Journal of Machine Learning Research 18 (2018) 1-50 Submitted 6/17; Published 5/18

Sketched Ridge Regression: Optimization Perspective, Statistical Perspective, and Model Averaging

Shusen Wang WSSATZJU@GMAIL.COM
International Computer Science Institute and Department of Statistics
University of California at Berkeley
Berkeley, CA 94720, USA

Alex Gittens GITTEA@RPI.EDU
Computer Science Department
Rensselaer Polytechnic Institute
Troy, NY 12180, USA

Michael W. Mahoney MMAHONEY@STAT.BERKELEY.EDU
International Computer Science Institute and Department of Statistics
University of California at Berkeley
Berkeley, CA 94720, USA

Editor: Mehryar Mohri

Abstract

We address the statistical and optimization impacts of the classical sketch and Hessian sketch used to approximately solve the Matrix Ridge Regression (MRR) problem. Prior research has quantified the effects of classical sketch on the strictly simpler least squares regression (LSR) problem. We establish that classical sketch has a similar effect upon the optimization properties of MRR as it does on those of LSR: namely, it recovers nearly optimal solutions. By contrast, Hessian sketch does not have this guarantee; instead, the approximation error is governed by a subtle interplay between the "mass" in the responses and the optimal objective value.

For both types of approximation, the regularization in the sketched MRR problem results in significantly different statistical properties from those of the sketched LSR problem. In particular, there is a bias-variance trade-off in sketched MRR that is not present in sketched LSR. We provide upper and lower bounds on the bias and variance of sketched MRR; these bounds show that classical sketch significantly increases the variance, while Hessian sketch significantly increases the bias. Empirically, sketched MRR solutions can have risks that are higher by an order-of-magnitude than those of the optimal MRR solutions.

We establish theoretically and empirically that model averaging greatly decreases the gap between the risks of the true and sketched solutions to the MRR problem. Thus, in parallel or distributed settings, sketching combined with model averaging is a powerful technique that quickly obtains near-optimal solutions to the MRR problem while greatly mitigating the increased statistical risk incurred by sketching.

Keywords: Randomized Linear Algebra, Matrix Sketching, Ridge Regression

1. Introduction

Regression is one of the most fundamental problems in machine learning. The simplest and most thoroughly studied regression model is least squares regression (LSR). Given features $\mathbf{X} = [\mathbf{x}_1^T; \ldots, \mathbf{x}_n^T] \in \mathbb{R}^{n \times d}$ and responses $\mathbf{y} = [y_1, \ldots, y_n]^T \in \mathbb{R}^n$, the LSR problem $\min_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2$ can be solved in $\mathcal{O}(nd^2)$ time using the QR decomposition or in $\mathcal{O}(ndt)$ time using accelerated gradient descent algorithms. Here, t is the number of iterations, which depends on the initialization, the condition number of $\mathbf{X}^T\mathbf{X}$, and the stopping criterion.

This paper considers the $n \gg d$ problem, where there is much redundancy in $\mathbf{X}$. Matrix sketching, as used in the paradigm of Randomized Linear Algebra (RLA) (Mahoney, 2011; Woodruff, 2014; Drineas and Mahoney, 2016), aims to reduce the size of $\mathbf{X}$ while limiting information loss; the sketching operation can consist of sampling a subset of the rows of $\mathbf{X}$, or forming linear combinations of the rows of $\mathbf{X}$. Either operation is modeled mathematically by multiplication with a sketching matrix $\mathbf{S}$ to form the sketch $\mathbf{S}^T\mathbf{X}$. The sketching matrix $\mathbf{S} \in \mathbb{R}^{n \times s}$ satisfies $d < s \ll n$ so that $\mathbf{S}^T\mathbf{X}$ generically has the same rank but much fewer rows as $\mathbf{X}$. Sketching has been used to speed up LSR (Drineas et al., 2006b, 2011; Clarkson and Woodruff, 2013; Meng and Mahoney, 2013; Nelson and Nguyên, 2013) by solving the sketched LSR problem $\min_{\mathbf{w}} \|\mathbf{S}^T\mathbf{X}\mathbf{w} - \mathbf{S}^T\mathbf{y}\|_2^2$ instead of the original LSR problem. Solving sketched LSR costs either $\mathcal{O}(sd^2 + T_s)$ time using the QR decomposition or $\mathcal{O}(sdt + T_s)$ time using accelerated gradient descent algorithms, where t is as defined previously[1] and T_s is the time cost of sketching. For example, $T_s = \mathcal{O}(nd \log s)$ when $\mathbf{S}$ is the subsampled randomized Hadamard transform (Drineas et al., 2011), and $T_s = \mathcal{O}(nd)$ when $\mathbf{S}$ is a CountSketch matrix (Clarkson and Woodruff, 2013).

There has been much work in RLA on analyzing the quality of sketched LSR with different sketching methods and different objectives; see the reviews (Mahoney, 2011; Woodruff, 2014; Drineas and Mahoney, 2016) and the references therein. The concept of sketched LSR originated in the theoretical computer science literature, e.g., Drineas et al. (2006b, 2011), where the behavior of sketched LSR was first studied from an optimization perspective. Let $\mathbf{w}^\star$ be the optimal LSR solution and $\tilde{\mathbf{w}}$ be the solution to sketched LSR. This line of work established that if $s = \mathcal{O}(d/\epsilon + \text{poly}(d))$, then the objective value $\|\mathbf{X}\tilde{\mathbf{w}} - \mathbf{y}\|_2^2$ is at most $(1+\epsilon)$ times greater than $\|\mathbf{X}\mathbf{w}^\star - \mathbf{y}\|_2^2$. These works also bounded $\|\tilde{\mathbf{w}} - \mathbf{w}^\star\|_2^2$ in terms of the difference in the objective function values at $\tilde{\mathbf{w}}$ and $\mathbf{w}^\star$ and the condition number of $\mathbf{X}^T\mathbf{X}$.

A more recent line of work has studied sketched LSR from a statistical perspective: Ma et al. (2015); Raskutti and Mahoney (2016); Pilanci and Wainwright (2015); Wang et al. (2017c) considered statistical properties of sketched LSR such as the bias and variance. In particular, Pilanci and Wainwright (2015) showed that the solutions to sketched LSR have much higher variance than the optimal solutions.

Both of these perspectives are important and of practical interest. The optimization perspective is relevant when the approximate solution is used to initialize an (expensive) iterative optimization algorithm; the statistical perspective is relevant in machine learning and statistics applications where the approximate solution is directly used in lieu of the optimal solution.

1. The condition number of $\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X}$ is very close to that of $\mathbf{X}^T\mathbf{X}$, and thus the number of iterations t is almost unchanged.

In practice, regularized regression, e.g., ridge regression and LASSO, exhibit more attractive bias-variance trade-offs and generalization errors than vanilla LSR. Furthermore, the matrix generalization of LSR, where multiple responses are to be predicted, is often more useful than LSR. However, the properties of sketched regularized matrix regression are largely unknown. Hence, we consider the question: *how does our understanding of the optimization and statistical properties of sketched LSR generalize to sketched regularized regression problems?* We answer this question for the sketched matrix ridge regression (MRR) problem.

Recall that $\mathbf{X}$ is $n \times d$. Let $\mathbf{Y} \in \mathbb{R}^{n \times m}$ denote a matrix of corresponding responses. We study the MRR problem

$$\min_{\mathbf{W}} \left\{ f(\mathbf{W}) \triangleq \frac{1}{n} \left\| \mathbf{X}\mathbf{W} - \mathbf{Y} \right\|_F^2 + \gamma \|\mathbf{W}\|_F^2 \right\}, \tag{1}$$

which has optimal solution

$$\mathbf{W}^\star = (\mathbf{X}^T\mathbf{X} + n\gamma \mathbf{I}_d)^\dagger \mathbf{X}^T\mathbf{Y}. \tag{2}$$

Here, $(\cdot)^\dagger$ denotes the Moore-Penrose inversion operation. LSR is a special case of MRR, with $m = 1$ and $\gamma = 0$. The optimal solution $\mathbf{W}^\star$ can be obtained in $\mathcal{O}(nd^2 + nmd)$ time using a QR decomposition of $\mathbf{X}$. Sketching can be applied to MRR in two ways:

$$\mathbf{W}^{\mathrm{c}} = (\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X} + n\gamma \mathbf{I}_d)^\dagger (\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{Y}), \tag{3}$$

$$\mathbf{W}^{\mathrm{h}} = (\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X} + n\gamma \mathbf{I}_d)^\dagger \mathbf{X}^T\mathbf{Y}. \tag{4}$$

Following the convention of Pilanci and Wainwright (2015); Wang et al. (2017a), we call $\mathbf{W}^{\mathrm{c}}$ the **classical sketch** and $\mathbf{W}^{\mathrm{h}}$ the **Hessian sketch**. Table 1 lists the time costs of the three solutions to MRR.

Table 1: The time cost of the solutions to MRR. Here $T_s(\mathbf{X})$ and $T_s(\mathbf{Y})$ denote the time cost of forming the sketches $\mathbf{S}^T\mathbf{X} \in \mathbb{R}^{s \times d}$ and $\mathbf{S}^T\mathbf{Y} \in \mathbb{R}^{s \times m}$.

Solution	**Definition**	**Time Complexity**
Optimal Solution	(2)	$\mathcal{O}(nd^2 + nmd)$
Classical Sketch	(3)	$\mathcal{O}(sd^2 + smd) + T_s(\mathbf{X}) + T_s(\mathbf{Y})$
Hessian Sketch	(4)	$\mathcal{O}(sd^2 + nmd) + T_s(\mathbf{X})$

1.1 Main Results and Contributions

We summarize all of our upper bounds in Table 2. Our optimization analysis bounds the gap between the objective function values at the sketched and optimal solutions, while our statistical analysis quantifies the behavior of the bias and variance of the sketched solutions relative to those of the true solutions.

We first study classical and Hessian sketches from the **optimization perspective**. Theorems 1 and 2 show:

Table 2: A summary of our main results. In the table, $\mathbf{W}$ is the solution of classical/Hessian sketch with or without model averaging (mod. avg.); $\mathbf{W}^\star$ is the optimal solution; g is the number of models used in model averaging; and $\beta = \frac{\|\mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2 + n\gamma} \leq 1$, where γ is the regularization parameter. For conciseness, we take the sketching matrix $\mathbf{S} \in \mathbb{R}^{n \times s}$ to correspond to Gaussian projection, SRHT, or shrinkage leverage score sampling. Similar but more complex expressions hold for uniform sampling (with or without model averaging) and CountSketch (only without model averaging.) All the bounds hold with constant probability. The notation $\tilde{\mathcal{O}}$ conceals logarithmic factors.

	Classical Sketch		**Hessian Sketch**	
	w/o mod. avg.	**w/ mod. avg.**	**w/o mod. avg.**	**w/ mod. avg.**
$s =$	$\tilde{\mathcal{O}}(d/\epsilon)$		$\tilde{\mathcal{O}}(d/\epsilon)$	
$f(\mathbf{W}) - f(\mathbf{W}^\star) \leq$	$\beta\epsilon f(\mathbf{W}^\star)$	$\beta(\frac{\epsilon}{g} + \beta^2\epsilon^2) f(\mathbf{W}^\star)$	$\beta^2\epsilon\big[\frac{\|\mathbf{Y}\|_F^2}{n} - f(\mathbf{W}^\star)\big]$	$\beta^2(\frac{\epsilon}{g} + \epsilon^2)\big[\frac{\|\mathbf{Y}\|_F^2}{n} - f(\mathbf{W}^\star)\big]$
Theorems	Theorem 1	Theorem 7	Theorem 2	Theorem 8
$s =$	$\tilde{\mathcal{O}}(d/\epsilon^2)$		$\tilde{\mathcal{O}}(d/\epsilon^2)$	
$\frac{\mathsf{bias}(\mathbf{W})}{\mathsf{bias}(\mathbf{W}^\star)} \leq$	$1 + \epsilon$	$1 + \epsilon$	$(1+\epsilon)(1 + \frac{\epsilon\|\mathbf{X}\|_2^2}{n\gamma})$	$1 + \epsilon + \big(\frac{\epsilon}{\sqrt{g}} + \epsilon^2\big)\frac{\|\mathbf{X}\|_2^2}{n\gamma}$
$\frac{\mathsf{var}(\mathbf{W})}{\mathsf{var}(\mathbf{W}^\star)} \leq$	$(1+\epsilon)\frac{n}{s}$	$\frac{n}{s}\Big(\sqrt{\frac{1+\epsilon/g}{g}} + \epsilon\Big)^2$	$1 + \epsilon$	$1 + \epsilon$
Theorems	Theorem 5	Theorem 10	Theorem 6	Theorem 11

- Classical sketch achieves relative error in the objective value. With sketch size $s = \tilde{\mathcal{O}}(d/\epsilon)$, the sketched solution satisfies $f(\mathbf{W}^{\mathrm{c}}) \leq (1+\epsilon) f(\mathbf{W}^\star)$.
- Hessian sketch does not achieve relative error in the objective value. In particular, if $\frac{1}{n}\|\mathbf{Y}\|_F^2$ is much larger than $f(\mathbf{W}^\star)$, then $f(\mathbf{W}^{\mathrm{h}})$ can be far larger than $f(\mathbf{W}^\star)$.
- For both classical and Hessian sketch, the relative quality of approximation often improves as the regularization parameter γ increases (because β decreases).

We then study classical and Hessian sketch from the **statistical perspective**, by modeling $\mathbf{Y} = \mathbf{X}\mathbf{W}_0 + \mathbf{\Xi}$ as the sum of a true linear model and random noise, decomposing the risk $R(\mathbf{W}) = \mathbb{E}\|\mathbf{X}\mathbf{W} - \mathbf{X}\mathbf{W}_0\|_F^2$ into bias and variance terms, and bounding these terms. We draw the following conclusions (see Theorems 4, 5, 6 for the details):

- The bias of classical sketch can be nearly as small as that of the optimal solution. The variance is $\Theta\big(\frac{n}{s}\big)$ times that of the optimal solution; this bound is optimal. Therefore over-regularization[2] should be used to supress the variance. (As γ increases, the bias increases, and the variance decreases.)

2. For example, using a larger value of the regularization parameter γ than one would optimally choose for the unsketched problem.

- Since Hessian sketch uses the whole of $\mathbf{Y}$, the variance of Hessian sketch can be close to that of the optimal solution. However, Hessian sketch incurs a high bias, especially when $n\gamma$ is small compared to $\|\mathbf{X}\|_2^2$. This indicates that over-regularization is necessary for Hessian sketch to deliver solutions with low bias.

Our empirical evaluations bear out these theoretical results. In particular, in Section 4, we show in Figure 3 that even when the regularization parameter γ is fine-tuned, the risks of classical and Hessian sketch are worse than that of the optimal solution by an order of magnitude. This is an empirical demonstration of the fact that the near-optimal properties of sketch from the optimization perspective are much less relevant in a statistical setting than its sub-optimal statistical properties.

We propose to use **model averaging**, which averages the solutions of g sketched MRR problems, to attain lower optimization and statistical errors. Without ambiguity, we denote model-averaged classical and Hessian sketches by $\mathbf{W}^{\mathrm{c}}$ and $\mathbf{W}^{\mathrm{h}}$, respectively. Theorems 7, 8, 10, 11 establish the following results:

- Classical Sketch. Model averaging decreases the objective function value and the variance and does not increase the bias. Specifically, with the same sketch size s, model averaging ensures $\frac{f(\mathbf{W}^{\mathrm{c}})-f(\mathbf{W}^\star)}{f(\mathbf{W}^\star)}$ and $\frac{\mathsf{var}(\mathbf{W}^{\mathrm{c}})}{\mathsf{var}(\mathbf{W}^\star)}$ respectively decrease to almost $\frac{1}{g}$ of those of classical sketch without model averaging, provided that $s \gg d$. See Table 2 for the details.

- Hessian Sketch. Model averaging decreases the objective function value and the bias and does not increase the variance.

In the distributed setting, the feature-response pairs $(\mathbf{x}_1, \mathbf{y}_1), \cdots, (\mathbf{x}_n, \mathbf{y}_n) \in \mathbb{R}^d \times \mathbb{R}^m$ are divided among g machines. Assuming that the data have been shuffled randomly, each machine contains a sketch of the MRR constructed by uniformly sampling rows from the data set without replacement. We illustrate this procedure in Figure 1. In this setting, the model averaging procedure communicates the g local models only once to return the final estimate; this process has very low communication and latency costs, and suggests two further applications of classical sketch with model averaging:

- Model Averaging for Machine Learning. When a low-precision solution is acceptable, model averaging can be used in lieu of distributed numerical optimization algorithms requiring multiple rounds of communication. If $\frac{n}{g}$ is large enough compared to d and the row coherence of $\mathbf{X}$ is small, then "one-shot" model averaging has bias and variance comparable to the optimal solution.

- Model Averaging for Optimization. If a high-precision solution to MRR is required, then an iterative numerical optimization algorithm must be used. The cost of such algorithms heavily depends on the quality of the initialization.[3] A good initialization reduces the number of iterations needed to reach convergence. The averaged model

3. For example, the conjugate gradient method satisfies $\frac{\|\mathbf{W}^{(t)}-\mathbf{W}^\star\|_F^2}{\|\mathbf{W}^{(0)}-\mathbf{W}^\star\|_F^2} \leq \theta_1^t$ and stochastic block coordinate descent (Tu et al., 2016) satisfies $\frac{\mathbb{E}f(\mathbf{W}^{(t)})-f(\mathbf{W}^\star)}{f(\mathbf{W}^{(0)})-f(\mathbf{W}^\star)} \leq \theta_2^t$. Here $\mathbf{W}^{(t)}$ is the output of the t-th iteration; $\theta_1, \theta_2 \in (0,1)$ depend on the condition number of $\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d$ and some other factors.

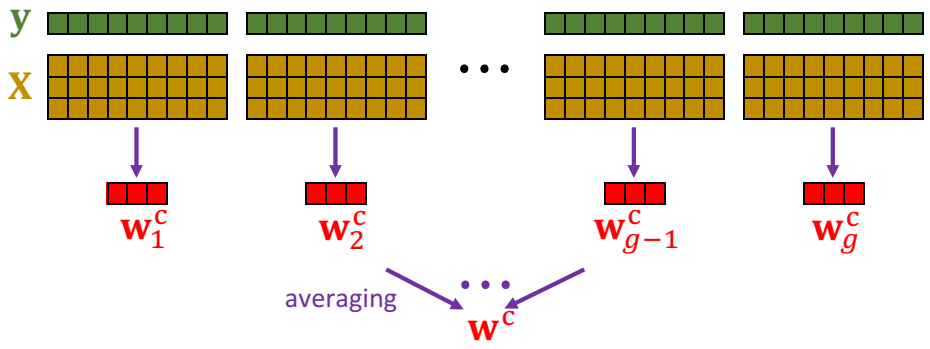


Figure 1: Using model averaging with the classical sketch in the distributed setting to approximately solve LSR.

is provably close to the optimal solution, so model averaging provides a high-quality initialization for more expensive algorithms.

1.2 Prior Work

The body of work on sketched LSR mentioned earlier (Drineas et al., 2006b, 2011; Clarkson and Woodruff, 2013; Meng and Mahoney, 2013; Nelson and Nguyên, 2013) shares many similarities with our results. However, the theories of sketched LSR developed from the optimization perspective do not obviously extend to MRR, and the statistical analysis of LSR and MRR differ: among other differences, LSR is unbiased while MRR is biased and therefore has a bias-variance tradeoff that must be considered.

Lu et al. (2013) has considered a different application of sketching to ridge regression: they assume $d \gg n$, reduce the number of features in $\mathbf{X}$ using sketching, and conduct statistical analysis. Our setting differs in that we consider $n \gg d$, reduce the number of samples by sketching, and allow for multiple responses.

The model averaging analyzed in this paper is similar in spirit to the AvgM algorithm of (Zhang et al., 2013). When classical sketch is used with uniform row sampling without replacement, our model averaging procedure is a special case of AvgM. However, our results do not follow from those of (Zhang et al., 2013). First, we make no assumption on the data, $\mathbf{X}$ and $\mathbf{Y}$, and the model (parameters), $\mathbf{W}$. Second, we study both the optimization objective, $\|\mathbf{X}\mathbf{W}^c - \mathbf{X}\mathbf{W}^\star\|_F^2$, and the statistical objective, $\mathbb{E}\|\mathbf{X}\mathbf{W}^c - \mathbf{X}\mathbf{W}_0\|_F^2$, where $\mathbf{W}^c$ is the average of the approximate solutions obtained used classical sketch, $\mathbf{W}_0$ is the unknown ground truth, and $\mathbf{W}^\star$ is the optimal solution based on the observed data; they studied solely the optimization objective. Third, our results apply to many other sketching ensembles than uniform sampling without replacement. Our results clearly indicate that the performance critically depends on the row coherence of $\mathbf{X}$; this dependence has not been explicitly captured in (Zhang et al., 2013). Zhang et al. (2015) studied a different statistical objective and their resulting bound has a higher-order of dependence on d and other parameters.

Iterative Hessian sketch has been studied in Pilanci and Wainwright (2015); Wang et al. (2017a,b). By way of comparison, all the algorithms in this paper are "one-shot" rather than iterative. This work has connections to the contemporary works (Avron et al., 2017; Thanei et al., 2017; Derezinski and Warmuth, 2017, 2018). Avron et al. (2017) studied classical sketch from the optimization perspective; Thanei et al. (2017) studied LSR with model averaging; Derezinski and Warmuth (2017, 2018) studied linear regression with volume sampling for experimental design.

1.3 Paper Organization

Section 2 defines our notation and introduces the sketching schemes we consider. Section 3 presents our theoretical results. Sections 4 and 5 conduct experiments to verify our theories and demonstrates the efficacy of model averaging. Section 6 sketches the proofs of our main results. Complete proofs are provided in the appendix.

2. Preliminaries

Throughout, we take $\mathbf{I}_n$ to be the $n \times n$ identity matrix and $\mathbf{0}$ to be a vector or matrix of all zeroes of the appropriate size. Given a matrix $\mathbf{A} = [a_{ij}]$, the i-th row is denoted by $\mathbf{a}_{i:}$, and the j-th column is denoted by $\mathbf{a}_{:j}$. The Frobenius and spectral norms of $\mathbf{A}$ are written as, respectively, $\|\mathbf{A}\|_F$ and $\|\mathbf{A}\|_2$. The set $\{1, 2, \cdots, n\}$ is written $[n]$. Let $\mathcal{O}$, Ω, and Θ be the standard asymptotic notation, and let $\tilde{\mathcal{O}}$ conceal logarithmic factors.

Throughout, we fix $\mathbf{X} \in \mathbb{R}^{n\times d}$ as our matrix of features. We set $\rho = \text{rank}(\mathbf{X})$ and write the SVD of $\mathbf{X}$ as $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{U}$, $\mathbf{\Sigma}$, $\mathbf{V}$ are respectively $n \times \rho$, $\rho \times \rho$, and $d \times \rho$ matrices. We let $\sigma_1 \geq \cdots \geq \sigma_\rho > 0$ be the singular values of $\mathbf{X}$. The Moore-Penrose inverse of $\mathbf{X}$ is defined by $\mathbf{X}^\dagger = \mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T$. The row leverage scores of $\mathbf{X}$ are $l_i = \|\mathbf{u}_{:i}\|_2^2$ for $i \in [n]$. The row coherence of $\mathbf{X}$ is $\mu(\mathbf{X}) = \frac{n}{\rho}\max_i \|\mathbf{u}_{:i}\|_2^2$. Throughout, we let μ be shorthand for $\mu(\mathbf{X})$. The notation defined in Table 3 is used throughout this paper.

Matrix sketching attempts to reduce the size of large matrices while minimizing the loss of spectral information that is useful in tasks like linear regression. We denote the process of sketching a matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ by $\mathbf{X}' = \mathbf{S}^T\mathbf{X}$. Here, $\mathbf{S} \in \mathbb{R}^{n\times s}$ is called a sketching matrix and $\mathbf{X}' \in \mathbb{R}^{s\times d}$ is called a sketch of $\mathbf{X}$. In practice, except for Gaussian projection (where the entries of $\mathbf{S}$ are i.i.d. sampled from $\mathcal{N}(0, 1/s)$), the sketching matrix $\mathbf{S}$ is not formed explicitly.

Matrix sketching can be accomplished by random sampling or random projection. **Random sampling** corresponds to sampling rows of $\mathbf{X}$ i.i.d. with replacement according to given row sampling probabilities $p_1, \cdots, p_m \in (0, 1)$. The corresponding (random) sketching matrix $\mathbf{S} \in \mathbb{R}^{n\times s}$ has exactly one non-zero entry, whose position indicates the index of the selected row in each column; in practice, this $\mathbf{S}$ is not explicitly formed. **Uniform sampling** fixes $p_1 = \cdots = p_n = \frac{1}{n}$. **Leverage score sampling** sets p_i proportional to the (exact or approximate (Drineas et al., 2012)) row leverage scores l_i of $\mathbf{X}$. In practice **shrinked leverage score sampling** can be a better choice than leverage score sampling (Ma et al.,

Table 3: The commonly used notation.

