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

Shusen Wang [1] **Alex Gittens** [2] **Michael W. Mahoney** [1]

Abstract

We address the statistical and optimization impacts of using classical sketch versus Hessian sketch to solve approximately the Matrix Ridge Regression (MRR) problem. Prior research has considered the effects of classical sketch on least squares regression (LSR), a strictly simpler 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. In 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 approximations, the regularization in the sketched MRR problem gives it significantly different statistical properties from 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 biases and variances of sketched MRR; these establish that the variance is significantly increased when classical sketches are used, while the bias is significantly increased when using Hessian sketches. 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 this gap. Thus, in the distributed setting, sketching combined with model averaging is a powerful technique that quickly obtains near-optimal solutions to the MRR problem while greatly mitigating the statistical risks incurred by sketching.

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 $\mathbf{X}$, and the stopping criterion.

This paper considers the $n \gg d$ problem, where there is much redundancy in $\mathbf{X}$. Matrix sketching, as used within Randomized Linear Algebra (RLA) (Mahoney, 2011; Woodruff, 2014), works by reducing the size of $\mathbf{X}$ without losing too much information; this operation can be modeled as taking actual rows or linear combinations of the rows of $\mathbf{X}$ with a sketching matrix $\mathbf{S}$ to form the sketch $\mathbf{S}^T\mathbf{X}$. Here $\mathbf{S} \in \mathbb{R}^{n\times s}$ 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., 2006; 2011; Clarkson & Woodruff, 2013; Meng & Mahoney, 2013; Nelson & 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, $T_{\mathrm{s}} = \mathcal{O}(nd \log s)$ 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 & 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;

[1] International Computer Science Institute and Department of Statistics, University of California at Berkeley, USA [2] Department of Computer Science, Rensselaer Polytechnic Institute, USA. Correspondence to: Shusen Wang <shusen@berkeley.edu>, Alex Gittens <gittea@rpi.edu>, Michael W. Mahoney <mmahoney@stat.berkeley.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

[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.

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 $\mathbf{S}^T\mathbf{Y} \in \mathbb{R}^{s\times m}$.

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

Woodruff, 2014) and the references therein.

The concept of sketched LSR originated in the theoretical computer science literature, e.g., (Drineas et al., 2006; 2011), where the behavior of sketched LSR was 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 function value $\big\|\mathbf{X}\tilde{\mathbf{w}} - \mathbf{y}\big\|_2^2$ is at most ϵ times worse than $\big\|\mathbf{X}\mathbf{w}^\star - \mathbf{y}\big\|_2^2$. These works also bounded $\|\tilde{\mathbf{w}} - \mathbf{w}^\star\|_2^2$ in terms of the difference in the objective function values 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 & Mahoney, 2016; Pilanci & Wainwright, 2015; Wang et al., 2016b) considered statistical properties of sketched LSR like the bias and variance. In particular, Pilanci & Wainwright (2015) showed that sketched LSR has much higher variance than the optimal solution.

Both of these perspectives are important and of practical interest. The optimization perspective is relevant when the data can be taken as deterministic values. The statistical perspective is relevant in machine learning and statistics applications where the data are random, and the regression coefficients are therefore themselves random variables.

In practice, regularized regression, e.g., ridge regression and LASSO, exhibit more attractive bias-variance tradeoffs 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, the question: *How, if at all, does our understanding of the optimization and statistical properties of sketched LSR generalize to sketched regularized regression?* We answer this question for sketched matrix ridge regression (MRR).

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

$$\min_{\mathbf{W}} \Big\{ f(\mathbf{W}) \triangleq \tfrac{1}{n}\big\|\mathbf{X}\mathbf{W} - \mathbf{Y}\big\|_F^2 + \gamma\|\mathbf{W}\|_F^2 \Big\}, \qquad (1)$$

which has optimal solution

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

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:

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

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

Following the convention of Pilanci & Wainwright (2015); Wang et al. (2016a), we call $\mathbf{W}^{\mathrm{c}}$ **classical sketch** and $\mathbf{W}^{\mathrm{h}}$ **Hessian sketch**, which approximate the **optimal solution** $\mathbf{W}^\star$. Table 1 lists the time costs of the three solutions to MRR.

