

This result can characterize the performance of sketching in the high SNR regime, where $\alpha \gg \sigma$. To understand the lower SNR regime, we need to study the variance, and thus we need to calculate

$$\mathrm{var} = \sigma^2\frac{1}{n}\mathbb{E}\left[\mathrm{tr}[\left(\frac{XRR^\top X^\top}{nd} + \lambda I_n\right)^{-2} XX^\top/n]\right] = \sigma^2\mathbb{E}\left[\mathrm{tr}[(\frac{W}{d} + \frac{\lambda}{\gamma}(\frac{G}{p})^{-1})^{-2}G^{-1}]\right]$$

where $G = XX^\top \sim \mathcal{W}_n(I_n, p)$ is a Wishart distribution, and $XRR^\top X^\top =_d G^{1/2}WG^{1/2}$, with $W \sim \mathcal{W}_n(I_n, r)$. This seems to be quite challenging, and we leave it to future work.

A.11 Results for primal Gaussian sketching

The statement requires some notions from free probability, see e.g., Voiculescu et al. (1992); Hiai & Petz (2006); Nica & Speicher (2006); Anderson et al. (2010); Couillet & Debbah (2011) for references .

Theorem A.2 (Bias of primal Gaussian sketch)**.** *Suppose X is an $n \times p$ standard Gaussian random matrix. Suppose also that L is a $d \times n$ matrix with i.i.d. $\mathcal{N}(0, 1/d)$ entries. Then the bias of primal sketch has the expression $MSE(\hat{\beta}_p) = \alpha^2 + \frac{\alpha^2}{\gamma}[\tau((a+b)^{-1}b(a+b)^{-1}b^{-1}) - 2\tau((a+b)^{-1})]$, where a and b two free random variables, that are freely independent in a non-commutative probability space, and τ is their trace. Specifically, the law of a is the MP law $F_{1/\xi}$ and $b = \frac{\lambda}{\gamma}\tilde{b}$, where the law of $\tilde{b}$ is the MP law $F_{1/\gamma}$.*

Proof of Theorem A.2. Note that

$$\mathrm{bias}^2 = \mathbb{E}\left[\left\|\left(X^\top L^\top LX/(nd) + \lambda I_p\right)^{-1}(X^\top X/n)\beta - \beta\right\|_2^2\right],$$

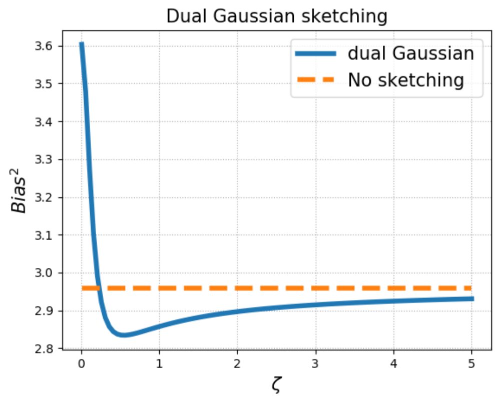

Figure 9: Dual Gaussian sketch improves MSE.

and $\left(X^\top L^\top LX/(nd) + \lambda I_p\right)^{-1} X^\top = X^\top (L^\top LXX^\top/(nd) + \lambda I_n)^{-1}$. Thus

$$\begin{aligned}
\text{bias}^2 &= \mathbb{E}\left[\left\|X^\top (\frac{L^\top LXX^\top}{nd} + \lambda I_n)^{-1}\frac{X}{n}\beta - \beta\right\|_2^2\right] \\
&= \alpha^2 + \frac{\alpha^2}{p}\mathbb{E}[\mathrm{tr}[(\frac{XX^\top L^\top L}{nd} + \lambda I_n)^{-1}\frac{XX^\top}{n}(\frac{L^\top LXX^\top}{nd} + \lambda I_n)^{-1}\frac{XX^\top}{n}] \\
&- 2\,\mathrm{tr}[(\frac{L^\top LXX^\top}{nd} + \lambda I_n)^{-1}\frac{XX^\top}{n}]].
\end{aligned}$$

First we find the l.s.d. of $(\frac{L^\top LXX^\top}{nd} + \lambda I_n)^{-1}\frac{XX^\top}{n}$. Write $W = L^\top L$, $G = XX^\top$. Then

$$(\frac{L^\top LXX^\top}{nd} + \lambda I_n)^{-1}\frac{XX^\top}{n} = (\frac{WG}{nd} + \lambda I_n)^{-1}\frac{G}{n} = G^{-1}(\frac{W}{d} + \lambda(\frac{G}{n})^{-1})^{-1}G,$$

which is similar to $(\frac{W}{d} + \lambda(\frac{G}{n})^{-1})^{-1}$. So it suffices to find the l.s.d. of $(\frac{W}{d} + \frac{\lambda}{\gamma}(\frac{G}{p})^{-1})^{-1}$.