Notation	Definition
$\mathbf{X} \in \mathbb{R}^{n\times d}$	each row is a data sample (feature vector)
$\mathbf{Y} \in \mathbb{R}^{n\times m}$	each row contains the corresponding responses
$\mathbf{U\Sigma V}^T$	the SVD of $\mathbf{X}$
ρ	the rank of $\mathbf{X}$
μ	the row coherence of $\mathbf{X}$
σ_i	the i-th largest singular value of $\mathbf{X}$
γ	the regularization parameter
β	$\beta = \frac{\|\mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2+n\gamma} \leq 1$
$\mathbf{S} \in \mathbb{R}^{n\times s}$	the sketching matrix
$\mathbf{W}^\star \in \mathbb{R}^{d\times m}$	the optimal solution (2)
$\mathbf{W}^{\mathrm{c}} \in \mathbb{R}^{d\times m}$	approximate solution obtained using the classical sketch (3)
$\mathbf{W}^{\mathrm{h}} \in \mathbb{R}^{d\times m}$	approximate solution obtained using the Hessian sketch (4)
$\mathbf{W}_0 \in \mathbb{R}^{d\times m}$	the unknown ground truth (in the statistical setting)

2015). The sampling probabilities of shrinked leverage score sampling are defined by $p_i = \frac{1}{2}\big(\frac{l_i}{\sum_{j=1}^n l_j} + \frac{1}{n}\big)$.[4]

The exact leverage scores are unnecessary in practice; constant-factor approximation to the leverage scores is sufficient. Leverage scores can be efficiently approximated by the algorithms of (Drineas et al., 2012). Let $l_1, \cdots, l_n$ be the true leverage scores. We denote the approximate leverages by $\widetilde{l}_1, \cdots, \widetilde{l}_n$ and require that they satisfy

$$\tilde{l}_q \in [l_q, \tau l_q] \quad \text{for all } \ q \in [n], \tag{5}$$

where $\tau \geq 1$ indicates the quality of approximation. We then use $p_q = \tilde{l}_q / \sum_j \tilde{l}_j$ as the sampling probabilities. One can obtain the same accuracies when using approximate leverage scores in place of the true leverage scores by increasing s by a factor of τ, so as long as τ is a small constant, the orders of the sketch sizes when using exact or approximate leverage score sampling are the same. Thus we do not distinguish between exact and approximate leverage scores in this paper. For shrinked leverage score sampling, we define the sampling probabilities

$$p_i = \frac{1}{2}\left(\frac{\tilde{l}_i}{\sum_{j=1}^n \tilde{l}_j} + \frac{1}{n}\right) \quad \text{for } \ i = 1, \ldots, n. \tag{6}$$

Gaussian projection is also well-known as the prototypical Johnson-Lindenstrauss transform (Johnson and Lindenstrauss, 1984). Let $\mathbf{G} \in \mathbb{R}^{n\times s}$ be a standard Gaussian matrix, i.e., each entry is sampled independently from $\mathcal{N}(0,1)$. The matrix $\mathbf{S} = \frac{1}{\sqrt{s}}\mathbf{G}$ is a Gaussian projection matrix. It takes $\mathcal{O}(nds)$ time to apply $\mathbf{S} \in \mathbb{R}^{n\times s}$ to any $n \times d$ dense matrix, which makes Gaussian projection computationally inefficient relative to other forms of sketching.

4. In fact, p_i can be any convex combination of $\frac{l_i}{\sum_{j=1}^n l_j}$ and $\frac{1}{n}$ (Ma et al., 2015). We use the weight $\frac{1}{2}$ for convenience; our conclusions extend in a straightforward manner to other weightings.

The **Subsampled randomized Hadamard transform (SRHT)** (Drineas et al., 2011; Lu et al., 2013; Tropp, 2011) is a more efficient alternative to Gaussian projection. Let $\mathbf{H}_n \in \mathbb{R}^{n\times n}$ be the Walsh-Hadamard matrix with $+1$ and -1 entries, $\mathbf{D} \in \mathbb{R}^{n\times n}$ be a diagonal matrix with diagonal entries sampled uniformly from $\{+1, -1\}$, and $\mathbf{P} \in \mathbb{R}^{n\times s}$ be the uniform row sampling matrix defined above. The matrix $\mathbf{S} = \frac{1}{\sqrt{n}}\mathbf{D}\mathbf{H}_n\mathbf{P} \in \mathbb{R}^{n\times s}$ is an SRHT matrix, and can be applied to any $n \times d$ matrix in $\mathcal{O}(nd \log s)$ time. In practice, the subsampled randomized Fourier transform (SRFT) (Woolfe et al., 2008) is often used in lieu of the SRHT, because the SRFT exists for all values of n, whereas $\mathbf{H}_n$ exists only for some values of n. Their performance and theoretical analyses are very similar.

CountSketch can be applied to any $\mathbf{X} \in \mathbb{R}^{n\times d}$ in $\mathcal{O}(nd)$ time (Charikar et al., 2004; Clarkson and Woodruff, 2013; Meng and Mahoney, 2013; Nelson and Nguyên, 2013; Pham and Pagh, 2013; Weinberger et al., 2009). Though more efficient to apply, CountSketch requires a larger sketch size than Gaussian projections, SRHT, and leverage score sampling to attain the same theoretical guarantees. Interested readers can refer to (Woodruff, 2014) for a detailed description of CountSketch. Unlike the other sketching methods mentioned here, model averaging with CountSketch may not be theoretically sound. See Remark 5 for further discussion.

3. Main Results

Sections 3.1 and 3.2 analyze sketched MRR from, respectively, the optimization and statistical perspectives. Sections 3.3 and 3.4 capture the impacts of model averaging on, respectively, the optimization and statistical properties of sketched MRR.

We described six sketching methods in Section 2. For simplicity, in this section, we refer to leverage score sampling, shrinked leverage score sampling, Gaussian projection, and SRHT as **the four sketching methods** while we refer to uniform sampling and CountSketch by name. Throughout, let μ be the row coherence of $\mathbf{X}$ and $\beta = \frac{\|\mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2 + n\gamma} \leq 1$.

3.1 Sketched MRR: Optimization Perspective

Theorem 1 shows that $f(\mathbf{W}^{\mathrm{c}})$, the objective value of classical sketch, is close to the optimal objective value $f(\mathbf{W}^\star)$, and that the approximation quality improves as the regularization parameter γ increases.

Theorem 1 (Classical Sketch) *Let $\beta = \frac{\|\mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2 + n\gamma} \leq 1$. For the four sketching methods with $s = \tilde{\mathcal{O}}\big(\frac{d}{\epsilon}\big)$, uniform sampling with $s = \mathcal{O}\big(\frac{\mu d \log d}{\epsilon}\big)$, and CountSketch with $s = \mathcal{O}\big(\frac{d^2}{\epsilon}\big)$, the inequality*

$$f(\mathbf{W}^c) - f(\mathbf{W}^\star) \quad \leq \quad \epsilon\beta\, f(\mathbf{W}^\star)$$

holds with probability at least 0.9. The uncertainty is with respect to the random choice of sketching matrix.

The corresponding guarantee for the performance of Hessian sketch is given in Theorem 2. It is weaker than the guarantee for classical sketch, especially when $\frac{1}{n}\|\mathbf{Y}\|_F^2$ is far larger than $f(\mathbf{W}^\star)$. If $\mathbf{Y}$ is nearly noiseless—$\mathbf{Y}$ is well-explained by a linear combination

of the columns of $\mathbf{X}$—and γ is small, then $f(\mathbf{W}^\star)$ is close to zero, and consequently $f(\mathbf{W}^\star)$ can be far smaller than $\frac{1}{n}\|\mathbf{Y}\|_F^2$. Therefore, in this case which is ideal for MRR, $f(\mathbf{W}^{\mathrm{h}})$ is not close to $f(\mathbf{W}^\star)$ and our theory suggests Hessian sketch does not perform as well as classical sketch. This is verified by our experiments (see Figure 2), which show that unless γ is large or a large portion of $\mathbf{Y}$ is outside the column space of $\mathbf{X}$, the ratio $\frac{f(\mathbf{W}^{\mathrm{h}})}{f(\mathbf{W}^\star)}$ can be large.

Theorem 2 (Hessian Sketch) *Let $\beta = \frac{\|\mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2+n\gamma} \leq 1$. For the four sketching methods with $s = \tilde{\mathcal{O}}\big(\frac{d}{\epsilon}\big)$, uniform sampling with $s = \mathcal{O}\big(\frac{\mu d \log d}{\epsilon}\big)$, and CountSketch with $s = \mathcal{O}(\frac{d^2}{\epsilon})$, the inequality*

$$f(\mathbf{W}^h) - f(\mathbf{W}^\star) \;\; \leq \;\; \epsilon\beta^2 \left(\tfrac{\|\mathbf{Y}\|_F^2}{n} - f(\mathbf{W}^\star)\right).$$

holds with probability at least 0.9. The uncertainty is with respect to the random choice of sketching matrix.

These two results imply that $f(\mathbf{W}^{\mathrm{c}})$ and $f(\mathbf{W}^{\mathrm{h}})$ can be close to $f(\mathbf{W}^\star)$. When this is the case, curvature of the objective function ensures that the sketched solutions $\mathbf{W}^{\mathrm{c}}$ and $\mathbf{W}^{\mathrm{h}}$ are close to the optimal solution $\mathbf{W}^\star$. Lemma 3 bounds the Mahalanobis distance $\|\mathbf{M}(\mathbf{W} - \mathbf{W}^\star)\|_F^2$. Here $\mathbf{M}$ is any non-singular matrix; in particular, it can be the identity matrix or $(\mathbf{X}^T\mathbf{X})^{1/2}$. Lemma 3 is a consequence of Lemma 25.

Lemma 3 (Mahalanobis Distance) *Let f be the objective function of MRR defined in (1), $\mathbf{W} \in \mathbb{R}^{d\times m}$ be arbitrary, and $\mathbf{W}^\star$ be the optimal solution defined in (2). For any non-singular matrix $\mathbf{M}$, the Mahalanobis distance satisfies*

$$\frac{1}{n}\big\|\mathbf{M}(\mathbf{W} - \mathbf{W}^\star)\big\|_F^2 \;\leq\; \frac{f(\mathbf{W}) - f(\mathbf{W}^\star)}{\sigma_{\min}^2\big[(\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{1/2}\mathbf{M}^{-1}\big]}.$$

By choosing $\mathbf{M} = (\mathbf{X}^T\mathbf{X})^{1/2}$, we can bound $\frac{1}{n}\|\mathbf{X}\mathbf{W} - \mathbf{X}\mathbf{W}^\star\|_F^2$ in terms of the difference in the objective values:

$$\tfrac{1}{n}\big\|\mathbf{X}\mathbf{W} - \mathbf{X}\mathbf{W}^\star\big\|_F^2 \;\; \leq \;\; \beta\big[f(\mathbf{W}) - f(\mathbf{W}^\star)\big],$$

where $\beta = \frac{\|\mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2+n\gamma} \leq 1$. With Lemma 3, we can directly apply Theorems 1 or 2 to bound $\frac{1}{n}\|\mathbf{X}\mathbf{W}^{\mathrm{c}} - \mathbf{X}\mathbf{W}^\star\|_F^2$ or $\frac{1}{n}\|\mathbf{X}\mathbf{W}^{\mathrm{h}} - \mathbf{X}\mathbf{W}^\star\|_F^2$.

3.2 Sketched MRR: Statistical Perspective

We consider the following fixed design model. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be the observed feature matrix, $\mathbf{W}_0 \in \mathbb{R}^{d\times m}$ be the true and unknown model, $\mathbf{\Xi} \in \mathbb{R}^{n\times m}$ contain unknown random noise, and

$$\mathbf{Y} \;=\; \mathbf{X}\mathbf{W}_0 + \mathbf{\Xi} \tag{7}$$

be the observed responses. We make the following standard weak assumptions on the noise:

$$\mathbb{E}[\mathbf{\Xi}] = \mathbf{0} \quad \text{and} \quad \mathbb{E}[\mathbf{\Xi}\mathbf{\Xi}^T] = \xi^2\mathbf{I}_n.$$

We observe $\mathbf{X}$ and $\mathbf{Y}$ and seek to estimate $\mathbf{W}_0$.

We can evaluate the quality of the estimate by the risk:

$$R(\mathbf{W}) \;=\; \tfrac{1}{n}\mathbb{E}\big\|\mathbf{X}\mathbf{W} - \mathbf{X}\mathbf{W}_0\big\|_F^2, \tag{8}$$

where the expectation is taken w.r.t. the noise $\mathbf{\Xi}$. We study the risk functions $R(\mathbf{W}^\star)$, $R(\mathbf{W}^{\mathrm{c}})$, and $R(\mathbf{W}^{\mathrm{h}})$ in the following.

Theorem 4 (Bias-Variance Decomposition) *We consider the data model described in this subsection. Let* $\mathbf{W}$ *be* $\mathbf{W}^\star$, $\mathbf{W}^c$, *or* $\mathbf{W}^h$, *as defined in* (2), (3), *or* (4), *respectively; then the risk function can be decomposed as*

$$R(\mathbf{W}) \;=\; \mathsf{bias}^2(\mathbf{W}) + \mathsf{var}(\mathbf{W}).$$

Recall the SVD of $\mathbf{X}$ *defined in Section 2:* $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. *The bias and variance terms can be written as*

$$\begin{aligned}
\mathsf{bias}(\mathbf{W}^\star) &= \gamma\sqrt{n}\Big\|(\mathbf{\Sigma}^2 + n\gamma\mathbf{I}_\rho)^{-1}\mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_0\Big\|_F,\\
\mathsf{var}(\mathbf{W}^\star) &= \tfrac{\xi^2}{n}\Big\|\big(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2}\big)^{-1}\Big\|_F^2,\\
\mathsf{bias}(\mathbf{W}^c) &= \gamma\sqrt{n}\Big\|\big(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger\mathbf{\Sigma}^{-1}\mathbf{V}^T\mathbf{W}_0\Big\|_F,\\
\mathsf{var}(\mathbf{W}^c) &= \tfrac{\xi^2}{n}\Big\|\big(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger\mathbf{U}^T\mathbf{S}\mathbf{S}^T\Big\|_F^2,\\
\mathsf{bias}(\mathbf{W}^h) &= \gamma\sqrt{n}\Big\|\Big(\mathbf{\Sigma}^{-2} + \tfrac{\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}-\mathbf{I}_\rho}{n\gamma}\Big)\big(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger\mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_0\Big\|_F,\\
\mathsf{var}(\mathbf{W}^h) &= \tfrac{\xi^2}{n}\Big\|\big(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger\Big\|_F^2.
\end{aligned}$$

The functions $\mathsf{bias}(\mathbf{W}^\star)$ and $\mathsf{var}(\mathbf{W}^\star)$ are deterministic. The randomness in $\mathsf{bias}(\mathbf{W}^{\mathrm{c}})$, $\mathsf{var}(\mathbf{W}^{\mathrm{c}})$, $\mathsf{bias}(\mathbf{W}^{\mathrm{h}})$, and $\mathsf{var}(\mathbf{W}^{\mathrm{h}})$ all arises from the sketching matrix $\mathbf{S}$.

Throughout this paper, we compare the bias and variance of classical sketch and Hessian sketch to those of the optimal solution $\mathbf{W}^\star$. We first study the bias, variance, and risk of $\mathbf{W}^\star$, which will help us understand the subsequent comparisons. We can assume that $\mathbf{\Sigma}^2 = \mathbf{V}^T\mathbf{X}^T\mathbf{X}\mathbf{V}$ is linear in n; this is reasonable because $\mathbf{X}^T\mathbf{X} = \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T$ and $\mathbf{V}$ is an orthogonal matrix.

- **Bias.** The bias of $\mathbf{W}^\star$ is independent of n and is increasing with γ. The bias is the price paid for using regularization to decrease the variance; for least squares regression, γ is zero, and the bias is zero.

- **Variance.** The variance of $\mathbf{W}^\star$ is inversely proportional to n. As n grows, the variance decreases to zero, and we must also decrease γ to ensure that the sum of the squared bias and variance decreases to zero.

- **Risk.** Note that $\mathbf{W}^\star$ is not the minimizer of $R(\cdot)$; $\mathbf{W}_0$ is the minimizer because $R(\mathbf{W}_0) = 0$. Nevertheless, because $\mathbf{W}_0$ is unknown, $\mathbf{W}^\star$ for a carefully chosen γ is a standard proxy for the exact minimizer in practice. It is thus highly interesting to compare the risk of MRR solutions obtained using sketching to to $R(\mathbf{W}^\star)$.

Theorem 5 provides upper and lower bounds on the bias and variance of solutions obtained using classical sketch. In particular, we see that that $\mathsf{bias}(\mathbf{W}^{\mathrm{c}})$ is within a factor of $(1 \pm \epsilon)$ of $\mathsf{bias}(\mathbf{W}^{\star})$. However, $\mathsf{var}(\mathbf{W}^{\mathrm{c}})$ can be $\Theta(\frac{n}{s})$ times worse than $\mathsf{var}(\mathbf{W}^{\star})$. The absolute value of $\mathsf{var}(\mathbf{W}^{\mathrm{c}})$ is inversely proportional to s, whereas the absolute value of $\mathsf{bias}(\mathbf{W}^{\mathrm{c}})$ is almost independent of s.

Theorem 5 (Classical Sketch) *Assume $s \leq n$. For Gaussian projection and SRHT sketching with $s = \tilde{\mathcal{O}}(\frac{d}{\epsilon^2})$, uniform sampling with $s = \mathcal{O}(\frac{\mu d \log d}{\epsilon^2})$, or CountSketch with $s = \mathcal{O}(\frac{d^2}{\epsilon^2})$, the inequalities*

$$\begin{aligned} 1-\epsilon &\leq \frac{\mathsf{bias}(\mathbf{W}^{c})}{\mathsf{bias}(\mathbf{W}^{\star})} \leq 1+\epsilon, \\ (1-\epsilon)\frac{n}{s} &\leq \frac{\mathsf{var}(\mathbf{W}^{c})}{\mathsf{var}(\mathbf{W}^{\star})} \leq (1+\epsilon)\frac{n}{s} \end{aligned}$$

hold with probability at least 0.9. For shrinked leverage score sampling with $s = \mathcal{O}(\frac{d \log d}{\epsilon^2})$, these inequalities, except for the lower bound on the variance, hold with probability at least 0.9. Here the randomness comes from the sketching matrix $\mathbf{S}$.

Remark 1 *To establish an upper (lower) bound on the variance, we need an upper (lower) bound on $\|\mathbf{S}\|_2^2$. There is no nontrivial upper nor lower bound on $\|\mathbf{S}\|_2^2$ for leverage score sampling, so the variance of leverage score sampling cannot be bounded. Shrinked leverage score sampling satisfies the upper bound $\|\mathbf{S}\|_2^2 \leq \frac{2n}{s}$; but $\|\mathbf{S}\|_2^2$ does not have a nontrivial lower bound, so there is no nontrivial lower bound on the variance of shrinked leverage score. Remark 4 explains the nonexistence of the relevant bounds on $\|\mathbf{S}\|_2^2$ for both variants of leverage score sampling.*

Theorem 6 establishes similar upper and lower bounds on the bias and variance of solutions obtained using Hessian sketch. The situation is the reverse of that with classical sketch: the variance of $\mathbf{W}^{\mathrm{h}}$ is close to that of $\mathbf{W}^{\star}$ if s is large enough, but as the regularization parameter γ goes to zero, $\mathsf{bias}(\mathbf{W}^{\mathrm{h}})$ becomes much larger than $\mathsf{bias}(\mathbf{W}^{\star})$. The theory suggest that Hessian sketch should be preferred over classical sketch when $\mathbf{Y}$ is very noisy, because Hessian sketch does not magnify the variance.

Theorem 6 (Hessian Sketch) *For the four sketching methods with $s = \tilde{\mathcal{O}}(\frac{d}{\epsilon^2})$, uniform sampling with $s = \mathcal{O}(\frac{\mu d \log d}{\epsilon^2})$, and CountSketch with $s = \mathcal{O}(\frac{d^2}{\epsilon^2})$, the inequalities*

$$\begin{aligned} \frac{\mathsf{bias}(\mathbf{W}^{h})}{\mathsf{bias}(\mathbf{W}^{\star})} &\leq (1+\epsilon)\left(1+\frac{\epsilon\|\mathbf{X}\|_2^2}{n\gamma}\right), \\ 1-\epsilon &\leq \frac{\mathsf{var}(\mathbf{W}^{h})}{\mathsf{var}(\mathbf{W}^{\star})} \leq 1+\epsilon \end{aligned}$$

hold with probability at least 0.9. Further assume that the ρ-th singular value of $\mathbf{X}$ satisfies $\sigma_\rho^2 \geq \frac{n\gamma}{\epsilon}$, then

$$\frac{\mathsf{bias}(\mathbf{W}^{h})}{\mathsf{bias}(\mathbf{W}^{\star})} \geq \frac{1}{1+\epsilon}\left(\frac{\epsilon\sigma_\rho^2}{n\gamma}-1\right)$$

with probability at least 0.9. Here the randomness is in the choice of sketching matrix $\mathbf{S}$.

The lower bound on the bias shows that the solution from Hessian sketch can exhibit a much higher bias than the optimal solution. The gap between $\mathsf{bias}(\mathbf{W}^{\mathrm{h}})$ and $\mathsf{bias}(\mathbf{W}^{\star})$ can be lessened by increasing the regularization parameter γ, but such over-regularization increases the baseline $\mathsf{bias}(\mathbf{W}^{\star})$ itself. It is also worth mentioning that unlike $\mathsf{bias}(\mathbf{W}^{\star})$ and $\mathsf{bias}(\mathbf{W}^{\mathrm{c}})$, $\mathsf{bias}(\mathbf{W}^{\mathrm{h}})$ is not monotonically increasing with γ, as is empirically verified in Figure 3.

In sum, our theory shows that the classical and Hessian sketches are not statistically comparable to the optimal solutions: classical sketch has too high a variance, and Hessian sketch has too high a bias for reasonable amounts of regularization. In practice, the regularization parameter γ should be tuned to optimize the prediction accuracy. Our experiments in Figure 3 show that even with carefully chosen γ, the risks of classical and Hessian sketch can be higher than the risk of the optimal solution by an order of magnitude. Formally speaking, $\min_\gamma R(\mathbf{W}^{\mathrm{c}}) \gg \min_\gamma R(\mathbf{W}^{\star})$ and $\min_\gamma R(\mathbf{W}^{\mathrm{h}}) \gg \min_\gamma R(\mathbf{W}^{\star})$ hold in practice.

Our empirical study in Figure 3 suggests classical and Hessian sketch both require over-regularization, i.e., setting γ larger than is best for the optimal solution $\mathbf{W}^{\star}$. Formally speaking, $\operatorname{argmin}_\gamma R(\mathbf{W}^{\mathrm{c}}) > \operatorname{argmin}_\gamma R(\mathbf{W}^{\star})$ and $\operatorname{argmin}_\gamma R(\mathbf{W}^{\mathrm{h}}) > \operatorname{argmin}_\gamma R(\mathbf{W}^{\star})$. Although this is the case for both types of sketching, the underlying explanations are different. Classical sketches have a high variance, so a large γ is required to supress their variance (the variance is non-increasing with γ). Hessian sketches magnify the bias when γ is small, so a reasonably large γ is necessary to lower their bias.

3.3 Model Averaging: Optimization Perspective

We consider model averaging as a method to increase the accuracy of sketched MRR solutions. The model averaging procedure is straightforward: one independently draws g sketching matrices $\mathbf{S}_1, \cdots, \mathbf{S}_g \in \mathbb{R}^{n\times s}$, uses these to form g sketched MRR solutions, denoted by $\{\mathbf{W}_i^{\mathrm{c}}\}_{i=1}^g$ or $\{\mathbf{W}_i^{\mathrm{h}}\}_{i=1}^g$, and averages these solutions to obtain the final estimate $\mathbf{W}^{\mathrm{c}} = \frac{1}{g}\sum_{i=1}^g \mathbf{W}_i^{\mathrm{c}}$ or $\mathbf{W}^{\mathrm{h}} = \frac{1}{g}\sum_{i=1}^g \mathbf{W}_i^{\mathrm{h}}$. Practical applications of model averaging are enumerated in Section 1.1.

Theorems 7 and 8 present guarantees on the optimization accuracy of using model averaging on classical/Hessian sketch solutions. We can contrast these with the guarantees provided for sketched MRR in Theorems 1 and 2. For classical sketch with model averaging, we see that when $\epsilon \le \frac{1}{g}$, the bound on $f(\mathbf{W}^{\mathrm{h}}) - f(\mathbf{W}^{\star})$ is proportional to ϵ/g. From Lemma 3 we see that the distance between $\mathbf{W}^{\mathrm{c}}$ and $\mathbf{W}^{\star}$ also decreases accordingly.

Theorem 7 (Classical Sketch with Model Averaging) *Let $\beta = \frac{\|\mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2+n\gamma} \le 1$. For the four methods, let $s = \tilde{\mathcal{O}}\big(\frac{d}{\epsilon}\big)$, and for uniform sampling, let $s = \mathcal{O}\big(\frac{\mu d \log d}{\epsilon}\big)$, then the inequality*

$$f(\mathbf{W}^{c}) - f(\mathbf{W}^{\star}) \quad \le \quad \beta\Big(\frac{\epsilon}{g} + \beta^2\epsilon^2\Big)\, f(\mathbf{W}^{\star})$$

holds with probability at least 0.8. Here the randomness comes from the choice of sketching matrices.