1.1. Main Results and Contributions

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

- Classical sketch achieves relative error in the objective value. With sketch size $s = \tilde{\mathcal{O}}(d/\epsilon)$, the objective 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 improves as the regularization parameter γ increases.

We then study classical and Hessian sketches from the **statistical perspective**, by modeling $\mathbf{Y} = \mathbf{X}\mathbf{W}_0 + \boldsymbol{\Xi}$ 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):

- The bias of the 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, i.e., large γ, should be used to supress the variance. (As γ increases, the bias increases, and the variance decreases.)
- Since $\mathbf{Y}$ is not sketched with Hessian sketch, the variance of Hessian sketch can be close to the optimal solution. However, Hessian sketch has 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 have low bias.

Our empirical evaluations bear out these theoretical results. In particular, in Section 4, we show in Figure 2 that even when the regularization parameter γ is fine-tuned, the risks of classical sketch 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 sketching 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 classical and Hessian sketches with model averaging by $\mathbf{W}^{\mathrm{c}}$ and $\mathbf{W}^{\mathrm{h}}$, respectively. Theorems 7, 8, 10, 11 give the following results:

- Classical Sketch. Assume the sketch size $s = \tilde{\mathcal{O}}(\frac{d}{\epsilon})$ and $\epsilon \leq \frac{1}{g}$; then the bound on $f(\mathbf{W}^{\mathrm{c}}) - f(\mathbf{W}^{\star})$ is proportional to $\frac{\epsilon}{g}$. Assume that $s = \tilde{\mathcal{O}}(\frac{d}{\epsilon^2})$ and $\epsilon^2 \leq \frac{1}{g}$; the bias does not increase; the variance bound is proportional to $\frac{1}{g}$.

- Hessian Sketch. Assume that $s = \tilde{\mathcal{O}}(\frac{d}{\epsilon})$ and $\epsilon \leq \frac{1}{g^2}$; then the bound on $f(\mathbf{W}^{\mathrm{h}}) - f(\mathbf{W}^{\star})$ is proportional to $\frac{\epsilon}{g^2}$. Assume that $s = \tilde{\mathcal{O}}(\frac{d}{\epsilon^2})$; the variance does not increase; if, additionally, $\epsilon \leq \frac{1}{g}$ and $n\gamma$ is much smaller than the squared spectral norm of $\mathbf{X}$, then the bias bound is proportional to $\frac{\epsilon}{g}$.

Note that classical sketch with uniform sampling and model averaging is very well known as bagging (Breiman, 1996) (or pasting (Breiman, 1999) or bootstrap aggregating). Different from bagging, our model averaging approach is not limited to uniform sampling.

Classical sketch with model averaging has three immediate applications. In the single-machine setting,

- Classical sketch with model averaging offers a way to improve the statistical performance in the presence of heavy noise. Assume the sketch size is $s = \tilde{\mathcal{O}}(\sqrt{nd})$. As g grows larger than $\frac{n}{s}$, the variance of the averaged solution can be even lower than the optimal solution. See Remark 1 for further discussion. This observation is in accordance with the observation that bagging reduces 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 constructed by uniformly sampled rows from the dataset without replacement. In this setting, the model averaging procedure will communicate the g local models only once to return the final estimate; this process has very low communication complexity and latency, and it suggests two further applications of classical sketch with model averaging:

- Model Averaging for Machine Learning. If a low-precision solution is acceptable, the averaged solution can be used in lieu of distributed numerical optimization algorithms requiring multiple rounds of communication. If $\frac{n}{g}$ is big 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 numerical optimization algorithms heavily depends on the quality of the initialization.[2] A good initialization saves lots of iterations. The averaged model 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., 2006; 2011; Clarkson & Woodruff, 2013; Meng & Mahoney, 2013; Nelson & 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 has a nontrivial bias 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, whereas they assumed $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are i.i.d. from an unknown distribution; second, our results apply to many other sketching ensembles than uniform sampling without replacement; and third, we provide both optimization and statistical perspectives, whereas they provide only a statistical perspective. Our results clearly indicate that the