By the definition, $W \sim \mathcal{W}_n(I_n, d)$, $G \sim \mathcal{W}_n(I_n, p)$, therefore the l.s.d. of W/d converges to the MP law $F_{1/\xi}$ and the l.s.d. of G/p converges to the MP law $F_{1/\gamma}$.

Also note that

$$(\frac{XX^\top L^\top L}{nd} + \lambda I_n)^{-1}\frac{XX^\top}{n}(\frac{L^\top LXX^\top}{nd} + \lambda I_n)^{-1}\frac{XX^\top}{n} = (\frac{W}{d} + \lambda nG^{-1})^{-1}G^{-1}(\frac{W}{d} + \lambda nG^{-1})^{-1}G.$$

We write $A = \frac{W}{d}$, $B = \frac{\lambda}{\gamma}(\frac{G}{p})^{-1}$. Then it suffices to find

$$\frac{\alpha^2}{p}\mathbb{E}\left[\mathrm{tr}[(A+B)^{-1}B(A+B)^{-1}B^{-1}]\right].$$

We will find an expression for this using free probability. For this we will need to use some series expansions. There are two cases, depending on whether the operator norm of BA^{-1} is less than or greater than unity, leading to different series expansions. We will work out below the first case, but the second case is similar and leads to the same answer.

$$\mathrm{tr}[(A+B)^{-1}B(A+B)^{-1}B^{-1}] = \mathrm{tr}[A^{-1}(I+BA^{-1})^{-1}BA^{-1}(I+BA^{-1})^{-1}B^{-1}]$$

Since the operator norm of BA^{-1} is less unity, we have the von Neumann series expansion

$$[I + BA^{-1}]^{-1} = \sum_{i=0}^{\infty}(-BA^{-1})^i,$$

then we have

$$\begin{aligned}
\operatorname{tr}[(A+B)^{-1}B(A+B)^{-1}B^{-1}] &= \sum_{i,j\geq 0}(-1)^{i+j}\operatorname{tr}[(BA^{-1})^{i+j+1}B^{-1}A^{-1}] \\
&= \sum_{i,j\geq 0}(-1)^{i+j}\operatorname{tr}[(A^{-1}B)^{i+j+1}A^{-1}B^{-1}].
\end{aligned}$$

Since A and B are asymptotically freely independent in the free probability space arising in the limit (e.g., Voiculescu et al., 1992; Hiai & Petz, 2006; Couillet & Debbah, 2011), and the polynomial $(a^{-1}b)^{i+j+1}a^{-1}b^{-1}$ involves an alternating sequence of a, b, we have

$$\frac{1}{n}\operatorname{tr}[(A^{-1}B)^{i+j+1}A^{-1}B^{-1}] \to \tau[(a^{-1}b)^{i+j+1}a^{-1}b^{-1}],$$

where a and the b are free random variables and τ is their law. Specifically, a is a free random variable with the MP law $F_{1/\xi}$ and b is $\frac{\lambda}{\gamma}\tilde{b}^{-1}$, where $\tilde{b}$ is a free r.v. with MP law $F_{1/\gamma}$. Moreover, they are freely independent.

Hence, we have

$$\begin{aligned}
\frac{1}{n}\operatorname{tr}[(A+B)^{-1}B(A+B)^{-1}B^{-1}] &\to \tau[\sum_{i\geq 0}(-1)^i(a^{-1}b)^{i+1}\sum_{j\geq 0}(-1)^j(a^{-1}b)^j a^{-1}b^{-1}] \\
&= \tau[(a^{-1}b)(1+a^{-1}b)^{-1}(1+a^{-1}b)^{-1}a^{-1}b^{-1}] \\
&= \tau[(a+b)^{-1}b(a+b)^{-1}b^{-1}].
\end{aligned}$$