For Hessian sketch with model averaging, if $\epsilon < \frac{1}{g}$, then the bound on $f(\mathbf{W}^{\mathrm{h}}) - f(\mathbf{W}^{\star})$ is proportional to $\frac{\epsilon}{g}$.

Theorem 8 (Hessian Sketch with Model Averaging) *Let* $\beta = \frac{\|\mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2 + n\gamma} \leq 1$. *For the four methods let* $s = \tilde{\mathcal{O}}\big(\frac{d}{\epsilon}\big)$, *and for uniform sampling let* $s = \mathcal{O}\big(\frac{\mu d \log d}{\epsilon}\big)$, *then the inequality*

$$f(\mathbf{W}^h) - f(\mathbf{W}^\star) \quad \leq \quad \beta^2 \left(\frac{\epsilon}{g} + \epsilon^2\right) \left(\frac{\|\mathbf{Y}\|_F^2}{n} - f(\mathbf{W}^\star)\right)$$

holds with probability at least 0.8. Here the randomness comes from the choice of sketching matrices.

3.4 Model Averaging: Statistical Perspective

Model averaging has the salutatory property of reducing the risks of the classical and Hessian sketches. Our first result conducts a bias-variance decomposition for the averaged solution of the sketched MRR problem.

Theorem 9 (Bias-Variance Decomposition) *We consider the fixed design model* (7). *Decompose the risk function defined in* (8) *as*

$$R(\mathbf{W}) \quad = \quad \mathsf{bias}^2(\mathbf{W}) + \mathsf{var}(\mathbf{W}).$$

The bias and variance terms are

$$\begin{aligned}
\mathsf{bias}\big(\mathbf{W}^c\big) &= \gamma\sqrt{n}\left\|\frac{1}{g}\sum_{i=1}^{g}\big(\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger\mathbf{\Sigma}^{-1}\mathbf{V}^T\mathbf{W}_0\right\|_F, \\
\mathsf{var}\big(\mathbf{W}^c\big) &= \frac{\xi^2}{n}\left\|\frac{1}{g}\sum_{i=1}^{g}\big(\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\right\|_F^2, \\
\mathsf{bias}\big(\mathbf{W}^h\big) &= \gamma\sqrt{n}\left\|\frac{1}{g}\sum_{i=1}^{g}\big(\mathbf{\Sigma}^{-2} + \tfrac{\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} - \mathbf{I}_\rho}{n\gamma}\big)\big(\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger\mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_0\right\|_F, \\
\mathsf{var}\big(\mathbf{W}^h\big) &= \frac{\xi^2}{n}\left\|\frac{1}{g}\sum_{i=1}^{g}\big(\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger\right\|_F^2.
\end{aligned}$$

Theorems 10 and 11 provide upper bounds on the bias and variance of averaged sketched MRR solutions for, respectively, classical sketch and Hessian sketch. We can contrast them with Theorems 5 and 6 to see the statistical benefits of model averaging. Theorem 10 shows that when $g \approx \frac{n}{s}$, classical sketch with model averaging yields a solution with comparable bias and variance to the optimal solution.

Theorem 10 (Classical Sketch with Model Averaging) *For the four sketching methods with* $s = \tilde{\mathcal{O}}\big(\frac{d}{\epsilon^2}\big)$, *or uniform sampling with* $s = \mathcal{O}\big(\frac{\mu d \log d}{\epsilon^2}\big)$, *the inequalities*

$$\frac{\mathsf{bias}(\mathbf{W}^c)}{\mathsf{bias}(\mathbf{W}^\star)} \leq 1 + \epsilon \qquad \textit{and} \qquad \frac{\mathsf{var}(\mathbf{W}^c)}{\mathsf{var}(\mathbf{W}^\star)} \leq \frac{n}{s}\left(\frac{\sqrt{1+\epsilon}}{\sqrt{h}} + \epsilon\right)^2,$$

where $h = \min\{g, \Theta(\frac{n}{s})\}$, *hold with probability at least 0.8. The randomness comes from the choice of sketching matrices.*

Theorem 11 shows that model averaging decreases the bias of Hessian sketch without increasing the variance. For Hessian sketch without model averaging, recall that $\mathsf{bias}(\mathbf{W}^{\mathrm{h}})$ is larger than $\mathsf{bias}(\mathbf{W}^\star)$ by a factor of $\mathcal{O}(\|\mathbf{X}\|_2^2/(n\gamma))$. Theorem 11 shows that model averaging significantly reduces the bias.

Theorem 11 (Hessian Sketch with Model Averaging) *For the four sketching methods with $s = \tilde{\mathcal{O}}\big(\frac{d}{\epsilon^2}\big)$, or uniform sampling with $s = \mathcal{O}\big(\frac{\mu d \log d}{\epsilon^2}\big)$, the inequalities*

$$\frac{\mathsf{bias}(\mathbf{W}^h)}{\mathsf{bias}(\mathbf{W}^\star)} \;\le\; 1 + \epsilon + \Big(\frac{\epsilon}{\sqrt{g}} + \epsilon^2\Big)\frac{\|\mathbf{X}\|_2^2}{n\gamma} \qquad \textit{and} \qquad \frac{\mathsf{var}(\mathbf{W}^h)}{\mathsf{var}(\mathbf{W}^\star)} \;\le\; 1 + \epsilon$$

hold with probability at least 0.8. Here the randomness comes from the choice of sketching matrices.

4. Experiments on Synthetic Data

We conduct experiments on synthetic data to verify our theory. Section 4.1 describes the data model and experiment settings. Sections 4.2 and 4.3 empirically study classical and Hessian sketch from the optimization and statistical perspectives, respectively, to verify Theorems 1, 2, 5, and 6. Sections 4.4 and 4.5 study model averaging from the optimization and statistical perspectives, respectively, to corroborate Theorems 7, 8, 10, and 11.

4.1 Settings

Following (Ma et al., 2015; Yang et al., 2016), we construct $\mathbf{X} = \mathbf{U}\mathsf{diag}(\boldsymbol{\sigma})\mathbf{V}^T \in \mathbb{R}^{n\times d}$ and $\mathbf{y} = \mathbf{X}\mathbf{w}_0 + \boldsymbol{\varepsilon} \in \mathbb{R}^n$ in the following way.

- We take $\mathbf{U}$ be the matrix of left singular vectors of $\mathbf{A} \in \mathbb{R}^{n\times d}$ which is constructed in the following way. (Note that $\mathbf{A}$ and $\mathbf{X}$ are different.) Let the rows of $\mathbf{A}$ be i.i.d. sampled from a multivariate t-distribution with covariance matrix $\mathbf{C}$ and $v = 2$ degree of freedom, where the (i,j)-th entry of $\mathbf{C} \in \mathbb{R}^{d\times d}$ is $2 \times 0.5^{|i-j|}$. Constructing $\mathbf{A}$ in this manner ensures that it has high row coherence.
- Let the entries of $\mathbf{b} \in \mathbb{R}^d$ be equally spaced between 0 and -6 and take $\sigma_i = 10^{b_i}$ for all $i \in [d]$.
- Let $\mathbf{V} \in \mathbb{R}^{d\times d}$ be an orthonormal basis for the column range of a $d \times d$ standard Gaussian matrix.
- Let $\mathbf{w}_0 = [\mathbf{1}_{0.2d}; 0.1\,\mathbf{1}_{0.6d}; \mathbf{1}_{0.2d}]$.
- Take the entries of $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ to be i.i.d. samples from the $\mathcal{N}(0, \xi^2)$ distribution.

This construction ensures that $\mathbf{X}$ has high row coherence, and its condition number is $\kappa(\mathbf{X}^T\mathbf{X}) = 10^{12}$. Let $\mathbf{S} \in \mathbb{R}^{n\times s}$ be any of the six sketching methods considered in this paper. We fix $n = 10^5$, $d = 500$, and $s = 5,000$. Since the sketching methods are randomized, we repeat each trial 10 times with idependent sketches and report averaged results.

4.2 Sketched MRR: Optimization Perspective

We seek to empirically verify Theorems 1 and 2 which study classical and Hessian sketches, respective, from the optimization perspective. In Figure 2, we plot the objective function value $f(\mathbf{w}) = \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \gamma\|\mathbf{w}\|_2^2$ against γ, under different settings of ξ (the standard deviation of the Gaussian noise added to the response). The black curves correspond to the optimal solution $\mathbf{w}^\star$; the color curves correspond to classical or Hessian sketch with different

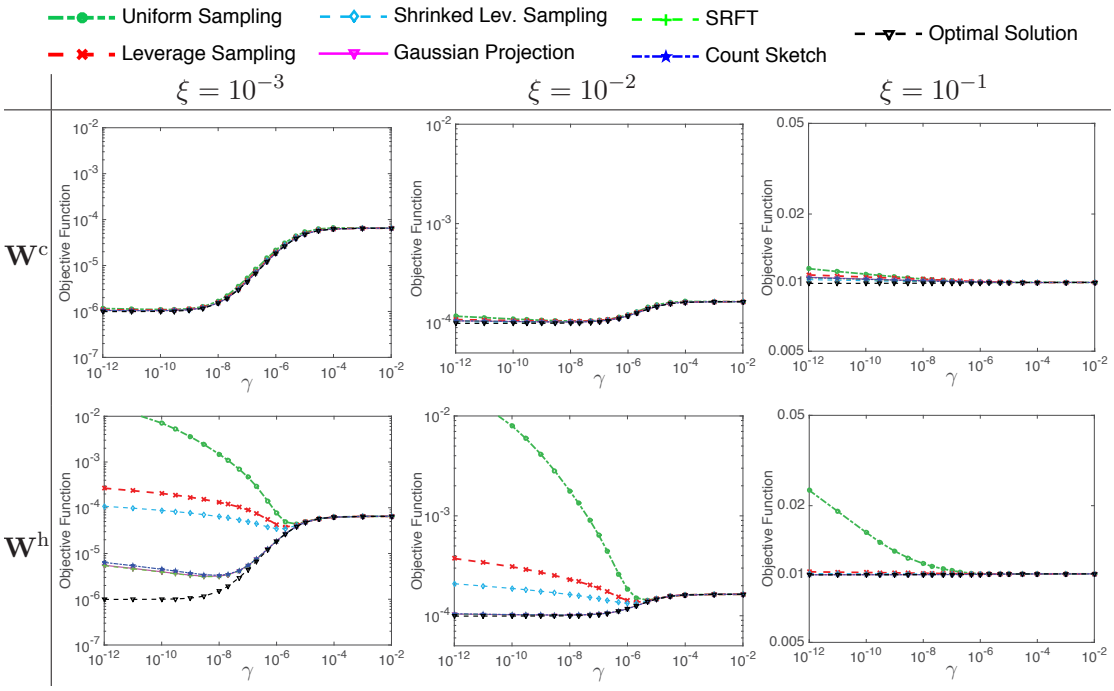


Figure 2: An empirical study of classical and Hessian sketch from the optimization perspective. The x-axis is the regularization parameter γ (log scale); the y-axis is the objective function values (log scale). Here ξ is the standard deviation of the Gaussian noise added to the response.

sketching methods. The results verify our theory: the objective value of the solution from the classical sketch, $\mathbf{w}^{\mathrm{c}}$, is always close to optimal; and the objective value of the solution from the Hessian sketch, $\mathbf{w}^{\mathrm{h}}$, is much worse than the optimal value when γ is small and $\mathbf{y}$ is mostly in the column space of $\mathbf{X}$.

4.3 Sketched MRR: Statistical Perspective

In Figure 3, we plot the analytical expressions for the squared bias, variance, and risk stated in Theorem 4 against the regularization parameter γ. Because these expressions involve the random sketching matrix $\mathbf{S}$, we randomly generate $\mathbf{S}$, repeat this procedure 10 times, and report the average of the computed squared biases, variances, and risks. We fix $\xi = 0.1$ (the standard deviation of the Gaussian noise). The results of this experiment match our theory: classical sketch magnified the variance, and Hessian sketch increased the bias. Even when γ is fine-tuned, the risks of classical and Hessian sketch can be much higher than those of the optimal solution. Our experiment also indicates that classical and Hessian sketch require setting γ larger than the best regularization parameter for the optimal solution $\mathbf{W}^{\star}$.

Classical and Hessian sketch do not outperform each other in terms of the risk. When variance dominates bias, Hessian sketch is better in terms of the risk; when bias dominates variance, classical sketch is preferable. In the experiment yielding Figure 3, Hessian sketch

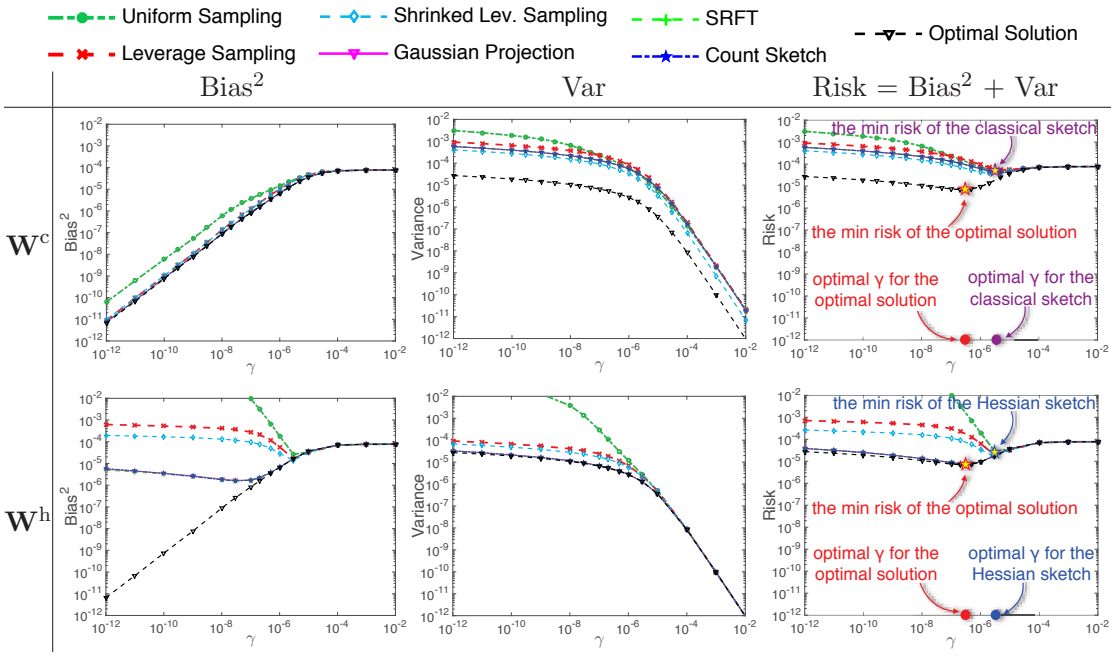


Figure 3: An empirical study of classical sketch and Hessian sketch from the statistical perspective. The x-axis is the regularization parameter γ (log-scale); the y-axes are respectively bias^2, variance, and risk (log-scale). We indicate the minimum risks and optimal choice of γ in the plots.

delivers lower risks than classical sketch. This is not generally true: if we use a smaller ξ (the standard deviation of the Gaussian noise), so that the variance is dominated by bias, then classical sketch results in lower risks than Hessian sketch.

4.4 Model Averaging: Optimization Objective

We consider different noise levels by setting $\xi = 10^{-2}$ or 10^{-1}, where ξ is defined in Section 4.1 as the standard deviation of the Gaussian noise in the response vector $\mathbf{y}$. We calculate the objective function values $f(\mathbf{w}^{\mathrm{c}}_{[g]})$ and $f(\mathbf{w}^{\mathrm{h}}_{[g]})$ for different settings of g, γ. We use different methods of sketching at the fixed sketch size $s = 5,000$.

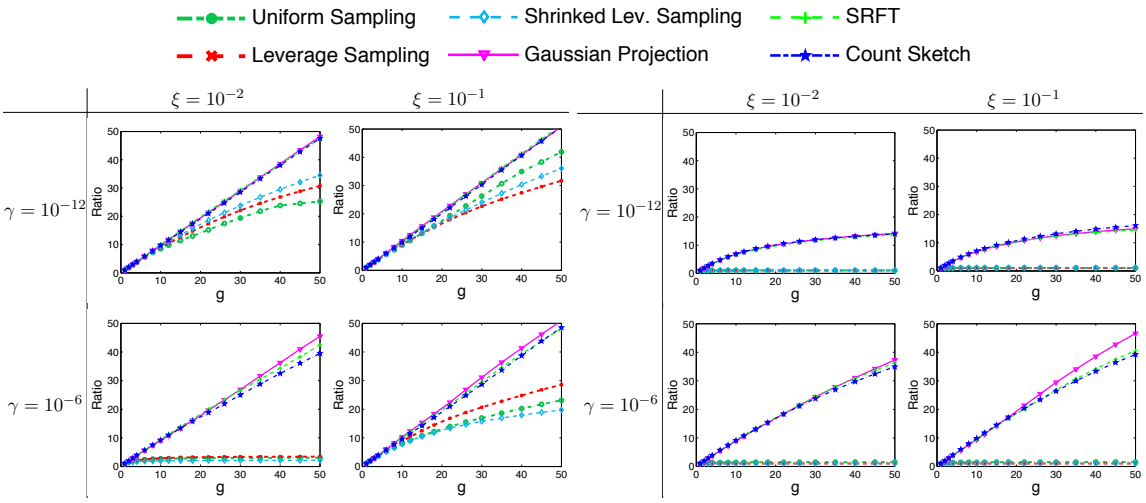
Theorem 7 indicates that for large s, e.g., Gaussian projection with $s = \tilde{\mathcal{O}}\big(\frac{d}{\epsilon}\big)$,

$$f\big(\mathbf{w}^{\mathrm{c}}_{[g]}\big) - f\big(\mathbf{w}^{\star}\big) \quad \leq \quad \beta\big(\tfrac{\epsilon}{g} + \beta^2\epsilon^2\big)\, f(\mathbf{w}^{\star}), \tag{9}$$

where $\beta = \frac{\|\mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2 + n\gamma} \leq 1$. In Figure 4(a) we plot the ratio

$$\frac{f(\mathbf{w}^{\mathrm{c}}_{[1]}) - f(\mathbf{w}^{\star})}{f(\mathbf{w}^{\mathrm{c}}_{[g]}) - f(\mathbf{w}^{\star})} \tag{10}$$

against g. Rapid growth of this ratio indicates that model averaging is highly effective. The results in Figure 4(a) indicate that model averaging significantly improves the accuracy



(a) Classical sketch with model averaging.

(b) Hessian sketch with model averaging.

Figure 4: An empirical study of model averaging from the optimization perspective. The x-axis is g, i.e., the number of models that are averaged. In 4(a), the y-axis is the ratio (log-scale) defined in (10). In 4(b), the y-axis is the ratio (log-scale) defined in (11). Here γ is the regularization parameter and ξ is the standard deviation of the Gaussian noise.

as measured by the objective function value. For the three random projection methods, the growth rate of this ratio is almost linear in g. In Figure 4(a), we observe that the regularization parameter γ affects the ratio (10). The ratio grows faster when $\gamma = 10^{-12}$ than when $\gamma = 10^{-6}$. This phenomenon is not explained by our theory.

Theorem 8 shows that for large sketch size s, e.g., Gaussian projection with $s = \tilde{\mathcal{O}}\left(\frac{d}{\epsilon}\right)$,

$$f(\mathbf{w}^{\mathrm{h}}) - f(\mathbf{w}^{\star}) \quad \leq \quad \beta^2 \left(\frac{\epsilon}{g} + \epsilon^2\right) \left(\frac{\|\mathbf{y}\|_2^2}{n} - f(\mathbf{w}^{\star})\right),$$

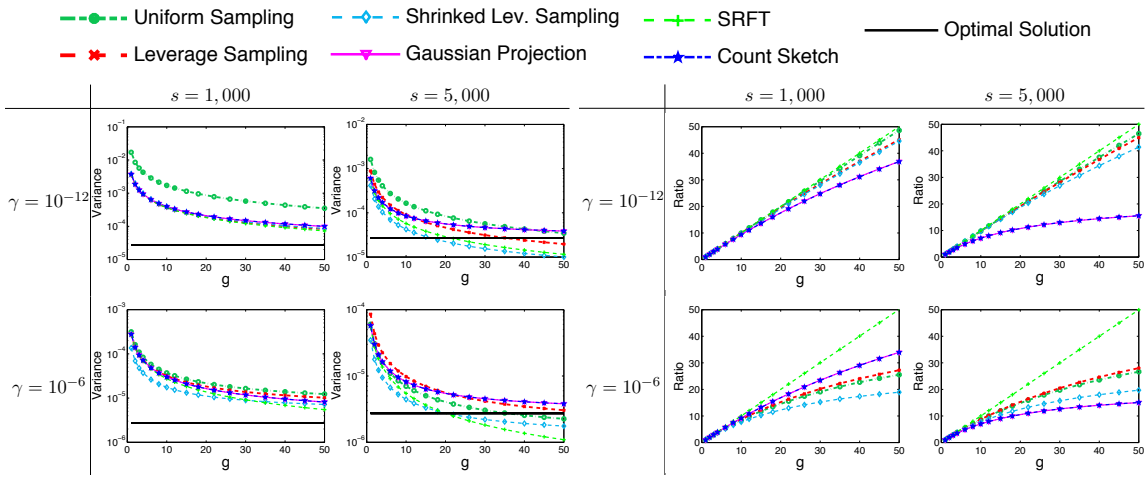
where $\beta = \frac{\|\mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2 + n\gamma} \leq 1$. In Figure 4(b), we plot the ratio

$$\frac{f(\mathbf{w}^{\mathrm{h}}_{[1]}) - f(\mathbf{w}^{\star})}{f(\mathbf{w}^{\mathrm{h}}_{[g]}) - f(\mathbf{w}^{\star})} \tag{11}$$

against g. Rapid growth of this ratio indicates that model averaging is highly effective. Our empirical results indicate that the growth rate of this ratio is moderately rapid for very small g and very slow for large g.

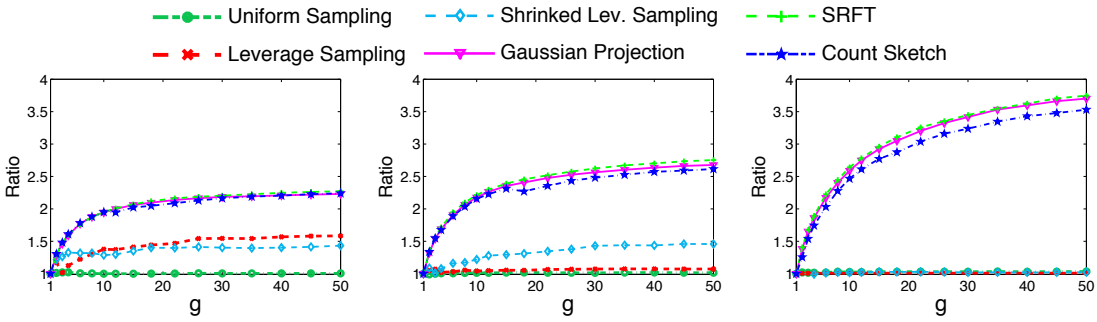
4.5 Model Averaging: Statistical Perspective

We empirically study model averaging from the statistical perspective. We calculate the bias and variance $\mathsf{bias}(\mathbf{w}^{\star})$, $\mathsf{var}(\mathbf{w}^{\star})$ of the optimal MRR solution according to Theorem 4



(a) The variance $\mathsf{var}(\mathbf{w}^{\mathrm{c}}_{[g]})$. (b) The ratio $\frac{\mathsf{var}(\mathbf{w}^{\mathrm{c}}_{[1]})}{\mathsf{var}(\mathbf{w}^{\mathrm{c}}_{[g]})}$.

Figure 5: An empirical study of the variance of classical sketch with model averaging. The x-axis is g, i.e., the number of models that are averaged. In 5(a), the y-axis is the variance $\mathsf{var}(\mathbf{w}^{\mathrm{c}}_{[g]})$ (log scale) defined in Theorem 9. In 5(b), the y-axis is the ratio $\frac{\mathsf{var}(\mathbf{w}^{\mathrm{c}}_{[1]})}{\mathsf{var}(\mathbf{w}^{\mathrm{c}}_{[g]})}$. Here γ is the regularization parameter and s is the sketch size.



(a) $s = 1,000$ (b) $s = 2,000$ (c) $s = 5,000$

Figure 6: An empirical study of the bias of Hessian sketch with model averaging. The x-axis is g, the number of models being averaged; the y-axis is the ratio (12).

and the bias and variance $\mathsf{bias}(\mathbf{w}^{\mathrm{c}}_{[g]})$, $\mathsf{var}(\mathbf{w}^{\mathrm{c}}_{[g]})$ and $\mathsf{bias}(\mathbf{w}^{\mathrm{h}}_{[g]})$, $\mathsf{var}(\mathbf{w}^{\mathrm{h}}_{[g]})$ of, respectively, the model averaged classical sketch solution and the model averaged Hessian sketch solution according to Theorem 9.