[2] For example, the conjugate gradient method satisfies $\frac{\|\mathbf{W}^{(t)} - \mathbf{W}^{\star}\|_F^2}{\|\mathbf{W}^{(0)} - \mathbf{W}^{\star}\|_F^2} \leq \theta_1^t$; the 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.

performance critically depends on the row coherence of $\mathbf{X}$; this dependence is not captured in (Zhang et al., 2013). For similar reasons, our work is different from the divide-and-conquer kernel ridge regression algorithm of (Zhang et al., 2015).

Iterative Hessian sketch has been studied by Pilanci & Wainwright (2015); Wang et al. (2016a). By way of comparison, all the algorithms in this paper are "one-shot" rather than iterative. Upon completion of this paper, we noticed the contemporary works (Avron et al., 2016; Thanei et al., 2017). Avron et al. (2016) studied classical sketch from the optimization perspective, and Thanei et al. (2017) studied LSR with model averaging.

1.3. Paper Organization

Section 2 defines our notation and introduces the sketching schemes we consider. Section 3 presents our theoretical results. Section 4 conducts experiments to verify our theories and demonstrates the usefulness of model averaging. Proofs of our claims and more empirical evaluations can be found in the technical report version (Wang et al., 2017).

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 $\mathbf{a}_{:j}$ denotes the j-th column. 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. Let $\tilde{\mathcal{O}}$ conceal logarithm factors.

Throughout, we fix $\mathbf{X} \in \mathbb{R}^{n\times d}$ as our matrix of features. We set $\rho = \mathrm{rank}(\mathbf{X})$ 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 $\mu(\mathbf{X})$.

Matrix sketching turns big matrices into smaller ones without losing too much information 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 per column, whose position indicates the index of the selected row; 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., 2015). The sampling probabilities of shrinked leverage score sampling are defined by $p_i = \frac{1}{2}\big(\frac{l_i}{\sum_{j=1}^n l_j} + \frac{1}{n}\big)$.[3]

Gaussian projection is also well-known as the prototypical Johnson-Lindenstrauss transform (Johnson & Lindenstrauss, 1984). Let $\mathbf{G} \in \mathbb{R}^{m\times s}$ 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 inefficient relative to other forms of sketching.

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 & Woodruff, 2013; Meng & Mahoney, 2013; Nelson & Nguyên, 2013; Pham & Pagh, 2013; Weinberger et al., 2009). Though more efficient to apply, CountSketch requires a bigger sketch size than Gaussian projections, SRHT, and leverage score sampling to attain the same theoretical guarantees. The readers can refer to (Woodruff, 2014) for a detailed description of CountSketch.

3. Main Results

Sections 3.1 and 3.2 analyze sketched MRR from, respectively, optimization and statistical perspectives. Sec-

[3] 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 simplicity; our conclusions extend in a straightforward manner to other weightings.

tions 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**; and we will mention explicitly uniform sampling and CountSketch. The notation defined in Table 2 are used throughout.

Table 2. 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
$\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$	the SVD of $\mathbf{X}$
μ	the row coherence of $\mathbf{X}$
γ	the regularization parameter
β	$\beta = \frac{\|\mathbf{X}\|_2^2}{\|\mathbf{X}\|_2^2 + n\gamma} \leq 1$
$\mathbf{S} \in \mathbb{R}^{n \times s}$	a sketching matrix

3.1. Sketched MRR: Optimization Perspective

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

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

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

holds with probability at least 0.9.

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, which show that unless γ is big 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)**.** *For the four sketching methods with $s = \tilde{\mathcal{O}}\big(\frac{\beta^2 d}{\epsilon}\big)$, uniform sampling with $s = \mathcal{O}\big(\frac{\mu\beta^2 d \log d}{\epsilon}\big)$, and CountSketch with $s = \mathcal{O}\big(\frac{\beta^2 d^2}{\epsilon}\big)$, the inequality*

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

holds with probability at least 0.9.