Therefore,

$$\begin{aligned}
\text{bias}^2 &\to \alpha^2 + \frac{\alpha^2}{\gamma}\left[\frac{1}{n}\operatorname{tr}[(A+B)^{-1}B(A+B)^{-1}B^{-1} - 2\operatorname{tr}[A+B]^{-1}]\right] \\
&= \alpha^2 + \frac{\alpha^2}{\gamma}[\tau((a+b)^{-1}b(a+b)^{-1}b^{-1}) - 2\tau((a+b)^{-1})].
\end{aligned}$$

□

A.12 Results for full sketching

The full sketch estimator projects down the entire data, and then does ridge regression on the sketched data. It has the form

$$\hat{\beta}_f = \left(X^\top L^\top LX/n + \lambda I_p\right)^{-1}\frac{X^\top L^\top LY}{n}.$$

We have

$$\begin{aligned}
\text{bias}^2 &= \alpha^2\lambda^2\int\frac{1}{\xi x+\lambda}dF_{\gamma/\xi}(x) = \alpha^2\lambda^2\frac{1}{\xi^2}\theta_2\left(\frac{\gamma}{\xi},\frac{\lambda}{\xi}\right) \\
\text{var} &= \sigma^2\gamma\left[\int\frac{1}{\xi x+\lambda}dF_{\gamma/\xi}(x) - \lambda\int\frac{1}{(\xi x+\lambda)^2}dF_{\gamma/\xi}(x)\right] \\
&= \sigma^2\gamma\left[\frac{1}{\xi}\theta_1\left(\frac{\gamma}{\xi},\frac{\lambda}{\xi}\right) - \lambda\frac{1}{\xi^2}\theta_2\left(\frac{\gamma}{\xi},\frac{\lambda}{\xi}\right)\right],
\end{aligned}$$

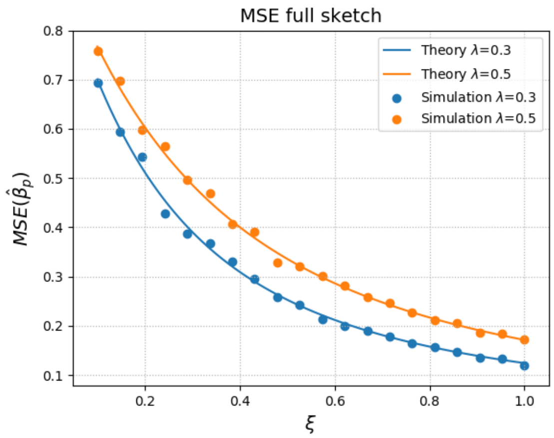

Figure 10: Simulation results for full sketch, with $n = 1000$, $\gamma = 0.1$. The simulation results are averaged over 30 independent experiments.

therefore

$$AMSE(\hat{\beta}_f) = \alpha^2\lambda^2\frac{1}{\xi^2}\theta_2\left(\frac{\gamma}{\xi},\frac{\lambda}{\xi}\right) + \sigma^2\gamma\left[\frac{1}{\xi}\theta_1\left(\frac{\gamma}{\xi},\frac{\lambda}{\xi}\right) - \lambda\frac{1}{\xi^2}\theta_2\left(\frac{\gamma}{\xi},\frac{\lambda}{\xi}\right)\right]$$

The optimal λ for full sketch is always $\lambda^* = \frac{\gamma\sigma^2}{\alpha^2}$, the same as ridge regression. Some simulation results are shown in Figure 10, and they show the expected shape (e.g., they decrease with ξ).

A.13 Numerical results

A.13.1 Dual orthogonal sketching

See Figure 11 for additional simulation results for dual orthogonal sketching.

A.13.2 Performance at a fixed regularization parameter

First we fix the regularization parameter at the optimal value for original ridge regression. The results are visualized in Figure 12. On the x axis, we plot the reduction in sample size m/n for primal sketch, and the reduction in dimension d/p for dual sketch. In this case, primal and dual sketch will increase both bias and variance, and empirically in the current case, dual sketch increases them more. So in this particular case, primal sketch is preferred.