4.5.1 Classical Sketch

Theorem 10 indicates that for large enough s, e.g., Gaussian projection with $s = \tilde{\mathcal{O}}\big(\frac{d}{\epsilon^2}\big)$, with high probability

$$\frac{\mathsf{bias}(\mathbf{w}^{\mathrm{c}}_{[g]})}{\mathsf{bias}(\mathbf{w}^\star)} \leq 1+\epsilon \qquad \text{and} \qquad \frac{\mathsf{var}(\mathbf{w}^{\mathrm{c}}_{[g]})}{\mathsf{var}(\mathbf{w}^\star)} \leq \frac{n}{s}\Big(\sqrt{\tfrac{1+\epsilon}{h}}+\epsilon\Big)^2,$$

where $h = \min\{g, \Theta(\frac{n}{s})\}$. This result implies that model averaging decreases the variance of classical sketch without significantly changing the bias. We conduct experiments to verify this point.

In Figure 5(a) we plot the variance $\mathsf{var}(\mathbf{w}^{\mathrm{c}}_{[g]})$ against g; the variance of the optimal solution $\mathbf{w}^\star$ is depicted for comparison. Clearly, the variance drops as g grows. In particular, when s is big ($s = 5,000$) and g exceeds $\frac{n}{s}$ ($= \frac{100{,}000}{5{,}000} = 20$), $\mathsf{var}(\mathbf{w}^{\mathrm{c}}_{[g]})$ can be even lower than $\mathsf{var}(\mathbf{w}^\star)$.

To more clearly decrease the impact of model averaging on the variance, in Figure 5(b) we plot the ratio $\frac{\mathsf{var}(\mathbf{w}^{\mathrm{c}}_{[1]})}{\mathsf{var}(\mathbf{w}^{\mathrm{c}}_{[g]})}$ against g. According to Theorem 10, this ratio grows linearly in g when s is at least $\tilde{\mathcal{O}}(dg)$, and otherwise is sublinear in g. This claim is verified by the empirical results in Figure 5(b).

When $\mathsf{bias}(\mathbf{w}^{\mathrm{c}}_{[g]})$ is plotted as a function of g, the curves are almost horizontal, indicating that, as expected, *the bias is insensitive to the number of models* g. We do not show such plots because these nearly horizontal curves are not interesting.

4.5.2 Hessian Sketch

Theorem 11 indicates that for large enough s, e.g., Gaussian projection with $s = \tilde{\mathcal{O}}\big(\frac{d}{\epsilon^2}\big)$, the inequalities

$$\frac{\mathsf{bias}(\mathbf{w}^{\mathrm{h}}_{[g]})}{\mathsf{bias}(\mathbf{w}^\star)} \;\leq\; 1+\epsilon+\Big(\tfrac{\epsilon}{\sqrt{g}}+\epsilon^2\Big)\tfrac{\|\mathbf{X}\|_2^2}{n\gamma} \qquad \text{and} \qquad \frac{\mathsf{var}(\mathbf{w}^{\mathrm{h}}_{[g]})}{\mathsf{var}(\mathbf{w}^\star)} \;\leq\; 1+\epsilon$$

hold with high probability. That is, model averaging improves the bias without affecting the variance. The bound

$$\frac{\mathsf{bias}(\mathbf{w}^{\mathrm{h}}_{[g]})-\mathsf{bias}(\mathbf{w}^\star)}{\mathsf{bias}(\mathbf{w}^\star)} \;\leq\; \epsilon+\Big(\tfrac{\epsilon}{\sqrt{g}}+\epsilon^2\Big)\tfrac{\|\mathbf{X}\|_2^2}{n\gamma}$$

indicates that if $n\gamma$ is much smaller than $\|\mathbf{X}\|_2^2$ and $\epsilon \leq \frac{1}{\sqrt{g}}$, or equivalently, s is at least $\tilde{\mathcal{O}}(dg)$, then the ratio is proportional to $\frac{\epsilon}{\sqrt{g}}$.

To verify Theorem 11, we set γ very small—$\gamma = 10^{-12}$—and vary s and g. In Figure 6 we plot the ratio

$$\frac{\mathsf{bias}(\mathbf{w}^{\mathrm{h}}_{[1]})-\mathsf{bias}(\mathbf{w}^\star)}{\mathsf{bias}(\mathbf{w}^{\mathrm{h}}_{[g]})-\mathsf{bias}(\mathbf{w}^\star)}, \tag{12}$$

by fixing $\gamma = 10^{-12}$ and varying s and g. The theory indicates that for large sketch size $s = \tilde{\mathcal{O}}(dg^2)$, this ratio should grow nearly linearly in g. Figure 6 shows that only for large s and very small g, the growth is near linear in g; this verifies our theory.

When we similarly plot $\mathsf{var}\big(\mathbf{w}^{\mathrm{h}}_{[g]}\big)$ against g, we observe that $\mathsf{var}\big(\mathbf{w}^{\mathrm{h}}_{[g]}\big)$ remains nearly unaffected as g grows from 1 to 50. Since the curves of the variance against g are almost horizontal lines, we do not show this plot in the paper.

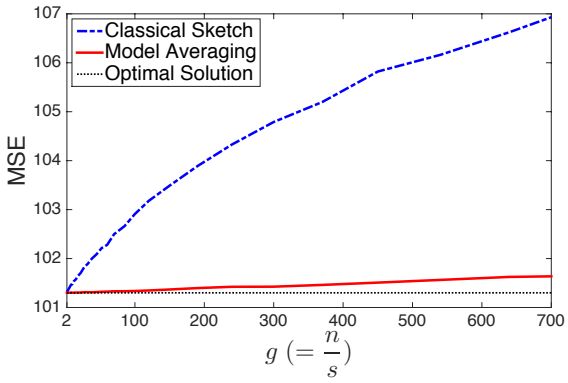


Figure 7: Prediction performance of classical sketch with and without model averaging on the Year Prediction data set. The x-axis is g, the number of data partitions, and the y-axis is the mean squared error (MSE) on the test set.

5. Model Averaging Experiments on Real-World Data

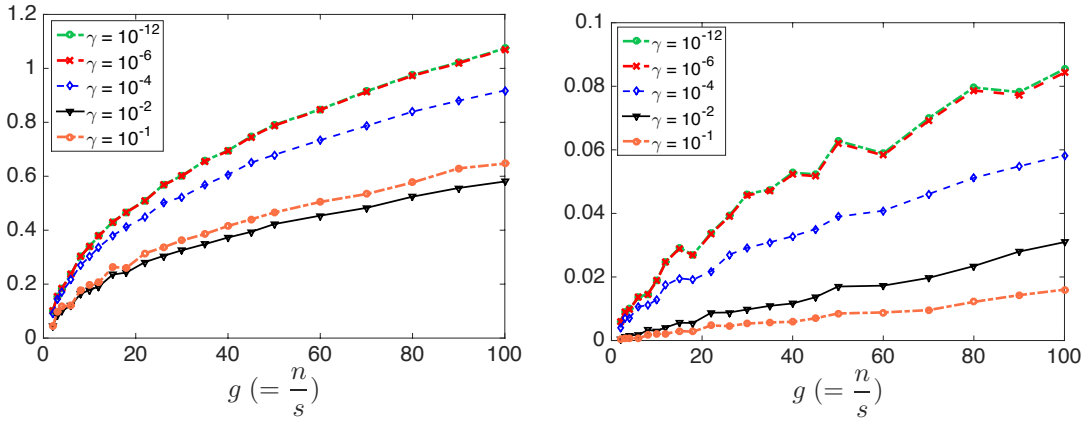
In Section 1 we mentioned that in the distributed setting where the feature-response pairs $(\mathbf{x}_1, \mathbf{y}_1), \cdots, (\mathbf{x}_n, \mathbf{y}_n) \in \mathbb{R}^{d \times m}$ are randomly and uniformly partitioned across g machines,[5] classical sketch with model averaging requires only one round of communication, and is therefore a communication-efficient algorithm that can be used to: (1) obtain an approximate solution of the MRR problem with risk comparable to a batch solution, and (2) obtain a low-precision solution of the MRR optimization problem that can be used as an initializer for more communication-intensive optimization algorithms. In this section, we demonstrate both applications.

We use the Million Song Year Prediction data set, which has $463,715$ training samples and $51,630$ test samples with 90 features and one response. We normalize the data by shifting the responses to have zero mean and scaling the range of each feature to $[-1, 1]$. We randomly partition the training data into g parts, which amounts to uniform row selection with sketch size $s = \frac{n}{g}$.

5.1 Prediction Error

We tested the prediction performance of sketched ridge regression by implementing classical sketch with model averaging in PySpark (Zaharia et al., 2010).[6] We ran our experiments using PySpark in local mode; the experiments proceeded in three steps: (1) use five-fold cross-validation to determine the regularization parameter γ; (2) learn the model $\mathbf{w}$ using the selected γ; and (3) use $\mathbf{w}$ to predict on the test set and record the mean squared errors (MSEs). These steps map cleanly onto the Map-Reduce programming model used by PySpark.

5. If the samples are i.i.d., then any deterministic partition is essentially a uniformly randomly distributed partition. Otherwise, we can invoke a Shuffle operation, which is supported by systems such as Apache Spark (Zaharia et al., 2010), to make the partitioning uniformly randomly distributed.
6. The code is available at https://github.com/wangshusen/SketchedRidgeRegression.git



(a) Classical sketch

(b) Classical sketch with model averaging

Figure 8: Optimization performance of classical sketch with and without model averaging. The x-axis is g, the number of data partitions, and the y-axis is the ratio $\frac{\|\mathbf{w}-\mathbf{w}^\star\|_2}{\|\mathbf{w}^\star\|_2}$.

In Figure 7, we plot the test MSE against $g = \frac{n}{s}$. As g grows, the sketch size $s = \frac{n}{g}$ decreases, so the performance of classical sketch deteriorates. However classical sketch with model averaging always has test MSE comparable to the optimal solution.

5.2 Optimization Error

We mentioned earlier that classical sketch with or without model averaging can be used to initialize optimization algorithms for solving MRR problems. If $\mathbf{w}$ is initialized with zero-mean random variables or deterministically with zeros, then $\mathbb{E}\big[\|\mathbf{w}-\mathbf{w}^\star\|_2/\|\mathbf{w}^\star\|_2\big] \geq 1$. Any $\mathbf{w}$ with the above ratio substantially smaller than 1 provides a better initialization. We implemented classical sketch with and without model averaging in Python and calculated the above ratio on the training set of the Year Prediction data set; to estimate the expectation, we repeated the procedure 100 times and report the average of the ratios.

In Figure 8, we plot the average of the ratio $\frac{\|\mathbf{w}-\mathbf{w}^\star\|_2}{\|\mathbf{w}^\star\|_2}$ against g for different settings of the regularization parameter γ. Clearly, classical sketch does not give a good initialization unless g is small (equivalently, the sketch size $s = \frac{n}{g}$ is large). In contrast, the averaged solution is always close to $\mathbf{w}^\star$.

6. Sketch of Proof

In this section, we outline the proofs of our main results. The complete details are provided in the appendix. Section 6.1 recaps several relevant properties of matrix sketching. Section 6.2 establishes certain properties of averages of sketches; these results are used to analyze the application of model averaging to the MRR problem. Sections 6.3 to 6.6 provide key structural results on sketched solutions to the MRR problem constructed with or without model averaging.

Our main results in Section 3 (Theorems 1, 2, 5, 6, 7, 8, 10, and 11) follow directly from the relevant properties of matrix sketching and the structural results for solutions to the

sketched MRR problem. Table 4 summarizes the dependency relationships among these theorems. For example, Theorem 1, which studies classical sketching from the optimization perspective, is one of our main theorems and is proven using Theorems 12 and 15.

Table 4: An overview of our results and their dependency relationships.

Main Theorems	Solution	Perspective	Prerequisites
Theorem 1	classical	optimization	Theorems 12 and 15
Theorem 2	Hessian	optimization	Theorems 12 and 16
Theorem 5	classical	statistical	Theorems 12, 13, 17, 18
Theorem 6	Hessian	statistical	Theorems 12 and 19
Theorem 7	classical, averaging	optimization	Theorems 14 and 20
Theorem 8	Hessian, averaging	optimization	Theorems 14 and 21
Theorem 10	classical, averaging	statistical	Theorems 14 and 22
Theorem 11	Hessian, averaging	statistical	Theorems 14 and 23

6.1 Properties of Matrix Sketching

Our analysis of the performance of solutions to the sketched MRR problem draws heavily on the three key properties defined in Assumption 1. Theorem 12 establishes that the six sketching methods considered in this paper indeed enjoy the three key properties under certain conditions. Finally, Theorem 13 establishes the lower bounds of $\|\mathbf{S}\|_2^2$ that are used to prove the lower bounds on the variance of sketched MRR solutions in Theorem 5.

Assumption 1 *Let $\eta, \epsilon \in (0,1)$ be fixed parameters. Let $\mathbf{B}$ be any fixed matrix of conformal shape, $\rho = \text{rank}(\mathbf{X})$, and $\mathbf{U} \in \mathbb{R}^{n\times\rho}$ be an orthonormal basis for the column span of $\mathbf{X}$. Let $\mathbf{S} \in \mathbb{R}^{n\times s}$ be a sketching matrix, where s depends on η and/or ϵ. Throughout this paper, we assume that $\mathbf{S}$ satisfies the following properties with a probability that depends on s:*

1.1 $\left\|\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} - \mathbf{I}_\rho\right\|_2 \leq \eta$ *(Subspace Embedding Property);*

1.2 $\left\|\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{B} - \mathbf{U}^T\mathbf{B}\right\|_F^2 \leq \epsilon\|\mathbf{B}\|_F^2$ *(Matrix Multiplication Property);*

1.3 When $s < n$, $\|\mathbf{S}\|_2^2 \leq \frac{\theta n}{s}$ for some constant θ (Bounded Spectral Norm Property).

The subspace embedding property requires that sketching preserves the inner products between the columns of a matrix with orthonormal columns. Equivalently, it ensures that the singular values of any sketched column-orthonormal matrix are all close to one. The subspace embedding property implies that, in particular, the squared norm of $\mathbf{Sx}$ is close to that of $\mathbf{x}$ for any n-dimensional vector in a fixed ρ-dimensional subspace. A dimension counting argument suggests that since $\mathbf{Sx}$ is an s-dimensional vector, its length must be scaled by a factor of $\sqrt{\frac{n}{s}}$ to ensure that this consequence of the subspace embedding property holds. The bounded spectral norm property requires that the spectral norm of $\mathbf{S}$ is not much larger than this rescaling factor of $\sqrt{\frac{n}{s}}$.

Remark 2 *The first two assumptions were identified in (Mahoney, 2011) and are the relevant structural conditions that allow strong results from the optimization perspective.*

Table 5: The two middle columns provide an upper bound on the sketch size s needed to satisfy the subspace embedding property and the matrix multiplication property, respectively, under the different sketching modalities considered; the right column lists the parameter θ with which the bounded spectral norm property holds. These properties hold with constant probability for the indicated values of s. Here τ is defined in (5) and reflects the quality of the approximation of the leverage scores of $\mathbf{U}$; μ is the row coherence of $\mathbf{U}$. For Gaussian projection and CountSketch, the small-o notation is a consequence of $s = o(n)$.

Sketching	**Subspace Embedding**	**Matrix Multiplication**	**Spectral Norm**
Leverage	$s = \mathcal{O}\big(\frac{\tau\rho}{\eta^2}\log\frac{\rho}{\delta_1}\big)$	$s = \mathcal{O}\big(\frac{\tau\rho}{\epsilon\delta_2}\big)$	$\theta = \infty$
Uniform	$s = \mathcal{O}\big(\frac{\mu\rho}{\eta^2}\log\frac{\rho}{\delta_1}\big)$	$s = \mathcal{O}\big(\frac{\mu\rho}{\epsilon\delta_2}\big)$	$\theta = 1$
Shrinked Leverage	$s = \mathcal{O}\big(\frac{\tau\rho}{\eta^2}\log\frac{\rho}{\delta_1}\big)$	$s = \mathcal{O}\big(\frac{\tau\rho}{\epsilon\delta_2}\big)$	$\theta = 2$
SRHT	$s = \mathcal{O}\big(\frac{\rho+\log n}{\eta^2}\log\frac{\rho}{\delta_1}\big)$	$s = \mathcal{O}\big(\frac{\rho+\log n}{\epsilon\delta_2}\big)$	$\theta = 1$
Gaussian Projection	$s = \mathcal{O}\big(\frac{\rho+\log(1/\delta_1)}{\eta^2}\big)$	$s = \mathcal{O}\big(\frac{\rho}{\epsilon\delta_2}\big)$	$\theta = 1 + o(1)$ w.h.p.
CountSketch	$s = \mathcal{O}\big(\frac{\rho^2}{\delta_1\eta^2}\big)$	$s = \mathcal{O}\big(\frac{\rho}{\epsilon\delta_2}\big)$	$\theta = 1 + o(1)$ w.h.p.

The third assumption is new, but Ma et al. (2015); Raskutti and Mahoney (2016) demonstrated that some sort of additional condition is necessary to obtain strong results from the statistical perspective.

Remark 3 *We note that $\mathbf{U}^T\mathbf{U} = \mathbf{I}_\rho$, and thus Assumption 1.1 can be expressed in the form of an approximate matrix multiplication bound (Drineas et al., 2006a). We call it the Subspace Embedding Property since, as first highlighted in Drineas et al. (2006b), this subspace embedding property is the key result necessary to obtain high-quality sketching algorithms for regression and related problems.*

Theorem 12 shows that the six sketching methods satisfy the three properties when s is sufficiently large. In particular, Theorem 12 shows that for all the sketching methods except leverage score sampling,[7] $\|\mathbf{S}\|_2^2$ has nontrivial upper bound. This is why Theorems 5 and 10 do not apply to leverage score sampling. This fact can also be viewed as a motivation to use shrinked leverage score sampling. We prove Theorem 12 in Appendix A.

Theorem 12 *Fix failure probability δ and error parameters η and ϵ; set the sketch size s as Table 5. Assumption 1.1 is satisfied with probability at least $1 - \delta_1$. Assumption 1.2 is satisfied with probability at least $1 - \delta_2$. Assumption 1.3 is satisfied either surely or with high probability (w.h.p.); the parameter θ is indicated in Table 5.*

Theorem 13 establishes lower bounds on $\|\mathbf{S}\|_2^2$, and will be applied to prove the lower bound on the variance of the classical sketch. From Table 6 we see that the lower bound for

7. If one leverage score approaches zero, then the corresponding sampling probability p_i goes to zero. By the definition of $\mathbf{S}$, the scale factor $\frac{1}{\sqrt{sp_i}}$ goes to infinity, which makes $\|\mathbf{S}\|_2^2$ unbounded. The shinked leverage score sampling avoids this problem and is thus a better choice than the leverage score sampling.

(shrinked) leverage score sampling is not interesting, because μ can be very large. This is why Theorem 5 does not provide a lower bound for shrinked leverage score sampling. We prove Theorem 13 in Appendix A.

Table 6: Lower bounds on ϑ for the sketching modalities (ϑ is defined in Theorem 13). The shrinked leverage score sampling is performed using the row leverage scores of a matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, and μ is the row coherence of $\mathbf{X}$.

Uniform	$\vartheta = 1$
Leverage	$\vartheta \geq \frac{1}{\mu}$
Shrinked Leverage	$\vartheta \geq \frac{2}{1+\mu}$
SRHT	$\vartheta = 1$
Gaussian Projection	$\vartheta \geq 1 - o(1)$ w.h.p.
CountSketch	$\vartheta \geq 1 - o(1)$ w.h.p.

Theorem 13 (Semidefinite Lower Bound on the Sketching Matrix) *When $s < n$, $\mathbf{S}^T\mathbf{S} \succeq \frac{\vartheta n}{s}\mathbf{I}_s$ holds either surely or with high probability (w.h.p.), where Table 6 provides the applicable ϑ for each sketching method.*

Remark 4 *Let $p_1, \cdots, p_n$ be an arbitrary set of sampling probabilities. By the definition of the associated sampling matrix $\mathbf{S} \in \mathbb{R}^{n\times s}$, the non-zero entries of $\mathbf{S}$ can be any of $\frac{1}{\sqrt{sp_i}}$, for $i \in [n]$.*

For leverage score sampling, since the smallest sampling probability can be zero or close, and the largest sampling probability can be close to one, $\|\mathbf{S}\|_2^2$ has no nontrivial upper or lower bound.[8] *It is because $\min_i p_i$ can be close to zero and $\max_i p_i$ can be large (close to one).*

For shrinked leverage score sampling, because $\min_i p_i$ is at least $\frac{1}{2n}$, $\|\mathbf{S}\|_2^2$ has a nontrivial upper bound; but as in the case of leverage score sampling, since $\max_i p_i$ can be large, there is no nontrivial lower bound on $\|\mathbf{S}\|_2^2$.

6.2 Matrix Sketching with Averaging

Assumptions 1.1 and 1.2 imply that sketching can be used to approximate certain matrix products, but what happens if we independently draw g sketches, use them to approximate the same matrix product, and then average the g results? Intuitively, averaging should lower the variance of the approximation without affecting its bias, and thus provide a better approximation of the true product.

To justify this intuition formally, let $\mathbf{S}_1, \cdots, \mathbf{S}_g \in \mathbb{R}^{n\times s}$ be sketching matrices and $\mathbf{A}$ and $\mathbf{B}$ be fixed conformal matrices. Then evidently

$$\frac{1}{g}\sum_{i=1}^{g} \mathbf{A}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{B} \;=\; \mathbf{A}^T\mathbf{S}\mathbf{S}^T\mathbf{B},$$

8. In our application, nontrivial bound means $\|\mathbf{S}\|_2^2$ is of order $\frac{n}{s}$.

where $\mathbf{S} = \frac{1}{\sqrt{g}}[\mathbf{S}_1, \cdots, \mathbf{S}_g] \in \mathbb{R}^{n \times gs}$ can be thought of as a sketching matrix formed by concatenating the g smaller sketching matrices. If $\mathbf{S}_1, \cdots, \mathbf{S}_g$ are all instance of column selection, SRHT, or Gaussian projection sketching matrices, then $\mathbf{S}$ is a larger instance of the same type of sketching matrix.[9]

To analyze the effect of model averaging on the solution to the sketched MRR problem, we make the following assumptions on the concatenated sketch matrix. Assumption 2.1 is the subspace embedding property, Assumption 2.2 is the matrix multiplication property, and Assumption 2.3 is the bounded spectral norm property.

Assumption 2 *Let $\eta, \epsilon \in (0,1)$ be fixed parameters. Let $\mathbf{B}$ be any fixed matrix of proper size, $\rho = \mathrm{rank}(\mathbf{X})$, and $\mathbf{U} \in \mathbb{R}^{n \times \rho}$ be an orthonormal basis for the column span of $\mathbf{X}$. Let $\mathbf{S}_1, \cdots, \mathbf{S}_g \in \mathbb{R}^{n \times s}$ be sketching matrices and $\mathbf{S} = \frac{1}{\sqrt{g}}[\mathbf{S}_1, \cdots, \mathbf{S}_g] \in \mathbb{R}^{n \times gs}$; here s depends on η and/or ϵ. Throughout this paper we assume that $\mathbf{S}$ and the $\mathbf{S}_i$ satisfy the following properties with a probability that depends on g and s:*

2.1 $\left\|\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} - \mathbf{I}_\rho\right\|_2 \leq \eta$ *for all* $i \in [g]$ *and* $\left\|\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} - \mathbf{I}_\rho\right\|_2 \leq \frac{\eta}{\sqrt{g}}$;

2.2 $\left(\frac{1}{g}\sum_{i=1}^g \left\|\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{B} - \mathbf{U}^T\mathbf{B}\right\|_F\right)^2 \leq \epsilon\|\mathbf{B}\|_F^2$ *and* $\left\|\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{B} - \mathbf{U}^T\mathbf{B}\right\|_F^2 \leq \frac{\epsilon}{g}\|\mathbf{B}\|_F^2$;

2.3 For some constant θ, $\|\mathbf{S}_i\|_2^2 \leq \frac{\theta n}{s}$ for all $i \in [g]$, and $\|\mathbf{S}\|_2^2 \leq \frac{\theta n}{gs}$ for $gs < n$.

Except in the case of leverage score sampling, when gs is comparable to or larger than n, $\|\mathbf{S}\|_2^2 = \Theta(1)$.

Theorem 14 establishes that random column selection, SRHT, and Gaussian projection matrices satisfy Assumptions 2.1, 2.2, and 2.3. We prove Theorem 14 in Appendix A.

Theorem 14 *Let $\mathbf{S}_1, \cdots, \mathbf{S}_g \in \mathbb{R}^{n \times s}$ be independent and identically distributed random sketching matrices that are either column selection, SRHT, or Gaussian projection matrices. Fix a failure probability δ and error parameters η and ϵ, then set the sketch size s as Table 5.*

Assumption 2.1 holds with probability at least $1 - (g+1)\delta_1$. Assumption 2.2 holds with probability at least $1 - 2\delta_2$. Assumption 2.3 is satisfied either surely or with high probability, with the parameter θ specified in Table 5.

In Theorem 12, Assumption 1.1 fails with probability at most δ_1. In contrast, in Theorem 14, the counterpart assumption fails with probability at most $(g+1)\delta_1$. However, this makes little difference in practice, because the dependence of s on δ_1 is logarithmic, so δ_1 can be set very small (recall Table 5) without increasing s significantly.