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 studies 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. *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:

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

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

$$\mathbf{Y} \;=\; \mathbf{X}\mathbf{W}_0 + \mathbf{\Xi} \tag{5}$$

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:

$$R(\mathbf{W}) \;=\; \tfrac{1}{n}\mathbb{E}\big\|\mathbf{X}\mathbf{W} - \mathbf{X}\mathbf{W}_0\big\|_F^2, \tag{6}$$

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), (4), respectively; then the risk function can be decomposed as*

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

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

$$\mathsf{bias}\big(\mathbf{W}^\star\big) \;=\; \gamma\sqrt{n}\Big\|(\mathbf{\Sigma}^2 + n\gamma\mathbf{I}_\rho)^{-1}\mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_0\Big\|_F,$$

$$\mathsf{var}\big(\mathbf{W}^\star\big) \;=\; \frac{\xi^2}{n}\Big\|\big(\mathbf{I}_\rho + n\gamma\mathbf{\Sigma}^{-2}\big)^{-1}\Big\|_F^2,$$

$$\mathsf{bias}\big(\mathbf{W}^c\big) \;=\; \gamma\sqrt{n}\Big\|\big(\mathbf{\Sigma}\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U}\mathbf{\Sigma} + n\gamma\mathbf{I}_\rho\big)^\dagger\mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_0\Big\|_F,$$

$$\mathsf{var}\big(\mathbf{W}^c\big) \;=\; \frac{\xi^2}{n}\Big\|\big(\mathbf{U}^T\mathbf{S}\mathbf{S}^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger\mathbf{U}^T\mathbf{S}\mathbf{S}^T\Big\|_F^2,$$

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

Theorem 5 provides upper and lower bounds on the bias and variance of the 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, $\mathsf{var}(\mathbf{W}^{\mathrm{c}})$ is $\Theta(\frac{n}{s})$ times worse than $\mathsf{var}(\mathbf{W}^\star)$.

Theorem 5 (Classical Sketch)**.** *For Gaussian projection and SRHT sketching with $s = \tilde{\mathcal{O}}(\frac{d}{\epsilon^2})$, uniform sampling with $s = \mathcal{O}(\mu\frac{d\log d}{\epsilon^2})$, 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,[4] *hold with probability at least 0.9.*

Theorem 6 establishes similar upper and lower bounds on the bias and variance of 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 $\mathsf{bias}(\mathbf{W}^\star)$.

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

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

hold with probability at least 0.9. Further assume that $\sigma_\rho^2 \geq \frac{n\gamma}{\epsilon}$. Then

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

holds with probability at least 0.9.

The lower bound on the bias shows that Hessian sketch can suffer from 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 2.

In sum, our theories show that 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 2 show that even with fine-tuned γ, the risks of classical and Hessian sketches 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 2 suggests classical and Hessian sketches both require over-regularization, i.e., setting γ larger than what 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 sketches, the underlying explanations are different. Classical sketch has a high variance, so a large γ is required to supress the variance (its variance is non-increasing with γ). Hessian sketch has very high bias when γ is small, so a reasonably large γ is necessary to lower its bias.

3.3. Model Averaging: Optimization Perspective

We consider model averaging as an approach to increasing 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 to combine the classical or 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 $\epsilon \leq \frac{1}{g}$, the bound on $f(\mathbf{W}^{\mathrm{h}}) - f(\mathbf{W}^\star)$ is proportional to ϵ/g. From Lemma 3 we can see that the distances from $\mathbf{W}^{\mathrm{c}}$ to $\mathbf{W}^\star$ also decreases accordingly.

Theorem 7 (Classical Sketch with Model Averaging)**.** *For the four methods, let $s = \tilde{\mathcal{O}}\big(\frac{\beta d}{\epsilon}\big)$; for uniform sampling, let $s = \mathcal{O}\big(\frac{\mu\beta d\log d}{\epsilon}\big)$. Then the inequality*

$$f(\mathbf{W}^c) - f(\mathbf{W}^\star) \quad\leq\quad \big(\tfrac{\epsilon}{g} + \beta^2\epsilon^2\big)\, f(\mathbf{W}^\star)$$

[4] For shrinked leverage score sampling, $\|\mathbf{S}\|_2^2$ does not enjoy nontrivial lower bound. This is why we do not have a lower bound on the variance.