A.13.3 Performance at the optimal regularization parameter

We find the optimal regularization parameter λ for primal and dual orthogonal sketching.a

Then we use the optimal regularization parameter for all settings, see Figure 13. Both primal and dual sketch increase the bias, but decrease the variance. It is interesting to note that, for equal parameters ξ and ζ, and in our particular case, dual sketch has smaller variance, but larger bias. So primal sketch is preferred bias or MSE is important, but dual sketch is more desired when one wants smaller variance. All in all, dual sketch has larger MSE than primal sketch in the current setting. It can also be seen that in this specific example, the

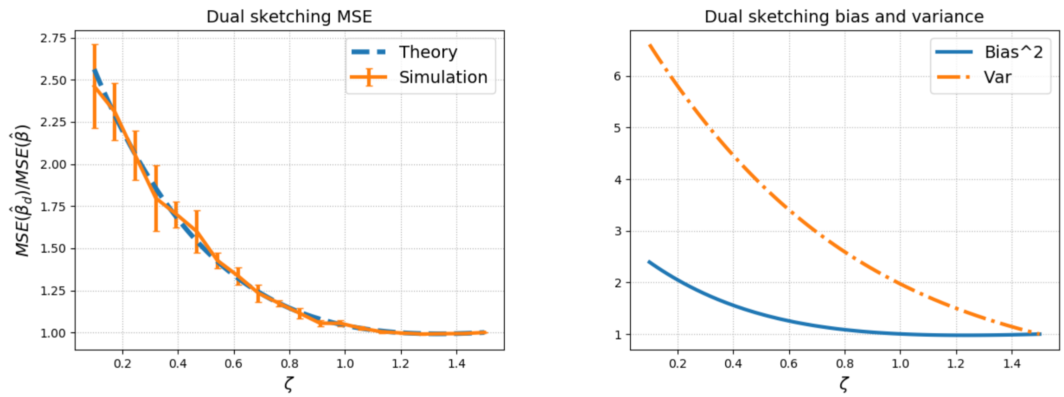

Figure 11: Dual orthogonal sketching with $\gamma = 1.5, \lambda = 1, \alpha = 3, \sigma = 1$. Left: MSE of dual sketching normalized by the MSE of ridge regression. The standard deviation is over 50 repetitions. Right: Bias and variance of dual sketching normalized by the bias and variance of ridge regression, respectively.

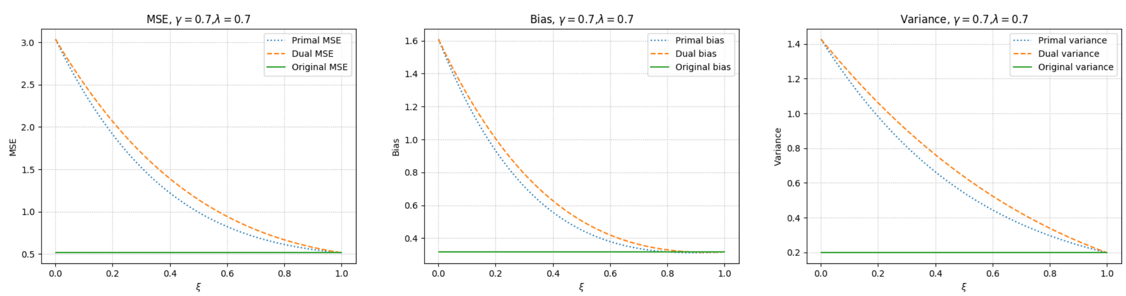

Figure 12: Fixed regularization parameter $\lambda = 0.7$, optimal for original ridge, in a setting where $\gamma = 0.7$, and $\alpha^2 = \sigma^2$.

optimal λ for primal sketch is smaller than that of dual sketch. However these results are hard to interpret, because there is no natural correspondence between the two parameters ξ and ζ.

A.14 Computational complexity

Since sketching is a method to reduce computational complexity, it is important to discuss how much computational efficiency we gain. Recall our three estimators

$$\hat{\beta} = \left(X^\top X/n + \lambda I_p\right)^{-1} X^\top Y/n = n^{-1} X^\top \left(XX^\top/n + \lambda I_n\right)^{-1} Y,$$
$$\hat{\beta}_p = \left(X^\top L^\top L X/n + \lambda I_p\right)^{-1} X^\top Y/n,$$
$$\hat{\beta}_d = n^{-1} X^\top \left(X R R^\top X^\top/n + \lambda I_n\right)^{-1} Y,$$

Their computational complexity, when computed in the usual way, is:

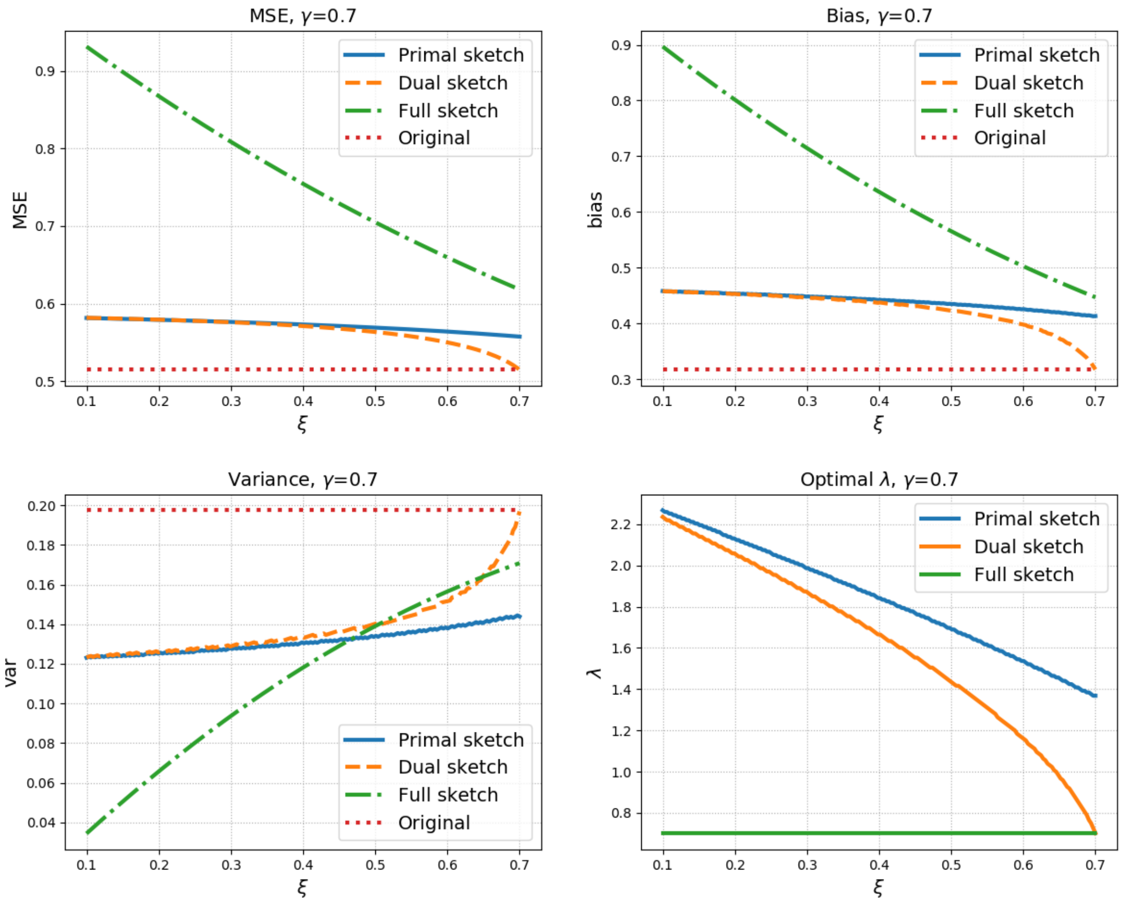

Figure 13: Primal and dual sketch at optimal λ. We take $\gamma = 0.7$ and let ξ range between 0.001 and 1, where for primal sketch $\xi = r/n$ while for dual sketch $\xi = d/p$.

- No sketch (Standard ridge): if $p < n$, computing $X^\top Y$ and $X^\top X$ requires $O(np)$ and $O(np^2)$ flops, then solving the linear equation $(X^\top X/n + \lambda I_p)\hat{\beta} = X^\top Y/n$ requires $O(p^3)$ flops by the LU decomposition. It is $O(np^2)$ flops in total.

  If $p > n$, we use the second formula for $\hat{\beta}$, and the total flops is $O(pn^2)$.
- Primal sketch: for the Hadamard sketch (and other sketches based on the FFT), computing LX by FFT requires $mp \log n$, computing $(LX)^\top LX$ requires mp^2, so the total flops is $O(p^3 + mp(\log n + p))$. So the primal sketch can reduce the computation cost only when $p < n$.
- Dual sketch: computing $XRR^\top X^\top$ requires $nd\,(\log p + n)$ flops by FFT, solving $(XRR^\top X^\top/n + \lambda I_n)^{-1}\,Y$ requires $O(n^3)$ flops, the matrix-vector multiplication of $X^\top$ and $(XRR^\top X^\top/n + \lambda I_n)^{-1}Y$ requires $O(np)$ flops, so the total flops is $O(n^3 + nd(\log p + n))$. Dual sketching can reduce the computation cost only when $p > n$.