Remark 5 *We do not know whether CountSketch enjoys the properties in Assumption 2. There are two difficulties in establishing this using the same route as is employed in our proof of Theorem 12 for other sketching methods. First, the concatenation of multiple CountSketch matrices is not a CountSketch matrix. Second, the probability that a CountSketch matrix does not have the subspace embedding property is constant, rather than exponentially small.*

9. CountSketch sketching matrices does not have this property. If $\mathbf{S}_i \in \mathbb{R}^{n \times s}$ is a CountSketch matrix, then it has only one non-zero entry in each row. In contrast, $\mathbf{S} \in \mathbb{R}^{n \times gs}$ has g non-zero entries in each row.

6.3 Sketched MRR: Optimization Perspective

The randomness in the performance of the classical and Hessian sketch is entirely due to the choice of random sketching matrix. We now assume that the randomly sampled sketching matrices are "nice" in that they satisfy the assumptions just introduced, and state deterministic results on the optimization performance of the classical and Hessian sketches.

Theorem 15 holds under the subspace embedding property and the matrix multiplication property (Assumptions 1.1 and 1.2), and quantifies the suboptimality of the classical sketch. We prove this result in Appendix B.

Theorem 15 (Classical Sketch) *Let Assumptions 1.1 and 1.2 hold for the sketching matrix* $\mathbf{S} \in \mathbb{R}^{n \times s}$. *Let* η *and* ϵ *be defined in Assumption 1, and let* $\alpha = \frac{2\max\{\epsilon, \eta^2\}}{1-\eta}$ *and* $\beta = \frac{\|\mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2 + n\gamma}$, *then*

$$f(\mathbf{W}^c) - f(\mathbf{W}^\star) \quad \leq \quad \alpha\beta f(\mathbf{W}^\star).$$

Theorem 16 holds under the subspace embedding property (Assumption 1.1), and quantifies the suboptimality of the Hessian sketch. We prove this result in Appendix B.

Theorem 16 (Hessian Sketch) *Let Assumption 1.1 hold for the sketching matrix* $\mathbf{S} \in \mathbb{R}^{n \times s}$. *Let* η *be defined in Assumption 1 and* $\beta = \frac{\|\mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2 + n\gamma}$, *then*

$$f(\mathbf{W}^h) - f(\mathbf{W}^\star) \quad \leq \quad \frac{\eta^2\beta^2}{(1-\eta)^2}\left(\frac{\|\mathbf{Y}\|_F^2}{n} - f(\mathbf{W}^\star)\right).$$

6.4 Sketched MRR: Statistical Perspective

Similarly, we assume that the randomly sampled sketching matrices are nice, and state deterministic results on the bias and variance of the classical and Hessian sketches.

Theorem 17 holds under the subspace embedding property (Assumption 1.1) and the bounded spectral norm property (Assumption 1.3), and bounds the bias and variance of the classical sketch. Specifically, it shows that the bias of the classical sketch is close to that of the optimal solution, but that the variance may be much larger. We prove this result in Appendix C.

Theorem 17 (Classical Sketch) *Let* η *and* θ *be defined in Assumption 1. Under Assumption 1.1, it holds that*

$$\tfrac{1}{1+\eta} \;\leq\; \tfrac{\mathsf{bias}(\mathbf{W}^c)}{\mathsf{bias}(\mathbf{W}^\star)} \;\leq\; \tfrac{1}{1-\eta}.$$

Further assume $s \leq n$; *under Assumptions 1.1 and 1.3, it holds that*

$$\tfrac{\mathsf{var}(\mathbf{W}^c)}{\mathsf{var}(\mathbf{W}^\star)} \;\leq\; \tfrac{(1+\eta)}{(1-\eta)^2}\tfrac{\theta n}{s}.$$

Theorem 18 establishes a lower bound on the variance of the classical sketch. We prove this result in Appendix C.

Theorem 18 (Lower Bound on the Variance) *Under Assumption 1.1 and the additional assumption that* $\mathbf{S}^T\mathbf{S} \succeq \frac{\vartheta n}{s}\mathbf{I}_s$*, it holds that*

$$\frac{\mathsf{var}(\mathbf{W}^c)}{\mathsf{var}(\mathbf{W}^\star)} \quad \geq \quad \frac{1-\eta}{(1+\eta)^2}\frac{\vartheta n}{s}.$$

Theorem 19 holds under the subspace embedding property (Assumption 1.1), and quantifies the bias and variance of the Hessian sketch. We prove this result in Appendix C.

Theorem 19 (Hessian Sketch) *Let* η *be defined in Assumption 1, take* $\rho = \mathrm{rank}(\mathbf{X})$*, and let* $\sigma_1 \geq \cdots \geq \sigma_\rho$ *be the singular values of* $\mathbf{X}$*. Under Assumption 1.1, it holds that*

$$\begin{aligned}\frac{\mathsf{bias}(\mathbf{W}^h)}{\mathsf{bias}(\mathbf{W}^\star)} \;&\leq\; \frac{1}{1-\eta}\Big(1 + \frac{\eta\sigma_1^2}{n\gamma}\Big),\\ \frac{1}{1+\eta} \;&\leq\; \frac{\mathsf{var}(\mathbf{W}^h)}{\mathsf{var}(\mathbf{W}^\star)} \;\leq\; \frac{1}{1-\eta}.\end{aligned}$$

Further assume that $\sigma_\rho^2 \geq \frac{n\gamma}{\eta}$*. Then*

$$\frac{\mathsf{bias}(\mathbf{W}^h)}{\mathsf{bias}(\mathbf{W}^\star)} \;\geq\; \frac{1}{1+\eta}\Big(\frac{\eta\sigma_\rho^2}{n\gamma} - 1\Big).$$

6.5 Model Averaging: Optimization Perspective

Theorem 20 holds under the subspace embedding property (Assumption 2.1) and the matrix multiplication property (Assumption 2.2). We prove this result in Appendix D.

Theorem 20 (Classical Sketch with Model Averaging) *Let* η *and* ϵ *be defined in Assumption 2, and let* $\alpha = 2\big(\frac{1}{\sqrt{g}} + 2\beta\eta\big)^2 \max\big\{\epsilon, \eta^2\big\}$ *and* $\beta = \frac{\|\mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2 + n\gamma} \leq 1$*. Under Assumption 2.1 and 2.2, we have that*

$$f(\mathbf{W}^c) - f(\mathbf{W}^\star) \quad \leq \quad \alpha\beta f(\mathbf{W}^\star).$$

Theorem 21 holds under the subspace embedding property (Assumption 2.1), and is proven in Appendix D.

Theorem 21 (Hessian Sketch with Model Averaging) *Let* η *be defined in Assumption 2, and let* $\alpha = \big(\frac{\eta}{\sqrt{g}} + \frac{\eta^2}{1-\eta}\big)$ *and* $\beta = \frac{\|\mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2 + n\gamma} \leq 1$*. Under Assumption 2.1, we have that*

$$f(\mathbf{W}^h) - f(\mathbf{W}^\star) \quad \leq \quad \alpha^2\beta^2\Big(\frac{1}{n}\|\mathbf{Y}\|_F^2 - f(\mathbf{W}^\star)\Big).$$

6.6 Model Averaging: Statistical Perspective

Theorem 22 requires the subspace embedding property (Assumption 2.1). In addition, to bound the variance, the spectral norms of $\mathbf{S}_1, \cdots, \mathbf{S}_g$ and $\mathbf{S} = \frac{1}{\sqrt{g}}[\mathbf{S}_1, \cdots, \mathbf{S}_g]$ must be bounded (Assumption 2.3). This result shows that model averaging decreases the variance of the classical sketch without increasing its bias. We prove this result in Appendix E.

Theorem 22 (Classical Sketch with Model Averaging) *Under Assumption 2.1, it holds that*

$$\frac{\mathsf{bias}(\mathbf{W}^c)}{\mathsf{bias}(\mathbf{W}^\star)} \;\leq\; \frac{1}{1-\eta}.$$

Under Assumptions 2.1 and 2.3, it holds that

$$\frac{\mathsf{var}(\mathbf{W}^c)}{\mathsf{var}(\mathbf{W}^\star)} \;\leq\; \frac{\theta n}{s}\left(\frac{\sqrt{1+\eta/\sqrt{g}}}{\sqrt{h}} + \frac{\eta\sqrt{1+\eta}}{1-\eta}\right)^2.$$

Here η and θ are defined in Assumption 2 and $h = \min\{g,\, \frac{n}{s}\big(1-o(1)\big)\}$,

Theorem 23 requires the subspace embedding property (Assumption 2.1), and shows that model averaging decreases the bias of the Hessian sketch without increasing its variance. We prove this result in Appendix E.

Theorem 23 (Hessian Sketch with Model Averaging) *Under Assumption 2.1, it holds that:*

$$\frac{\mathsf{bias}(\mathbf{W}^h)}{\mathsf{bias}(\mathbf{W}^\star)} \;\leq\; \frac{1}{1-\eta} + \Big(\frac{\eta}{\sqrt{g}} + \frac{\eta^2}{1-\eta}\Big)\frac{\|\mathbf{X}\|_2^2}{n\gamma},$$
$$\frac{\mathsf{var}(\mathbf{W}^h)}{\mathsf{var}(\mathbf{W}^\star)} \;\leq\; \frac{1}{1-\eta}.$$

Here η is defined in Assumption 2.

7. Conclusions

We studied sketched matrix ridge regression (MRR) from the optimization and statistical perspectives. Using classical sketch, by taking a large enough sketch, one can obtain an ϵ-accurate approximate solution. Counterintuitively and in contrast to classical sketch, the relative error of Hessian sketch increases as the responses $\mathbf{Y}$ are better approximated by linear combinations of the columns of $\mathbf{X}$. Both classical and Hessian sketches can have statistical risks that are worse than the risk of the optimal solution by an order of magnitude.

We proposed the use of model averaging to attain better optimization and statistical properties. We have shown that model averaging leads to substantial improvements in the theoretical error bounds, suggesting applications in distributed optimization and machine learning. We also empirically verified its practical benefits.

Our fixed-design statistical analysis has limitations. We have shown that the classical sketch and Hessian sketch can significantly increase the in-sample statistical risk, which implies large training error, and that model averaging can alleviate such problems. However, our statistical results are not directly applicable to an unseen test sample. We conjecture that the generalization error can be bounded by following the random design analysis of Hsu et al. (2014), which is left as future work.

Acknowledgments

We thank the anonymous reviewers and Serena Ng for their helpful suggestions. We thank the Army Research Office and the Defense Advanced Research Projects Agency for partial support of this work.

Appendix A. Properties of Matrix Sketching: Proofs

In Section A.1 we prove Theorem 12. In Section A.2, we prove Theorem 13. In Section A.3 we prove Theorem 14.

A.1 Proof of Theorem 12

We prove that the six sketching methods considered in this paper satisfy the three key properties. In Section A.1.1 we show the six sketching methods satisfy Assumptions 1.1 and 1.2. In section A.1.2 we show the six sketching methods satisfy Assumption 1.3.

A.1.1 Proof of Assumptions 1.1 and 1.2

For uniform sampling, leverage score sampling, Gaussian projection, SRHT, and CountSketch, the subspace embedding property and matrix multiplication property have been established by the previous works (Drineas et al., 2008, 2011; Meng and Mahoney, 2013; Nelson and Nguyên, 2013; Tropp, 2011; Woodruff, 2014). See also (Wang et al., 2016b) for a summary.

In the following we prove only that **shrinked leverage score sampling** satisfies assumptions 1.1 and 1.2. We cite the following lemma from (Wang et al., 2016a); this lemma was first established in the works (Drineas et al., 2008; Gittens, 2011; Woodruff, 2014).

Lemma 24 (Wang et al. (2016a)) *Let $\mathbf{U} \in \mathbb{R}^{n\times\rho}$ be a fixed matrix with orthonormal columns. Let the column selection matrix $\mathbf{S} \in \mathbb{R}^{n\times s}$ sample s columns according to probabilities $p_1, p_2, \cdots, p_n$. Assume $\alpha \geq \rho$ and*

$$\max_{i\in[n]} \frac{\|\mathbf{u}_{i:}\|_2^2}{p_i} \;\leq\; \alpha.$$

When $s \geq \alpha\frac{6+2\eta}{3\eta^2}\log(\rho/\delta_1)$, it holds that

$$\mathbb{P}\Big\{\big\|\mathbf{I}_\rho - \mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}\big\|_2 \;\geq\; \eta\Big\} \;\leq\; \delta_1.$$

When $s \geq \frac{\alpha}{\epsilon\delta_2}$, it holds that

$$\mathbb{E}\big\|\mathbf{U}\mathbf{B} - \mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{B}\big\|_F^2 \;\leq\; \delta_2\epsilon\|\mathbf{B}\|_F^2;$$

as a consequence of Markov's inequality, it holds that

$$\mathbb{P}\Big\{\big\|\mathbf{U}\mathbf{B} - \mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{B}\big\|_F^2 \;\geq\; \epsilon\|\mathbf{B}\|_F^2\Big\} \;\leq\; \delta_2.$$

Here the expectation and probability are with respect to the randomness in $\mathbf{S}$.

Now we apply the above lemma to analyze **shrinked leverage score sampling**. Given the approximate shrinked leverage scores defined in (5), the sampling probabilities satisfy

$$p_i \;=\; \tfrac{1}{2}\big(\tfrac{1}{n} + \tfrac{\tilde{l}_i}{\sum_{q=1}^n \tilde{l}_q}\big) \;\geq\; \tfrac{\|\mathbf{u}_{i:}\|_2^2}{2\tau\rho}.$$

Here $\tilde{l}_i$ and τ are defined in (5). Thus for all $i \in [n]$, $\frac{\|\mathbf{u}_{i:}\|_2^2}{p_i} \leq 2\tau\rho$. We can then apply Lemma 24 to show that Assumption 1.1 holds with probability at least $1-\delta_1$ when $s \geq 2\tau\rho\frac{6+2\eta}{3\eta^2}\log\frac{\rho}{\delta_1}$ and that Assumption 1.2 holds with probability at least $1-\delta_2$ when $s \geq \frac{2\tau\rho}{\epsilon\delta_2}$.

A.1.2 Proof of Assumption 1.3

For uniform sampling (without replacement) and SRHT, when $s < n$, it is easy to show that $\mathbf{S}^T\mathbf{S} = \frac{n}{s}\mathbf{I}_s$, and thus $\|\mathbf{S}\|_2^2 = \frac{n}{s}$. Let $\{p_i^{\mathrm{s}}\}$ and $\{p_i^{\mathrm{u}}\}$ be the sampling probabilites of shrinked leverage score sampling and uniform sampling, respectively. Obviously $p_i^{\mathrm{s}} \geq \frac{1}{2}p_i^{\mathrm{u}}$. Thus for shrinked leverage score sampling, $\|\mathbf{S}\|_2^2 \leq \frac{2n}{s}$.

The greatest singular value of a standard Gaussian matrix $\mathbf{G} \in \mathbb{R}^{n\times s}$ is at most $\sqrt{n}+\sqrt{s}+t$ with probability at least $1-2e^{-t^2/2}$ (Vershynin, 2012). Thus a Gaussian projection matrix $\mathbf{S}$ satisfies

$$\|\mathbf{S}\|_2^2 \;=\; \frac{1}{s}\|\mathbf{G}\|_2^2 \;\leq\; \frac{(\sqrt{n}+\sqrt{s}+t)^2}{s}$$

with probability at least $1-2e^{-t^2/2}$.

If $\mathbf{S}$ is the CountSketch matrix, then each row of $\mathbf{S}$ has exactly one nonzero entry, either 1 or -1. Because the columns of $\mathbf{S}$ are orthogonal to each other, it holds that

$$\|\mathbf{S}\|_2^2 \;=\; \max_{i\in[s]}\|\mathbf{s}_{:i}\|_2^2 \;=\; \max_{i\in[s]}\mathrm{nnz}(\mathbf{s}_{:i}).$$

The problem of bounding $\mathrm{nnz}(\mathbf{s}_{:i})$ is equivalent to assigning n balls into s bins uniformly at random and bounding the number of balls in the bins. Patrascu and Thorup (2012) showed that for $s \ll n$, the maximal number of balls in any bin is at most $n/s + \mathcal{O}\big(\sqrt{n/s}\log^c n\big)$ with probability at least $1-\frac{1}{n}$, where $c = \mathcal{O}(1)$. Thus

$$\|\mathbf{S}\|_2^2 \;=\; \max_{i\in[s]}\mathrm{nnz}(\mathbf{s}_{:i}) \;\leq\; \frac{n}{s} + \mathcal{O}\bigg(\frac{\sqrt{n}\log^c n}{\sqrt{s}}\bigg) \;=\; \frac{n}{s}\big(1+o(1)\big)$$

holds with probability at least $1-\frac{1}{n}$.

A.2 Proof of Theorem 13

For uniform sampling (without replacement) and SRHT, it holds that $\mathbf{S}^T\mathbf{S} = \frac{n}{s}\mathbf{I}_s$.

For non-uniform sampling with probabilities $p_1, \cdots, p_n$, (with $\sum_i p_i = 1$), let $p_{\max} = \max_i p_i$. The smallest entry in $\mathbf{S}$ is $\frac{1}{\sqrt{sp_{\max}}}$, and thus $\mathbf{S}^T\mathbf{S} \succeq \frac{1}{sp_{\max}}\mathbf{I}_s$. For leverage score sampling, $p_{\max} = \frac{\mu}{n}$. For shrinked leverage score sampling, $p_{\max} = \frac{1+\mu}{2n}$. The lower bound on $\|\mathbf{S}\|_2^2$ is thus established.

The smallest singular value of any $n \times s$ standard Gaussian matrix $\mathbf{G}$ is at least $\sqrt{n}-\sqrt{s}-t$ with probability at least $1-2e^{-t^2/2}$ (Vershynin, 2012). Thus if $\mathbf{S} = \frac{1}{\sqrt{s}}\mathbf{G}$ is the

Gaussian projection matrix, the smallest eigenvalue of $\mathbf{S}^T\mathbf{S}$ is $(1-o(1))\frac{n}{s}$ with probability very close to one.

If $\mathbf{S}$ is the CountSketch matrix, then each row of $\mathbf{S}$ has exactly one nonzero entry, either 1 or -1. Because the columns of $\mathbf{S}$ are orthogonal to each other, it holds that

$$\sigma_{\min}^2(\mathbf{S}) \;=\; \min_{i\in[s]} \|\mathbf{s}_{:i}\|_2^2 \;=\; \min_{i\in[s]} \mathrm{nnz}(\mathbf{s}_{:i}).$$

The problem of bounding $\mathrm{nnz}(\mathbf{s}_{:i})$ is equivalent to assigning n balls into s bins uniformly at random and bounding the number of balls in the bins. Standard concentration arguments imply that each bin has at least $\frac{n}{s}(1-o(1))$ balls w.h.p., and hence $\sigma_{\min}^2(\mathbf{S}) \geq \frac{n}{s}(1-o(1))$ w.h.p.

A.3 Proof of Theorem 14

Assumption 2.1. By Theorem 12 and the union bound, we have that $\left\|\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T-\mathbf{I}_\rho\right\|_2 \leq \eta$ hold simultaneously for all $i \in [g]$ with probability at least $1-g\delta_1$. Because $\mathbf{S} \in \mathbb{R}^{n\times gs}$ is the same type of sketching matrix, it follows from Theorem 12 that $\left\|\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}-\mathbf{I}_\rho\right\|_2 \leq \frac{\eta}{\sqrt{g}}$ holds with probability at least $1-\delta_1$.

Assumption 2.2. By the same proof of Theorem 12, we can easily show that

$$\mathbb{E}\left\|\mathbf{U}^T\mathbf{B} - \mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{B}\right\|_F^2 \;\leq\; \delta_2\epsilon\,\|\mathbf{B}\|_F^2,$$

where $\mathbf{B}$ is any fixed matrix and the expectation is taken w.r.t. $\mathbf{S}$. It follows from Jensen's inequality that

$$\left(\mathbb{E}\left\|\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{B} - \mathbf{U}^T\mathbf{B}\right\|_F\right)^2 \;\leq\; \mathbb{E}\left\|\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{B} - \mathbf{U}^T\mathbf{B}\right\|_F^2 \;\leq\; \delta_2\epsilon\left\|\mathbf{B}\right\|_F^2.$$

It follows that

$$\frac{1}{g}\sum_{i=1}^{g}\mathbb{E}\left\|\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{B} - \mathbf{U}^T\mathbf{B}\right\|_F \;\leq\; \sqrt{\delta_2\epsilon}\left\|\mathbf{B}\right\|_F,$$

and thus

$$\Big(\frac{1}{g}\sum_{i=1}^{g}\mathbb{E}\left\|\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{B} - \mathbf{U}^T\mathbf{B}\right\|_F\Big)^2 \;\leq\; \delta_2\epsilon\left\|\mathbf{B}\right\|_F^2.$$

It follows from Markov's bound that

$$\mathbb{P}\bigg\{\Big(\frac{1}{g}\sum_{i=1}^{g}\left\|\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{B} - \mathbf{U}^T\mathbf{B}\right\|_F\Big)^2 \;\leq\; \epsilon\left\|\mathbf{B}\right\|_F^2\bigg\} \;\geq\; 1-\delta_2.$$

Because $\mathbf{S} \in \mathbb{R}^{n\times gs}$ is the same type of sketching matrix, it follows from Theorem 12 that $\left\|\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{B} - \mathbf{U}^T\mathbf{B}\right\|_F^2 \leq \frac{\epsilon}{g}\|\mathbf{B}\|_F^2$ holds with probability at least $1-\delta_2$.

Assumption 2.3. Theorem 12 shows that $\|\mathbf{S}_i\|_2^2$ can be bounded either surely or w.h.p. (assuming n is large enough). Because $g \ll n$, $\|\mathbf{S}_i\|_2^2$ can be bounded simultaneously for all $i \in [g]$ either surely or w.h.p.

Suppose $sg < n$. Because $\mathbf{S} \in \mathbb{R}^{n \times gs}$ is the same type of sketching matrix, it follows from Theorem 12 that $\|\mathbf{S}\|_2^2 \leq \frac{\theta n}{gs}$ holds either surely or w.h.p.

Suppose $sg \geq n$. It is not hard to show that uniform sampling, shrinked leverage score sampling, and SRHT satisfy $\|\mathbf{S}\|_2 = \Theta(1)$ w.h.p. Previously we have shown that a random Gaussian projection matrix $\mathbf{S} \in \mathbb{R}^{n \times sg}$ satisfies

$$\|\mathbf{S}\|_2^2 \;\leq\; \big(1 + o(1)\big)\, \frac{(\sqrt{n} + \sqrt{gs})^2}{gs}$$

w.h.p. Hence for $sg \geq n$, $\|\mathbf{S}\|_2^2 \leq 4 + o(1)$ w.h.p.

Appendix B. Sketched MRR from the Optimization Perspective: Proofs

In Section B.1 we establish a key lemma. In Section B.2 we prove Theorem 15. In Section B.3 we prove Theorem 16.

B.1 Key Lemma

Recall that the objective function of the matrix ridge regression (MRR) problem is

$$f(\mathbf{W}) \;\triangleq\; \frac{1}{n}\big\|\mathbf{X}\mathbf{W} - \mathbf{Y}\big\|_F^2 + \gamma \|\mathbf{W}\|_F^2.$$

The optimal solution is $\mathbf{W}^\star = \operatorname{argmin}_{\mathbf{W}} f(\mathbf{W})$. The following is the key lemma for understanding the difference between the objective value at $\mathbf{W}^\star$ and any arbitrary $\mathbf{W}$.

Lemma 25 *For any matrix $\mathbf{W}$ and any nonsingular matrix $\mathbf{M}$ of proper size, it holds that*

$$\begin{aligned}
f(\mathbf{W}) &= \frac{1}{n}\mathrm{tr}\Big[\mathbf{Y}^T\mathbf{Y} - (2\mathbf{W}^\star - \mathbf{W})^T(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_n)\mathbf{W}\Big], \\
f(\mathbf{W}^\star) &= \frac{1}{n}\Big[\big\|\mathbf{Y}^\perp\big\|_F^2 + n\gamma\big\|\big(\mathbf{\Sigma}^2 + n\gamma\mathbf{I}_\rho\big)^{-1/2}\mathbf{U}^T\mathbf{Y}\big\|_F^2\Big], \\
f(\mathbf{W}) - f(\mathbf{W}^\star) &= \frac{1}{n}\Big\|(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{1/2}(\mathbf{W} - \mathbf{W}^\star)\Big\|_F^2, \\
\Big\|\mathbf{M}^{-1}(\mathbf{W} - \mathbf{W}^\star)\Big\|_F^2 &\leq \sigma_{\min}^{-2}\Big[(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{1/2}\mathbf{M}\Big]\,\Big\|(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{1/2}(\mathbf{W} - \mathbf{W}^\star)\Big\|_F^2.
\end{aligned}$$

Here $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the SVD and $\mathbf{Y}^\perp = \mathbf{Y} - \mathbf{X}\mathbf{X}^\dagger\mathbf{Y}$.