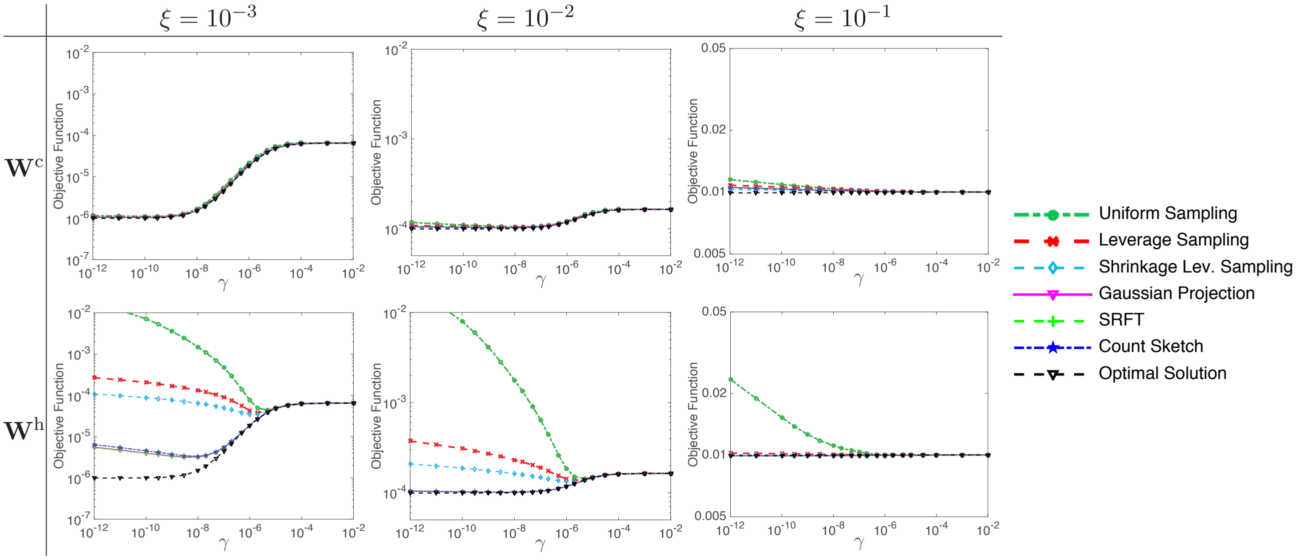

Figure 1. Empirical study of classical sketch and Hessian sketch from optimization perspective. The x-axis is the regularization parameter γ (log-scale); the y-axis is the objective function values (log-scale).

holds with probability at least 0.8.

For Hessian sketch with model averaging, if $\frac{\epsilon}{\beta^2} \leq \frac{1}{g^2}$, then the bound on $f(\mathbf{W}^{\mathrm{h}}) - f(\mathbf{W}^\star)$ is proportional to $\frac{\epsilon}{g^2}$.

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

$$f(\mathbf{W}^h) - f(\mathbf{W}^\star) \quad \leq \quad \big(\tfrac{\epsilon}{g^2} + \tfrac{\epsilon^2}{\beta^2}\big)\,\big(\tfrac{\|\mathbf{Y}\|_F^2}{n} - f(\mathbf{W}^\star)\big).$$

holds with probability at least 0.8.

3.4. Model Averaging: Statistical Perspective

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

Theorem 9 (Bias-Variance Decomposition)**.** *We consider the fixed design model* (5)*. The risk function defined in* (6) *can be decomposed as*

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

The bias and variance terms are

$$\mathsf{bias}\big(\mathbf{W}^c\big) = \gamma\sqrt{n}\Big\| \frac{1}{g}\sum_{i=1}^{g}\big(\mathbf{\Sigma}\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U}\mathbf{\Sigma} + n\gamma\mathbf{I}_\rho\big)^\dagger \mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_0\Big\|_F,$$

$$\mathsf{var}\big(\mathbf{W}^c\big) = \frac{\xi^2}{n}\Big\|\frac{1}{g}\sum_{i=1}^{g}\big(\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\Big\|_F^2,$$

$$\mathsf{bias}\big(\mathbf{W}^h\big) = \gamma\sqrt{n}\Big\|\frac{1}{g}\sum_{i=1}^{g}\big(\mathbf{\Sigma}^{-2} + \frac{\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} - \mathbf{I}_\rho}{n\gamma}\big) \cdot \big(\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger\mathbf{\Sigma}\mathbf{V}^T\mathbf{W}_0\Big\|_F,$$

$$\mathsf{var}\big(\mathbf{W}^h\big) = \frac{\xi^2}{n}\Big\|\frac{1}{g}\sum_{i=1}^{g}\big(\mathbf{U}^T\mathbf{S}_i\mathbf{S}_i^T\mathbf{U} + n\gamma\mathbf{\Sigma}^{-2}\big)^\dagger\Big\|_F^2.$$

Theorems 10 and 11 provide upper bounds on the bias and variance of model-averaged sketched MRR 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 (Classical Sketch with Model Averaging)**.** *For shrinked leverage score sampling, Gaussian projection, SRHT 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*

$$\frac{\mathsf{bias}(\mathbf{W}^c)}{\mathsf{bias}(\mathbf{W}^\star)} \leq 1+\epsilon,$$

$$\frac{\mathsf{var}(\mathbf{W}^c)}{\mathsf{var}(\mathbf{W}^\star)} \leq \frac{n}{s}\Big(\sqrt{\tfrac{1+\epsilon/g}{g}} + \epsilon\Big)^2$$

hold with probability at least 0.8.

Remark 1. *From this result, we see that if $\epsilon \leq \frac{1}{\sqrt{g}}$, then the variance is proportional to $\frac{1}{g}$. If g and s are at least*

$$g = \mathcal{O}\Big(\frac{n}{s}\Big) \quad \textit{and} \quad s = \tilde{\mathcal{O}}\big(\sqrt{nd}\big),$$

then the risk $R(\mathbf{W}^c)$ is close to $R(\mathbf{W}^\star)$. If g and s are larger, then the variance $\mathsf{var}(\mathbf{W}^c)$ *can even be even lower than* $\mathsf{var}(\mathbf{W}^\star)$.

Theorem 11 shows that model averaging decreases the bias of Hessian sketch without increasing the variance. For Hessian sketch without model averaging, recall that $\mathsf{bias}(\mathbf{W}^{\mathrm{h}})$