Proof Let $\mathbf{U}$ be the left singular vectors of $\mathbf{X}$. The objective value $f(\mathbf{W})$ can be written as

$$\begin{aligned}
f(\mathbf{W}) &= \frac{1}{n}\big\|\mathbf{X}\mathbf{W} - \mathbf{Y}\big\|_F^2 + \gamma\big\|\mathbf{W}\big\|_F^2 \\
&= \frac{1}{n}\mathrm{tr}\Big[\mathbf{Y}^T\mathbf{Y} - (2\mathbf{W}^\star - \mathbf{W})^T(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_n)\mathbf{W}\Big],
\end{aligned}$$

so

$$\begin{aligned}
f(\mathbf{W}^\star) &= \frac{1}{n}\mathrm{tr}\Big[\mathbf{Y}^T\Big(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{-1}\mathbf{X}^T\Big)\mathbf{Y}\Big] \\
&= \frac{1}{n}\mathrm{tr}\Big[\mathbf{Y}^T\Big(\mathbf{I}_n - \mathbf{U}(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2})^{-1}\mathbf{U}^T\Big)\mathbf{Y}\Big] \\
&= \frac{1}{n}\mathrm{tr}\Big[\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{U}\mathbf{U}^T\mathbf{Y} + \mathbf{Y}^T\mathbf{U}\mathbf{U}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{U}(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2})^{-1}\mathbf{U}^T\mathbf{Y}\Big] \\
&= \frac{1}{n}\Big\{\mathrm{tr}\Big[\mathbf{Y}^T(\mathbf{I}_n - \mathbf{U}\mathbf{U}^T)\mathbf{Y}\Big] + n\gamma\cdot\mathrm{tr}\Big[\mathbf{Y}^T\mathbf{U}\big(\mathbf{\Sigma}^2 + n\gamma\mathbf{I}_\rho\big)^{-1}\mathbf{U}^T\mathbf{Y}\Big]\Big\} \\
&= \frac{1}{n}\Big[\big\|\mathbf{Y}^\perp\big\|_F^2 + n\gamma\big\|\big(\mathbf{\Sigma}^2 + n\gamma\mathbf{I}_\rho\big)^{-1/2}\mathbf{U}^T\mathbf{Y}\big\|_F^2\Big].
\end{aligned}$$

The difference in the objective values is therefore

$$\begin{aligned}
f(\mathbf{W}) - f(\mathbf{W}^\star) &= \frac{1}{n}\mathrm{tr}\Big[(\mathbf{W} - \mathbf{W}^\star)^T(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)(\mathbf{W} - \mathbf{W}^\star)\Big] \\
&= \frac{1}{n}\Big\|(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{1/2}(\mathbf{W} - \mathbf{W}^\star)\Big\|_F^2.
\end{aligned}$$

Because $\sigma_{\min}(\mathbf{A})\|\mathbf{B}\|_F \leq \|\mathbf{A}\mathbf{B}\|_F$ holds for any nonsingular $\mathbf{A}$ and any $\mathbf{B}$, it holds for any nonsingular matrix $\mathbf{M}$ that

$$\begin{aligned}
\sigma_{\min}^2\Big[(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{1/2}\mathbf{M}\Big]\Big\|\mathbf{M}^{-1}(\mathbf{W} - \mathbf{W}^\star)\Big\|_F^2 &\leq \Big\|(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{1/2}\mathbf{M}\mathbf{M}^{-1}(\mathbf{W} - \mathbf{W}^\star)\Big\|_F^2 \\
&= \Big\|(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{1/2}(\mathbf{W} - \mathbf{W}^\star)\Big\|_F^2.
\end{aligned}$$

The last claim in the lemma follows from the above inequality. ■

B.2 Proof of Theorem 15

Proof Let $\rho = \mathrm{rank}(\mathbf{X})$, $\mathbf{U} \in \mathbb{R}^{n\times\rho}$ be the left singular vectors of $\mathbf{X}$, and $\mathbf{Y}^\perp = \mathbf{Y} - \mathbf{X}\mathbf{X}^\dagger\mathbf{Y} = \mathbf{Y} - \mathbf{U}\mathbf{U}^T\mathbf{Y}$. It follows from the definition of $\mathbf{W}^\star$ and $\mathbf{W}^{\mathrm{c}}$ that

$$\mathbf{W}^{\mathrm{c}} - \mathbf{W}^\star = (\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{-1}\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{Y} - (\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{-1}\mathbf{X}^T\mathbf{Y}.$$

It follows that

$$\begin{aligned}
&(\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X} + n\gamma\mathbf{I}_d)(\mathbf{W}^{\mathrm{c}} - \mathbf{W}^\star) \\
&= \mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{Y}^\perp + \mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X}\mathbf{X}^\dagger\mathbf{Y} - (\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X} + n\gamma\mathbf{I}_d)(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{-1}\mathbf{X}^T\mathbf{Y} \\
&= \mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{Y}^\perp - n\gamma\mathbf{X}^\dagger\mathbf{Y} + (\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X} + n\gamma\mathbf{I}_d)\big[\mathbf{X}^\dagger - (\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{-1}\mathbf{X}^T\big]\mathbf{Y} \\
&= \mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{Y}^\perp - n\gamma\mathbf{X}^\dagger\mathbf{Y} + n\gamma(\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X} + n\gamma\mathbf{I}_d)(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{-1}\mathbf{X}^\dagger\mathbf{Y} \\
&= \mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{Y}^\perp + n\gamma(\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X} - \mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{-1}\mathbf{X}^\dagger\mathbf{Y}.
\end{aligned}$$

It follows that

$$(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{-1/2}(\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X} + n\gamma\mathbf{I}_d)(\mathbf{W}^{\mathrm{c}} - \mathbf{W}^\star) = \mathbf{A} + \mathbf{B}, \tag{13}$$

where

$$\begin{aligned}\mathbf{A} &= \big[(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^{1/2}\big]^\dagger\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{Y}^\perp \;=\; \mathbf{V}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\mathbf{\Sigma}\mathbf{U}\mathbf{S}\mathbf{S}^T\mathbf{Y}^\perp,\\ \mathbf{B} &= n\gamma\big[(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^{1/2}\big]^\dagger(\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X}-\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^\dagger\mathbf{X}^\dagger\mathbf{Y}\\ &= n\gamma\mathbf{V}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\mathbf{\Sigma}(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}-\mathbf{I}_\rho)\mathbf{\Sigma}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1}\mathbf{\Sigma}^{-1}\mathbf{U}^T\mathbf{Y}\\ &= n\gamma\mathbf{V}\mathbf{\Sigma}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}-\mathbf{I}_\rho)(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1}\mathbf{U}^T\mathbf{Y}.\end{aligned}$$

It follows from (13) that

$$\begin{aligned}&(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^{1/2}\big(\mathbf{W}^{\mathrm{c}}-\mathbf{W}^\star\big)\\ &= \big[(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^{-1/2}(\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X}+n\gamma\mathbf{I}_d)(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^{-1/2}\big]^\dagger\big(\mathbf{A}+\mathbf{B}\big).\end{aligned}$$

By Assumption 1.1, we have that

$$(1-\eta)(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d) \;\preceq\; (\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X}+n\gamma\mathbf{I}_d) \;\preceq\; (1+\eta)(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d).$$

It follows that

$$\Big\|\big[(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^{-1/2}(\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X}+n\gamma\mathbf{I}_d)(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^{-1/2}\big]^\dagger\Big\|_2 \;\le\; \frac{1}{1-\eta}.$$

Thus

$$\Big\|(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^{1/2}\big(\mathbf{W}^{\mathrm{c}}-\mathbf{W}^\star\big)\Big\|_F^2 \;\le\; \frac{1}{1-\eta}\Big\|\mathbf{A}+\mathbf{B}\Big\|_F^2 \;\le\; \frac{2}{1-\eta}\Big(\big\|\mathbf{A}\big\|_F^2+\big\|\mathbf{B}\big\|_F^2\Big).$$

Lemma 25 shows

$$f\big(\mathbf{W}^{\mathrm{c}}\big)-f\big(\mathbf{W}^\star\big)=\frac{1}{n}\Big\|(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^{1/2}(\mathbf{W}^{\mathrm{c}}-\mathbf{W}^\star)\Big\|_F^2\le\frac{2}{n(1-\eta)}\Big(\big\|\mathbf{A}\big\|_F^2+\big\|\mathbf{B}\big\|_F^2\Big). \quad (14)$$

We respectively bound $\|\mathbf{A}\|_F^2$ and $\|\mathbf{B}\|_F^2$ in the following. It follows from Assumption 1.2 and $\mathbf{U}^T\mathbf{Y}^\perp=\mathbf{0}$ that

$$\begin{aligned}\|\mathbf{A}\|_F^2 &= \Big\|\mathbf{V}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\mathbf{\Sigma}\mathbf{U}\mathbf{S}\mathbf{S}^T\mathbf{Y}^\perp\Big\|_F^2\\ &\le \big\|(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\mathbf{\Sigma}\big\|_2^2\,\big\|\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{Y}^\perp-\mathbf{U}^T\mathbf{Y}^\perp\big\|_F^2\\ &\le \epsilon\big\|(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\mathbf{\Sigma}\big\|_2^2\,\big\|\mathbf{Y}^\perp\big\|_F^2.\end{aligned}$$

By the definition of $\mathbf{B}$, we have

$$\begin{aligned}&\|\mathbf{B}\|_F^2 \;\le\; n^2\gamma^2\big\|\mathbf{\Sigma}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}-\mathbf{I}_\rho)(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1}\mathbf{U}^T\mathbf{Y}\big\|_F^2\\ &\le\; n^2\gamma^2\big\|\mathbf{\Sigma}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}-\mathbf{I}_\rho)(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\big\|_2^2\big\|(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\mathbf{U}^T\mathbf{Y}\big\|_F^2\\ &=\; n^2\gamma^2\big\|\mathbf{\Sigma}\mathbf{N}\big\|_2^2\big\|(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\mathbf{U}^T\mathbf{Y}\big\|_F^2,\end{aligned}$$

where we define $\mathbf{N}=(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}-\mathbf{I}_\rho)(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}$. By Assumption 1.1, we have

$$-\eta(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1} \;\preceq\; \mathbf{N} \;\preceq\; \eta(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1}.$$

It follows that

$$\begin{aligned}
\|\mathbf{B}\|_F^2 &\leq n^2\gamma^2\big\|\mathbf{\Sigma N}^2\mathbf{\Sigma}\big\|_2\big\|(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\mathbf{U}^T\mathbf{Y}\big\|_F^2 \\
&\leq \eta^2 n^2\gamma^2\big\|\mathbf{\Sigma}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-2}\mathbf{\Sigma}\big\|_2\big\|(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\mathbf{U}^T\mathbf{Y}\big\|_F^2 \\
&= \eta^2 n^2\gamma^2\big\|(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1}\mathbf{\Sigma}\big\|_2^2\big\|(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\mathbf{U}^T\mathbf{Y}\big\|_F^2 \\
&= \eta^2 n\gamma\big\|(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\mathbf{\Sigma}\big\|_2^2\big\|(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\mathbf{U}^T\mathbf{Y}\big\|_F^2.
\end{aligned}$$

The last equality follows from the fact that $\|(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\|_2 \leq (n\gamma)^{-1/2}$. It follows that

$$\begin{aligned}
\|\mathbf{A}\|_F^2+\|\mathbf{B}\|_F^2 &\leq \max\left\{\epsilon,\eta^2\right\}\Big\|(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_d)^{-1}\mathbf{\Sigma}\Big\|_2\Big[\big\|\mathbf{Y}^\perp\big\|_F^2+n\gamma\big\|(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_d)^{-1/2}\mathbf{U}^T\mathbf{Y}\big\|_F^2\Big] \\
&\leq \max\left\{\epsilon,\eta^2\right\}\frac{\sigma_{\max}^2}{\sigma_{\max}^2+n\gamma}\Big[\big\|\mathbf{Y}^\perp\big\|_F^2+n\gamma\big\|(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_d)^{-1/2}\mathbf{U}^T\mathbf{Y}\big\|_F^2\Big] \\
&\leq \max\left\{\epsilon,\eta^2\right\}\beta n f(\mathbf{W}^\star). \qquad (15)
\end{aligned}$$

The last inequality follows from Lemma 25. The claimed result now follows from (15) and (14). ■

B.3 Proof of Theorem 16

Proof By the definition of $\mathbf{W}^{\mathrm{h}}$ and $\mathbf{W}^\star$, we have

$$\begin{aligned}
&(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^{1/2}\big(\mathbf{W}^{\mathrm{h}}-\mathbf{W}^\star\big) \\
&= (\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^{1/2}\Big[(\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X}+n\gamma\mathbf{I}_d)^\dagger-(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^\dagger\Big]\mathbf{X}^T\mathbf{Y} \\
&= \mathbf{V}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{1/2}\Big[(\mathbf{\Sigma}\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}\mathbf{\Sigma}+n\gamma\mathbf{I}_\rho)^\dagger-(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1}\Big]\mathbf{\Sigma}\mathbf{U}^T\mathbf{Y}.
\end{aligned}$$

It follows from Assumption 1.1 that $\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}$ has full rank, and thus

$$\begin{aligned}
&(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^{1/2}\big(\mathbf{W}^{\mathrm{h}}-\mathbf{W}^\star\big) \\
&= \mathbf{V}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{1/2}\Big[(\mathbf{\Sigma}\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}\mathbf{\Sigma}+n\gamma\mathbf{I}_\rho)^{-1}-(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1}\Big]\mathbf{\Sigma}\mathbf{U}^T\mathbf{Y} \\
&= \mathbf{V}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{1/2}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1}(\mathbf{\Sigma}^2-\mathbf{\Sigma}\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}\mathbf{\Sigma})(\mathbf{\Sigma}\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}\mathbf{\Sigma}+n\gamma\mathbf{I}_\rho)^{-1}\mathbf{\Sigma}\mathbf{U}^T\mathbf{Y} \\
&= \mathbf{V}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\mathbf{\Sigma}(\mathbf{I}_\rho-\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U})\mathbf{\Sigma}(\mathbf{\Sigma}\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}\mathbf{\Sigma}+n\gamma\mathbf{I}_\rho)^{-1}\mathbf{\Sigma}\mathbf{U}^T\mathbf{Y},
\end{aligned}$$

where the second equality follow from $\mathbf{M}^{-1}-\mathbf{N}^{-1}=\mathbf{N}^{-1}(\mathbf{N}-\mathbf{M})\mathbf{M}^{-1}$. We define

$$(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^{1/2}\big(\mathbf{W}^{\mathrm{h}}-\mathbf{W}^\star\big) = \mathbf{VABC},$$

where

$$\begin{aligned}
\mathbf{A} &= (\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\mathbf{\Sigma}(\mathbf{I}_\rho-\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U})\mathbf{\Sigma}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}, \\
\mathbf{B} &= (\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{1/2}(\mathbf{\Sigma}\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}\mathbf{\Sigma}+n\gamma\mathbf{I}_\rho)^{-1}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{1/2}, \\
\mathbf{C} &= (\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\mathbf{\Sigma}\mathbf{U}^T\mathbf{Y}.
\end{aligned}$$

It follows from Assumption 1.1 that

$$\begin{aligned}\|\mathbf{A}\|_2 &\leq \eta\Big\|(\boldsymbol{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\Big\|_2 \leq \eta\beta,\\ \|\mathbf{B}\|_2 &\leq (1-\eta)^{-1}.\end{aligned}$$

It holds that

$$\begin{aligned}\big\|\mathbf{C}\big\|_F^2 &\leq \Big\|(\boldsymbol{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\boldsymbol{\Sigma}\mathbf{U}^T\mathbf{Y}\Big\|_F^2\\ &= \Big[\operatorname{tr}\Big(\mathbf{Y}^T\mathbf{U}\mathbf{U}^T\mathbf{Y}\Big)-n\gamma\operatorname{tr}\Big(\mathbf{Y}^T\mathbf{U}(\boldsymbol{\Sigma}^2+n\gamma\mathbf{I}_d)^{-1}\mathbf{U}^T\mathbf{Y}\Big)\Big]\\ &= \Big[-\operatorname{tr}\Big(\mathbf{Y}^T(\mathbf{I}_d-\mathbf{U}\mathbf{U}^T)\mathbf{Y}\Big)-n\gamma\operatorname{tr}\Big(\mathbf{Y}^T\mathbf{U}(\boldsymbol{\Sigma}^2+n\gamma\mathbf{I}_d)^{\dagger}\mathbf{U}^T\mathbf{Y}\Big)+\operatorname{tr}\big(\mathbf{Y}^T\mathbf{Y}\big)\Big]\\ &= \Big(-nf(\mathbf{W}^\star)+\big\|\mathbf{Y}\big\|_F^2\Big),\end{aligned}$$

where the last equality follows from Lemma 25. It follows from Lemma 25 that

$$\begin{aligned}f(\mathbf{W}^{\mathrm{h}})-f(\mathbf{W}^\star) &= \frac{1}{n}\big\|(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^{1/2}\big(\mathbf{W}^{\mathrm{h}}-\mathbf{W}^\star\big)\big\|_F^2\\ &= \frac{1}{n}\big\|\mathbf{A}\mathbf{B}\mathbf{C}\big\|_F^2 \leq \frac{\eta^2\beta^2}{(1-\eta)^2}\Big(\frac{1}{n}\big\|\mathbf{Y}\big\|_F^2-f(\mathbf{W}^\star)\Big).\end{aligned}$$

■

Appendix C. Sketched MRR from the Statistical Perspective: Proofs

In Section C.1 we prove Theorem 4. In Section C.2 we prove Theorem 17. In Section C.3 we prove Theorem 18. In Section A.2 we prove Theorem 13. In Section C.4 we prove Theorem 19. Recall that the fixed design model is $\mathbf{Y}=\mathbf{X}\mathbf{W}_0+\boldsymbol{\Xi}$ where $\boldsymbol{\Xi}$ is random, $\mathbb{E}\boldsymbol{\Xi}=0$, and $\mathbb{E}[\boldsymbol{\Xi}\boldsymbol{\Xi}^T]=\xi^2\mathbf{I}_n$.

C.1 Proofs of Theorem 4

We prove Theorem 4 in the following. In the proof we exploit several identities. The Frobenius norm and matrix trace satisfy

$$\|\mathbf{A}-\mathbf{B}\|_F^2 \;=\; \operatorname{tr}\big[(\mathbf{A}-\mathbf{B})(\mathbf{A}-\mathbf{B})^T\big) \;=\; \operatorname{tr}(\mathbf{A}\mathbf{A}^T)+\operatorname{tr}(\mathbf{B}\mathbf{B}^T)-2\operatorname{tr}(\mathbf{A}\mathbf{B}^T)$$

for any conformal matrices $\mathbf{A}$ and $\mathbf{B}$. The trace is linear, and thus for any fixed $\mathbf{A}$ and $\mathbf{B}$ and conformal random matrix $\boldsymbol{\Psi}$,

$$\mathbb{E}\big[\operatorname{tr}(\mathbf{A}\boldsymbol{\Psi}\mathbf{B})\big] \;=\; \operatorname{tr}\big[\mathbf{A}(\mathbb{E}\boldsymbol{\Psi})\mathbf{B}\big],$$

where the expectation is taken with respect to $\boldsymbol{\Psi}$.

Proof It follows from the definition of the optimal solution $\mathbf{W}^\star$ in (2) that

$$\begin{aligned}
\mathbf{X}\mathbf{W}^\star &= \mathbf{X}(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^\dagger\mathbf{X}^T(\mathbf{X}\mathbf{W}_0+\mathbf{\Xi}) \\
&= \mathbf{U}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1}\mathbf{\Sigma}^3\mathbf{V}^T\mathbf{W}_0+\mathbf{U}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1}\mathbf{\Sigma}^2\mathbf{U}^T\mathbf{\Xi} \\
&= \mathbf{U}\Big[\mathbf{I}_\rho-n\gamma(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1}\Big]\mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_0+\mathbf{U}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1}\mathbf{\Sigma}^2\mathbf{U}^T\mathbf{\Xi} \\
&= \mathbf{X}\mathbf{W}_0-n\gamma\mathbf{U}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1}\mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_0+\mathbf{U}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1}\mathbf{\Sigma}^2\mathbf{U}^T\mathbf{\Xi}.
\end{aligned}$$

Since $\mathbb{E}[\mathbf{\Xi}]=\mathbf{0}$ and $\mathbb{E}[\mathbf{\Xi}\mathbf{\Xi}^T]=\xi^2\mathbf{I}_n$, it holds that

$$\begin{aligned}
R(\mathbf{W}^\star) &= \frac{1}{n}\mathbb{E}\big\|\mathbf{X}\mathbf{W}^\star-\mathbf{X}\mathbf{W}_0\big\|_F^2 \\
&= \frac{1}{n}\Big\|-n\gamma(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1}\mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_0+(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1}\mathbf{\Sigma}^2\mathbf{U}^T\mathbf{\Xi}\Big\|_F^2 \\
&= n\gamma^2\Big\|(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1}\mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_0\Big\|_F^2+\frac{\xi^2}{n}\Big\|(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1}\mathbf{\Sigma}^2\Big\|_F^2.
\end{aligned}$$

This exposes expressions for the bias and variance of the optimal solution $\mathbf{W}^\star$.

We now decompose the risk function $R\big(\mathbf{W}^{\mathrm{c}}\big)$. It follows from the definition of $\mathbf{W}^{\mathrm{c}}$ in (3) that

$$\begin{aligned}
\mathbf{X}\mathbf{W}^{\mathrm{c}} &= \mathbf{X}(\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X}+n\gamma\mathbf{I}_d)^\dagger\mathbf{X}^T\mathbf{S}\mathbf{S}^T(\mathbf{X}\mathbf{W}_0+\mathbf{\Xi}) \\
&= \mathbf{U}\mathbf{\Sigma}\big(\mathbf{\Sigma}\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}\mathbf{\Sigma}+n\gamma\mathbf{I}_d\big)^\dagger\mathbf{\Sigma}\big(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_0+\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{\Xi}\big) \\
&= \mathbf{U}(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}+n\gamma\mathbf{\Sigma}^{-2})^{-1}\Big[(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}+n\gamma\mathbf{\Sigma}^{-2})\mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_0-n\gamma\mathbf{\Sigma}^{-1}\mathbf{V}^T\mathbf{W}_0+\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{\Xi}\Big] \\
&= \mathbf{X}\mathbf{W}_0+\mathbf{U}(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}+n\gamma\mathbf{\Sigma}^{-2})^{-1}\big(-n\gamma\mathbf{\Sigma}^{-1}\mathbf{V}^T\mathbf{W}_0+\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{\Xi}\big).
\end{aligned}$$

Since $\mathbb{E}[\mathbf{\Xi}]=\mathbf{0}$ and $\mathbb{E}[\mathbf{\Xi}\mathbf{\Xi}^T]=\xi^2\mathbf{I}_n$, it follows that

$$\begin{aligned}
R\big(\mathbf{W}^{\mathrm{c}}\big) &= \frac{1}{n}\mathbb{E}\big\|\mathbf{X}\mathbf{W}^{\mathrm{c}}-\mathbf{X}\mathbf{W}_0\big\|_F^2 \\
&= \frac{1}{n}\Big\|-n\gamma(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}+n\gamma\mathbf{\Sigma}^{-2})^{-1}\mathbf{\Sigma}^{-1}\mathbf{V}^T\mathbf{W}_0+(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}+n\gamma\mathbf{\Sigma}^{-2})^{-1}\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{\Xi}\Big\|_F^2 \\
&= n\gamma^2\Big\|(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}+n\gamma\mathbf{\Sigma}^{-2})^{-1}\mathbf{\Sigma}^{-1}\mathbf{V}^T\mathbf{W}_0\Big\|_F^2+\frac{\xi^2}{n}\Big\|(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}+n\gamma\mathbf{\Sigma}^{-2})^{-1}\mathbf{U}^T\mathbf{S}\mathbf{S}^T\Big\|_F^2.
\end{aligned}$$

This exposes expressions for the bias and variance of the approximate solution $\mathbf{W}^{\mathrm{c}}$.

We now decompose the risk function $R\big(\mathbf{W}^{\mathrm{h}}\big)$. It follows from the definition of $\mathbf{W}^{\mathrm{h}}$ in (4) that

$$\begin{aligned}
\mathbf{X}\mathbf{W}^{\mathrm{h}}-\mathbf{X}\mathbf{W}_0 &= \mathbf{X}(\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X}+n\gamma\mathbf{I}_n)^\dagger\mathbf{X}^T(\mathbf{X}\mathbf{W}_0+\mathbf{\Xi})-\mathbf{X}\mathbf{W}_0 \\
&= \mathbf{X}(\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X}+n\gamma\mathbf{I}_d)^\dagger\mathbf{X}^T\mathbf{X}\mathbf{W}_0-\mathbf{X}\mathbf{W}_0+\mathbf{X}(\mathbf{X}^T\mathbf{S}\mathbf{S}^T\mathbf{X}+n\gamma\mathbf{I}_d)^\dagger\mathbf{X}^T\mathbf{\Xi} \\
&= \mathbf{U}\big[(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}+n\gamma\mathbf{\Sigma}^{-2})^{-1}-\mathbf{I}_\rho^{-1}\big]\mathbf{U}^T\mathbf{X}\mathbf{W}_0+\mathbf{U}(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}^T+n\gamma\mathbf{\Sigma}^{-2})^\dagger\mathbf{U}^T\mathbf{\Xi} \\
&= \mathbf{U}\big(\mathbf{I}_\rho-\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}-n\gamma\mathbf{\Sigma}^{-2}\big)\big(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}+n\gamma\mathbf{\Sigma}^{-2}\big)^{-1}\mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_0 \\
&\quad+\mathbf{U}(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}^T+n\gamma\mathbf{\Sigma}^{-2})^\dagger\mathbf{U}^T\mathbf{\Xi},
\end{aligned}$$

where the last equality follows from the fact that $\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{B}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{A}^{-1}$ for any conformal nonsingular matrices $\mathbf{A}$ and $\mathbf{B}$. Since $\mathbb{E}[\boldsymbol{\Xi}] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{\Xi}\boldsymbol{\Xi}^T] = \xi^2\mathbf{I}_n$, it follows that

$$R\big(\mathbf{W}^{\mathrm{h}}\big) \;=\; \mathsf{bias}^2\big(\mathbf{W}^{\mathrm{h}}\big) + \mathsf{var}\big(\mathbf{W}^{\mathrm{h}}\big),$$

where

$$\begin{aligned}
\mathsf{bias}^2\big(\mathbf{W}^{\mathrm{h}}\big) &= \frac{1}{n}\Big\|\big(n\gamma\boldsymbol{\Sigma}^{-2} + \mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} - \mathbf{I}_\rho\big)\big(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\boldsymbol{\Sigma}^{-2}\big)^{-1}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{W}_0\Big\|_F^2, \\
\mathsf{var}\big(\mathbf{W}^{\mathrm{h}}\big) &= \frac{\xi^2}{n}\Big\|\big(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\boldsymbol{\Sigma}^{-2}\big)^{-1}\Big\|_F^2.
\end{aligned}$$

This exposes expressions for the bias and variance of $\mathbf{W}^{\mathrm{h}}$. ∎

C.2 Proof of Theorem 17

Proof Assumption 1.1 ensures that $(1-\eta)\mathbf{I}_\rho \preceq \mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} \preceq (1+\eta)\mathbf{I}_\rho$. It follows that

$$(1-\eta)\big(\mathbf{I}_\rho + n\gamma\boldsymbol{\Sigma}^{-2}\big) \;\preceq\; \mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\boldsymbol{\Sigma}^{-2} \;\preceq\; (1+\eta)\big(\mathbf{I}_\rho + n\gamma\boldsymbol{\Sigma}^{-2}\big).$$

The bias term can be written as

$$\begin{aligned}
\mathsf{bias}^2\big(\mathbf{W}^{\mathrm{c}}\big) &= n\gamma^2\big\|\big(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\boldsymbol{\Sigma}^{-2}\big)^\dagger\boldsymbol{\Sigma}^{-1}\mathbf{V}^T\mathbf{W}_0\big\|_F^2 \\
&= n\gamma^2\,\mathrm{tr}\Big(\mathbf{W}_0^T\mathbf{V}\boldsymbol{\Sigma}^{-1}\big[(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\boldsymbol{\Sigma}^{-2})^\dagger\big]^2\boldsymbol{\Sigma}^{-1}\mathbf{V}^T\mathbf{W}_0\Big) \\
&\le \tfrac{n\gamma^2}{(1-\eta)^2}\big\|\big(\mathbf{I}_\rho + n\gamma\boldsymbol{\Sigma}^{-2}\big)^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{V}^T\mathbf{W}_0\big\|_F^2 \\
&= \tfrac{n\gamma^2}{(1-\eta)^2}\big\|\big(\boldsymbol{\Sigma}^2 + n\gamma\mathbf{I}_\rho\big)^{-1}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{W}_0\big\|_F^2 \\
&= \tfrac{1}{(1-\eta)^2}\,\mathsf{bias}^2(\mathbf{W}^\star).
\end{aligned}$$

We can analogously show $\mathsf{bias}^2(\mathbf{W}^{\mathrm{c}}) \ge \frac{1}{(1+\eta)^2}\mathsf{bias}^2(\mathbf{W}^\star)$.

Let $\mathbf{B} = \big(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\boldsymbol{\Sigma}^{-2}\big)^\dagger\mathbf{U}^T\mathbf{S} \in \mathbb{R}^{\rho\times s}$. By Assumption 1.1, it holds that

$$(1-\eta)\big[\big(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\boldsymbol{\Sigma}^{-2}\big)^2\big]^\dagger \;\preceq\; \mathbf{B}\mathbf{B}^T \;\preceq\; (1+\eta)\big[\big(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\boldsymbol{\Sigma}^{-2}\big)^2\big]^\dagger.$$

Applying Assumption 1.1 again, we obtain

$$(1-\eta)^2\big(\mathbf{I}_\rho + n\gamma\boldsymbol{\Sigma}^{-2}\big)^2 \;\preceq\; \big(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\boldsymbol{\Sigma}^{-2}\big)^2 \;\preceq\; (1+\eta)^2\big(\mathbf{I}_\rho + n\gamma\boldsymbol{\Sigma}^{-2}\big)^2.$$

Note that both sides are nonsingular. Combining the above two equations, we have

$$\tfrac{1-\eta}{(1+\eta)^2}\big(\mathbf{I}_\rho + n\gamma\boldsymbol{\Sigma}^{-2}\big)^{-2} \;\preceq\; \mathbf{B}\mathbf{B}^T \;\preceq\; \tfrac{1+\eta}{(1-\eta)^2}\big(\mathbf{I}_\rho + n\eta\boldsymbol{\Sigma}^{-2}\big)^{-2}.$$

Taking the trace of all the terms, we obtain

$$\tfrac{1-\eta}{(1+\eta)^2} \;\le\; \tfrac{\|\mathbf{B}\|_F^2}{\|(\mathbf{I}_\rho + n\gamma\boldsymbol{\Sigma}^{-2})^{-1}\|_F^2} \;\le\; \tfrac{1+\eta}{(1-\eta)^2}.$$

The variance term can be written as

$$\begin{aligned}\mathsf{var}\big(\mathbf{W}^{\mathrm{c}}\big) &= \tfrac{\xi^2}{n}\big\|\mathbf{B}\mathbf{S}^T\big\|_F^2 \;\le\; \tfrac{\xi^2}{n}\big\|\mathbf{B}\big\|_F^2\,\big\|\mathbf{S}\big\|_2^2 \\ &\le \tfrac{\xi^2(1+\eta)}{n(1-\eta)^2}\big\|\big(\mathbf{I}_\rho + n\gamma\boldsymbol{\Sigma}^{-2}\big)^{-1}\big\|_F^2\,\big\|\mathbf{S}\big\|_2^2 \\ &= \tfrac{(1+\eta)\|\mathbf{S}\|_2^2}{(1-\eta)^2}\mathsf{var}(\mathbf{W}^\star).\end{aligned}$$

The upper bound on the variance follows from Assumption 1.3. ■

C.3 Proof of Theorem 18

Proof Let $\mathbf{B} = \big(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\boldsymbol{\Sigma}^{-2}\big)^\dagger\mathbf{U}^T\mathbf{S} \in \mathbb{R}^{\rho\times s}$. In the proof of Theorem 5 we show that

$$\mathsf{var}\big(\mathbf{W}^{\mathrm{c}}\big) \;=\; \tfrac{\xi^2}{n}\big\|\mathbf{B}\mathbf{S}^T\big\|_F^2.$$

If $\mathbf{S}^T\mathbf{S} \succeq \frac{\vartheta n}{s}\mathbf{I}_s$, then it holds that

$$\mathsf{var}\big(\mathbf{W}^{\mathrm{c}}\big) \;=\; \tfrac{\xi^2}{n}\big\|\mathbf{B}\mathbf{S}^T\big\|_F^2 \;\ge\; \tfrac{\vartheta n}{s}\tfrac{\xi^2}{n}\big\|\mathbf{B}\big\|_F^2 \;\ge\; \tfrac{\vartheta n}{s}\tfrac{1-\eta}{(1+\eta)^2}\mathsf{var}(\mathbf{W}^\star).$$

This establishes the lower bounds on the variance. ■

C.4 Proof of Theorem 19

Proof Theorem 4 shows that

$$\begin{aligned}\mathsf{bias}\big(\mathbf{W}^{\mathrm{h}}\big) &= \gamma\sqrt{n}\left\|\Big(\boldsymbol{\Sigma}^{-2} + \tfrac{\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}-\mathbf{I}_\rho}{n\gamma}\Big)\big(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\boldsymbol{\Sigma}^{-2}\big)^\dagger\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{W}_0\right\|_F \\ &= \gamma\sqrt{n}\big\|\mathbf{A}\boldsymbol{\Sigma}^2\mathbf{B}\big\|_F \;\le\; \gamma\sqrt{n}\big\|\mathbf{A}\boldsymbol{\Sigma}^2\big\|_2\big\|\mathbf{B}\big\|_F, \\ \mathsf{var}\big(\mathbf{W}^{\mathrm{h}}\big) &= \frac{\xi^2}{n}\Big\|\big(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\boldsymbol{\Sigma}^{-2}\big)^\dagger\Big\|_F^2,\end{aligned}$$

where we define

$$\begin{aligned}\mathbf{A} &= \boldsymbol{\Sigma}^{-2} + \tfrac{\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}-\mathbf{I}_\rho}{n\gamma}, \\ \mathbf{B} &= \boldsymbol{\Sigma}^{-2}\big(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\boldsymbol{\Sigma}^{-2}\big)^\dagger\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{W}_0.\end{aligned}$$

We first analyze the bias. It follows from Assumption 1.1 that

$$\boldsymbol{\Sigma}^{-2}\big(\mathbf{I}_\rho - \tfrac{\eta}{n\gamma}\boldsymbol{\Sigma}^2\big) \;\preceq\; \mathbf{A} \;\preceq\; \boldsymbol{\Sigma}^{-2}\big(\mathbf{I}_\rho + \tfrac{\eta}{n\gamma}\boldsymbol{\Sigma}^2\big). \tag{16}$$

Since $\big(\mathbf{I}_\rho - \frac{\eta}{n\gamma}\boldsymbol{\Sigma}^2\big)^2 \preceq \big(\mathbf{I}_\rho + \frac{\eta}{n\gamma}\boldsymbol{\Sigma}^2\big)^2 \preceq \big(1 + \frac{\eta\sigma_1^2}{n\gamma}\big)^2\mathbf{I}_\rho$, it follows that

$$\mathbf{A}^2 \;\preceq\; \boldsymbol{\Sigma}^{-4}\big(\mathbf{I}_\rho + \tfrac{\eta}{n\gamma}\boldsymbol{\Sigma}^2\big)^2 \;\preceq\; \big(1 + \tfrac{\eta\sigma_1^2}{n\gamma}\big)^2\boldsymbol{\Sigma}^{-4}.$$

Thus

$$\big\|\mathbf{A}\mathbf{\Sigma}^2\big\|_2^2 \;=\; \big\|\mathbf{\Sigma}^2\mathbf{A}^2\mathbf{\Sigma}^2\big\|_2 \;\leq\; \Big(1+\tfrac{\eta\sigma_1^2}{n\gamma}\Big)^2.$$

It follows from Assumption 1.1 that

$$\begin{aligned}&(1+\eta)^{-1}\big(\mathbf{I}_\rho+n\gamma\mathbf{\Sigma}^{-2}\big)^{-1} \;\preceq\; \big((1+\eta)\mathbf{I}_\rho+n\gamma\mathbf{\Sigma}^{-2}\big)^{-1}\\ &\preceq\; \big(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}+n\gamma\mathbf{\Sigma}^{-2}\big)^{\dagger} \;\preceq\; \big((1-\eta)\mathbf{I}_\rho+n\gamma\mathbf{\Sigma}^{-2}\big)^{-1} \;\preceq\; (1-\eta)^{-1}\big(\mathbf{I}_\rho+n\gamma\mathbf{\Sigma}^{-2}\big)^{-1}.\end{aligned}$$

Thus

$$\begin{aligned}\mathbf{B}^T\mathbf{B} &= \mathbf{W}_0^T\mathbf{V}\mathbf{\Sigma}^3\big(\mathbf{\Sigma}^{-2}(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}+n\gamma\mathbf{\Sigma}^{-2})^{\dagger}\mathbf{\Sigma}^{-2}\big)^2\mathbf{\Sigma}^3\mathbf{V}^T\mathbf{W}_0\\ &\preceq (1-\eta)^{-2}\mathbf{W}_0^T\mathbf{V}\mathbf{\Sigma}^3\big(\mathbf{\Sigma}^{-2}(\mathbf{I}_\rho+n\gamma\mathbf{\Sigma}^{-2})^{-1}\mathbf{\Sigma}^{-2}\big)^2\mathbf{\Sigma}^3\mathbf{V}^T\mathbf{W}_0\\ &= (1-\eta)^{-2}\mathbf{W}_0^T\mathbf{V}\mathbf{\Sigma}\big(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho\big)^{-2}\mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_0. \qquad (17)\end{aligned}$$

It follows that

$$\|\mathbf{B}\|_F^2 \;=\; \mathrm{tr}\big(\mathbf{B}^T\mathbf{B}\big) \;\leq\; (1-\eta)^{-2}\big\|\big(\mathbf{\Sigma}^{-2}+n\gamma\mathbf{I}_\rho\big)^{-1}\mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_0\big\|_F^2 \;=\; \tfrac{\mathsf{bias}^2(\mathbf{W}^\star)}{n\gamma^2(1-\eta)^2},$$

where the last equality follows from the definition of $\mathsf{bias}(\mathbf{W}^\star)$. By the definition of $\mathbf{A}$ and $\mathbf{B}$, we have

$$\mathsf{bias}^2\big(\mathbf{W}^{\mathrm{h}}\big) \;\leq\; \gamma^2 n\,\big\|\mathbf{A}\mathbf{\Sigma}^2\big\|_2^2\|\mathbf{B}\|_F^2 \;=\; \tfrac{1}{(1-\eta)^2}\,\big(1+\tfrac{\eta\sigma_1^2}{n\gamma}\big)^2\,\mathsf{bias}^2\big(\mathbf{W}^\star\big).$$

Thus, the upper bound on $\mathsf{bias}\big(\mathbf{W}^{\mathrm{h}}\big)$ is established.

Using the same $\mathbf{A}$ and $\mathbf{B}$, we can also show that

$$\mathsf{bias}\big(\mathbf{W}^{\mathrm{h}}\big) \;=\; \gamma\sqrt{n}\big\|\mathbf{A}\mathbf{\Sigma}^2\mathbf{B}\big\|_F \;\geq\; \gamma\sqrt{n}\;\sigma_{\min}\big(\mathbf{A}\mathbf{\Sigma}^2\big)\,\big\|\mathbf{B}\big\|_F.$$

Assume that $\sigma_\rho^2 \geq \frac{n\gamma}{\eta}$. It follows from (16) that

$$\mathbf{A}^2 \;\succeq\; \big(\tfrac{\eta\sigma_\rho^2}{n\gamma}-1\big)^2\mathbf{\Sigma}^{-4}.$$

Thus

$$\sigma_{\min}^2\big(\mathbf{A}\mathbf{\Sigma}^2\big) \;=\; \sigma_{\min}\big(\mathbf{\Sigma}^2\mathbf{A}^2\mathbf{\Sigma}^2\big) \;\geq\; \big(\tfrac{\eta\sigma_\rho^2}{n\gamma}-1\big)^2.$$

It follows from (17) that

$$\mathbf{B}^T\mathbf{B} \;\succeq\; (1+\eta)^{-2}\mathbf{W}_0^T\mathbf{V}\mathbf{\Sigma}\big(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho\big)^{-2}\mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_0.$$

Thus

$$\|\mathbf{B}\|_F^2 \;=\; \mathrm{tr}\big(\mathbf{B}^T\mathbf{B}\big) \;\geq\; (1+\eta)^{-2}\big\|\big(\mathbf{\Sigma}^{-2}+n\gamma\mathbf{I}_\rho\big)^{-1}\mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_0\big\|_F^2 \;=\; \tfrac{1}{n\gamma^2(1+\eta)^2}\,\mathsf{bias}^2\big(\mathbf{W}^\star\big).$$

In sum, we obtain

$$\mathsf{bias}^2\big(\mathbf{W}^{\mathrm{h}}\big) \;\geq\; \gamma^2 n\;\sigma_{\min}^2\big(\mathbf{A}\mathbf{\Sigma}^2\big)\,\big\|\mathbf{B}\big\|_F^2 \;=\; (1+\eta)^{-2}\,\big(\tfrac{\eta\sigma_\rho^2}{n\gamma}-1\big)^2\,\mathsf{bias}^2\big(\mathbf{W}^\star\big).$$

Thus, the lower bound on $\mathsf{bias}\big(\mathbf{W}^{\mathrm{h}}\big)$ is established.

It follows from Assumption 1.1 that

$$(1+\eta)^{-1}\left(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2}\right)^{-1} \preceq \left(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2}\right)^{-1} \preceq (1-\eta)^{-1}\left(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2}\right)^{-1}.$$

It follows from Theorem 4 that

$$\begin{aligned}\mathsf{var}\big(\mathbf{W}^{\mathrm{h}}\big) &= \tfrac{\xi^2}{n}\left\|\left(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2}\right)^{-1}\right\|_F^2 \\ &\in \tfrac{1}{1\mp\eta}\tfrac{\xi^2}{n}\left\|\left(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2}\right)^{-1}\right\|_F^2 \\ &= \tfrac{1}{1\mp\eta}\mathsf{var}\big(\mathbf{W}^\star\big).\end{aligned}$$

This concludes the proof. ■

Appendix D. Model Averaging from the Optimization Perspective: Proofs

In Section D.1 we prove Theorem 20. In Section D.2 we prove Theorem 21.

D.1 Proof of Theorem 20

Proof By Lemma 25, we only need to show that $\|(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{1/2}(\mathbf{W}^{\mathrm{c}} - \mathbf{W}^\star)\|_F^2 \leq n\alpha\beta f(\mathbf{W}^\star)$. In the proof, we define $\rho = \mathrm{rank}(\mathbf{X})$ and let $\sigma_1 \geq \cdots \geq \sigma_\rho$ be the singular values of $\mathbf{X}$.

In the proof of Theorem 15 we show that

$$\begin{aligned}&(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{1/2}\big(\mathbf{W}_i^{\mathrm{c}} - \mathbf{W}^\star\big) \\ &= \left[(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{-1/2}(\mathbf{X}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{X} + n\gamma\mathbf{I}_d)(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{-1/2}\right]^\dagger\big(\mathbf{A}_i + \mathbf{B}_i\big) \\ &= \mathbf{C}_i^\dagger\big(\mathbf{A}_i + \mathbf{B}_i\big),\end{aligned}$$

where

$$\begin{aligned}\mathbf{A}_i &= \mathbf{V}(\mathbf{\Sigma}^2 + n\gamma\mathbf{I}_\rho)^{-1/2}\mathbf{\Sigma}\mathbf{U}\mathbf{S}_i\mathbf{S}_i^T\mathbf{Y}^\perp, \\ \mathbf{B}_i &= n\gamma\mathbf{V}\mathbf{\Sigma}(\mathbf{\Sigma}^2 + n\gamma\mathbf{I}_\rho)^{-1/2}(\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} - \mathbf{I}_\rho)(\mathbf{\Sigma}^2 + n\gamma\mathbf{I}_\rho)^{-1}\mathbf{U}^T\mathbf{Y} \\ \mathbf{C}_i &= \left[(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{1/2}\right]^\dagger\big(\mathbf{X}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{X} + n\gamma\mathbf{I}_d\big)\left[(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{1/2}\right]^\dagger \\ &= \mathbf{V}(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2})^{-1/2}(\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2})(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2})^{-1/2}\mathbf{V}^T \\ &= \mathbf{V}\mathbf{V}^T + \mathbf{V}(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2})^{-1/2}(\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} - \mathbf{I}_\rho)(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2})^{-1/2}\mathbf{V}^T.\end{aligned}$$

By Assumption 2.1, we have that $\mathbf{C}_i \succeq \big(1 - \frac{\eta\,\sigma_{\max}^2}{\sigma_{\max}^2 + n\gamma}\big)\mathbf{V}\mathbf{V}^T$. Since $\eta \leq 1/2$, it follows that $\mathbf{C}_i^\dagger \preceq \big(1 + \frac{2\eta\,\sigma_{\max}^2}{\sigma_{\max}^2 + n\gamma}\big)\mathbf{V}\mathbf{V}^T$. Let $\mathbf{C}_i^\dagger = \mathbf{V}\mathbf{V}^T + \mathbf{V}\mathbf{\Delta}_i\mathbf{V}^T$. It holds that $\mathbf{\Delta}_i \preceq \frac{2\eta\,\sigma_{\max}^2}{\sigma_{\max}^2 + n\gamma}\,\mathbf{V}\mathbf{V}^T \preceq$

$2\eta\beta\mathbf{V}\mathbf{V}^T$. By definition, $\mathbf{W}^{\mathrm{c}} = \frac{1}{g}\sum_{i=1}^g \mathbf{W}_i^{\mathrm{c}}$. It follows that

$$\begin{aligned}
\Big\|(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^{1/2}(\mathbf{W}_i^{\mathrm{c}}-\mathbf{W}^\star)\Big\|_F &= \Big\|\frac{1}{g}\sum_{i=1}^g \mathbf{C}_i^\dagger(\mathbf{A}_i+\mathbf{B}_i)\Big\|_F \\
&\le \Big\|\frac{1}{g}\sum_{i=1}^g(\mathbf{A}_i+\mathbf{B}_i)\Big\|_F + \Big\|\frac{1}{g}\sum_{i=1}^g \mathbf{V}\boldsymbol{\Delta}_i\mathbf{V}^T(\mathbf{A}_i+\mathbf{B}_i)\Big\|_F \\
&\le \Big\|\frac{1}{g}\sum_{i=1}^g \mathbf{A}_i\Big\|_F + \Big\|\frac{1}{g}\sum_{i=1}^g \mathbf{B}_i\Big\|_F + \frac{1}{g}\sum_{i=1}^g \|\boldsymbol{\Delta}_i\|_2\Big(\|\mathbf{A}_i\|_F+\|\mathbf{B}_i\|_F\Big) \\
&\le \Big\|\frac{1}{g}\sum_{i=1}^g \mathbf{A}_i\Big\|_F + \Big\|\frac{1}{g}\sum_{i=1}^g \mathbf{B}_i\Big\|_F + 2\eta\beta\frac{1}{g}\sum_{i=1}^g\Big(\|\mathbf{A}_i\|_F+\|\mathbf{B}_i\|_F\Big). \qquad (18)
\end{aligned}$$

By Assumption 2.3, we have that

$$\frac{1}{g}\sum_{i=1}^g \|\mathbf{A}_i\|_F = \|(\boldsymbol{\Sigma}^2+n\gamma\mathbf{I}_d)^{-1/2}\boldsymbol{\Sigma}\|_2 \cdot \frac{1}{g}\sum_{i=1}^g \|\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{Y}^\perp\|_F \le \sqrt{\tfrac{\epsilon\,\sigma_{\max}^2}{\sigma_{\max}^2+n\gamma}}\|\mathbf{Y}^\perp\|_F.$$

We apply Assumption 2.1 and follow the proof of Theorem 15 to show that

$$\|\mathbf{B}_i\|_F^2 \le \eta^2 n\gamma\tfrac{\sigma_{\max}^2}{\sigma_{\max}^2+n\gamma}\Big\|(\boldsymbol{\Sigma}^2+n\gamma\mathbf{I}_d)^{-1/2}\mathbf{U}^T\mathbf{Y}\Big\|_F^2.$$

It follows that

$$\begin{aligned}
&\frac{1}{g}\sum_{i=1}^g\Big(\|\mathbf{A}_i\|_F+\|\mathbf{B}_i\|_F\Big) \\
&\le \max\Big\{\sqrt{\epsilon},\eta\Big\}\sqrt{\tfrac{\sigma_{\max}^2}{\sigma_{\max}^2+n\gamma}}\Big(\|\mathbf{Y}^\perp\|_F+\sqrt{n\gamma}\|(\boldsymbol{\Sigma}^2+n\gamma\mathbf{I}_d)^{-1/2}\mathbf{U}^T\mathbf{Y}\|_F\Big) \\
&\le \max\Big\{\sqrt{\epsilon},\eta\Big\}\sqrt{\beta}\sqrt{2\|\mathbf{Y}^\perp\|_F^2+2n\gamma\|(\boldsymbol{\Sigma}^2+n\gamma\mathbf{I}_d)^{-1/2}\mathbf{U}^T\mathbf{Y}\|_F^2} \\
&= \max\big\{\sqrt{\epsilon},\eta\big\}\sqrt{\beta}\,\sqrt{2n\,f(\mathbf{W}^\star)}. \qquad (19)
\end{aligned}$$

Here the equality follows from Lemma 25. Let $\mathbf{S} = \frac{1}{g}[\mathbf{S}_1,\cdots,\mathbf{S}_g] \in \mathbb{R}^{n\times sg}$. We have that

$$\begin{aligned}
\frac{1}{g}\sum_{i=1}^g \mathbf{A}_i &= \mathbf{V}(\boldsymbol{\Sigma}^2+n\gamma\mathbf{I}_d)^{-1/2}\boldsymbol{\Sigma}\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{Y}^\perp, \\
\frac{1}{g}\sum_{i=1}^g \mathbf{B}_i &= n\gamma\mathbf{V}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^2+n\gamma\mathbf{I}_d)^{-1/2}(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}-\mathbf{I}_\rho)(\boldsymbol{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1}\mathbf{U}^T\mathbf{Y}.
\end{aligned}$$

Applying Assumptions 2.1 and 2.2, we use the same techniques as in the above to obtain

$$\begin{aligned}
&\Big\|\frac{1}{g}\sum_{i=1}^g \mathbf{A}_i\Big\|_F + \Big\|\frac{1}{g}\sum_{i=1}^g \mathbf{B}_i\Big\|_F \le \sqrt{2\Big\|\frac{1}{g}\sum_{i=1}^g \mathbf{A}_i\Big\|_F^2 + 2\Big\|\frac{1}{g}\sum_{i=1}^g \mathbf{B}_i\Big\|_F^2} \\
&\le \max\Big\{\tfrac{\sqrt{\epsilon}}{\sqrt{g}},\tfrac{\eta}{\sqrt{g}}\Big\}\sqrt{\tfrac{\sigma_{\max}^2}{\sigma_{\max}^2+n\gamma}}\sqrt{2n\,f(\mathbf{W}^\star)} = \max\big\{\sqrt{\epsilon},\eta\big\}\tfrac{\sqrt{\beta}}{\sqrt{g}}\sqrt{2n\,f(\mathbf{W}^\star)}. \qquad (20)
\end{aligned}$$

It follows from (18), (19), and (20) that

$$\begin{aligned}
&\left\|(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^{1/2}(\mathbf{W}_i^{\mathrm{c}}-\mathbf{W}^\star)\right\|_F \\
&\leq \left\|\frac{1}{g}\sum_{i=1}^{g}\mathbf{A}_i\right\|_F+\left\|\frac{1}{g}\sum_{i=1}^{g}\mathbf{B}_i\right\|_F+2\eta\beta\,\frac{1}{g}\sum_{i=1}^{g}\left(\|\mathbf{A}_i\|_F+\|\mathbf{B}_i\|_F\right) \\
&\leq \left[\tfrac{1}{\sqrt{g}}\max\left\{\sqrt{\epsilon},\eta\right\}+2\beta\eta\cdot\max\left\{\sqrt{\epsilon},\eta\right\}\right]\sqrt{\beta}\sqrt{2n\,f(\mathbf{W}^\star)} \\
&= \max\left\{\sqrt{\epsilon},\eta\right\}\cdot\left(\tfrac{1}{\sqrt{g}}+2\beta\eta\right)\sqrt{\beta}\sqrt{2n\,f(\mathbf{W}^\star)} \\
&= \sqrt{\alpha\beta n\,f(\mathbf{W}^\star)}.
\end{aligned}$$

This concludes our proof. ■

D.2 Proof of Theorem 21

Proof By Lemma 25, we only need to show that $\left\|(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^{1/2}\big(\mathbf{W}^{\mathrm{h}}-\mathbf{W}^\star\big)\right\|_F^2 \leq \alpha^2\beta^2\big(-nf(\mathbf{W}^\star)+\|\mathbf{Y}\|_F^2\big)$.