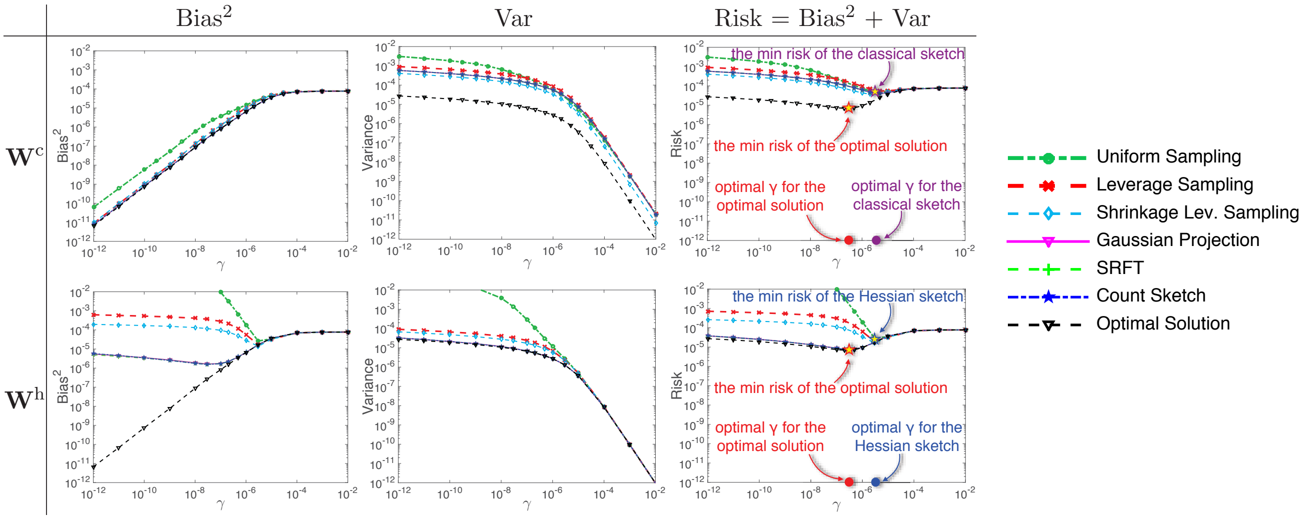

Figure 2. Empirical study of classical sketch and Hessian sketch from statistical perspective. The x-axis is the regularization parameter γ (log-scale); the y-axes are respectively bias^2, variance, and risk (log-scale). We annotate the minimum risks and optimal γ in the plots.

is larger than bias$(\mathbf{W}^\star)$ by a factor of $\mathcal{O}(\|\mathbf{X}\|_2^2/(n\gamma))$. Theorem 11 shows that model averaging reduces this ratio by a factor of $\frac{\epsilon}{g}$ when $\epsilon \leq \frac{1}{g}$.

Theorem 11 (Hessian Sketch with Model Averaging)**.** *For the four 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*

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

hold with probability at least 0.8.

4. Sketched Ridge Regression Experiments

Following (Ma et al., 2015; Yang et al., 2016), we constructed $\mathbf{X} \in \mathbb{R}^{n\times d}$ to have condition number $\kappa(\mathbf{X}^T\mathbf{X}) = 10^{12}$ and high row coherence, fixed $\mathbf{w}_0 = [\mathbf{1}_{0.2d}; 0.1\mathbf{1}_{0.6d}; \mathbf{1}_{0.2d}]$, and set $\mathbf{y} = \mathbf{X}\mathbf{w}_0 + \boldsymbol{\varepsilon} \in \mathbb{R}^n$, where the entries of $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ were i.i.d. sampled from $\mathcal{N}(0, \xi^2)$. The details of this data model are given in the technical report version (Wang et al., 2017). 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 $s = 5{,}000$. Because the analytical expressions involve the random sketching matrix $\mathbf{S}$, we randomly generate $\mathbf{S}$, repeat this procedure 10 times, and report the averaged results.

In Figure 1, 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 noise intensity ξ. The results verify our theory: classical sketch $\mathbf{w}^{\mathrm{c}}$ is always close to optimal; Hessian sketch $\mathbf{w}^{\mathrm{h}}$ is much worse than the optimal when γ is small and $\mathbf{y}$ is mostly in the column space of $\mathbf{X}$.

We conducted experiments on synthetic data to verify Theorems 5 and 6 and to show the effects of classical and Hessian sketching on the bias and variance. We set the noise intensity to be $\xi = 0.1$. In Figure 2, we plot the analytical expressions for the squared bias, variance, and risk stated in Theorem 4 against the regularization parameter γ. The results of this experiment match our theory: classical sketch magnified the variance, and Hessian sketch increased the bias. Even if γ is fine-tuned, the risks of classical and Hessian sketches can be much higher than those of the optimal solution.[5] Our experiments also indicate that classical and Hessian sketches require setting γ larger than the best regularization parameter for the optimal solution $\mathbf{W}^\star$.

5. Conclusions

We studied sketched matrix ridge regression (MRR) from 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.

[5] In the experiment yielding Figure 2, Hessian sketch had lower risk than classical sketch. This is not generally true: if we used a smaller ξ, so that the variance is dominated by bias, then classical sketch results in lower risks than Hessian sketch.

Acknowledgements

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

References

Avron, Haim, Clarkson, Kenneth L., and Woodruff, David P. Sharper bounds for regression and low-rank approximation with regularization. *arXiv preprint arXiv:1611.03225*, 2016.