In the proof of Theorem 2 we show that

$$(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^{1/2}\big(\mathbf{W}_i^{\mathrm{h}}-\mathbf{W}^\star\big) \;=\; \mathbf{V}\mathbf{A}_i\mathbf{B}_i\mathbf{C},$$

where

$$\begin{aligned}
\mathbf{A}_i &= (\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\mathbf{\Sigma}(\mathbf{I}_\rho-\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U})\mathbf{\Sigma}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}, \\
\mathbf{B}_i &= (\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{1/2}(\mathbf{\Sigma}\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U}\mathbf{\Sigma}+n\gamma\mathbf{I}_\rho)^{-1}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{1/2}, \\
\mathbf{C} &= (\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1/2}\mathbf{\Sigma}\mathbf{U}^T\mathbf{Y}.
\end{aligned}$$

It follows from Assumption 2.1 that for all $i\in[g]$,

$$\tfrac{1}{1+\eta}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1} \;\preceq\; (\mathbf{\Sigma}\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U}\mathbf{\Sigma}+n\gamma\mathbf{I}_\rho)^{-1} \;\preceq\; \tfrac{1}{1-\eta}(\mathbf{\Sigma}^2+n\gamma\mathbf{I}_\rho)^{-1}.$$

We let $\mathbf{B}_i=\mathbf{I}_\rho+\mathbf{\Delta}_i$. Thus $-\frac{\eta}{1+\eta}\mathbf{I}_\rho\preceq\mathbf{\Delta}_i\preceq\frac{\eta}{1-\eta}\mathbf{I}_\rho$. It follows that

$$\begin{aligned}
&(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^{1/2}\big(\mathbf{W}^{\mathrm{h}}-\mathbf{W}^\star\big) \;=\; \frac{1}{g}\sum_{i=1}^{g}(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^{1/2}\big(\mathbf{W}_i^{\mathrm{h}}-\mathbf{W}^\star\big) \\
&= \frac{1}{g}\sum_{i=1}^{g}\mathbf{V}\mathbf{A}_i(\mathbf{I}_\rho+\mathbf{\Delta}_i)\mathbf{C} \;=\; \frac{1}{g}\sum_{i=1}^{g}\mathbf{V}\mathbf{A}_i\mathbf{C}+\frac{1}{g}\sum_{i=1}^{g}\mathbf{V}\mathbf{A}_i\mathbf{\Delta}_i\mathbf{C}.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\left\|(\mathbf{X}^T\mathbf{X}+n\gamma\mathbf{I}_d)^{1/2}\big(\mathbf{W}^{\mathrm{h}}-\mathbf{W}^\star\big)\right\|_F \;\leq\; \left\|\frac{1}{g}\sum_{i=1}^{g}\mathbf{A}_i\right\|_2\left\|\mathbf{C}\right\|_F+\frac{1}{g}\sum_{i=1}^{g}\|\mathbf{A}_i\|_2\|\mathbf{\Delta}_i\|_2\|\mathbf{C}\|_F \\
&\leq \left\|\frac{1}{g}\sum_{i=1}^{g}\mathbf{A}_i\right\|_2\left\|\mathbf{C}\right\|_F+\frac{\eta}{1-\eta}\Big(\frac{1}{g}\sum_{i=1}^{g}\|\mathbf{A}_i\|_2\Big)\|\mathbf{C}\|_F. \qquad (21)
\end{aligned}$$

Let $\mathbf{S} = \frac{1}{g}[\mathbf{S}_1, \cdots, \mathbf{S}_g] \in \mathbb{R}^{n \times gs}$. It follows from the definition of $\mathbf{A}_i$ that

$$\begin{aligned}
\big\|\mathbf{A}_i\big\|_2 &= \Big\|(\boldsymbol{\Sigma}^2 + n\gamma\mathbf{I}_\rho)^{-1/2}\boldsymbol{\Sigma}(\mathbf{I}_\rho - \mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U})\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^2 + n\gamma\mathbf{I}_\rho)^{-1/2}\Big\|_2 \\
&\leq \eta\Big\|(\boldsymbol{\Sigma}^2 + n\gamma\mathbf{I}_\rho)^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^2 + n\gamma\mathbf{I}_\rho)^{-1/2}\Big\|_2 = \eta\tfrac{\sigma_{\max}^2}{\sigma_{\max}^2 + n\gamma} = \eta\beta, \\
\Big\|\frac{1}{g}\sum_{i=1}^g \mathbf{A}_i\Big\|_2 &= \Big\|(\boldsymbol{\Sigma}^2 + n\gamma\mathbf{I}_\rho)^{-1/2}\boldsymbol{\Sigma}(\mathbf{I}_\rho - \mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U})\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^2 + n\gamma\mathbf{I}_\rho)^{-1/2}\Big\|_2 \\
&\leq \tfrac{\eta}{\sqrt{g}}\big\|(\boldsymbol{\Sigma}^2 + n\gamma\mathbf{I}_\rho)^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}^2 + n\gamma\mathbf{I}_\rho)^{-1/2}\big\|_2 = \tfrac{\eta}{\sqrt{g}}\tfrac{\sigma_{\max}^2}{\sigma_{\max}^2 + n\gamma} = \tfrac{\eta\beta}{\sqrt{g}}.
\end{aligned}$$

It follows from (21) that

$$\begin{aligned}
&\Big\|(\mathbf{X}^T\mathbf{X} + n\gamma\mathbf{I}_d)^{1/2}\big(\mathbf{W}^{\mathrm{h}} - \mathbf{W}^\star\big)\Big\|_F \\
&\leq \Big(\tfrac{\eta}{\sqrt{g}} + \tfrac{\eta^2}{1-\eta}\Big)\beta\big\|\mathbf{C}\big\|_F \\
&\leq \Big(\tfrac{\eta}{\sqrt{g}} + \tfrac{\eta^2}{1-\eta}\Big)\beta\sqrt{-nf(\mathbf{W}^\star) + \|\mathbf{Y}\|_F^2},
\end{aligned}$$

where the latter inequality follows from the proof of Theorem 16. This concludes the proof. ■

Appendix E. Model Averaging from the Statistical Perspective: Proofs

In Section E.1 we prove Theorem 22. In Section E.2 we prove Theorem 23.

E.1 Proof of Theorem 22

Proof The bound on $\mathsf{bias}\big(\mathbf{W}^{\mathrm{c}}\big)$ can be shown in the same way as the proof of Theorem 17.

We prove the bound on $\mathsf{var}\big(\mathbf{W}^{\mathrm{c}}\big)$ in the following. It follows from Assumption 2.1 that

$$(1+\eta)^{-1}(\mathbf{I}_\rho + n\gamma\boldsymbol{\Sigma}^{-2})^{-1} \preceq (\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} + n\gamma\boldsymbol{\Sigma}^{-2})^\dagger \preceq (1-\eta)^{-1}(\mathbf{I}_\rho + n\gamma\boldsymbol{\Sigma}^{-2})^{-1}.$$

Let

$$(\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} + n\gamma\boldsymbol{\Sigma}^{-2})^\dagger = (\mathbf{I}_\rho + n\gamma\boldsymbol{\Sigma}^{-2})^{-1/2}(\mathbf{I}_\rho + \boldsymbol{\Delta}_i)(\mathbf{I}_\rho + n\gamma\boldsymbol{\Sigma}^{-2})^{-1/2}.$$

It holds that

$$-\frac{\eta}{1+\eta}\mathbf{I}_\rho \preceq \boldsymbol{\Delta}_i \preceq \frac{\eta}{1-\eta}\mathbf{I}_\rho.$$

By the definition of $\mathsf{var}(\mathbf{W}^{\mathrm{c}})$ in Theorem 9, we have that

$$\begin{aligned}
&\sqrt{\mathsf{var}\big(\mathbf{W}^{\mathrm{c}}\big)} \\
&= \frac{\xi}{\sqrt{n}}\bigg\|\frac{1}{g}\sum_{i=1}^{g}(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2})^{-1}\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T + \frac{1}{g}\sum_{i=1}^{g}(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2})^{-1/2}\mathbf{\Delta}_i(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2})^{-1/2}\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\bigg\|_F \\
&\leq \frac{\xi}{\sqrt{n}}\Big(\Big\|(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2})^{-1}\mathbf{U}^T\mathbf{S}\mathbf{S}^T\Big\|_F + \frac{1}{g}\sum_{i=1}^{g}\Big\|(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2})^{-1/2}\mathbf{\Delta}_i(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2})^{-1/2}\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\Big\|_F\Big) \\
&\leq \frac{\xi}{\sqrt{n}}\big\|(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2})^{-1}\big\|_F\Big(\big\|\mathbf{U}^T\mathbf{S}\big\|_2\big\|\mathbf{S}\big\|_2 + \frac{\eta}{1-\eta}\frac{1}{g}\sum_{i=1}^{g}\big\|\mathbf{U}^T\mathbf{S}_i\big\|_2\big\|\mathbf{S}_i\big\|_2\Big) \\
&= \sqrt{\mathsf{var}\big(\mathbf{W}^{\star}\big)}\Big(\big\|\mathbf{U}^T\mathbf{S}\big\|_2\big\|\mathbf{S}\big\|_2 + \frac{\eta}{1-\eta}\frac{1}{g}\sum_{i=1}^{g}\big\|\mathbf{U}^T\mathbf{S}_i\big\|_2\big\|\mathbf{S}_i\big\|_2\Big).
\end{aligned}$$

Under Assumption 2.1, we have that $\|\mathbf{S}_i^T\mathbf{U}\|_2^2 \leq 1+\eta$ and $\|\mathbf{S}^T\mathbf{U}\|_2^2 \leq 1+\frac{\eta}{\sqrt{g}}$. It follows that

$$\sqrt{\frac{\mathsf{var}\big(\mathbf{W}^{\mathrm{c}}\big)}{\mathsf{var}\big(\mathbf{W}^{\star}\big)}} \;\leq\; \sqrt{1+\frac{\eta}{\sqrt{g}}}\big\|\mathbf{S}\big\|_2 + \frac{\eta\sqrt{1+\eta}}{1-\eta}\frac{1}{g}\sum_{i=1}^{g}\big\|\mathbf{S}_i\big\|_2.$$

Now the desired result follows from Assumption 2.3. ■

E.2 Proof of Theorem 23

Proof The bound on $\mathsf{var}\big(\mathbf{W}^{\mathrm{h}}\big)$ can be established in the same way as Theorem 19.

We prove the bound on $\mathsf{bias}\big(\mathbf{W}^{\mathrm{h}}\big)$ in the following. Let

$$(\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2})^{\dagger} = (\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2})^{-1/2}(\mathbf{I}_\rho + \mathbf{\Delta}_i)(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2})^{-1/2}.$$

Under Assumption 2.1, we have that $\mathbf{\Delta}_i \preceq \frac{\eta}{1-\eta}\mathbf{I}_\rho$. It follows from Theorem 9 that

$$\begin{aligned}
\mathsf{bias}\big(\mathbf{W}^{\mathrm{h}}\big) \;&=\; \gamma\sqrt{n}\bigg\|\frac{1}{g}\sum_{i=1}^{g}\Big(\mathbf{\Sigma}^{-2} + \tfrac{\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U}-\mathbf{I}_\rho}{n\gamma}\Big)(\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2})^{\dagger}\mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_0\bigg\|_F \\
&\leq\; \gamma\sqrt{n}\bigg\|\frac{1}{g}\sum_{i=1}^{g}\Big(\mathbf{\Sigma}^{-2} + \tfrac{\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U}-\mathbf{I}_\rho}{n\gamma}\Big)(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2})^{-1}\mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_0\bigg\|_F \\
&\quad + \gamma\sqrt{n}\bigg\|\frac{1}{g}\sum_{i=1}^{g}\Big(\mathbf{\Sigma}^{-2} + \tfrac{\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U}-\mathbf{I}_\rho}{n\gamma}\Big)(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2})^{-1/2}\mathbf{\Delta}_i(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2})^{-1/2}\mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_0\bigg\|_F \\
&\triangleq\; \gamma\sqrt{n}(A+B),
\end{aligned}$$

where

$$\begin{aligned} A &= \Big\| \frac{1}{g}\sum_{i=1}^{g} \Big(\boldsymbol{\Sigma}^{-2} + \tfrac{\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U}-\mathbf{I}_\rho}{n\gamma} \Big) \big(\mathbf{I}_\rho + n\gamma\boldsymbol{\Sigma}^{-2}\big)^{-1}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{W}_0 \Big\|_F \\ &= \Big\| \Big(\boldsymbol{\Sigma}^{-2} + \tfrac{\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}-\mathbf{I}_\rho}{n\gamma} \Big) \big(\mathbf{I}_\rho + n\gamma\boldsymbol{\Sigma}^{-2}\big)^{-1}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{W}_0 \Big\|_F, \\ B &= \Big\| \frac{1}{g}\sum_{i=1}^{g} \Big(\boldsymbol{\Sigma}^{-2} + \tfrac{\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U}-\mathbf{I}_\rho}{n\gamma} \Big) \big(\mathbf{I}_\rho + n\gamma\boldsymbol{\Sigma}^{-2}\big)^{-1/2}\boldsymbol{\Delta}_i\big(\mathbf{I}_\rho + n\gamma\boldsymbol{\Sigma}^{-2}\big)^{-1/2}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{W}_0 \Big\|_F \\ &\le \frac{1}{g}\sum_{i=1}^{g} \Big\| \Big(\boldsymbol{\Sigma}^{-2} + \tfrac{\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U}-\mathbf{I}_\rho}{n\gamma} \Big) \big(\mathbf{I}_\rho + n\gamma\boldsymbol{\Sigma}^{-2}\big)^{-1/2}\boldsymbol{\Delta}_i\big(\mathbf{I}_\rho + n\gamma\boldsymbol{\Sigma}^{-2}\big)^{-1/2}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{W}_0 \Big\|_F. \end{aligned}$$

It follows from Assumption 2.1 that $\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}-\mathbf{I}_\rho$ is semidefinitely bounded between $\pm\frac{\eta}{\sqrt{g}}\mathbf{I}_\rho$. Thus

$$\Big(1 - \tfrac{\eta\sigma_{\max}^2}{n\gamma\sqrt{g}}\Big)\boldsymbol{\Sigma}^{-2} \preceq \boldsymbol{\Sigma}^{-2} + \tfrac{\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}-\mathbf{I}_\rho}{n\gamma} \preceq \Big(1 + \tfrac{\eta\sigma_{\max}^2}{n\gamma\sqrt{g}}\Big)\boldsymbol{\Sigma}^{-2}.$$

It follows that

$$\begin{aligned} A &= \Big\| \Big(\boldsymbol{\Sigma}^{-2} + \tfrac{\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}-\mathbf{I}_\rho}{n\gamma} \Big) \big(\mathbf{I}_\rho + n\gamma\boldsymbol{\Sigma}^{-2}\big)^{-1}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{W}_0 \Big\|_F \\ &\le \Big(1 + \tfrac{\eta\sigma_{\max}^2}{n\gamma\sqrt{g}}\Big) \Big\| \big(\boldsymbol{\Sigma}^2 + n\gamma\mathbf{I}_\rho\big)^{-1}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{W}_0 \Big\|_F. \end{aligned}$$

Similar to the proof of Theorem 19, we can show that

$$\begin{aligned} B &\le \Big(1 + \frac{\eta\sigma_{\max}^2}{n\gamma}\Big) \cdot \frac{1}{g}\sum_{i=1}^{g} \Big\| \boldsymbol{\Sigma}^{-2}\big(\mathbf{I}_\rho + n\gamma\boldsymbol{\Sigma}^{-2}\big)^{-1/2}\boldsymbol{\Delta}_i\big(\mathbf{I}_\rho + n\gamma\boldsymbol{\Sigma}^{-2}\big)^{-1/2}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{W}_0 \Big\|_F \\ &\le \frac{\eta}{1-\eta}\Big(1 + \frac{\eta\sigma_{\max}^2}{n\gamma}\Big) \cdot \Big\| \big(\boldsymbol{\Sigma}^2 + n\gamma\mathbf{I}_\rho\big)^{-1}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{W}_0 \Big\|_F. \end{aligned}$$

Hence

$$\begin{aligned} &\mathsf{bias}\big(\mathbf{W}^{\mathrm{h}}\big) \le \gamma\sqrt{n}\big(A + B\big) \\ &\le \Big[\tfrac{1}{1-\eta} + \Big(\tfrac{\eta}{\sqrt{g}} + \tfrac{\eta^2}{1-\eta} \Big) \tfrac{\sigma_{\max}^2}{n\gamma} \Big] \gamma\sqrt{n} \Big\| \big(\boldsymbol{\Sigma}^2 + n\gamma\mathbf{I}_\rho\big)^{-1}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{W}_0 \Big\|_F \\ &= \Big[\tfrac{1}{1-\eta} + \Big(\tfrac{\eta}{\sqrt{g}} + \tfrac{\eta^2}{1-\eta} \Big) \tfrac{\sigma_{\max}^2}{n\gamma} \Big] \mathsf{bias}\big(\mathbf{W}^\star\big). \end{aligned}$$

Here the equality follows from Theorem 4. ■

References

Haim Avron, Kenneth L. Clarkson, and David P. Woodruff. Sharper Bounds for Regularized Data Fitting. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, volume 81, pages 27:1–27:22, Dagstuhl, Germany, 2017. Schloss Dagstuhl.

Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding Frequent Items in Data Streams. *Theoretical Computer Science*, 312(1):3–15, 2004.

Kenneth L. Clarkson and David P. Woodruff. Low Rank Approximation and Regression in Input Sparsity Time. In *Annual ACM Symposium on Theory of Computing (STOC)*, 2013.

Michal Derezinski and Manfred K. Warmuth. Unbiased estimates for linear regression via volume sampling. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Michal Derezinski and Manfred K. Warmuth. Subsampling for ridge regression via regularized volume sampling. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Petros Drineas and Michael W. Mahoney. RandNLA: Randomized Numerical Linear Algebra. *Communications of the ACM*, 59(6):80–90, 2016.

Petros Drineas, Ravi Kannan, and Michael W. Mahoney. Fast Monte Carlo Algorithms for Matrices I: Approximating Matrix Multiplication. *SIAM Journal on Computing*, 36(1): 132–157, 2006a.

Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Sampling Algorithms for ℓ_2 Regression and Applications. In *Annual ACM-SIAM Symposium on Discrete Algorithm (SODA)*, 2006b.

Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Relative-Error CUR Matrix Decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, September 2008.

Petros Drineas, Michael W. Mahoney, S. Muthukrishnan, and Tamás Sarlós. Faster Least Squares Approximation. *Numerische Mathematik*, 117(2):219–249, 2011.

Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast Approximation of Matrix Coherence and Statistical Leverage. *Journal of Machine Learning Research*, 13:3441–3472, 2012.

Alex Gittens. The Spectral Norm Error of the Naive Nyström Extension. *arXiv preprint arXiv:1110.5305*, 2011.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600, 2014.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26(189-206), 1984.

Yichao Lu, Paramveer Dhillon, Dean P. Foster, and Lyle Ungar. Faster Ridge Regression via the Subsampled Randomized Hadamard Transform. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

Ping Ma, Michael W. Mahoney, and Bin Yu. A Statistical Perspective on Algorithmic Leveraging. *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.

Xiangrui Meng and Michael W. Mahoney. Low-Distortion Subspace Embeddings in Input-Sparsity Time and Applications to Robust Linear Regression. In *Annual ACM Symposium on Theory of Computing (STOC)*, 2013.

John Nelson and Huy L. Nguyên. OSNAP: Faster Numerical Linear Algebra Algorithms via Sparser Subspace Embeddings. In *IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, 2013.

Mihai Patrascu and Mikkel Thorup. The Power of Simple Tabulation-Based Hashing. *Journal of the ACM*, 59(3), 2012.

Ninh Pham and Rasmus Pagh. Fast and Scalable Polynomial Kernels via Explicit Feature Maps. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2013.

Mert Pilanci and Martin J. Wainwright. Iterative Hessian Sketch: Fast and Accurate Solution Approximation for Constrained Least-Squares. *Journal of Machine Learning Research*, pages 1–33, 2015.

Garvesh Raskutti and Michael W. Mahoney. A Statistical Perspective on Randomized Sketching for Ordinary Least-Squares. *Journal of Machine Learning Research*, 17(214): 1–31, 2016.

Gian-Andrea Thanei, Christina Heinze, and Nicolai Meinshausen. Random Projections For Large-Scale Regression. In *Big and Complex Data Analysis*. Springer, 2017.

Joel A. Tropp. Improved Analysis of the Subsampled Randomized Hadamard Transform. *Advances in Adaptive Data Analysis*, 3(01n02):115–126, 2011.

Stephen Tu, Rebecca Roelofs, Shivaram Venkataraman, and Benjamin Recht. Large Scale Kernel Learning using Block Coordinate Descent. *arXiv preprint arXiv:1602.05310*, 2016.

Roman Vershynin. *Introduction to the Non-Asymptotic Analysis of Random Matrices*, pages 210–268. Cambridge University Press, 2012.

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, 2017a.

Shusen Wang, Luo Luo, and Zhihua Zhang. SPSD Matrix Approximation via Column Selection: Theories, Algorithms, and Extensions. *Journal of Machine Learning Research*, 17(49):1–49, 2016a.

Shusen Wang, Zhihua Zhang, and Tong Zhang. Towards More Efficient SPSD Matrix Approximation and CUR Matrix Decomposition. *Journal of Machine Learning Research*, 17(210):1–49, 2016b.

Shusen Wang, Farbod Roosta-Khorasani, Peng Xu, and Michael W. Mahoney. GIANT: Globally Improved Approximate Newton Method for Distributed Optimization. *arXiv preprint arXiv:1709.03528*, 2017b.

Yining Wang, Adams Wei Yu, and Aarti Singh. On computationally tractable selection of experiments in measurement-constrained regression models. *Journal of Machine Learning Research*, 18(143):1–41, 2017c.

Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. Feature Hashing for Large Scale Multitask Learning. In *International Conference on Machine Learning (ICML)*, 2009.

David P. Woodruff. Sketching as a Tool for Numerical Linear Algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Franco Woolfe, Edo Liberty, Vladimir Rokhlin, and Mark Tygert. A Fast Randomized Algorithm for the Approximation of Matrices. *Applied and Computational Harmonic Analysis*, 25(3):335–366, 2008.

Jiyan Yang, Xiangrui Meng, and Michael W. Mahoney. Implementing Randomized Matrix Algorithms in Parallel and Distributed Environments. *Proceedings of the IEEE*, 104(1): 58–92, 2016.

Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker, and Ion Stoica. Spark: Cluster Computing with Working Sets. *HotCloud*, 10(10-10):95, 2010.

Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Communication-Efficient Algorithms for Statistical Optimization. *Journal of Machine Learning Research*, 14:3321–3363, 2013.

Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Divide and Conquer Kernel Ridge Regression: a Distributed Algorithm with Minimax Optimal Rates. *Journal of Machine Learning Research*, 16:3299–3340, 2015.