Breiman, Leo. Bagging predictors. *Machine Learning*, 24(2): 123–140, 1996.

Breiman, Leo. Pasting small votes for classification in large databases and on-line. *Machine Learning*, 36(1-2):85–103, 1999.

Charikar, Moses, Chen, Kevin, and Farach-Colton, Martin. Finding frequent items in data streams. *Theoretical Computer Science*, 312(1):3–15, 2004.

Clarkson, Kenneth L. and Woodruff, David P. Low rank approximation and regression in input sparsity time. In *Annual ACM Symposium on theory of computing (STOC)*, 2013.

Drineas, Petros, Mahoney, Michael W., and Muthukrishnan, S. Sampling algorithms for ℓ_2 regression and applications. In *Annual ACM-SIAM Symposium on Discrete Algorithm (SODA)*, 2006.

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

Drineas, Petros, Magdon-Ismail, Malik, Mahoney, Michael W., and Woodruff, David P. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13:3441–3472, 2012.

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

Lu, Yichao, Dhillon, Paramveer, Foster, Dean P, and Ungar, Lyle. Faster ridge regression via the subsampled randomized Hadamard transform. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

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

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

Meng, Xiangrui and Mahoney, Michael W. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Annual ACM Symposium on Theory of Computing (STOC)*, 2013.

Nelson, John and Nguyên, Huy L. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, 2013.

Pham, Ninh and Pagh, Rasmus. Fast and scalable polynomial kernels via explicit feature maps. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2013.

Pilanci, Mert and Wainwright, Martin J. Iterative Hessian sketch: Fast and accurate solution approximation for constrained least-squares. *Journal of Machine Learning Research*, pp. 1–33, 2015.

Raskutti, Garvesh and Mahoney, Michael W. A statistical perspective on randomized sketching for ordinary least-squares. *Journal of Machine Learning Research*, 17(214):1–31, 2016.

Thanei, Gian-Andrea, Heinze, Christina, and Meinshausen, Nicolai. Random projections for large-scale regression. *arXiv preprint arXiv:1701.05325*, 2017.

Tropp, Joel A. Improved analysis of the subsampled randomized Hadamard transform. *Advances in Adaptive Data Analysis*, 3 (01n02):115–126, 2011.

Tu, Stephen, Roelofs, Rebecca, Venkataraman, Shivaram, and Recht, Benjamin. Large scale kernel learning using block coordinate descent. *arXiv preprint arXiv:1602.05310*, 2016.

Wang, Jialei, Lee, Jason D, Mahdavi, Mehrdad, Kolar, Mladen, and Srebro, Nathan. Sketching meets random projection in the dual: A provable recovery algorithm for big and high-dimensional data. *arXiv preprint arXiv:1610.03045*, 2016a.

Wang, Shusen, Gittens, Alex, and Mahoney, Michael W. Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. *arXiv preprint arXiv:1702.04837*, 2017.

Wang, Yining, Yu, Adams Wei, and Singh, Aarti. Computationally feasible near-optimal subset selection for linear regression under measurement constraints. *arXiv preprint arXiv:1601.02068*, 2016b.

Weinberger, Kilian, Dasgupta, Anirban, Langford, John, Smola, Alex, and Attenberg, Josh. Feature hashing for large scale multitask learning. In *International Conference on Machine Learning (ICML)*, 2009.

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

Woolfe, Franco, Liberty, Edo, Rokhlin, Vladimir, and Tygert, Mark. A fast randomized algorithm for the approximation of matrices. *Applied and Computational Harmonic Analysis*, 25 (3):335–366, 2008.

Yang, Jiyan, Meng, Xiangrui, and Mahoney, Michael W. Implementing randomized matrix algorithms in parallel and distributed environments. *Proceedings of the IEEE*, 104(1):58–92, 2016.

Zhang, Yuchen, Duchi, John C., and Wainwright, Martin J. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321–3363, 2013.

Zhang, Yuchen, Duchi, John, and Wainwright, Martin. Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16:3299–3340, 2015.